Tracking the emergence of location-based spatial representations

Sam C. Berens[1,*], Bárður H. Joensen[1], and Aidan J. Horner[1,2*]

1. Department of Psychology, University of York, UK

2. York Biomedical Research Institute, University of York, UK

* Corresponding authors: sam.berens@york.ac.uk & aidan.horner@york.ac.uk

Pages: 23, Figures: 4

Abstract

Scene-selective regions of the human brain form allocentric representations of locations in our environment. These representations are independent of heading direction and allow us to know where we are regardless of our direction of travel. However, we know little about how these location-based representations are formed. Using fMRI representational similarity analysis, we tracked the emergence of location-based representations in scene-selective brain regions. We estimated patterns of activity for two distinct scenes, taken before and after participants learnt they were from the same location. During a learning phase, we presented participants with two types of panoramic videos: (1) an overlap video condition displaying two distinct scenes (0^o and 180^o) from the same location, and (2) a no-overlap video displaying two distinct scenes from different locations (that served as a control condition). In the parahippocampal cortex (PHC) and retrosplenial cortex (RSC), representations of scenes from the same location became more similar to each other only after they had been shown in the overlap condition, suggesting the emergence of location-based viewpoint-independent representations. Whereas location-based representations emerged in the PHC regardless of subsequent behaviour, RSC representations only emerged for locations where participants could behaviourally identify the two scenes as belonging to the same location. The results demonstrate that we can track the emergence of location-based representations in the PHC and RSC in a single fMRI session and suggest that the RSC plays a key role in using such representations to locate ourselves in space.

Introduction

Rapidly learning the spatial layout of a new environment is a critical function that supports flexible cognition. This ability is thought to be underpinned by the emergence of spatial representations in scene-selective brain regions, including location-based representations that signal where we are irrespective of our current heading direction. Given we are unable to sample all possible viewpoints from a given location simultaneously, the formation of location-based representations requires the integration of scenes across differing viewpoints. Despite evidence for the existence of location-based, viewpoint-independent, representations in scene-selective regions (Marchette, Vass, Ryan, & Epstein, 2015; Robertson, Hermann, Mynick, Kravitz, & Kanwisher, 2016; Vass & Epstein, 2013), we know little about how such representations emerge.

Models of spatial navigation suggest that distinct brain regions are responsible for supporting allocentric (viewpoint-independent) and egocentric (viewpoint-dependent) representations of our environment (Byrne, Becker, & Burgess, 2007; Julian, Keinath, Marchette, & Epstein, 2018). Specifically, the parahippocampal cortex (PHC) is thought to encode allocentric spatial representations related to navigational landmarks (Burgess, Becker, King, & O'Keefe, 2001; Epstein, Patai, Julian, & Spiers, 2017), and spatial context more broadly (Epstein & Vass, 2014). In contrast, the parietal lobe is thought to support egocentric representations of specific viewpoints that underpin route planning (Byrne et al., 2007; Calton & Taube, 2009). To enable efficient route planning, a transformation between allocentric and egocentric representations is thought to occur in the retrosplenial cortex, cueing allocentric representations from egocentric inputs and vice versa (Bicanski & Burgess, 2018; Byrne et al., 2007).

In support of these models, human fMRI studies using representational similarity analyses (RSA) have found evidence for location-based, viewpoint-independent, representations (henceforth referred to as "location-based representations") in a network of brain regions including the PHC and RSC (Marchette, Vass, Ryan, & Epstein, 2014; Robertson et al., 2016; Vass & Epstein, 2013). However, little is known about how such representations are formed. First, we don't know whether location-based representations can emerge rapidly (i.e., in the course of a single fMRI session). Second, without tracking their formation, it is difficult to determine exactly what information is being represented. For instance, shared representations across viewpoints may relate to long-term semantic knowledge that is invoked when seeing different views of a well-known location (see Marchette, Ryan, & Epstein, 2017).

Here, we test whether location-based representations of novel environments can be learnt by integrating visual information across different scenes. Multivariate patterns of BOLD activity were recorded as participants passively observed a number of scenes depicting different views of novel locations. Subsequently, using an experimental manipulation introduced by Robertson et al. (2016), participants watched videos that showed these scenes as part of a wider panorama. Half of the videos allowed participants to learn the spatial relationship between two scenes from the same location (overlap condition). The remaining videos acted as a control by presenting scenes from different locations (no-overlap condition). Following the videos, we again recorded patterns of activity for each of the scenes. Whereas Roberston et al. (2016) only assessed scene representations following video presentation, we also scanned before and during video presentation. This allowed us to track the *emergence* of location-based representations using representational similarity analyses (RSA), as well as assess neural activity when these representations were being formed.

We show that patterns in the PHC and RSC become more similar between scenes following the presentation of the video panoramas. This increase in similarity is specific to the 'overlap' video condition, where the scenes from the same location were presented. No increase in pattern similarity was seen for the 'no-overlap' condition, where two scenes from different locations were presented. Importantly, whereas location-based representations in the PHC emerged regardless of behavioural performance, representations in the RSC emerged only in instances where participants could (following scanning) identify that the scenes came from the same location. Thus, in support of computational models of spatial navigation, the RSC appears to play a critical role in translating allocentric representations in the medial temporal lobe into more behaviourally-relevant egocentric representations in the parietal cortex.

Methods

Participants

Twenty-eight, right handed participants were recruited from the University of York, UK. These participants had no prior familiarity with the locations used as stimuli in the experiment (see below). All participants gave written informed consent and were reimbursed for their time. Participants had either normal or corrected-to-normal vision and reported no history of neurological or psychiatric illness. Data from five participants could not be included in the final sample due to: problems with fMRI data acquisition (1 participant), excess of motion related artefacts in the imaging data (3 participants), and a failure to respond during one of the in-scanner tasks (1 participant). As such, analyses included 23 participants (10 males) with a mean age of 21.96 years (*SD* = 3.22). The study was approved by a local research ethics committee at the University of York.

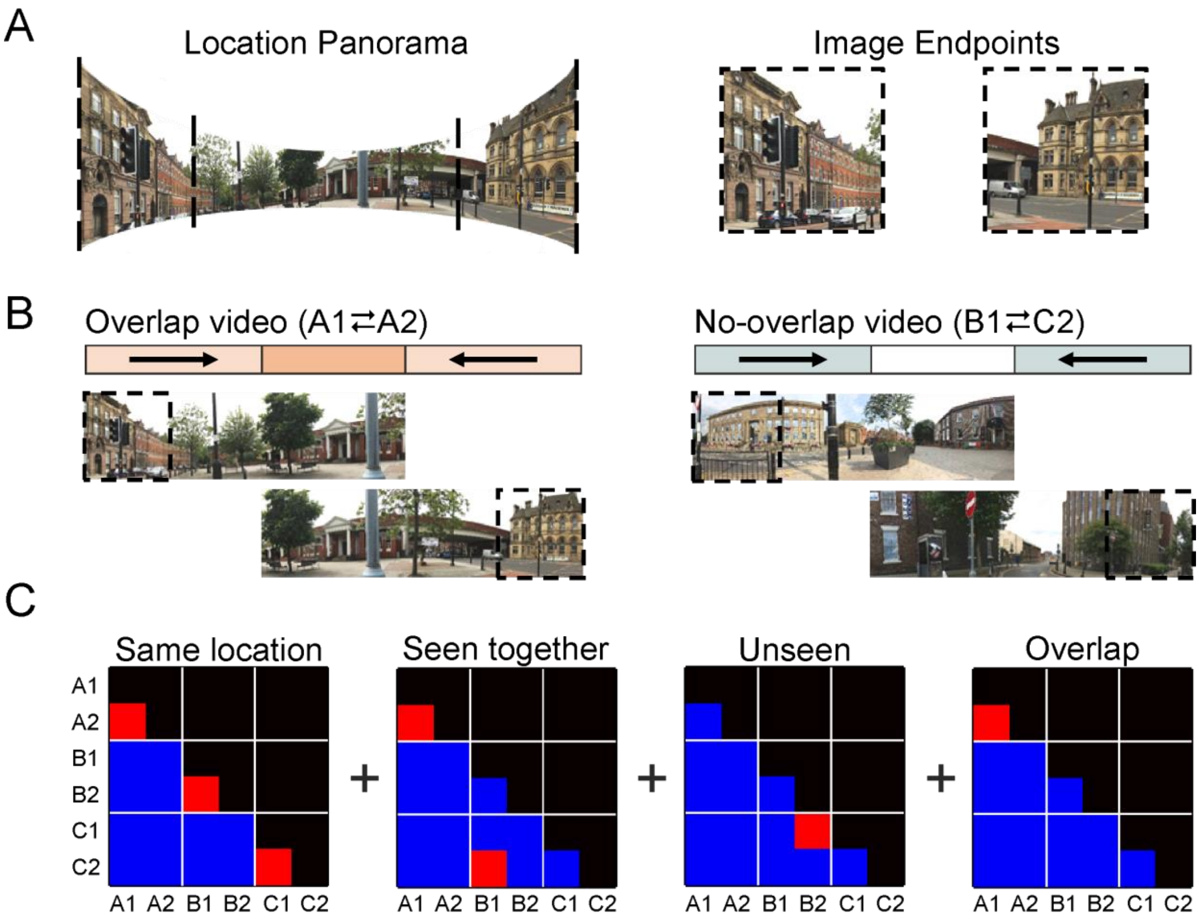

Figure 1. Stimuli used, and analyses performed, during the in-scanner tasks. **(A)** An example location panorama with 2 endpoint images. Single endpoints were show during the in-scanner target detection task. As in Roberston et al. (2016), full panoramas were never shown as whole images but were presented during the in-scanner videos. **(B)** Depiction of the 2 video conditions: overlap vs no-overlap videos. Overlap videos showed camera pans from each endpoint of a given panorama (denoted A1 and A2) to the centre of that panorama. The central overlap allowed participants to learn a spatially coherent representation that included both A1 and A2. No-overlap videos involved pans from endpoints B1 and C2 (taken from different panoramas) meaning that there was no visual overlap. **(C)** Similarity contrast matrices used to model changes in representational similarity between endpoints (i.e., between A1, A2, B1, B2, C1, and C2). Red squares indicate positively weighted correlations and blue squares indicate negatively weighted correlations (matrices are scaled to sum to 0). From left to right, the matrices account for the representational similarity of endpoints: (1) from the same location regardless of video condition, (2) that were seen in the same video (including overlap and no-overlap videos), (3) in the unseen condition specifically, and (4) in the overlap condition specifically. Linear combinations of these matrices, along with their interactions with a session regressor, accounted for each RSA effect across all experimental conditions.

Stimuli

We generated 12 panoramic images of different urban locations from the City of Sunderland, and Middlesbrough town centre, UK (Figure 1; https://osf.io/cgy97; also see Roberston et al. (2016)). These panoramas spanned a 210° field-of-view horizontally but were restricted in the vertical direction to limit the appearance of proximal features (< 2 meters from camera). Throughout the experiment, 24 'endpoint images' displaying 30° scenes taken from either end of each panorama were shown (i.e., centred at 0° and 180°; Figure 1A). These images were shown both inside and outside of the scanner to assess participants' spatial knowledge of the depicted locations and for the representational similarity analysis (see below).

Endpoints were also shown in a number of videos (see https://osf.io/cgy97). In *overlap videos*, images A1 and A2 (taken from opposite ends of the same panorama) were presented such that their spatial relationship could be inferred (Figure 1B). Here, a camera panned from each endpoint to the centre of the panorama showing that A1 and A2 belonged to the same location. In contrast, a *no-overlap video* featured endpoints from two unrelated panoramas (images B1 and C2). Again, these videos showed an end-to-centre camera pan from each image. However, since there was no visual overlap between the video segments, observers could only infer that endpoints B1 and C2 belonged to different locations. The no-overlap condition acted as a control condition, ensuring endpoints B1 and C2 were seen in a similar video to endpoints in the overlap condition (A1 and A2), with the same overall exposure and temporal proximity. To ensure that the occurrence of a visual overlap was easily detectable, all videos alternated the end-to-centre sweep from each endpoint over two repetitions.

Pairs of endpoints from the same panorama were grouped into sets of 3. The first pair in each set were assigned to the overlap video condition (A1 and A2). Two endpoints from different panoramas were assigned to the no-overlap video condition (B1 and C2). The remaining endpoints belonged to an *unseen video* condition as they were not shown during any video (B2 and C1). These assignments were counterbalanced across participants such that each image appeared in all 3 conditions an equal number of times. The order of camera pans during videos (e.g. A1 first vs A2 first) was also counterbalanced both within and across participants. Analyses showed the visual similarity of image endpoints was matched across experimental conditions as measured by the Gist descriptor (Oliva & Torralba, 2001) and local correlations in luminance and colour information (https://osf.io/un5gr). Pilot data revealed that participants could not reliably identify which endpoints belonged to the same location without having seen the videos (https://osf.io/ev5ry).

Procedure

Prior to entering the scanner, participants performed a behavioural task to assess their ability to infer which image endpoints were from the same location. Once in the scanner, they undertook a functional localiser task to identify scene-selective regions of the parahippocampal cortex (PHC) and retrosplenial cortex (RSC). They were then shown each image endpoint multiple times (performing a low-level attentional task) to assess baseline representational similarity between each image endpoint (i.e., prior to learning). Overlap and No-overlap videos were then presented, with participants instructed to identify whether the endpoints in each video belonged to the same location or not. Following this video learning phase, each image endpoint was again presented multiple times to assess post-learning representational similarity between image endpoints. Finally, outside the scanner, participants performed the same pre-scanner behavioural task to assess the extent to which participants had learnt which image endpoints belonged to the same location (and a further test of associative memory, see below).

Pre-/Post-scanner tasks

Participants were tested on their ability to identify which endpoints belonged to the same location both before and after scanning (both outside of the scanner). On each trial, one endpoint surrounded by a red box was presented for 3 seconds. Following this, 5 other endpoints were displayed in a random sequence, each shown alongside a number denoting the order of appearance (i.e., 1-5; Figure 2A; 2 seconds per image, 500 ms inter-stimulus interval). One image in the sequence (the target) was taken from the same panorama as the cue. The remaining 4 endpoints (lures) belonged to panoramas in the same set of stimuli. As such, if endpoint B1 was presented as the cue, B2 would be the target, and A1, A2, C1, and C2 would be lures (i.e., a 5-alternative forced choice, 5-AFC). After the 5 alternatives had been shown, participants were prompted to select the target using a numeric key press (1-5). Across 24 trails, each endpoint was used as a cue image.

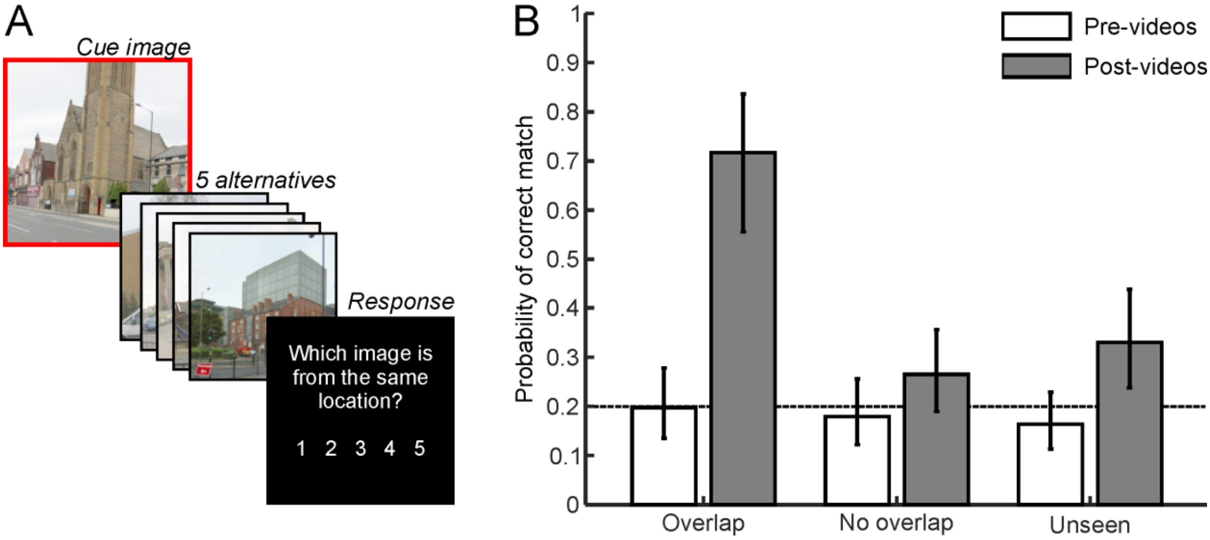

Figure 2. Behavioural task and results **(A)** A schematic illustration of the pre- and post-video behavioural task. One endpoint was first presented as a cue (enclosed by red-box), followed by 5 numbered alternatives. Participants were then prompted to select which one of the alternatives belonged to the same location as the cue. **(B)** Performance on the pre- and post-video behavioural tasks plotted by video condition. Error bars represent 95% confidence intervals and the dashed line at p = .2 reflects chance level.

Following scanning, and the second block of the location identity task describe above, participants were also asked to identify which images appeared together in the same video. Note: this is slightly different to the previous task since participants could have known that endpoints B1 and C2 appeared in the same video, despite not knowing which endpoints were from the same location (i.e., B2 and C1 respectively). Using a similar procedure to that described above, endpoints from either the overlap or no-overlap video conditions were cued and participants were asked to select the appropriate endpoint from the 5 alternatives in the same set.

In-scanner tasks

Functional localiser

Before the main experimental task, participants undertook a functional localiser scan with the purpose of identifying four scene selective regions of interest (ROIs) - in particular, the left and right parahippocampal cortex (PHC), and the left and right retrosplenial cortex (RSC). This involved presenting 4 blocks of scene images (coasts, mountains, streets, and woodlands) interleaved with 4 blocks of face images (male and female). In each block, 10 unique images were shown in quick succession with a display time of 700 ms per image and an inter-stimulus interval of 200 ms. Blocks were separated with a 9 second inter-block interval and their running order was counterbalanced across participants. The scene images used here were different to those in the main experiment and none were repeated during the localiser itself. All images were shown in greyscale and were presented with a visual angle of ~14°. To ensure localiser images were being attended to, participants were tasked with detecting an odd-ball target that was superimposed onto one of the images in each block. The target was a small red dot with a 3-pixel radius. When this was seen, participants were required to respond with a simple button press as quickly as possible (mean detection performance: *d'* = 3.116, *SD* = 0.907).

Presentation of endpoint images

Participants were shown all 24 endpoint images during an event-related functional imaging task. The task was optimised to measures multivariate patterns of BOLD activity specific to individual endpoints and was run both before and after participants had seen the panoramic videos (session 1: pre-videos; session 2: post-videos). All endpoints were presented 9 times for both the pre- and post-video functional run. Images were displayed for 2.5 seconds with an inter-stimulus interval of 2 seconds. The order of stimuli in each functional run was optimised to facilitate the decoding of unique BOLD patterns across endpoints. No image was presented on successive trails to avoid large adaptation effects and the design included 12 null events in each functional run (i.e., 10% of all events). Like the functional localiser, participants were tasked with detecting an odd-ball target that was superimposed onto a small proportion of the images. Here, the target was a group of 3 small red dots (3-pixel radius, < 0.2°), with each dot drawn at a random position on the image. Targets were present on 1 out of every 9 trials such that 8 repetitions of each endpoint image were target free (target trails were not used to estimate BOLD patterns). As above, participants were required to respond to these targets with a simple button press (mean detection performance, *d'*, was 3.362, *SD* = 0.642, pre-videos, and 3.659, *SD* = 0.485, post-videos).

Panoramic video task

Participants watched all video clips from the overlap and no-overlap video conditions whilst being scanned. Each video lasted a total of 20 seconds and was followed by a 10 second rest period. In the first 3 seconds of this rest period, participants were prompted to indicate whether each video segment depicted scenes from the same or different locations. Responses were recorded with a left/right button press. This question was asked to ensure that participants were attending to the visual overlap across segments (mean discrimination performance: *d'* = 3.220, *SD* = 0.373). All videos were repeated 3 times in a pseudorandom order to allow for sufficient learning. Prior to entering the scanner, participants were asked to remember which endpoints were seen together in the same video, even if they appeared in a no-overlap video. Participants were told that a test following the scan would assess their knowledge of this.

MRI acquisition

All functional and structural volumes were acquired on a 3 Tesla Siemens MAGNETOM Prisma scanner equipped with a 64-channel phased array head coil. T2*-weighted scans were acquired with echo-planar imaging (EPI), 35 axial slices (approximately 0° to AC-PC line; interleaved) and the following parameters; repetition time = 2000 ms, echo time = 30 ms, flip angle = 80°, slice thickness = 3 mm, in-plane resolution = 3 × 3 mm. The number of volumes acquired during (a) the functional localiser, (b) the video task, and (c) each run of the endpoint presentation task was 75, 363, and 274 respectively. To allow for T1 equilibrium, the first 3 EPI volumes were acquired prior to the task starting and then discarded. Subsequently, a field map was captured to allow the correction of geometric distortions caused by field inhomogeneity (see the MRI pre-processing section below). Finally, for purposes of co-registration and image normalization, a whole-brain T1-weighted structural scan was acquired with a 1mm^3 resolution using a magnetization-prepared rapid gradient echo pulse sequence.

MRI pre-processing

Image pre-processing was performed in SPM12 (www.fil.ion.ucl.ac.uk/spm). This involved spatially realigning all EPI volumes to the first image in the time series. At the same time, images were corrected for field inhomogeneity based geometric distortions (as well as the interaction between motion and such distortions) using the Realign and Unwarp algorithms in SPM (Andersson, Hutton, Ashburner, Turner, & Friston, 2001; Hutton et al., 2002). For the RSA, multivariate BOLD patterns of interest were taken as *t*-statistics from a first-level general linear model (GLM) of unsmoothed EPI data in native space. Aside from regressors of interest, each first-level GLM included a set of nuisance regressors: 6 affine motion parameters, their first-order derivatives, and regressors censoring periods of excessive motion (rotations > 1°, and translations > 1mm). For the analyses of univariate BOLD activations, EPI data were warped to MNI space with transformation parameters derived from structural scans (using the DARTEL toolbox; Ashburner, 2007). Subsequently, the EPI data were spatially smoothed with an isotropic 8 mm FWHM Gaussian kernel prior to GLM analysis.

We also generated four binary masks per participant to represent each ROI in native space. To do this, a first-level GLM of the functional localiser data modelled BOLD responses to scene and face stimuli presented during the localiser task. Each ROI was then defined as the conjunction between a "scene > face" contrast and an anatomical mask of each region that had been warped to native space (left/right PHC sourced from: Tzourio-Mazoyer et al., 2002; left/right RSC sourced from: Julian, Fedorenko, Webster, & Kanwisher, 2012). Note, given previous research implicating the occipital place area (OPA) as a critical scene-selective region (e.g., Marchette et al., 2015; Robertson et al., 2016), we also attempted to identify this functional region in each participant. However, we were only able to identify the OPA bilaterally in 6 (out of 23) participants, nor did we see the OPA in a group-level analysis. As such, this region was not included as a region of interest in subsequent analyses.

Representational similarity analyses

Visual representations of specific endpoints

We first examined whether the passive viewing of endpoint images evoked stimulus specific visual representations in each of our four ROIs (left and right PHC and RSC). Multivariate BOLD responses to the endpoints were estimated for session 1 (pre-videos) and session 2 (post-videos) separately. We then computed the similarity of these responses across sessions by correlating BOLD patterns in session 1 with patterns in session 2. The resulting correlation coefficients were Fisher-transformed and entered as a dependent variable into a mixed-effects regression model with random effects for subjects and endpoints. The main predictor of interest was a fixed effect that contrasted correlations between like endpoints (A1-A1, B1-B1) with correlations between different endpoints (i.e., A1-A2, A1-B1 etc.).

As well as running this analysis in each ROI, we performed a complementary searchlight analysis to detect endpoint-specific representations in other brain regions. Here, BOLD pattern similarity was computed at each point in the brain using spherical searchlights with a 3-voxel radius (the mean number of voxels per searchlight was 105.56). Fisher-transformed correlations for same- verses different-endpoints were contrasted at the first-level before running a group-level random effects analysis.

Location-based memory representations

We next tested our principle hypothesis in each ROI: whether representations of endpoints A1 and A2 became more similar to one another as a result of watching the overlap videos. Here, we generated a mixed-effects regression model to compare Fisher-transformed similarity estimates between endpoints of the same set (see Figure 1C). One fixed-effect predictor accounted for similarity changes from session 1 to session 2. A second accounted for similarity differences between the overlap endpoints (i.e., A1-A2) and all other endpoint correlations ('Overlap' matrix in figure 1C). As such, the interaction between these two predictors (*Session*Overlap*) coded the RSA effect of interest.

The model also included predictors to account for changes in similarity between: (*i)* endpoints from the same location (A1-A2, B1-B2, C1-C2), (*ii)* endpoints shown in the same video (A1-A2, B1-C2), and (*iii)* endpoints that were not shown in any video (C1-B2). Together, these predictors ensured that variance loading onto the *Session*Overlap* effect was properly attributable to the learning of spatially coherent representations rather than some combination of other factors (e.g. same location + seen in same video). Predictors relating to *session* and each video condition (overlap, no-overlap, unseen) constituted a 2x3 factorial structure and so were therefore tested with a *Session*Condition F*-test. The model also included a behavioural covariate specifying whether participants were able to match endpoints A1 to A2 in the post-scanner task (mean centred with 3-levels: 0, 1 or 2 correct responses per pair). This examined whether changes in representational similarity were dependent on participants ability to identify that endpoints from the overlap condition belonged to the same location after scanning (i.e., a 3-way interaction; *Session*Overlap*Behaviour*). Finally, random effects in the model accounted for statistical dependencies across endpoints, sessions, and subjects.

Finally, we also ran a complementary searchlight analysis that tested for RSA effects outside of our *a priori* ROIs (searchlight radius: 3-voxel). Here, a first-level analysis contrasted Fisher-transformed correlations for overlap endpoints vs all other endpoint combinations. A group-level random effects analysis then compared these similarity contrasts between sessions to test the *Session*Overlap* interaction.

Statistical inference

All *p*-vales are reported as two-tailed statistics. Corrections for multiple comparisons across our four regions of interest are made for each *a priori* hypothesis. Additionally, we report whole-brain effects from searchlight and mass univariate analyses when they survive family-wise error corrected thresholds ($p < .05$ FWE) at the cluster level (cluster defining threshold: $p < .001$ uncorrected). When key significance tests failed to reject the null hypothesis, we performed a complementary Bayesian analysis to examine whether the null was statistically preferred over the alternative hypothesis. In each case, a Bayes factor in favour of the null hypothesis (BF_{01}) was computed with a Cauchy prior centred at zero (i.e., no effect) and a scale parameter, *r*, of $\sqrt{2}/2$. Bayes factors greater than 3 are taken as evidence in favour of the null hypothesis (Kass & Raftery, 1995).

Results

Behavioural performance

We first analysed behavioural responses to the pre- and post-scanner tasks to determine (a) whether participants were able to identify which endpoints belonged to the same location, and (b) whether performance increased as a result of watching the overlap videos. A generalised linear mixed-effects analysis modelled correct vs incorrect matches between cue and target endpoints as a function of session (pre- vs post- videos) and experimental condition (overlap, no-overlap, and unseen). As such, the model constituted a 2 x 3 factorial design with random intercepts and slopes for both participants and stimuli.

The results, displayed in Figure 2B, revealed significant main effects of session ($F_{1,\ 1098}$ = 47.302, p < .001), and condition ($F_{2,\ 1098}$ = 6.500, p = .002), as well an interaction between them ($F_{2,\ 1098}$ = 11.231, p < .001). The interaction indicated that performance was at chance level across all conditions before the videos (min p = 0.249, BF_{01} = 2.531), but substantially increased in the overlap video condition having seen the videos (t_{1098} = 6.867, p < .001; post-video > pre-video). This increase was not seen in the no-overlap condition (t_{1098} = 1.761, p = .079), however a significant increase was seen in the unseen condition (t_{1098} = 3.159, p = .002). The performance increases in the control conditions (only significant in the unseen condition) were likely the result of participants being able to exclude overlap endpoints as non-target alternatives in the 5-AFC test (i.e., a recall-to-reject strategy, disregarding A1 and A2 when cued with either B1, B2, C1 or C2). Consistent with this, session 2 performance in the no-overlap and unseen conditions was not significantly different from chance level in a 3-AFC test (0.33; as opposed to 0.2 in a 5-AFC test; no-overlap: t_{1098} = -1.494, p = .135, BF_{01} = 1.729; unseen: t_{1098} = -0.054, p = .957, BF_{01} = 4.567). Nonetheless, performance in the overlap condition did significantly differ from this adjusted chance level (t_{1098} = 4.514, p < .001).

Furthermore, participants increased ability to match endpoints in the overlap condition was not characteristic of a general tendency to match endpoints that appeared in the same video (i.e., selecting B1 when cued with C2). This was evident since matches between no-overlap endpoints were not more likely in session 2 compared with session 1 (t_{366} = 0.646, p = .519, BF_{01} = 3.785). In contrast, performance increases in the overlap condition (i.e., the post-video > pre-video effect reported above) were significantly larger than this general effect of matching all endpoints that appeared in the same video ($t_{949.20} = 5.027$, $p < .001$; *d.f.* computed via the Welch–Satterthwaite approximation). Additionally, participants were unable to explicitly match no-overlap endpoints shown in the same video during the final behavioural task (comparison to 0.2 chance level: t_{334} = -0.467, p = .641, BF_{01} = 4.141). In sum, participants rapidly learnt which scenes were from the same location, however this was only seen in the overlap condition (and not in the no overlap, or unseen, conditions).

Visual representations of specific endpoints

The mixed-effects model examining representational similarity across sessions revealed that correlations between like endpoints were greater than correlations between different endpoints in the left PHC (t_{13246} = 3.277, p = .004), and the right RSC (t_{13246} = 2.566, p = .041). This effect was not significant in the right PHC (t_{13246} = 1.474, p = .562, BF_{01} = 1.773), or the left RSC (t_{13246} = 1.815, p = .278, BF_{01} = 1.122). The searchlight analysis that tested this effect across the whole brain revealed representations in one large cluster that peaked in the right occipital lobe (area V1; t_{22} = 11.50, $p < .001$, k = 5202) and extended into the areas V2, V3, V4, and the fusiform gyri bilaterally. Three smaller clusters were also detected in the right Precuneus (t_{22} = 4.64, p = .011, k = 44), right inferior parietal lobule (t_{22} = 4.40, p = .028, k = 37), and right RSC (t_{22} = 4.32, p = .025, k = 38). Unthresholded statistical maps of these effects are available at https://neurovault.org/collections/4819.

Location-based memory representations

The mixed-effects model examining representational similarity between different endpoints revealed a significant *Session**Condition* interaction in the right PHC ($F_{2,\ 2746}$ = 6.402, p = .007, p-value corrected for multiple comparisons; Figure 3A). Post-hoc tests showed that this effect was driven by increased BOLD pattern similarity across sessions for endpoints in the overlap condition (t_{2746} = 2.854, p = .004), but not for any other condition (no-overlap: t_{2746} = 0.678, p = .498, BF_{01} = 3.714; unseen: t_{2746} = -0.584, p = .559, BF_{01} = 3.917). The *Session**Condition* interaction was not significant in any other ROI, including the right RSC ($F's_{2,\ 2746}$ < 2.014, $p's$ > .133, uncorrected). However, we saw a significant *Session**Overlap**Behaviour* interaction in the right RSC (t_{2746} = 2.530, p = .046, p-value corrected for multiple comparisons; Figure 3D). This suggests that the RSC encoded viewpoint-independent representations only when the spatial relationships between viewpoints could be retrieved later. No other ROIs showed a significant *Session**Condition**Behaviour* interaction (*t*'s < 0.867, $p's$ > .385, $BF_{01}'s$ > 3.263).

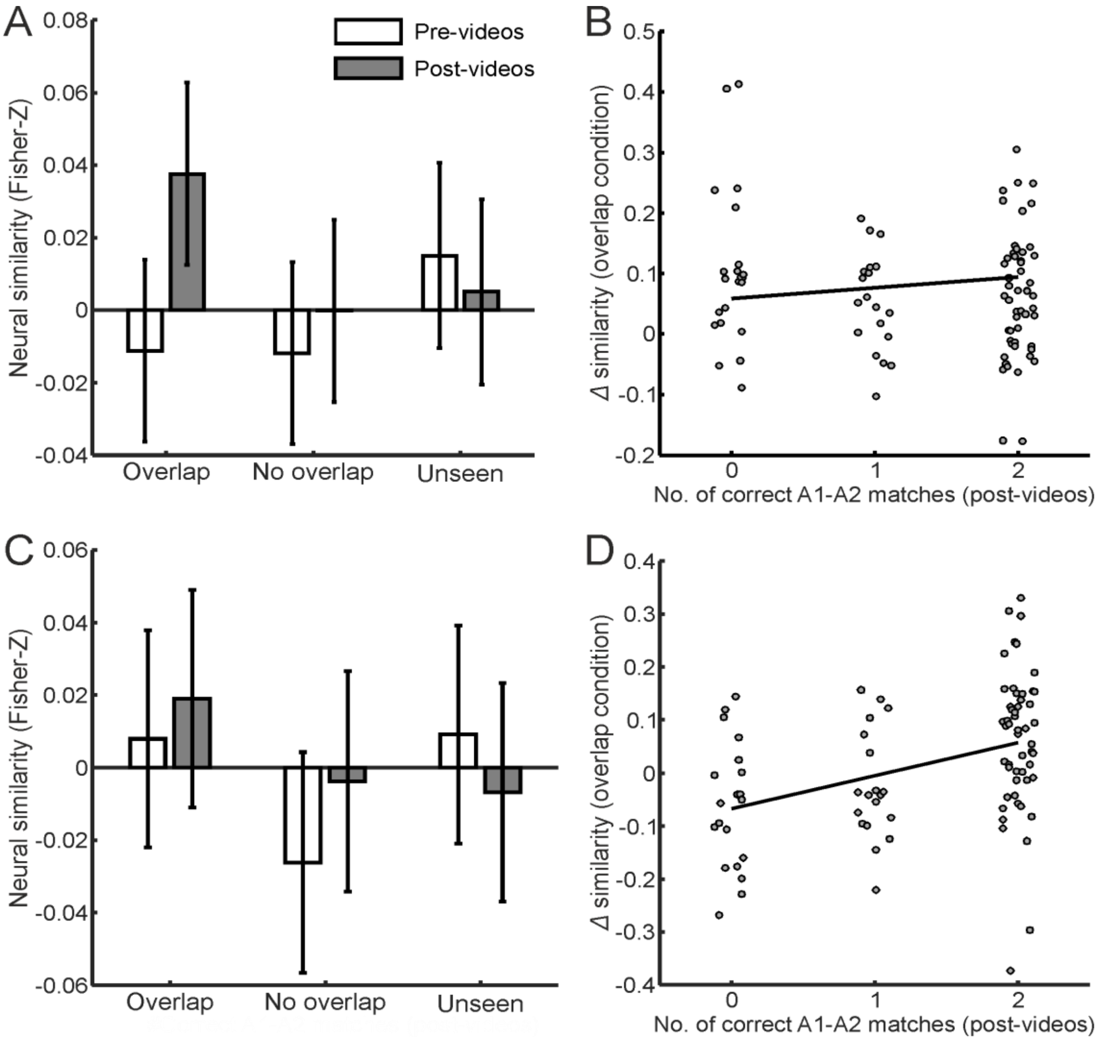

Figure 3. Results of the representational similarity analyses in the right parahippocampal cortex (PHC, top row) and right retrosplenial cortex (RSC, bottom row). **A:** PHC similarity estimates of scenes in the pre- and post-video sessions, plotted by experimental condition. There was a significant pre- to post increase in similarity estimates in the overlap condition (t_{2746} = 2.854, *p* = .004) that was not present in the no overlap and unseen conditions (t_{2746} = 0.678, *p* = .498, and t_{2746} = -0.584, *p* = .559 respectively). **B:** In the PHC, pre-video to post-video changes in representational similarity for the overlap condition plotted against the number of correct matches between overlap endpoints in the post-video behavioural task. This association was not significant (t_{2746} = 0.876, *p* = .386). **C & D:** Same as panels A and B but for the RSC region of interest. The RSC showed no overall similarity increases in any of the experimental conditions (t_{2746} = 0.539, t_{2746} = 1.086, and t_{2746} = -0.776 for the overlap, no overlap and unseen conditions respectively, all *p's* > .277). Nonetheless, there was a significant association between behavioural performance and similarity increases in the overlap condition (t_{2746} = 2.530, *p* = .011). All bars plot baseline corrected similarity estimates having subtracted out correlations between non-associated endpoints that were only accounted for by the intercept term in the mixed-effects model (e.g. A1-B1, A1-B2, etc.). Error bars indicate 95% confidence intervals.

The searchlight analysis that tested for a *Session*Overlap* interaction across the whole brain revealed one small cluster in the right inferior occipital gyrus (area V4; t_{21} = 4.78, p_{FWE} = .010, *k* = 38). However, when BOLD similarity in the cluster was modelled with the full mixed-effects analysis described above, the *Session*Overlap* effect was found to not be significant (t_{2746} = 1.532, *p* = .126, uncorrected, BF_{01} = 1.649). Model parameter estimates suggested that the searchlight effect was likely driven by below baseline BOLD similarity in the overlap condition prior to scanning, 95% CI: [-0.127, -0.0105].

Differentiating the PHC and RSC

We next assessed whether there was evidence for dissociable roles of the right PHC and RSC, given that both represented location-based information but were differently associated with behavioural performance. Specifically, we assessed whether the presence of location-based representations was significantly more associated with subsequent behaviour in the RSC than the PHC. This would imply that the RSC plays a greater role in guiding subsequent behaviour than the PHC. We therefore tested whether the *Session*Overlap*Behaviour* (3-way) effect was larger in the RSC than the PHC. A comparison of the effect size did show evidence for such a dissociation (t_{2746} = 3.535*,* p < .001).

This implies that the right PHC exhibited increased pattern similarity between A1 and A2 endpoints even when those endpoints were not subsequently remembered as belonging to the same location. We directly tested this by re-running the RSA having excluded A1/A2 pairs that were consistently remembered as belonging to the same location (i.e., having 2 correct responses during the post-scanner test). Despite these exclusions, the *Session*Overlap* interaction remained significant in the right PHC (t_{1190} = 2.504, p = .012). In contrast, the RSC only showed increased pattern similarity when the endpoints were consistently remembered as belonging to the same location. Re-running the RSA on these remembered pairs alone yielded a *Session*Overlap* effect that was not statistically significant in the RSC, but was in the expected direction (t_{1550} = 1.748, p = .081, BF_{01} = 1.234). This lack of sensitivity is likely attributable to two outlying data-points that had standardised residuals of -3.191 and -2.705, values that were considerably larger than all other residuals in the model (see figure 3D). Excluding these outliers yielded a significant *Session*Overlap* effect (t_{1548} = 2.183, p = .029).

In sum, we saw an increase in pattern similarity in the PHC and RSC for scenes from different viewpoints when shown in the same overlap video. This increased pattern similarity was only seen in the overlap condition, where scene endpoints were from the same spatial location. Importantly, we saw a dissociation between PHC and RSC. Whereas the PHC showed increased pattern similarity regardless of subsequent behaviour, the RSC only showed increased pattern similarity when participants were subsequently able to identify those scenes as belonging to the same location (outside of the scanner).

Univariate responses to endpoints

We investigated whether each of our ROIs produced univariate BOLD activations consistent with a *Session*Condition* interaction, or a 3-way interaction with behaviour. No such effects were found; all *F's* < 1.253, *p's* > .265. Furthermore, a mass univariate analysis testing for these effects at the whole brain level yielded no significant activations.

Univariate responses to videos

In a final analysis, we investigated whether univariate BOLD responses to the video clips differed between the overlap and no-overlap conditions. Here, first-level subtractions (overlap > no-overlap) were performed for each participant and the resulting contrast estimates were entered into a group-level random effects analysis. This revealed two clusters showing a significantly greater BOLD response to overlap videos (Figure 4, hot colours). The largest of these peaked in the medial prefrontal cortex and extended into the anterior cingulate, left frontal pole, and left middle frontal gyrus (t_{22} = 5.66, p_{FWE} < .001, k = 674). The second cluster peaked in the left supramarginal gyrus (t_{22} = 4.84, p_{FWE} = .008, k = 154), adjacent to a smaller, sub-threshold effect in the left angular gyrus.

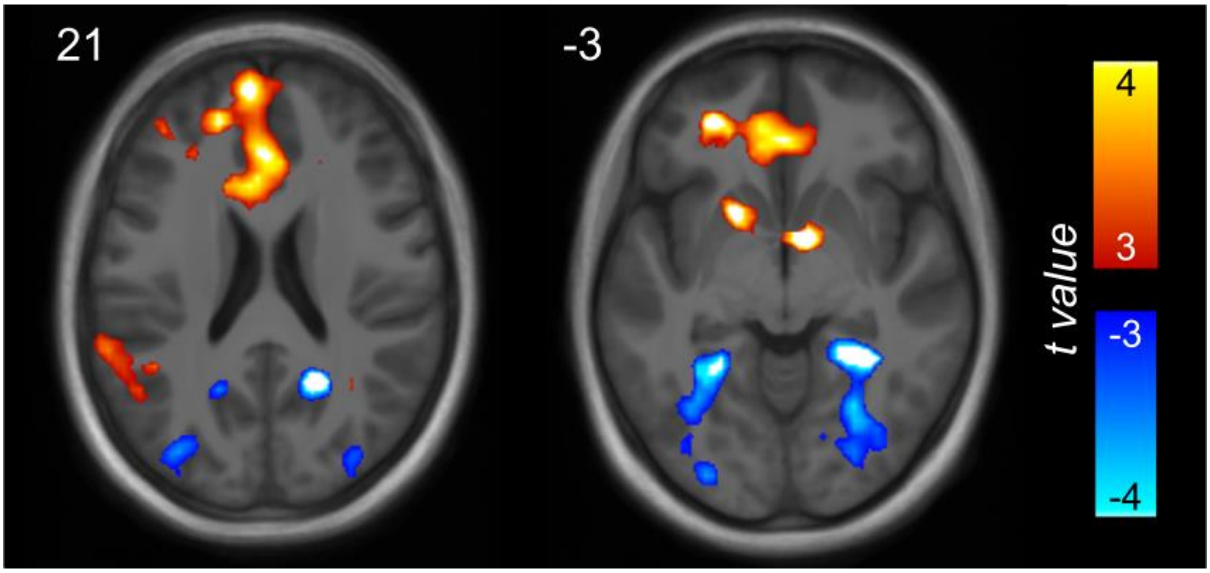

Figure 4. Univariate BOLD effects showing differences in activity between the two video conditions (thresholded at $t_{22} > 3$, $p < .003$ uncorrected). Hot colours indicate areas showing a greater response to overlap vs no-overlap videos. Cool colours indicate areas showing a greater response to no-overlap vs overlap videos. An unthresholded statistical map of this contrast is available at https://neurovault.org/collections/4819.

No effects for the reverse contrast (i.e., no-overlap > overlap) reached statistical significance at the whole-brain level. However, a small volume correction for the parahippocampal and retrosplenial cortices bilaterally revealed two significant clusters (Figure 4, cool colours). These were found in the right retrosplenial cortex (t_{22} = -4.95, p_{FWE} = .024, k = 32) and right parahippocampal cortex (t_{22} = -4.87, p_{FWE} = .022, k = 34, extending into the fusiform gyrus). Subthreshold effects in the left retrosplenial and parahippocampal cortices were also evident. These effects were also evident in an analysis that contrasted overlap and no-overlap BOLD estimates averaged across each ROI in native space. Here, both the right PHC and right RSC showed greater mean BOLD activity in the no-overlap video condition relative to the overlap condition (t_{42} = 3.638, p = .003 and t_{42} = 3.499, p = .004, respectively; corrected for multiple comparisons). Effects in the left PHC and left RSC were below threshold and considerably weaker (t_{42} = 1.828, p = .299 and t_{42} = 2.212, p = .130, respectively).

In sum, we saw greater activity in the medial prefrontal cortex during the overlap relative to no overlap videos. In contrast, the PHC and RSC, showed greater activity during the no-overlap relative to overlap videos. In other words, the medial posterior regions that showed *increased* pattern similarity following presentation of the overlap video showed *decreased* activity when participants were watching the videos.

Discussion

We show that scene-selective brain regions rapidly learn spatial representations of novel environments by integrating information across different viewpoints. Once participants observed the spatial relationship between different viewpoints from a given location, the right PHC and RSC maintained location-based, viewpoint-independent, representations. Whereas these location-based representations in PHC appeared regardless of whether participants could identify which scenes were from the same location, representations in RSC only emerged for scene pairs that participants could subsequently identify as being from the same location.

The results provide further evidence that PHC and RSC support spatial representations that are not solely driven by visual features in a scene (Marchette et al., 2015; Robertson et al., 2016; Vass & Epstein, 2013; c.f. Watson, Hartley, & Andrews, 2017). In particular, the RSC effect replicates that of Roberston et al. (2016) using a similar panoramic video manipulation, although we only see the effect when subsequent behaviour is taken into account. Further, they demonstrate: (1) that location-based representations can emerge rapidly (in a single scanning session) and (2) that PHC and RSC are dissociable in terms of their behavioural relevance. Although we were specifically interested in the emergence of viewpoint-independent spatial representations, a similar approach could be used to track the emergence of viewpoint-independent representations of other stimulus categories (e.g., objects or faces), opening the door to understanding how such representations are formed across the visual system.

These results demonstrate that we can track, using fMRI, the emergence of location-based representations in a single scanning session. The firing fields of place cells have been shown to emerge rapidly in the rodent hippocampus (Monaco, Rao, Roth, & Knierim, 2014). When placed on a circular track, individual locations where rats engaged in head-scanning behaviour (i.e., attentive, exploratory behaviour) on one run were associated with consistent place fields on the next run through the same location. Our results are consistent with this rapid emergence, providing evidence for location-based representations after only three learning exposures to the videos.

We also found that right RSC only exhibited location-based representations when participants were able to identify which scenes belonged to that location in a post-scanner test (PHC representations emerged regardless of subsequent behaviour). This implies that the ability to identify differing scenes as from the same location is more dependent on representations in RSC than PHC. Computational models hold that a network of medial posterior and temporal regions (including the PHC and RSC) perform complementary functions in support of spatial navigation and imagery (Bicanski & Burgess, 2018; Byrne et al., 2007). Specifically, PHC is thought to represent allocentric information related to the spatial geometry of the environment. Conversely, the posterior parietal cortex supports egocentric representations that allows the organism to actively navigate. The RSC transforms allocentric representations in the MTL into egocentric representations in the parietal cortex (and vice versa). Critically, the models predict that spatial navigation and planning is carried out in an egocentric reference frame. Thus, the RSC is critical to the translation of allocentric, to more behaviourally-relevant egocentric information.

Our behavioural task required participants to match distinct scenes from the same location. This task likely requires a transformation from the presented egocentric viewpoint, to an allocentric representation (ego-to-allocentric). Subsequently, the allocentric representation allows for the retrieval of the associated viewpoint from the same location (allo-to-egocentric). Consequently, RSC is likely to be more tightly coupled to behaviour relative to the PHC, as shown in the present data. This is because the RSC is required for both the ego-to-allocentric, and allo-to-egocentric, mappings necessary for the task. If only the allo-to-egocentric mapping is disrupted, the PHC can still show viewpoint-independence (because the ego-to-allocentric transformation can still occur). However, if the ego-to-allocentric mapping is disrupted, both the PHC and RSC will fail to show viewpoint-independence. Thus, failure to encode either direction of mapping between these reference frames will disrupt RSC representations (and behaviour), whereas the PHC can still demonstrate viewpoint-independence under some circumstances (even when there is no behavioural evidence for allocentric learning).

A related possibility is that participants engaged in active imagery of the associated scenes during the passive viewing task for specific locations, leading to subsequent improvements in behaviour for these locations. Note, that the task did not explicitly require memory retrieval; participants had to attend to the images and respond to odd-ball targets. As such, the activation of these representations do not appear to depend on any task-specific memory demands *per se*. It is possible that the retrieval of PHC representations (i.e., ego-to-allocentric mapping) occurs relatively automatically, consistent with the proposal that allocentric representations in the MTL are automatically updated during self-motion in an environment (Bicanski & Burgess, 2018; Byrne et al., 2007). However, the retrieval of associated egocentric representations (i.e., allo-to-egocentric mapping) may not occur automatically during passive viewing of scenes, consistent with the observation that viewpoint-independent representations in the RSC are abolished when participants engage in a task that prevents them from active retrieval of spatial information (Marchette et al., 2015). Importantly, both of the above accounts are consistent with the proposal that the RSC plays a critical role in mapping between allocentric and egocentric spatial representations.

Although we have provided evidence for location-based representations in both the PHC and RSC, it is not clear precisely what form such representations take. The PHC has been proposed to represent several complementary spatial representations, including geometric information regarding one's location in relation to bearings and distances to environmental features (e.g., boundaries; Park, Brady, Greene, & Oliva, 2011). The representations that we observed in PHC may reflect enriched spatial representations relating specific scenes to environmental features outside the current field-of-view. Also consistent with our results, PHC may represent spatial contexts more broadly (Epstein & Vass, 2014). The experimental manipulation used here could be modified to learn novel locations in the same spatial context, potentially dissociating between the above accounts. A further proposal is that viewpoint-independent representations in the PHC reflect prominent landmarks that are visible across viewpoints (Marchette et al., 2015). While this proposal yields similar predictions to the above accounts, it is perhaps less able to account for our finding of shared representations of views that did not contain any of the same landmarks.

RSC representations may reflect the retrieval of spatial or conceptual information associated with the environment (Marchette et al., 2015). Further evidence suggests that the RSC contains multiple viewpoint dependent and independent (Vass & Epstein, 2013), as well as local and global (Jacob et al., 2017; Marchette et al., 2014), spatial representations. This multitude of representations fits with the proposed role of the RSC as a transformation circuit, mapping between allocentric and egocentric representations. The heterogeneity of representations, relative to the PHC, may also be a further reason why we did not see clear evidence for location-based representations without taking behaviour into account. Regardless of the exact nature of such representations, our results provide clear evidence that we can track their emergence in both PHC and RSC.

We also examined activity during learning of new spatial relationships (i.e., when participants were watching the videos). BOLD activations in medial posterior brain regions were greater for no-overlap videos (depicting a spatially incoherent panorama) compared to overlap videos (depicting a complete panorama). This effect perhaps reflects greater fMRI adaptation during the overlap videos since they presented the central viewpoint of the panorama more frequently than no-overlap videos (Figure 1). However, it is interesting that the same cortical regions that showed *increased* pattern similarity following presentation of the overlap video showed *decreased* activity when participants were watching the videos. This underlines the complex relationship between univariate activity during learning, and resultant changes in patterns of activity following learning.

Additionally, we found that mPFC showed greater BOLD response in the overlap than no-overlap condition. This may reflect a mnemonic integration process that guides the learning of viewpoint-independent representations. Similar effects in mPFC have been observed in tasks that require integrating overlapping memories to support inference and generalisation (Milivojevic, Vicente-Grabovetsky, & Doeller, 2015; Schlichting, Mumford, & Preston, 2015). Indeed, mPFC has been suggested to operate as a mnemonic resonator, detecting new information that is congruent with previously learnt material so that it can be integrated into a generalised representation (van Kesteren, Ruiter, Fernández, & Henson, 2012). Our results are broadly in line with this proposal, where mPFC may be detecting the presence of overlapping spatial information during the overlap videos, resulting in the integration of previously learnt representations into more coherent viewpoint-independent representations in posterior medial regions (i.e., PHC and RSC). Despite this, our results do not exclude the possibility that mPFC activations reflect dis-inhibition from medial-posterior inputs (which showed reduced activity), or attentional differences related to the behavioural task.

We have shown that brain regions in the scene network, specifically right PHC and RSC, rapidly learn representations of novel environments by integrating information across different viewpoints. They appear to be relatively viewpoint-independent in that they become active regardless of which part of an environment is in the current field-of-view. We show that PHC and RSC have dissociable roles, with RSC playing a critical role in translating allocentric representations into a behavioural-relevant egocentric reference frame. Finally, our experimental approach allows for tracking the emergence of viewpoint-independent representations across the visual system.

Acknowledgments

We thank all the staff at the York Neuroimaging Centre (YNiC) for their assistance in running this project. We are also grateful to Tim Andrews and Tom Hartley for early discussions regarding experimental design. AJH is funded by the Wellcome Trust (204277/Z/16/Z) and ESRC (ES/R007454/1). BHJ is funded a PhD studentship awarded by the Department of Psychology, University of York.

References

Andersson, J. L. R., Hutton, C., Ashburner, J., Turner, R., & Friston, K. (2001). Modeling geometric deformations in EPI time series. *NeuroImage*, *13*(5), 903–919.

Ashburner, J. (2007). A fast diffeomorphic image registration algorithm. *NeuroImage*, *38*(1), 95–113.

Bicanski, A., & Burgess, N. (2018). A neural-level model of spatial memory and imagery. *ELife*, (7).

Burgess, N., Becker, S., King, J. A., & O'Keefe, J. (2001). Memory for events and their spatial context: Models and experiments. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *356*(1413), 1493–1503.

Byrne, P., Becker, S., & Burgess, N. (2007). Remembering the past and imagining the future: A neural model of spatial memory and imagery. *Psychological Review*, *114*(2), 340–375.

Calton, J. L., & Taube, J. S. (2009). Where am I and how will I get there from here? A role for posterior parietal cortex in the integration of spatial information and route planning. *Neurobiology of Learning and Memory*, *91*(2), 186–196.

Epstein, R. A., Patai, E. Z., Julian, J. B., & Spiers, H. J. (2017). The cognitive map in humans: Spatial navigation and beyond. *Nature Neuroscience*, *20*(11), 1504–1513.

Epstein, R. A., & Vass, L. K. (2014). Neural systems for landmark-based wayfinding in humans. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1635).

Hutton, C., Bork, A., Josephs, O., Deichmann, R., Ashburner, J., & Turner, R. (2002). Image distortion correction in fMRI: A quantitative evaluation. *NeuroImage*, *16*(1), 217–240.

Jacob, P. Y., Casali, G., Spieser, L., Page, H., Overington, D., & Jeffery, K. (2017). An independent, landmark-dominated head-direction signal in dysgranular retrosplenial cortex. *Nature Neuroscience*, *20*(2), 173–175.

Julian, J. B., Fedorenko, E., Webster, J., & Kanwisher, N. (2012). An algorithmic method for functionally defining regions of interest in the ventral visual pathway. *NeuroImage*, *60*(4), 2357–2364.

Julian, J. B., Keinath, A. T., Marchette, S. A., & Epstein, R. A. (2018). The Neurocognitive Basis of Spatial Reorientation. *Current Biology*, *28*(17), R1059–R1073.

Kass, R. E., & Raftery, A. E. (1995). Bayes Factors Bayes Factors. *Journal of the American Statistical Association ISSN:*, *90*(430), 773–795.

Marchette, S. A., Ryan, J., & Epstein, R. A. (2017). Schematic representations of local environmental space guide goal-directed navigation. *Cognition*, *158*, 68–80.

Marchette, S. A., Vass, L. K., Ryan, J., & Epstein, R. A. (2014). Anchoring the neural compass: Coding of local spatial reference frames in human medial parietal lobe. *Nature Neuroscience*, *17*(11), 1598–1606.

Marchette, S. A., Vass, L. K., Ryan, J., & Epstein, R. A. (2015). Outside Looking In: Landmark Generalization in the Human Navigational System. *Journal of Neuroscience*, *35*(44), 14896–14908.

Milivojevic, B., Vicente-Grabovetsky, A., & Doeller, C. F. (2015). Insight reconfigures hippocampal-prefrontal memories. *Current Biology*, *25*(7), 821–830.

Monaco, J. D., Rao, G., Roth, E. D., & Knierim, J. J. (2014). Attentive scanning behavior drives one-trial potentiation of hippocampal place fields. *Nature Neuroscience*, *17*(5), 725–731.

Oliva, A., & Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, *42*(3), 145–175.

Park, S., Brady, T. F., Greene, M. R., & Oliva, A. (2011). Disentangling Scene Content from Spatial Boundary: Complementary Roles for the Parahippocampal Place Area and Lateral Occipital Complex in Representing Real-World Scenes. *Journal of Neuroscience*, *31*(4), 1333–1340.

Robertson, C. E., Hermann, K. L., Mynick, A., Kravitz, D. J., & Kanwisher, N. (2016). Neural Representations Integrate the Current Field of View with the Remembered 360° Panorama in Scene-Selective Cortex. *Current Biology*, *26*(18), 2463–2468.

Schlichting, M. L., Mumford, J. A., & Preston, A. R. (2015). Learning-related representational changes reveal dissociable integration and separation signatures in the hippocampus and prefrontal cortex. *Nature Communications*, *6*, 1–10.

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., … Joliot, M. (2002). Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *NeuroImage*, *15*(1), 273–289.

Van Kesteren, M. T. R., Ruiter, D. J., Fernández, G., & Henson, R. N. (2012). How schema and novelty augment memory formation. *Trends in Neurosciences*, *35*(4), 211–219.

Vass, L. K., & Epstein, R. A. (2013). Abstract Representations of Location and Facing Direction in the Human Brain. *Journal of Neuroscience*, *33*(14), 6133–6142.

Watson, D. M., Hartley, T., & Andrews, T. J. (2017). Patterns of response to scrambled scenes reveal the importance of visual properties in the organization of scene-selective cortex. *Cortex*, *92*, 162–174.