Comparison of view-based and reconstruction-based models of human navigational strategy

Luise Gootjes-Dreesbach
School of Psychology and Clinical Language Sciences, University of Reading, Reading, United Kingdom

Lyndsey C. Pickup
Optellum Ltd., Oxford Centre for Innovation, Oxford, United Kingdom

Andrew W. Fitzgibbon
Microsoft Research Ltd., Cambridge, United Kingdom

Andrew Glennerster
School of Psychology and Clinical Language Sciences, University of Reading, Reading, United Kingdom

There is good evidence that simple animals, such as bees, use view-based strategies to return to a familiar location, whereas humans might use a 3-D reconstruction to achieve the same goal. Assuming some noise in the storage and retrieval process, these two types of strategy give rise to different patterns of predicted errors in homing. We describe an experiment that can help distinguish between these models. Participants wore a head-mounted display to carry out a homing task in immersive virtual reality. They viewed three long, thin, vertical poles and had to remember where they were in relation to the poles before being transported (virtually) to a new location in the scene from where they had to walk back to the original location. The experiment was conducted in both a rich-cue scene (a furnished room) and a sparse scene (no background and no floor or ceiling). As one would expect, in a rich-cue environment, the overall error was smaller, and in this case, the ability to separate the models was reduced. However, for the sparse-cue environment, the view-based model outperforms the reconstruction-based model. Specifically, the likelihood of the experimental data is similar to the likelihood of samples drawn from the view-based model (but assessed under both models), and this is not true for samples drawn from the reconstruction-based model.

Introduction

Theories about navigation and "homing" in animals fall into two broad categories. On the one hand, the individual could make a mental map, record the "home" location on the map, and then return to it. An alternative is for the individual to remember the sensory data at the home location (e.g., the views from that point) and to try to return to a similar sensory state. In this paper, we make quantitative models of both types of strategy and compare their ability to predict human performance on a homing task.

Both 3-D reconstruction and view-based homing have been advocated as models to explain navigation behavior in animals, including humans, and both have been implemented in robots. The idea of a reconstruction or "cognitive map" has a long history (Tolman, 1948), and it has been argued that this is instantiated in the hippocampus and surrounding cortex, including "place" and "grid" cells (Hafting, Fyhn, Molden, Moser, & Moser, 2005; Jacobs et al., 2013; O'Keefe & Nadel, 1978). The proposal is that sensory information from a variety of different sensory modalities is integrated in a common allocentric map. This requires that information from a range of senses is transformed into an allocentric frame despite starting in different coordinate frames (e.g., proprioceptive, vestibular, visual, or auditory; Andersen, Snyder, Bradley, & Xing, 1997; Burgess, 2006; Burgess, Jeffery, & O'Keefe, 1999; McNaughton, Battaglia, Jensen, Moser, & Moser, 2006; Mou, McNamara, Rump, & Xiao, 2006; Snyder, Grieve, Brotchie, & Andersen, 1998). Indeed, our ability to integrate information from several senses has often been cited as evidence of the brain's ability to build a *multimodal* cognitive map (Tcheang, Bülthoff, & Burgess, 2011) as has people's ability to take an appropriate novel short cut between two points (path integration; Schinazi, Nardi, Newcombe, Shipley, & Epstein, 2013).

Citation: Gootjes-Dreesbach, L., Pickup, L. C., Fitzgibbon, A. W., & Glennerster, A. (2017). Comparison of view-based and reconstruction-based models of human navigational strategy. *Journal of Vision*, *17*(9):11, 1–19, doi:10.1167/17.9.11.

doi: 10.1167/17.9.11 Received November 9, 2016; published August 16, 2017 ISSN 1534-7362 Copyright 2017 The Authors

On the other hand, there is an extensive literature supporting the use of view-based strategies for navigation, certainly in simple animals, such as ants and bees, but also in humans. A classic study by Cartwright and Collett (1983) showed that bees returning to a feeding site flew so that, when landmarks around the feeder were altered, they matched the retinal image as closely as possible to the image they had learned rather than flying to the correct 3-D location. Similar evidence exists for ants (Graham & Cheng, 2009; Graham & Collett, 2002; Lent, Graham, & Collett, 2010; Wehner & Räber, 1979). Mallot and colleagues (Franz, Schölkopf, Mallot, & Bülthoff, 1998b; Gillner & Mallot, 1998) have proposed that human navigation may be based on an internal representation linking actions with the sensory consequences of those actions. The representation was a "graph," i.e., a set of nodes connected by edges (the operations that connect the nodes). In this case, the nodes were views of the scene, and the edges were actions, such as rotation or translation, of the observer. A representation of this sort is more tolerant to inconsistencies in performance across tasks than an all-purpose allocentric model, and this appears to be true of humans, too. For example, adding junctions and turns increases people's estimate of the distance between two points (Sadalla & Magel, 1980; Sadalla & Staplin, 1980), depth judgments can be intransitive (Svarverud, Gilson, & Glennerster, 2012), judgments of the directions between visible points can be inconsistent with any 3-D interpretation (Koenderink, van Doorn, Kappers, & Lappin, 2002). An allocentric 3-D map and a graph of views may be two extreme cases on a spectrum. Chrastil and Warren (2014) have argued for an intermediate representation, a "labeled" graph, on which some distance and angle information is included to describe the separation of local features or locations, but there is no globally consistent map.

Computer vision and robotics have implemented both types of approach although the 3-D reconstruction solution is far more common. In 3-D reconstruction techniques, image features are tracked and their location computed in a world-based coordinate system as the camera moves. Modern systems do this for every pixel or even finer and in real time (2d3 Ltd., 2003; Hartley & Zisserman, 2004; Meilland & Comport, 2013; Newcombe & Davison, 2010; Whelan, Leutenegger, Salas-moreno, Glocker, & Davison, 2015). This differs from biological proposals in a number of ways, including the lack of intermediate coordinate frames between an image and the allocentric 3-D reconstruction, such as head-centered or other egocentric frames. There have also been robotic implementations of navigation that have avoided 3-D representations altogether (Ni, Kannan, Criminisi, & Winn, 2009; Zhu et al., 2016).

In this paper, we compare the ability of a view-based model (Pickup, Fitzgibbon, Gilson, & Glennerster, 2011) and a 3-D reconstruction model (Pickup, Fitzgibbon, & Glennerster, 2013) and predict human performance in a "homing" task. With no noise, both models predict that participants would return to the location that they viewed at the start, but once noise is introduced, the models predict different patterns of errors and a different dependence of the errors on scene structure. Throughout the paper, we refer to the true "home" location as the "goal" point and the position they actually return to as the "end" point. For the reconstruction-based model, errors are assumed to arise from Gaussian noise in the image centered on the true projection for any point. This results in errors in the reconstructed location of the scene points and hence errors in the homing task. For the view-based model, the important noise is assumed to be quite different. "Features" are calculated from one or more images, for example, the angle or disparity between two points. For more detail about the "features" we use, see "The view-based model" in Modeling details. In the modeling we describe in this paper, we assume that different noise is applied to each "feature." This is quite different from the type of noise we assume in the reconstruction-based model, and so the pattern of homing errors predicted by the models is also quite different.

In order to distinguish the models, we generated simple environments in which the spatial distribution of predicted errors would be different for the two models (even when the most likely location predicted by both models was the same, namely the true "goal" location). To do this, we showed a simple scene comprised of three vertical poles (see Figure 2) and varied the position of one of the poles relative to the other two. As Figure 1 shows, this can have a dramatic effect on the distribution of estimates of the "goal" location. These example trials were specifically collected to allow the varying spatial distribution of errors to be displayed easily and are not part of the main data set (see Methods for details).

Overview

We describe the homing task (Methods) and how some of the conditions were arranged to maximize differences between the predictions of the two models. In Experimental results, we show homing locations for the participants, and in Modeling details, we introduce the two models. We describe how the parameters are estimated to optimize the reconstruction- and view-based classes of models, respectively. In Model comparison, we compare the two models using the sparse-scene data. We argue that comparing the likelihood of the data under the two models is not the most robust way to distinguish between the models. Therefore, in addition to comparing the likelihoods of

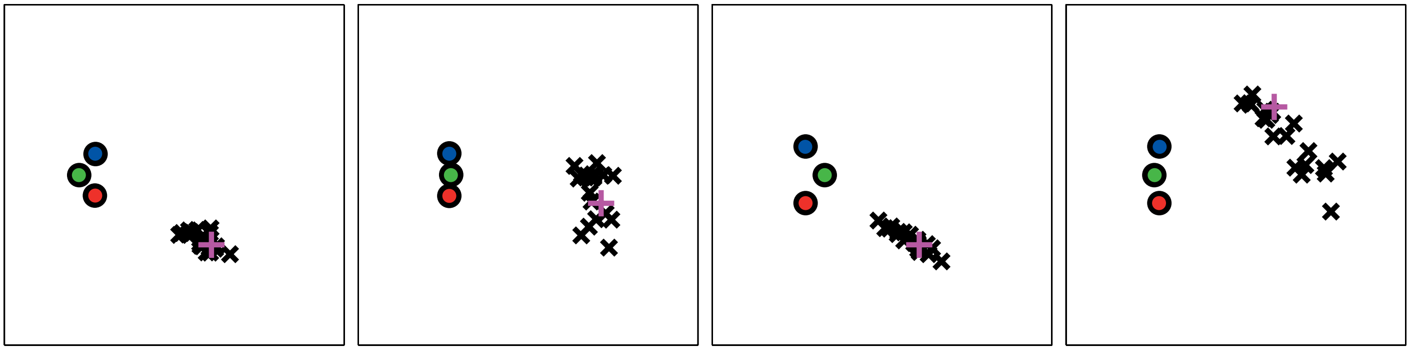

Figure 1. Illustration of different distributions of errors. Raw data from four "demonstration" conditions (see Introduction; data taken from Pickup et al., 2013). The colored dots (red, green, and blue; left side of each subplot) show the locations of the three vertical poles. The magenta "+" indicates the location to which participants were trying get ("goal point"). Crosses indicate the "end points," i.e., the actual locations at which participants reported having reached the goal point.

the two models, we sample from each and compare properties of participants' data to those of the samples drawn from the two models. The rich-cue data is less suitable for model comparison as it varies less with scene structure (see section on Navigation in a rich-cue environment).

Methods

Participants

Some of the data were gathered in the Department of Physiology, Anatomy and Genetics in Oxford (four participants, S1–S4) and some were collected in the School of Psychology and Clinical Language Sciences in Reading (four participants, S5–S8). All participants had normal or corrected-to-normal vision (6/6 or better) as well as normal stereopsis (60 arcsec or better in the TNO). Participants were naïve to the experimental purpose and design with the exception of S6 who was an author (LGD). The study received approval of the research ethics committees of both the University of Oxford and the University of Reading.

Virtual reality display

For both parts of the experiment, an nVisor SX111 (NVIS, Reston, VA) head-mounted display was used. This headset has a wide field of view (102°) and a large binocular overlap (50°), which means that, typically, all three landmarks in the sparse environment could be viewed simultaneously in stereo from the goal location. The headset was fitted with retro-reflective markers mounted on a rigid wire frame that could be tracked by infrared Vicon cameras. The Vicon tracking system (Tracker Version 2, Vicon, Oxford, UK) reported the six degrees of freedom head position and orientation to the graphics PC at 240 Hz. The stimuli were rendered on a desktop PC running Linux and proprietary OpenGL software. The system had a total latency of less than 40 ms. Further details of the setup and its calibration can be found in Gilson, Fitzgibbon, and Glennerster (2011). The experimental space (3.5 m by 3.5 m) was tracked using nine Vicon cameras in the Oxford setup (MX3 and T20s) and in Reading (3.1 m by 3.5 m) using 14 cameras (MX3s and T20s).

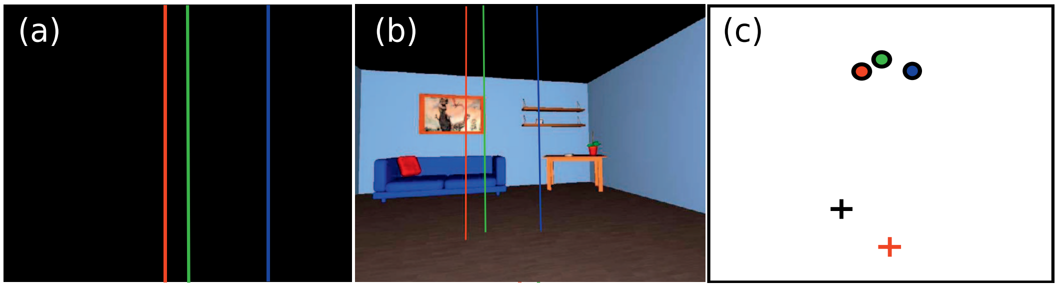

Figure 2. An example trial of the homing task used for modeling. View of the three-pole stimuli in (a) sparse-cue and (b) rich-cue environments. (c) Top-down view of the virtual room showing the three colored poles; the black "+" shows the goal point. Participants had to return to this location after being transported to a new location, marked by the red "+."

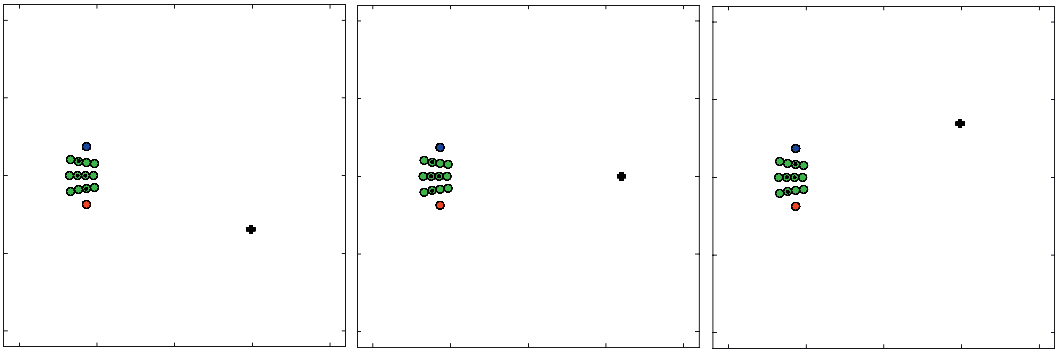

Figure 3. Stimulus configurations. The 36 pole configurations for interval 1 of the experiment. Each of the three boxes represents a 4 × 4 m area of the virtual reality space. The red, green, and blue dots represent pole positions, and the black circles represent the centers of the interval 1 viewing zones. The 12 green pole locations marked with a black dot were included twice: once as a "normal" condition and once with the green pole moving between intervals 1 and 2 by 4.0–4.5 cm. This makes 48 conditions in all. The marks on the side of the plot indicate a distance of 1 m.

Stimuli

Two types of scenes were shown in virtual reality, a sparse scene consisting of only three long poles and a rich scene containing the same poles set in a virtual room with furniture and familiar objects (see Figure 2). The three thin, vertical poles were always one pixel wide on the screens independent of viewing distance. In the sparse condition, the relative positions of the poles could not be determined by looking up or down as no ground plane was shown. The background was black. The red and blue poles always appeared in the same position in the virtual room, and the green pole changed position between trials (Pickup et al., 2011; Pickup et al., 2013) and sometimes between intervals (see "Optimisation for model comparison" in Methods). The full set of green pole positions are shown in Figure 3. The stimuli were viewed from within three possible 20 cm × 80 cm viewing zones, in which participants could freely move to gain motion parallax. The angle between the red and blue poles as seen from the center points of each of the viewing zones was constant at 18°.

The rich scene added a ground plane, walls, and objects. The objects were a couch, picture, table, and bookshelf positioned behind the poles and were visible from all three possible viewing zones. The ground plane had a repeated wooden texture; all other objects were rendered without texture. The light source was a point light attached to the participant's head location. No shadows were enabled, so movement of the light source was undetectable in the sparse scene. In the rich scene, movement of the light source did cause a slight change in the luminosity of surfaces, mainly the walls, but participants focused on the three poles in the center of the room.

Procedure

Each trial followed a two-interval pattern. At the beginning of each trial, participants guided themselves to a specific location in the virtual room. A yellow box drawn in one eye's image showed a plan view of the viewing zone that participants had to enter to start the trial, and a red cross showed their own location. This meant that they could walk to the viewing zone while still wearing the headset. When their cyclopean point was within the zone, the image changed, and they saw the three poles in front of them.

In this first interval, the poles were visible provided that participants moved within the 20 cm × 80 cm viewing box (otherwise the scene went blank). In the virtual room, this box was centered at one of the three starting locations shown in Figure 3 with the long axis of the viewing box at right angles to the line joining the center of the viewing box to the point midway between the red and the blue poles. The size of the viewing box was chosen to allow for lateral movement. Participants were allowed to view the stimulus for as long as they wished but generally did so for a few seconds. When participants were happy that they had remembered their location with respect to the poles, they pressed a button on a hand-held controller.

On the button press, the participant's current location within the viewing box was marked as the "goal point" for the trial (the position to which they would then try to return). In contrast to the data in Figure 1, participants were free to choose the goal point themselves within the viewing zone. After a blank interstimulus interval of 0.5 s, they were instantaneously transported to another location (by moving the virtual room). Hence, although their view of the poles changed between intervals, participants' physical location in the real world remained the same. Note that, for the data shown in Figure 1, an extra interval was inserted into the protocol (interval 1b) as the participant was shown a fixed view of the scene that was repeated across multiple trials (Pickup et al., 2013), but in the current paper, we used two intervals, and hence, there was a slightly different goal point on every trial.

On some trials, an additional small movement of the green pole was introduced between intervals (described in more detail in "Optimisation for model comparison" in Methods). The participants' task in interval 2 was to navigate back to the "goal" location and to press a button when they believed they had reached that point. This location was then marked as the "end point" for that trial. If the task were carried out perfectly, the coordinates of this end point would exactly match the coordinates of the goal point when described in the original frame of reference.

In the sparse-cue condition, each participant completed 336 trials in total, corresponding to seven complete repetitions of the 48 conditions shown in Figure 3, which were spread out over several experimental sessions. Twelve out of the 48 conditions were repeated conditions but with a small change in position of the green pole between intervals as described above and in "Optimisation for model comparison" in Methods. Participants S5–S8 also completed an additional full set of 336 trials using the rich-cue environment. As in the sparse-cue condition, the scene remained the same between intervals 1 and 2 (other than any small movement of the green pole). Over the course of the experimental sessions, participants alternated between rich-cue and sparse-cue blocks.

Optimization for model comparison

The purpose of changing the position of the green pole between intervals on some trials was to elicit navigational errors that were maximally discriminative in relation to the two models that we examined (analogous to the picking of cue values in an adaptive psychometric procedure to provide the most informative response on each trial, e.g., Watson & Pelli, 1983).

To find the pole configuration and specific shift in the green pole position that would have the largest effect on model predictions, a set of possible pole locations was created, and the green pole was shifted by 4.0–4.5 cm either in depth or laterally between intervals 1 and 2. These shift magnitudes were chosen to be small enough that participants in a pilot study did not report noticing that there had been any shift during an informal interview afterward but large enough that the differences in model predictions were as pronounced as possible. For each possible condition, we then evaluated the predicted end point distributions under the view-based model and the reconstruction-based model using typical parameterizations in each case that gave predictions about the likely locations of participants' end points (e.g., as shown in Figure 11). This allowed us to select the 12 cases with the biggest difference between view-based and reconstruction-based predictions. Our measure of difference was the Kullback–Leibler divergence (Kullback & Leibler, 1951) between the two distributions of predicted end points.

After the main experiment had been completed, we tested four participants (S1–S4, tested in Oxford) on 96 trials (the final two repetitions of the experiment) using a forced-choice paradigm to determine whether the green pole had moved between intervals. Participants carried out the task exactly as before, navigating back to the goal location, but were asked to press one button to indicate that the green pole had moved and another if it had not (movement on 50% of trials).

Experimental results

Sparse-cue scene

Figure 4 shows data from the sparse-cue condition. It plots in plan view, for all participants, the location of the poles, including all the possible locations of the green pole (shown in more detail in Figure 3), the locations of the goal points (black), and the locations of the "end" points (red and blue crosses for conditions in which the green pole did or did not move between intervals, respectively). As explained in the Methods, the goal point was determined by the location at which the participant left the viewing zone in interval 1, so it was slightly different on every trial—hence, the spread of the black dots. We did not analyze the distribution of goal points chosen by participants, and this distribution may well not have been random within the viewing zone, e.g., if participants found some configurations easier to memorize than others (the raw data is available in the Supplementary Material). This means that Figure 4 cannot illustrate the spread of end points and how it is affected by the relative landmark positions in the same way as Figure 1. Instead, we leave the quantitative analysis of the end point data to the comparison of models, in which every end point from every trial is related to the prediction of the two models tailor-made for that particular configuration of poles and that particular goal location, which is unique for each trial. Nevertheless, even by plotting all the homing errors for one condition relative to the correct location, as shown in Figure 5, there is a strong indication of systematic biases in the distribution of homing errors. When the data from all trials are combined together as in Figure 4, some general observations can be made. For example, certain observers tend to move system-

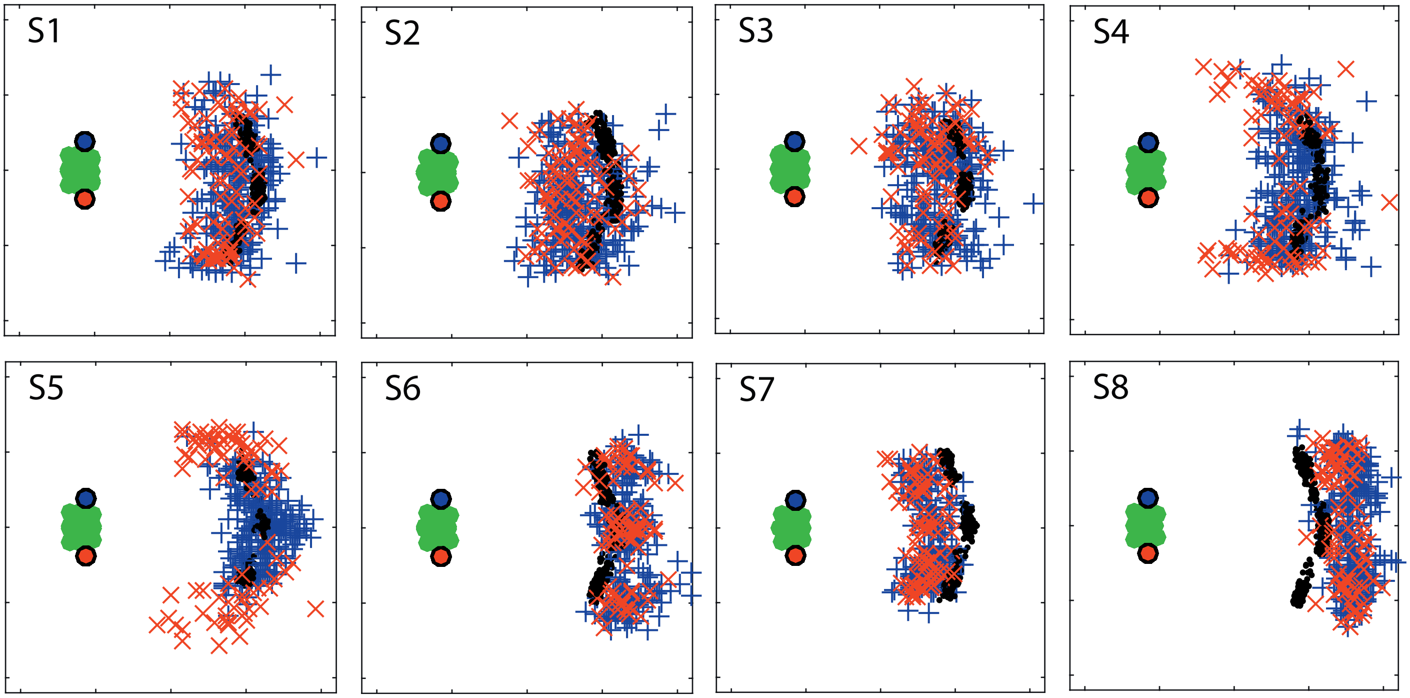

Figure 4. Homing in a sparse-cue environment. Blue "+" symbols show the "end points" in the green-pole-static conditions, red "×" symbols show end points from the green-pole-moving conditions, and black dots show the goal points. Note that all conditions are plotted together, accounting for the three clusters of goal points, which correspond to the three viewing zone locations, and also for all green pole locations.

atically closer to the poles (or, for other observers, further from the poles) than the true goal location.

Of the four participants we tested to see whether they could notice the green pole moving, two had a d' of less than one (0.29 and 0.58), one had a d' of 1.46, and one had an infinite d', i.e., this participant (S4) was correct on all 96 trials. Unlike the participants S1–S4, participants S5–S8 were simply asked about whether they saw the green pole move between intervals at the very end of the experiment, and participants S6, S7, and S8 said they did not. S5 noticed the movement, and this participant shows a distinctly different navigation behavior in the green-pole-moving condition than in the green-pole-static conditions. In "Model comparison results" in Model comparison, we show a reanalysis of the data from S4 and S5, excluding all the green-pole-moving trials. The reanalysis shows that the conclusions about model comparison were not affected.

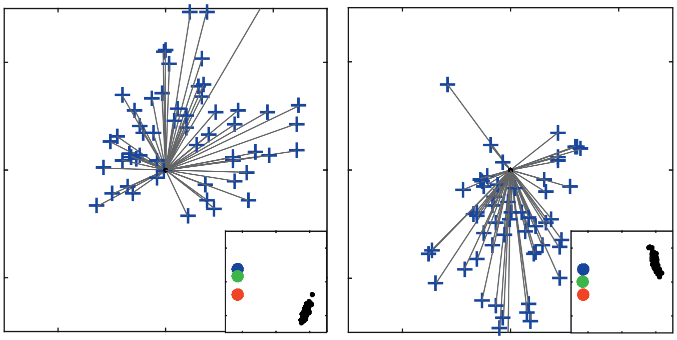

Figure 5. Systematic biases in homing errors. For two example conditions from the sparse-cue experiment (both with a static green pole position between intervals), end point locations are plotted relative to the goal point (shown at the center of the plot). The plan view of pole position and original goal point spread are shown in the insets. Ticks are separated by 1 m, and the orientation of the plot is the same as the insets.

Rich-cue scene

Figure 6 shows the data for the rich-cue scene. Comparing Figures 6 and 4, the spread of end points in Figure 6 appears smaller. Confirming this, permutation tests showed that the root-mean-square error of all four participants who carried out both experiments was significantly lower in the rich-cue versus the sparse-cue conditions (all $ps < 0.001$).

Modeling details

In this section, we describe details of the reconstruction-based and view-based models of homing; then in the next section on Model comparison, we compare

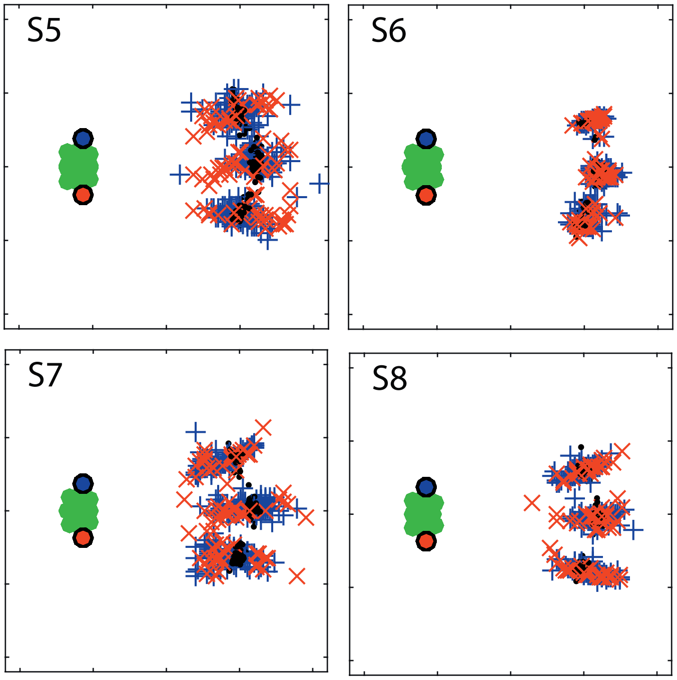

Figure 6. Homing in a rich-cue environment. End points in the green-pole-static (blue "+") and green-pole-moving (red "×") conditions. The goal points are shown as black dots. This experiment was carried out by four observers as shown. As in Figure 4, data for all the conditions are plotted together.

the data from Experimental results with the predictions of both models.

The reconstruction-based model

The reconstruction-based model creates a metric reconstruction of the world given two "cameras" or viewpoints at known locations in the "viewing zone" from which participants viewed the scene. It assumes that the image location of rays from objects in the scene (the three poles) is known up to some degree of precision (σ). σ as well as the separation between the cameras were free parameters when Pickup et al. (2013) optimized this reconstruction model using their data (but the rotation of the cameras was constrained so that they always faced the green pole). They found that two cameras placed at the maximum permitted separation (80 cm) best explained the navigation data in that paper (and a σ of 0.0128 times the assumed focal length of the camera, see Pickup et al., 2013, for details). The logic of allowing such a wide baseline is that participants could move from side to side in the viewing zone up to a maximum of 80 cm, and this provided useful motion parallax, so the models should have access to this information too. The image error associated with each feature (σ) results in a spread in the estimate of the location of the corresponding pole as shown in Figure 7.

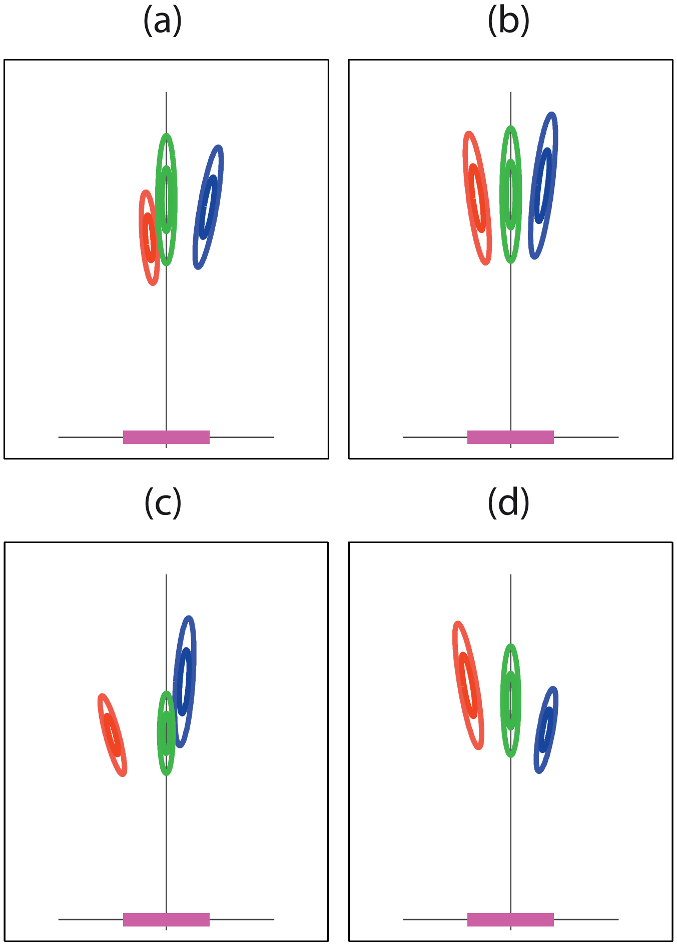

Figure 7. Examples from the reconstruction-based model. The pole configurations match those in Figure 1. Each plot shows a model built up in egocentric coordinates, in which the "forward" direction is taken as being toward the green pole. The ellipses illustrate the uncertainty (covariance of the model) around each pole estimate. The magenta strip shows the 0.8-m-wide region over which the set of views used for the reconstruction is taken. This matches the width of the start zone in interval 1 of the homing task. Reproduced from Pickup et al. (2013) under the terms of the creative commons attribution license.

In the reconstruction model used in this paper, these parameters (camera location and σ) were fixed as above, but one parameter was allowed to vary to best fit the data for each participant (λ), defined in Equation 1 below. So, the steps required to determine the likelihood of a given data set under the reconstruction model were as follows:

1. A metric reconstruction was carried out from a series of views in interval 1. The views were taken from a line within the viewing zone, orthogonal to

a line from the center of the viewing zone to the center point between the two outer poles.

2. From these views, we obtained a Gaussian representation of each pole's location in egocentric 2-D coordinates.
3. The difference between the current location's egocentric pole representation and the goal point's egocentric representation was found using the Bhattacharyya distance *d*. Details of how this was calculated are given in Pickup et al. (2013), and Figure 7 provides an illustration: If panels a and b were overlaid, there would be quite a large overlap between ellipses of the same color whereas the overlap between panels a and c would be much smaller and the Bhattacharyya distance correspondingly larger. Bhattacharyya distance is a standard measure of distance between probability distributions. Each pole was treated independently. An alternative to this approach that takes the relationship between poles into consideration is described in Pickup et al. (2013), but it bears some similarities to the view-based model, and we opted to test the extremes of the reconstruction- to view-based spectrum.
4. The (un-normalized) *likelihood* of the location matching the goal point was taken to be

$$L \propto \exp\{-\lambda d\}, \quad (1)$$

where λ is the additional weighting parameter, which we allowed to vary across participants. It determines how quickly the likelihood should decay with the magnitude of the Bhattacharyya distance, *d* (Pickup et al., 2013).

5. To turn the likelihoods into full probabilities that can be compared across trials, we ensured that the integral of the likelihood function across the (x, y) plane is unity. For this, we estimated the integral

$$Z = \int_x \int_y \exp\{\lambda d(x,y)\} dxdy \quad (2)$$

where $d(x, y)$ is the Bhattacharyya distance between the goal point's egocentric representation and the egocentric representation built at the point (x, y).

6. The (normalized) probability of an end point under the model is therefore

$$L = \frac{1}{Z}\exp\{-\lambda d\}. \quad (3)$$

7. The total likelihood of the data set was found by multiplying together the normalized probabilities for each data point.

Examples of the reconstruction-based end point likelihood maps are shown on the left hand side of Figure 11. For each participant, we found the value of λ that maximized the total likelihood of that participant's training data. Different values for λ indicate individual differences in sensitivity to divergence between the representations being compared. One participant may be very sensitive to the change in representations, and another may not. For the eight participants in these experiments, the values of λ were 0.195, 0.102, 0.120, 0.167 (0.208 in the reanalysis described in "Model comparison results" in Model comparison), 0.4644 (0.476 in the reanalysis), 0.241, 0.118, and 0.144, respectively.

The view-based model

A view-based model is defined using image "features" based on the visual angle between two poles or the relative disparity between them. These image measurements are assumed to be made at the "home" location and then compared with the same measurements at the current location until the best match is achieved. There are potentially a very large number of measures that could be used as the basis of these features even for a scene as sparse as the one we used with three poles, for example, the angular separation between any pair of poles, the ratio between angular separations, or the relative disparity between any pair of poles. Disparity was calculated assuming that observers faced the green pole using a fixed interocular distance (7.9 cm for the participants collected in Oxford and 7.4 cm for those collected in Reading, based on the calibration of the head-mounted displays). Pickup et al. (2011) tested a large combination of such measures and identified two that resulted in the features best able to account for the homing data in that paper. In this paper, we adopted these previously identified features, just as we used the previously identified parameters in the reconstruction model ("The reconstruction-based model" in Modeling details). The features were based on (a) the angle between the outer poles (red and blue) and (b) the disparity between the green pole and its nearest neighbor divided by the angle between the green pole and its nearest neighbor. Specifically, the features we used are the difference between the above quantities viewed from the goal and end points, calculated as

$$f_{\mathrm{A}} = \frac{\phi_\gamma^G - \phi_\gamma^E}{\phi_\gamma^G}, \quad (4)$$

$$f_{\mathrm{B}} = \frac{\delta_\alpha^G}{\phi_\alpha^G} - \frac{\delta_\alpha^E}{\phi_\alpha^E}. \quad (5)$$

where the superscript "G" means "viewed from the goal location," and "E" means "viewed from the participant's end point."

ϕ_γ and ϕ_α refer to the largest and smallest angles between pairs of poles:

$$\phi_\alpha = \min(\theta_{rg}, \theta_{gb}, \theta_{rb}), \quad (6)$$

$$\phi_\gamma = \max(\theta_{rg}, \theta_{gb}, \theta_{rb}). \quad (7)$$

where the three monocular measurements available from a view of our three-pole stimuli are θ_{rg}, θ_{gb}, and θ_{rb}, which are the three visual angles between the poles (red–green, green–blue, and red–blue, respectively). The relative magnitudes of angles in the scene is important because changes in a small angle have much more of an impact on performance than the same changes in a wide angle. Finally, δ_α refers to the disparity between the green pole and its nearest neighbor (see Pickup et al., 2011).

Having fixed the features on the basis of previously published data, the only adjustment of the view-based model per participant was to find the best-fitting values of the standard deviation for each feature. Figure 8 shows the two features in the sparse-cue data set gathered from each participant as well as the 2-D Gaussian distributions that resulted from the training data. The mean and covariance of this Gaussian are a full description of the model. This allows for interaction between the two features, which is near zero for all participants.

Most participants show slight biases, i.e., the peak of the distribution does not coincide with the origin, especially in the proportional γ error (x-axis). This can also be observed in Figures 4 and 6, in which responses are, on the whole, either slightly closer or further away from the three poles than the goal point, which can be accounted for in this model by allowing the mean proportional error in γ (feature f_A) to be slightly higher than zero.

As with the reconstruction-based models, these view-based models have to be normalized, and it is not sufficient that the Gaussian in feature space is normalized because the transform between feature space and (x, y) space is not area-preserving. As before, we find the *total likelihood* of the data under the view-based model by multiplying together the probabilities of all the individual end points recorded for a given participant. An example of the view-based model results being transformed back into a room coordinate frame is shown on the right hand side of Figure 11.

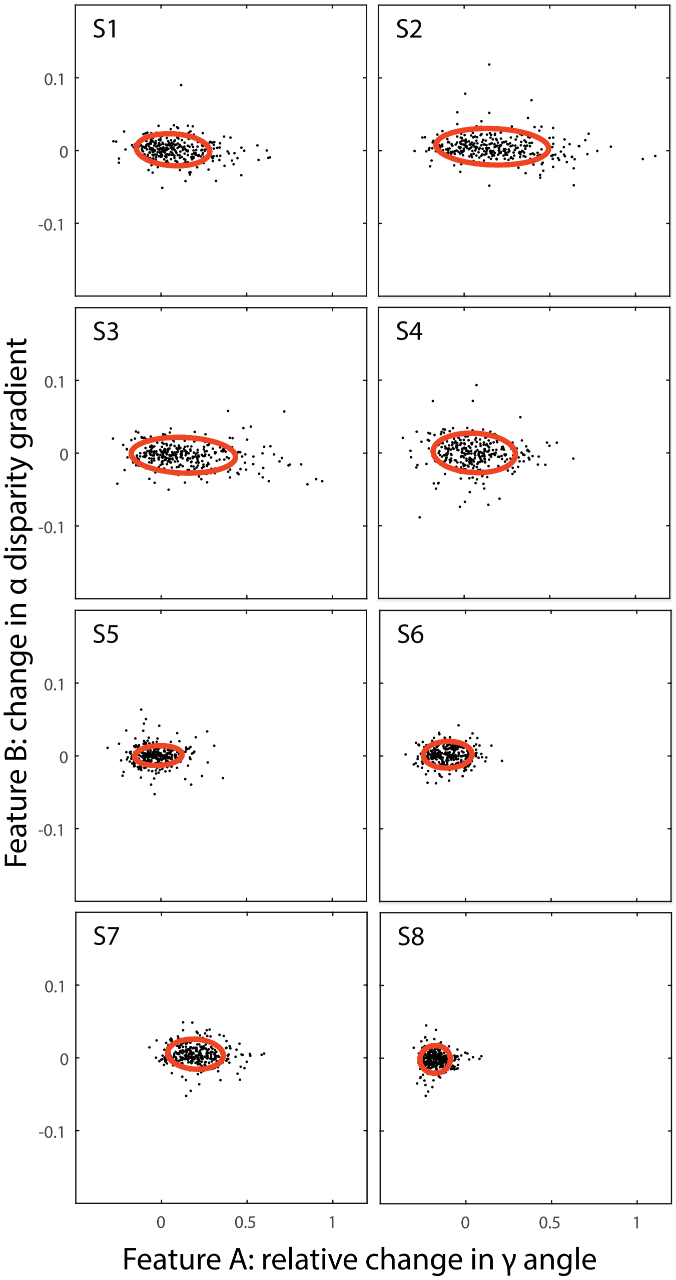

Figure 8. Data from the sparse-cue conditions plotted in feature error space. The full sets of data collected from the eight participants are shown here projected into our 2-D feature space. In this case, the horizontal axis feature is the proportional error in the red–blue angle, γ, and the vertical axis feature is the disparity gradient for the pair of poles spanned by the smallest visual angle (Equation 5). Both of these features are dimensionless as they are ratios. Red ellipses indicate the 2-D Gaussians fitted to the training subset of the point cloud (with the ellipse indicating one standard deviation).

Model comparison

In this section, we describe the methods we used to compare the two models, and then, in "Model comparison results" in Model comparison, we show the comparison. An illustration of the principle underlying the model comparison is shown in Figure 9, which replots the data from Figure 1 but now showing the predictions of the two types of model. As we pointed out in the Introduction, these data were collected

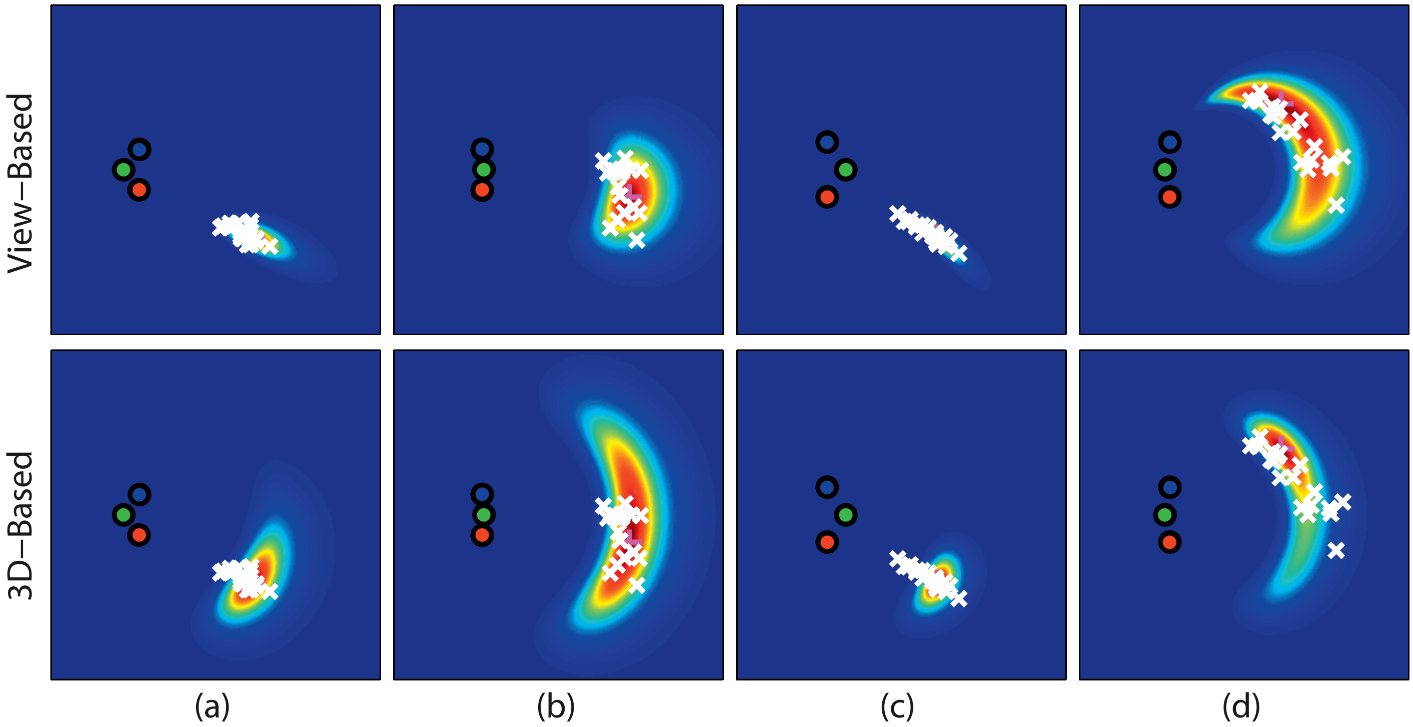

Figure 9. Illustration of model predictions. A comparison between view-based and reconstruction-based models of homing. Raw data from the four "demonstration" conditions shown in Figure 1. The colored dots again show the locations of the three vertical poles, the white markers participants' end positions, and the magenta "+" indicates the true goal location. Top row: Predictions of end point location likelihood under a view-based model. Bottom row: Predictions of end point location likelihood under a reconstruction-based model. Columns a through d show four different configurations of the poles and goal points, which led to very different patterns of errors by the participant.

separately from the main data in this paper and are for illustration only (Pickup et al., 2013), but they show, for example, how the elongated distribution of end points in all four panels match the shape of the model predictions of the view-based model better than the 3-D reconstruction-based model. One might think that all that is required is to compute the ratio of likelihoods of the data under each model (Bayes factor). For each participant, we plot likelihoods for each model but also demonstrate how sampling from each model can provide a more detailed picture of the differences between the models and how the data compare to samples taken from each model. This constitutes a principled method for determining which model should be preferred.

Methods for model comparison

When considering all the data from one participant, the likelihood of the whole set of data under both models can be compared to the likelihood of a random sample from both of the models (in which a "sample" consists of the same number of "test" conditions that the participant carried out, i.e., 84). Measuring a test statistic of data under two models and comparing it to samples drawn from both models is far more informative that simply comparing the likelihood of the data under each model. Of course, we can do the random sampling many times over and produce a distribution of likelihoods of the samples under both models. Then we can see whether the experimental data plausibly come from one or the other of the sampling distributions.

Figure 10 illustrates this point for a very simple case of two "models" that are simply 1-D Gaussian distributions with a different mean and standard deviation. The likelihood of a sample under the red Gaussian "model" is similar whether the sample was drawn from the red probability density function (PDF) or from the blue PDF. One can see that this might happen from looking at the left panel in Figure 10. The values of x would be quite different for the two samples, but the height of the red curve over all the samples (total likelihood) could well be similar. This intuition is confirmed in the right panel in Figure 10, which shows on the y-axis that the likelihood of samples under the red Gaussian model is very similar whether the samples originate from the red Gaussian model (red dots) or from the blue Gaussian model (blue dots). Of course, the reverse is not true. Sampling from the red distribution yields a very large number of

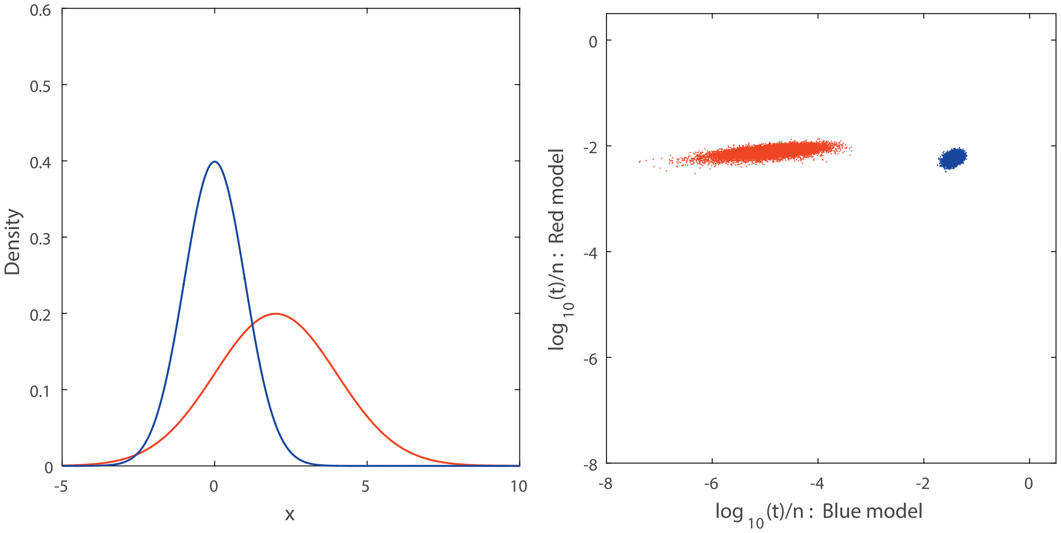

Figure 10. Simple example of model comparison. The left panel shows two normal distributions, which define the two models we want to compare in this example. We sample from each model in turn to create simulated data sets and evaluate the likelihood of the samples under each model (t_{data}). Results of the comparison are plotted in the right panel, showing the mean log likelihood of data sampled from the red and blue models. The red model has similar likelihoods for data generated using either model whereas the likelihoods under the blue model are quite distinct and allow the underlying distribution to be identified.

samples that are extremely unlikely under the blue model, pulling down the total likelihoods and making the cloud of samples from the red PDF quite different from those drawn from the blue PDF when assessed under the blue model. The view-based and reconstruction models that we examine show a similar pattern. For each model, we create a reference distribution derived from simulated data sets that are sampled directly from the model predictions. This distribution tells us what kind of likelihoods we would expect if the model used to create the simulated data sets was the "true" underlying model. Hence, we can say whether the likelihood of the experimental data under that model is "typical" or "untypical" in relation to the random samples or, at least, whether it is more typical of one model over another.

For our experimental data, Figure 11 shows the sampling process used to generate the simulated data sets. The two plots on the top row show the predictions of each model in two example conditions. The plots on the bottom row show 1,000 sampled end points from those distributions. The model gives likelihood maps for each of the goal point–pole position combinations, and by taking samples of hypothetical end points from each of the test set condition likelihood maps, we can combine samples from several conditions into a full simulated data set.

For each participant, we trained both a reconstruction-based model and a view-based model using the 252 green-pole-static conditions (sparse-cue, seven repetitions of 36 conditions). For the reconstruction-based model, this training determined λ (the one free parameter in that model), and for the view-based model, the training determined the mean and covariance matrix of the errors in feature space (which specify the view-based model). The remaining 84 green-pole-moving trials (seven repetitions of 12 different conditions) were used as the "test set" over which the sampling distribution was evaluated, yielding the mean log likelihood of the test data as a single scalar value, which we call "t_{data}." These trials had been chosen in advance as being especially discriminative between the two models (see "Optimisation for model comparison" in Methods). We also show additional analyses for participants S4 and S5 because of concerns that they may have noticed the difference between green-pole-moving and green-pole-static trials and changed their strategy as a result (discussed in "Optimisation for model comparison" in Methods). In this case, we avoided the green-pole-moving trials in the testing phase: The model was trained on two thirds of the green-pole-static data (random split) and tested on the other third (Figure 13).

In each simulated experiment, we drew one sample from one of the models for each of the 84 conditions that make up a complete test set, i.e., we had 84 "end point" locations. For each of these sets of 84 simulated trials, we could calculate the test statistic (t_{data}) under each of the models, i.e., both under the model that was used to generate the simulated data set *and* under the rival model. These two values of t_{data} give rise to a single point plotted at the relevant coordinate in Figure 12, color-coded according to the model from which the sample was drawn. We repeated this many times (10^4 independent samples) for each model under consideration and for each participant.

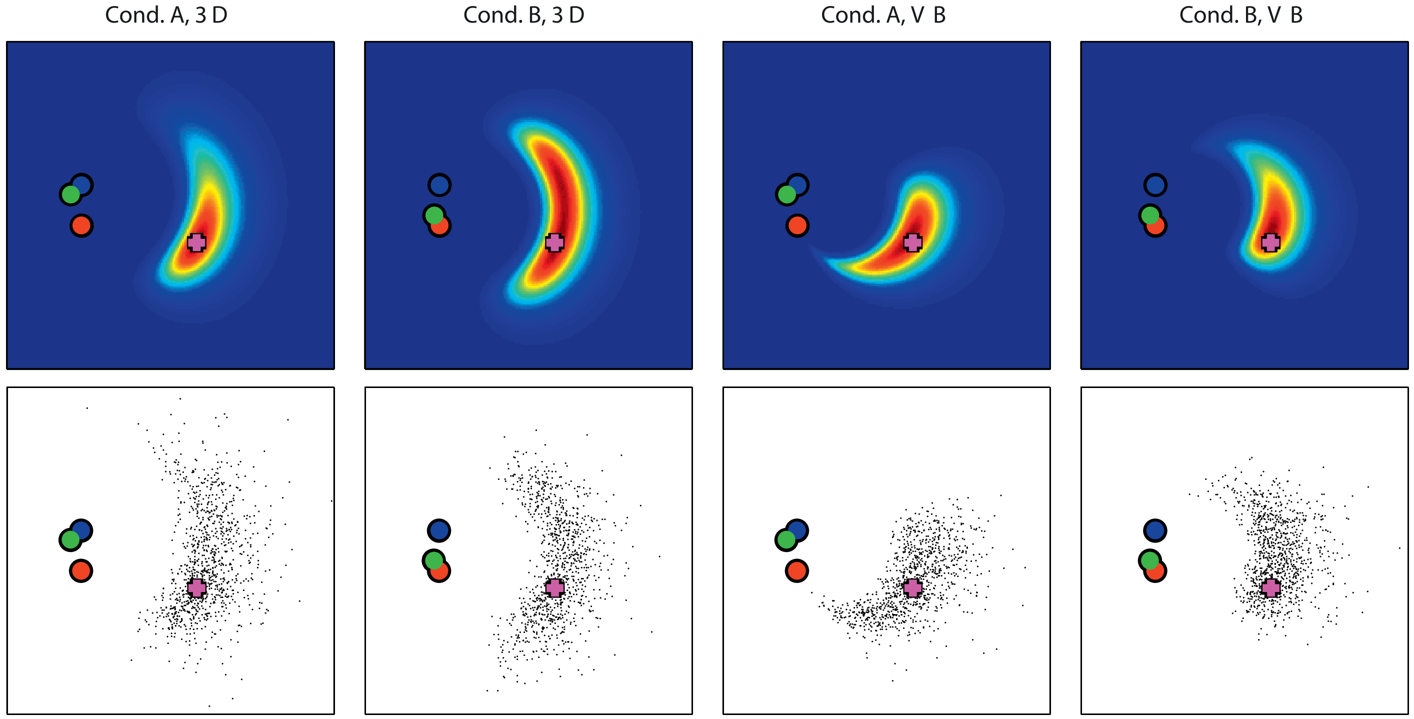

Figure 11. Sampling from the models. Examples of sampling from the end point distributions described by one reconstruction-based model (left half of the figure) and one view-based model (right half of the figure). There are two green pole configurations, labeled "A" and "B." The top row shows the end point probability distributions according to the two models, and the bottom row shows samples from these distributions (1,000 points are shown per plot). Each box above represents a 5 × 5 m area in virtual space. The red, green, and blue dots represent pole locations in interval 1, and the magenta marker represents the "goal point" location, i.e., the final viewing location for interval 1 of the experiment. Sampling was carried out using a simple rejection sampler for both models.

Model comparison results

Figure 12 shows the results of the model comparison for all eight participants in the sparse-cue condition. As in the right hand panel of Figure 10, each axis plots the mean log likelihood, $\log(t)/n$, under the view-based model and the reconstruction model of samples taken from each model. The red cloud of dots shows likelihoods of simulated data drawn from the reconstruction model, blue dots shows the same for the view-based model. Now we can see where the data fall on this plot. The magenta dot shows the likelihoods of the actual data for each participant under both models.

For all participants except S5, the experimental data has a higher total likelihood under the view-based model than under the reconstruction-based model. Also, inspecting the marginal distributions in Figure 12, it is clear that the experimental data (shown as magenta circles) lie at the extremes of the distributions of likelihoods for the reconstruction model (shown in red) and, in fact, the experimental data are significantly different from the samples drawn from the reconstruction model for all participants (all $ps < 0.05$). Using the same criteria, the data are significantly different from the simulated view-based data sets for only three out of eight participants (S4, S5, and S6). This suggests that the view-based model may be preferable to the reconstruction-based one. However, as discussed above, the real differences between the models emerge when the likelihoods of the data under both models are considered together. In this case, it is clear from inspection of Figure 12 that for all participants except S5, the experimental data (magenta circles) are more similar to the samples from the view-based model (blue dots) than they are to those from the reconstruction model (red dots).

S4 and S5 were discussed previously (Experimental results) as they may have detected that the green pole sometimes moved and changed their strategy on these trials. Figure 13 shows a reanalysis for these participants using only the green-pole-static trials, i.e., avoiding any trials in which the green pole moved. For participant S5 in particular, the data are now much more similar to the samples drawn from the view-based model.

A quantitative version of this informal inspection is to grow a circle out from the experimental data for each participant (magenta dot in Figure 12) and to collect a cumulative count of the number of samples from each model (blue or red dots) that fall within the circle as the radius increases. This is illustrated in Figure 14, which

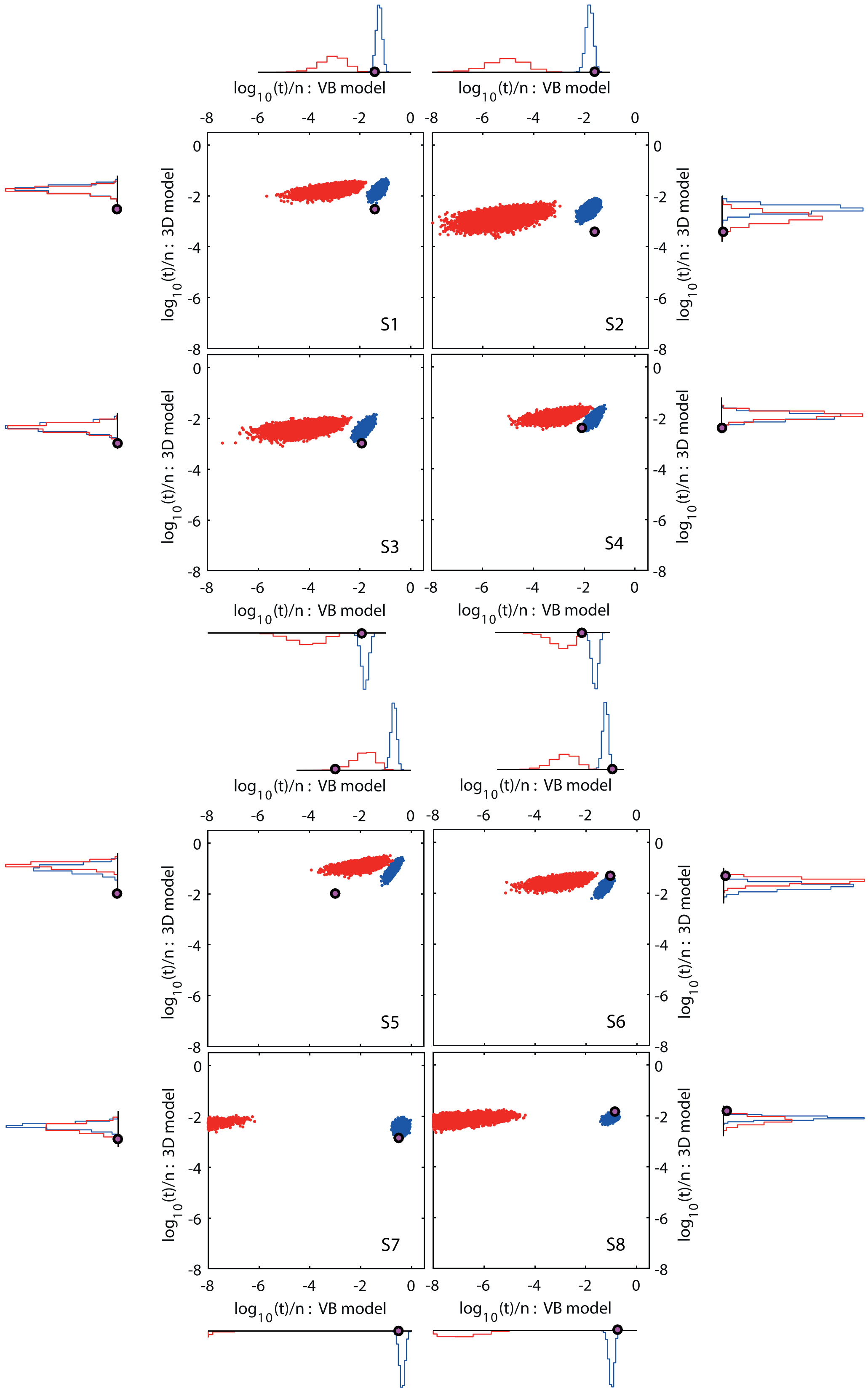

Figure 12. Results of model comparison in the sparse-cue conditions. The magenta marker shows the likelihood (t_{data}) of the data under the two models for each participant. The red dots show the same but for samples drawn from the reconstruction model; similarly, the blue dots show the distribution we would expect to see under the view-based model. The marginal histograms show the distribution of likelihoods of each type of simulated data set under each model.

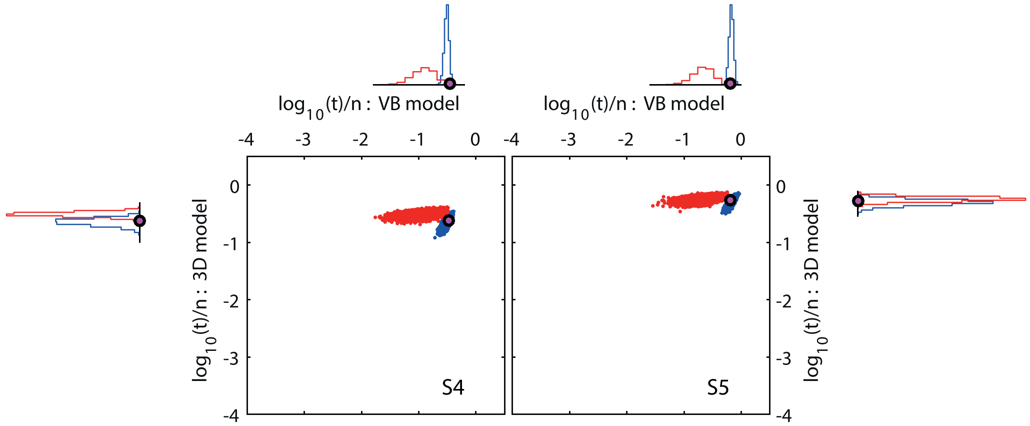

Figure 13. Reanalysis for two participants, excluding green-pole-moving trials. The possibility that participants S4 and S5 may have used a different strategy on green-pole-moving trials prompted a reanalysis of their data excluding these trials. Two thirds of the green-pole-static data were used for training and the remaining third for testing. Otherwise, the plots are the same as for Figure 12.

shows the number of t_{data} points from simulated data sets for both models that fall within a certain radius around the real data. Data from all participants are combined in this plot. There are always more samples from view-based models for every radius in the combined plot. Specifically, we calculate the ratio of reconstruction-based simulations relative to all data sets that fall within a given radius; we then increase the radius up to the point at which it includes either all of the samples from one model or all of the samples from the other model. Before that point, the ratio never exceeds 0.01 for all participants except S4 ($ratio =$ 0.4112), S5 ($ratio = 1$), and S6 ($ratio = 0.1592$). After reanalysis, this ratio drops to 0.384 for S4 and to 0.169 for S5 (note that Figure 14 shows the reanalyzed data). If the two models provided equally good descriptions of the data, one would expect (on average) a similar number of samples from each model to lie within the tested region, whatever its radius. If, instead, almost all the samples turn out to come from one model rather than the other, as we have found here (Figure 14), this is strong evidence to prefer that model over its rival.

We consider the implication of these results in more detail in the Discussion. For now, it is worth looking back at the biases in end points shown in Figure 4 and noting that one of the reasons that the view-based model performs better is that it is able to model this tendency of participants to systematically overestimate or underestimate the relative distance to the landmarks (Figure 4).

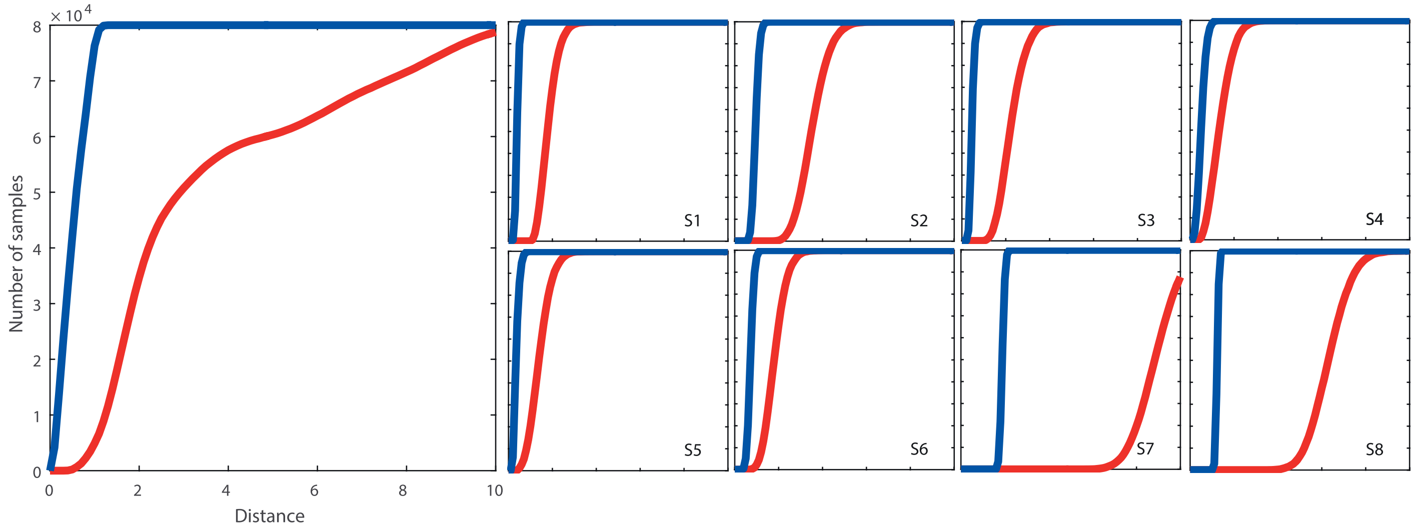

Figure 14. Number of model samples in the neighborhood of the data. This shows the cumulative number of samples (across all participants) that fall within a given radius of the real data likelihoods as shown in Figures 12 and 13. The blue and red curves show this for samples from the view-based and reconstruction-based models (blue and red points in previous figures, respectively).

Navigation in a rich-cue environment

In the sparse-room condition, a limited number of features were available to participants, which made modeling tractable. Although this method uses an environment that would not be encountered in a real-life situation, it allowed us to draw distinctions between the two types of models.

Adding a large number of additional points makes this approach unfeasible. For the sparse environment, we originally investigated 19 possible view-based features in describing the spatial relationship between the three available landmarks (Pickup et al., 2011). The number of possible combinations of features increases exponentially in a rich-cue environment, making it unfeasible to systematically determine which ones give rise to models that best fit the data. Additional complications arise from the fact that visible features are now spread through three dimensions rather than two.

Nevertheless, it is still possible to demonstrate the effect that extra scene features have, in general, on both view-based and 3-D models. Figure 15 shows how exactly the same view-based model that we used for the sparse-cue data gives rise to a much smaller range of predicted navigation errors when the range of depths of features is increased. This shrinkage occurs without altering the model and even when the number of features remains the same. A similar diminution of the range of predicted errors is evident for the 3-D reconstruction model when more features are added to the scene (Figure 15, bottom row). The parameters of the model are the same as for the sparse-cue scene, but now the likelihood of a location matching the goal point depends on the Bhattacharyya distance between the Gaussian distributions of more features than before, and the features are more widely distributed in space. This shrinkage in the predicted range of navigation errors agrees qualitatively with the data on rich-cue environments, in which RMS errors reduced significantly compared to the sparse-cue scene ("Rich-cue scene" in Experimental results). However, any attempt to make a quantitative comparison of the models using data from the rich scene would face severe challenges—not least in attempting to identify which of the many possible features might be used by observers.

Discussion

In this paper, we have compared the ability of two different classes of models to explain the pattern of errors that participants make when they carry out a homing task in virtual reality. We showed that a model based on view-based features, such as the visual angle separating two landmarks, outperforms a model based on recreating the full 3-D coordinates of landmarks in the scene (Pickup et al., 2011; Pickup et al., 2013). The data from eight individual participants support this conclusion. The data have higher likelihoods under the view-based model than the 3-D reconstruction model (the location of all but one of the magenta points in Figures 12 and 13 is below the line of equality), but more significantly, the likelihood of the data under the view-based model is similar to that of a random sample drawn from the view-based model (Figures 12 through 14). This is not true for the 3-D reconstruction model: The likelihood of the data under the 3-D reconstruction model is *not* similar to the likelihood of a random sample taken from the 3-D reconstruction model.

We have also carried out the same homing task in a rich-cue environment, i.e., a virtual room with a floor, walls, and furniture in addition to the three poles that were the only visible features in the sparse environment.

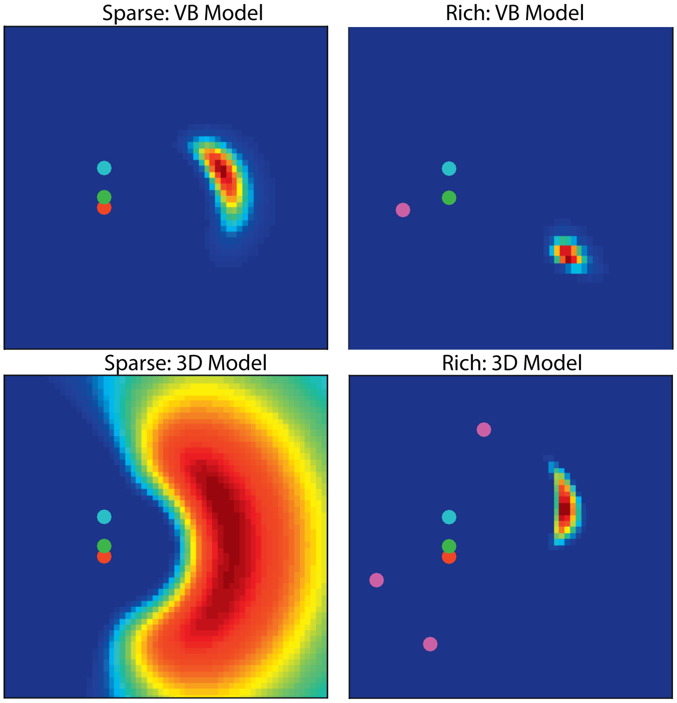

Figure 15. The effect of a richer cue scene. Both the view-based model (top) and the 3-D reconstruction model (bottom) show a constriction of the range of predicted end point locations when the scene is changed. Top right: The purple pole has a greater depth in relation to the green and blue poles, and this has the effect of tightening the likelihood distribution even without any change in the model. Bottom right: Adding more poles also restricts the range of the likelihood distribution for the reconstruction-based model. Model parameterization is taken from participant S5 for both the view-based model and the reconstruction model (λ).

Unsurprisingly, all participants in the rich-cue environment showed lower RMS errors when returning to the goal location than they did in the sparse-cue condition (Figure 6). After all, there are now many more cues available that specify the goal location, and some of these will change very rapidly with changes in the observer's location. This means that a view-based model will predict a narrower spread of "end points" than it does in a sparse-cue environment even without any change to the model. Similarly, a 3-D reconstruction model will predict a narrower spread of errors in the rich-cue environment because it depends on the overlap of the Gaussian error distributions around a larger number of features spread throughout the room. As Figure 15 shows, these factors mean that both models predict a shrinkage of the pattern of errors, qualitatively in line with the shrinkage that participants show. This narrow range of errors makes it hard to distinguish the predicted pattern of errors for the 3-D reconstruction and view-based models in a rich-cue environment, and we have not attempted to do so here. Any systematic comparison would need to take account of a very large range of potential features for the view-based model and possibly include eye tracking to improve any estimate of the features observers were using. In the case of the 3-D reconstruction algorithm, the Bhattacharyya distance for comparing current and stored reconstructions would have to be related to the 3-D Gaussian spread of errors around each feature rather than 2-D Gaussians projected onto a plane as we have used. Nevertheless, our assumption is that the visual system does not switch between view-based and reconstruction strategies depending on the complexity of the scene. We have shown that a view-based strategy provides a better fit to the data in a sparse-cue environment, and we have not found evidence that it switches to a different approach in a rich-cue environment.

Tasks similar to the one used in the present paper have been used to investigate navigational strategy in humans. Waller, Loomis, Golledge, and Beall (2001) conducted an experiment that bears a strong resemblance to our homing task as the participants were asked to return to a previously viewed goal location in a simple virtual reality environment that consisted of only three landmarks (like our experiment, these were vertical poles, but in this case, the poles had a fixed physical radius of 12 cm). The authors distinguished between two hypotheses: Either participants used the distance of landmarks or their bearings. However, because participants were inside the triangle made up of the three landmarks for most of the time, the angle between landmarks was often too large to see both at once, and so, understandably, this cue was found to be relatively ineffective. This was not the case in the current experiment, in which all three landmarks were visible at the same time. Here, the angle between landmarks was one of the most successful features. Foo, Warren, Duchon, and Tarr (2005) compared navigation in a landmark-free environment to that in which landmarks were available. They found that participants could not find novel shortcuts when landmarks were removed (i.e., in a plain desert world), suggesting that participants did not form an accurate cognitive map of the surroundings in this case. As in the homing task used here, the authors also tested navigation on trials in which landmarks were deliberately displaced, finding that participants adjusted their route in line with the displacement. The task used by Foo et al. differs from our simple homing because it can, in theory, be solved by triangulation using two previously learned routes.

Homing requires participants to correctly identify their position in a coordinate frame of some kind and to relate that to a different position at a later time. Described at this very general level, "homing" could encompass a wide variety of tasks, from threading a needle up to large-scale navigation. Each of these tasks could be described within a view-based framework or a 3-D, Cartesian one. For example, the 2-D image vector joining the end of a thread to the eye of a needle, recorded in two binocular views, is very similar to the novel view of the scene at the beginning of interval 2 in our experiment: The task is to change that view until the "goal" image is reached when the thread and the needle coincide. Many other tasks have been described in a similar way: Wilkie, Wann, and Allison (2008) suggest view-based information as the basis for steering a bicycle through a series of obstacles. McBeath, Shaffer, and Kaiser (1995) suggest that catching in baseball involves tracking the angle between the ball and background objects, and similar strategies have been discussed in relation to making a cup of tea (Land, Mennie, & Rusted, 1999). Chains of goal locations, or "nodes" linked together, allow an agent to follow a route, and there are principled methods for dividing up a scene into discrete nodes to form a graph (Franz, Schölkopf, Mallot, & Bülthoff, 1998a), which may be laid down in long-term memory (Röhrich, Hardiess, & Mallot, 2014). A 3-D reconstruction is conceptually far simpler. In theory, a world-based, 3-D model could underlie actions from the smallest (threading a needle) to the largest (navigation), all using reconstruction in the same coordinate frame. Computer vision demonstrates that this is technically feasible (Cummins & Newman, 2011; Davison, 2003).

We have compared two very distinct approaches to building representations of a point in space and compared this to a stored representation of that point. The experimental setup was specifically designed to maximize the difference between the two models in this context. However, some authors have suggested intermediate models that include locally defined 3-D coordinate frames for restricted regions of space with links between these metric reconstructions that are looser and not necessarily defined in a 3-D coordinate frame (Chrastil & Warren, 2014; Mallot & Basten, 2009; Meilinger, 2008). On a smaller scale, Pickup et al. (2013) examined a model to represent the layout of a scene based on the relative 3-D location of pairs of features. This has many properties in common with a view-based model using the relative *2-D* location of image features and could be described as a hybrid or intermediate model between a view-based and a 3-D reconstruction model. Models that include view- and reconstruction-based components might perform better in capturing the pattern of navigation errors that we have observed, but we have only explored the two extremes here.

Conclusion

We have shown that, at least for a simple environment, it is possible to contrast the predictions of two models of homing behavior based either on matching views or on building a 3-D reconstruction of the scene. In this simple case, there is clear evidence in favor of a view-based model. For a richly textured environment, we cannot distinguish between the models. An argument based purely on grounds of parsimony, however, would favor the same model applying in this case, too.

Keywords: motion parallax, *virtual reality*, *stereopsis*, *homing*, *view-based*, *navigation*

Acknowledgments

Supported by EPSRC EP/K011766/1, Dstl and EPSRC EP/N019423/1 and AHRC AH/N006011/1.

Commercial relationships: none.

Corresponding author: Luise Gootjes-Dreesbach.
Email: E.L.Gootjes-Dreesbach@reading.ac.uk.
Address: School of Psychology and Clinical Language Sciences, University of Reading, Reading, United Kingdom.

References

2d3 Ltd. (2003). Boujou: Automated camera tracking [computer software]. Retrieved from www.2d3.com

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

Burgess, N. (2006). Spatial memory: How egocentric and allocentric combine. *Trends in Cognitive Sciences*, *10*, 551–557.

Burgess, N., Jeffery, K. J., & O'Keefe, J. (1999). *The hippocampal and parietal foundations of spatial cognition*. Oxford, UK: Oxford University Press.

Cartwright, B. A., & Collett, T. S. (1983). Landmark learning in bees: Experiments and models. *Journal of Comparative Physiology*, *151*, 521–543.

Chrastil, E. R., & Warren, W. H. (2014). From cognitive maps to cognitive graphs. *PLoS One*, *9*(11), e112544.

Cummins, M., & Newman, P. (2011). Appearance-only slam at large scale with fab-map 2.0. *The International Journal of Robotics Research*, *30*(9), 1100–1123.

Davison, A. J. (2003). Real-time simultaneous localisation and mapping with a single camera. In: *Proceedings. Ninth IEEE International Conference on computer vision* (pp. 1403–1410). New York: IEEE.

Foo, P., Warren, W. H., Duchon, A., & Tarr, M. J. (2005). Do humans integrate routes into a cognitive map? Map- versus landmark-based navigation of novel shortcuts. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(2), 195–215, doi:10.1037/0278-7393.31.2.195.

Franz, M. O., Schölkopf, B., Mallot, H. A., & Bülthoff, H. H. (1998a). Learning view graphs for robot navigation. *Autonomous Robots*, *5*(1), 111–125, doi:10.1145/267658.267687.

Franz, M. O., Schölkopf, B., Mallot, H. A., & Bülthoff, H. H. (1998b). Where did I take that snapshot? Scene-based homing by image matching. *Biological Cybernetics*, *79*, 191–202, doi:10.1007/s004220050470.

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

Gilson, S. J., Fitzgibbon, A. W., & Glennerster, A. (2011). An automated calibration method for non-see-through head mounted displays. *Journal of Neuroscience Methods*, *199*(2), 328–335, doi:10.1016/j.jneumeth.2011.05.011.

Graham, P., & Cheng, K. (2009). Which portion of the natural panorama is used for view-based navigation in the Australian desert ant? *Journal of Comparative Physiology A*, *195*(7), 681–689.

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

Hafting, T., Fyhn, M., Molden, S., Moser, M. B., & Moser, E. I. (2005, June 19). Microstructure of a spatial map in the entorhinal cortex. *Nature*, *436*, 801–806.

Hartley, R. I., & Zisserman, A. (2004). *Multiple view geometry in computer vision* (2nd ed.). Cambridge, UK: Cambridge University Press.

Jacobs, J., Weidemann, C. T., Miller, J. F., Solway, A., Burke, J. F., Wei, X. X., . . . Kahana, M. J. (2013). Direct recordings of grid-like neuronal activity in human spatial navigation. *Nature Neuroscience*, *16*(9), 1188–1190.

Koenderink, J. J., van Doorn, A. J., Kappers, A. M., & Lappin, J. S. (2002). Large-scale visual frontoparallels under full-cue conditions, *Perception*, *31*(12), 1467–1475.

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, *22*(1), 79–86.

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

Lent, D. D., Graham, P., & Collett, T. S. (2010). Image-matching during ant navigation occurs through saccade-like body turns controlled by learned visual features. *Proceedings of the National Academy of Sciences, USA*, *107*(37), 16348–16353.

Mallot, H. A., & Basten, K. (2009). Embodied spatial cognition: Biological and artificial systems. *Image and Vision Computing*, *27*(11), 1658–1670.

McBeath, M. K., Shaffer, D. M., & Kaiser, M. K. (1995, Apr 28). How baseball outfielders determine where to run to catch fly balls. *Science*, *268*, 569–573.

McNaughton, B. L., Battaglia, F. P., Jensen, O., Moser, E. I., & Moser, M. B. (2006). Path integration and the neural basis of the "cognitive map." *Nature Reviews Neuroscience*, *7*, 663–678.

Meilinger, T. (2008). The network of reference frames theory: A synthesis of graphs and cognitive maps. In C. Freska, N. S. Newcombe, P. Gärdenfors, & S. Wölfl (Eds.), *Spatial Cognition VI, LNAI 5248* (pp. 344–360). Berlin Heidelberg, Germany: Springer-Verlag.

Meilland, M., & Comport, A. I. (2013). Super-resolution 3D tracking and mapping. In *Robotics and Automation (ICRA), 2013 IEEE International Conference on, IEEE* (pp. 5717–5723). New York: IEEE.

Mou, W., McNamara, T. P., Rump, B., & Xiao, C. (2006). Roles of egocentric and allocentric spatial representations in locomotion. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *32*(6), 1274–1290.

Newcombe, R. A., & Davison, A. J. (2010). Live dense reconstruction with a single moving camera. Presented at IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 13–18, 2010, San Francisco, CA.

Ni, K., Kannan, A., Criminisi, A., & Winn, J. (2009). Epitomic location recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *31*(12), 2158–2167.

O'Keefe, J., & Nadel, L. (1978). *The hippocampus as a cognitive map*. Oxford, UK: Oxford University Press.

Pickup, L. C., Fitzgibbon, A. W., Gilson, S. J., & Glennerster, A. (2011). View-based modelling of human visual navigation errors. Presented at the IEEE IVMSP Workshop, June 16–17, 2011, Ithaca, NY.

Pickup, L. C., Fitzgibbon, A. W., & Glennerster, A. (2013). Modelling human visual navigation using multi-view scene reconstruction. *Biological Cybernetics*, *107*(4), 449–464.

Röhrich, W. G., Hardiess, G., & Mallot, H. A. (2014). View-based organization and interplay of spatial working and long-term memories. *PLoS One*, *9*(11), e112793.

Sadalla, E. K., & Magel, S. G. (1980). The perception of traversed distance. *Environment and Behavior*, *12*(1), 65–79.

Sadalla, E. K., & Staplin, L. J. (1980). The perception of traversed distance intersections. *Environment and Behavior*, *12*(2), 167–182.

Schinazi, V. R., Nardi, D., Newcombe, N. S., Shipley, T. F., & Epstein, R. A. (2013). Hippocampal size predicts rapid learning of a cognitive map in humans. *Hippocampus*, *23*(6), 515–528.

Snyder, L. H., Grieve, K. L., Brotchie, P., & Andersen, R. A. (1998, August 27). Separate body- and world-referenced representations of visual space in parietal cortex. *Nature*, *394*, 887–891.

Svarverud, E., Gilson, S., & Glennerster, A. (2012). A demonstration of "broken" visual space. *PLoS One*, *7*(3), e33782, doi: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. *Proceedings of the National Academy of Sciences, USA*, *108*(3), 1152–1157, doi:10.1073/pnas.1011843108.

Tolman, E. C. (1948). Cognitive maps in rats and men. *Psychological Review*, *55*(4), 189–208, doi:10.1037/h0061626.

Waller, D., Loomis, J. M., Golledge, R. G., & Beall, A. C. (2001). Place learning in humans: The role of distance and direction information. *Spatial Cognition and Computation*, *2*, 333–354, doi:10.1023/A:1015514424931.

Watson, A. B., & Pelli, D. G. (1983). Quest: A Bayesian adaptive psychometric method. *Perception & Psychophysics*, *33*(2), 113–120.

Wehner, R., & Räber, F. (1979). Visual spatial memory in desert ants, cataglyphis fortis (hymenoptera, formicidae). *Experientia*, *35*, 1569–1571.

Whelan, T., Leutenegger, S., Salas-moreno, R. F., Glocker, B., & Davison, A. J. (2015). ElasticFusion: Dense SLAM without a pose graph. In L. E. Kavraki, D. Hsu, & J. Buchli (Eds.), *Proceedings of robotics: Science and systems*. Presented July 13–17, 2015, Rome, Italy, doi:10.15607/RSS.2015.XI.001.

Wilkie, R. M., Wann, J. P., & Allison, R. S. (2008). Active gaze, visual look-ahead, and locomotor control. *Journal of Experimental Psychology: Human Perception and Performance*, *34*(5), 1150–1164.

Zhu, Y., Mottaghi, R., Kolve, E., Lim, J. J., Gupta, A., Fei-Fei, L., & Farhadi, A. (2016). Target-driven visual navigation in indoor scenes using deep reinforcement learning. CoRR abs/1609.05143, URL http://arxiv.org/abs/1609.05143.