Towards Spatial Memory of a Humanoid Robot

Laxmi R. Iyer, Lukáš Gajdošech, and Branislav Zigo

Faculty of Mathematics, Physics and Informatics, Comenius University
Bratislava, Slovakia
{gajdosech, laxmi, zigo}@fmph.uniba.sk

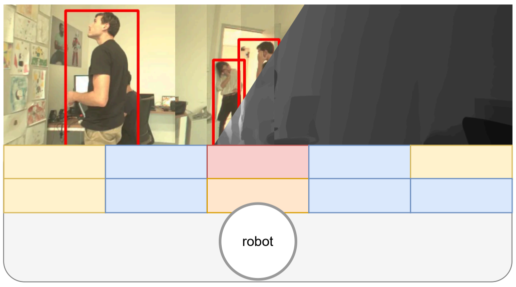

Fig. 1: Reconstructed RGB and depth panorama image with example spatial grid.

Abstract

In this work, we propose a novel approach to spatial memory in humanoid robotics. We imitate the allocentric spatial memory mechanism in humans by constructing a top-down spatial memory grid that estimates room occupancy and tracks human presence over time. Our method leverages recent visual foundational models. The robot captures a sequence of RGB images while rotating its neck joints, which are stitched into a panoramic view. Depth information is extracted and combined with person detection results to project spatial coordinates onto a 2D grid. This is inspired by the representation of space as an allocentric map in human spatial memory. Each grid cell maintains a confidence score reflecting the likelihood of human presence, decaying over time if the area is not observed. The decaying of confidence mimics forgetting in human memory. Our approach is evaluated using a fisheye camera for ground truth annotation, demonstrating robust occupancy estimation even in dynamic environments. Compared to traditional facial tracking and re-identification methods, our system provides enhanced adaptability and resilience to occlusions by relying on depth-aware spatial mapping.

1 Methodology

The robot captures a sequence of RGB images while rotating its neck joints, covering a horizontal field of view of 120°. These images are stitched into a panoramic view $\mathbf{I}_p$ by mapping yaw and pitch angles to pixel coordinates:

$$x = \left(\frac{\theta - \theta_{min}}{\theta_{max} - \theta_{min}} \right) W,$$

$$y = \left(1 - \frac{\phi - \phi_{min}}{\phi_{max} - \phi_{min}} \right) H,$$

where θ and ϕ are the neck yaw and pitch angles, and W and H are the panorama width and height. Depth estimation is performed using the DepthAnythingV2 model (Yang et al., 2024), or a similar foundational model for depth prediction, producing a depth map $\mathbf{D}$ for each image. Person detection is carried out using YOLO-World (Cheng et al., 2024), and segmentation is performed using the Segment Anything Model (SAM), providing pixel-level masks for each detected person. The stitched panorama $\mathbf{I}_p$ and depth map $\mathbf{D}$ are projected into a top-down spatial grid $\mathbf{G} \in \mathbb{R}^{M \times N}$, where each grid cell G_{ij} corresponds to a spatial bin in the room. Each grid cell maintains a confidence score $C_{ij}(t)$, which is initialized when a person is detected and decays over time if the area is not observed:

$$C_{ij}(t) = P_{ij} \left(0.5e^{-\lambda(t-t_0)} + 0.5 \right) +$$

$$(1 - P_{ij}) \left(-0.5e^{-\lambda(t-t_0)} + 0.5 \right),$$

where P_{ij} is the initial detection probability, t is the current time, t_0 is the timestamp of the last observation, and λ is the memory decay rate. The memory decay rate is based on psychological principles.

2 Spatial Memory in the Brain

There are a constellation of cells in the brain that are used in spatial memory, the most prominent being called *place cells*. Over fifty years ago, in 1971, the discovery of place cells by O'Keefe marked a breakthrough in spatial cognition. Now, the neural mechanisms underlying an understanding of space are relatively well understood.

There are two kinds of spatial representations in the brain - egocentric and allocentric. In an egocentric representational framework, locations are represented based on their relationship with the subject. For example, as a person, say John is present in the room and

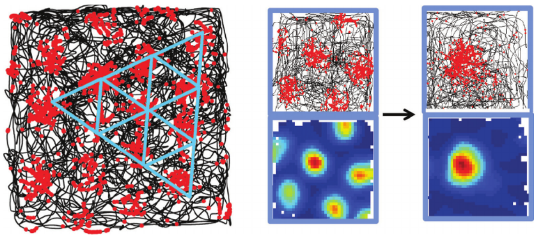

Fig. 2: Reproduced from May-Britt Moser a Moser (2024), Figure 1. Grid cells and place cells. (Left) A grid cell from the entorhinal cortex of the rat brain. The black trace shows the trajectory of a foraging rat in part of a 1.5-m-diameter-wide square enclosure. Spike locations of the grid cell are superimposed in red on the trajectory. Each red dot corresponds to one spike. Blue equilateral triangles have been drawn on top of the spike distribution to illustrate the regular hexagonal structure of the grid pattern. (Right) Grid cell and place cell. (Top) Trajectory with spike locations, as in the left part. (Bottom) Color coded rate map with red showing high activity and blue showing low activity. Grid cells are thought to provide much, but not all, of the entorhinal spatial input to place cells.

looks around, the representations are based on their relationship with John. As John moves, all the locations are updated in relation to John. If John and the objects move around, all relative positions are to be updated presenting a large computational overhead.

To combat this, there is another form of representation in the brain - the allocentric representation. Here, distances are encoded with respect to other objects independent of the observer. In the hippocampus, each place cell represents a particular area in space such that a local population of place cells together encode the environment that the rat is in. Place cells do not just encode place information but are able to link place to a variety of other features in the environment. Figure 2 shows the place cells corresponding to a maze that the rat is in. In addition, the brain has grid cells in the endorhinal cortex which also encode place in a different manner which is out of the scope of this paper, head direction cells which fire when the animal is facing a particular direction, and boundary cells, cells that fire when an animal is in the boundary of an environment. Together these cells can provide a means for an animal to navigate a path.

Our grid occupancy matrix is inspired by the place cells in the hippocampus. Of course, the place cells are used for mapping the self to the map, but as they are also used for mapping other features, we feel it is a good estimation. As seen in Figure 2, the place cells roughly encode a grid occupancy matrix of the environment.

3 Evaluation

The proposed spatial memory approach will be evaluated using a wide, fisheye camera, which will record the scene from the back of the room. Ground truth annotations will be created manually by labeling the positions of individuals at different times, providing precise occupancy data. This evaluation will be part of our future work focusing on the whole pipeline for multiparty conversational agent.

Traditional tracking techniques, such as ConfTrack, utilize Kalman filters for multi-person tracking, relying on bounding boxes and identity re-identification (Jung et al., 2024). These methods often struggle with occlusions and require consistent visibility of individuals. Conversely, EyeEcho tracks facial expressions using acoustic signals and specialized hardware, limiting its applicability in dynamic and crowded environments (Li et al., 2024). Our method's depth-aware spatial mapping and confidence decay model allow for robust occupancy estimation without relying on consistent facial visibility, making it more versatile.

4 Conclusion

Our proposed approach introduces a cognitive-inspired allocentric spatial memory model intended to improve occupancy estimation and robustness to occlusions in dynamic environments. Future implementation and thorough evaluation of this spatially-aware mechanism could significantly advance the natural interaction capabilities of humanoid robots in shared spaces.

References

Cheng, T., Song, L., Ge, Y., Liu, W., Wang, X. a Shan, Y. (2024). Yolo-world: Real-time open-vocabulary object detection. *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16901–16911.

Jung, H., Kang, S., Kim, T. a Kim, H. (2024). Conf-track: Kalman filter-based multi-person tracking by utilizing confidence score of detection box. *Winter Conference on Applications of Computer Vision (WACV)*, pp. 6569–6578.

Li, K., Zhang, R., Chen, S., Chen, B., Sakashita, M., Guimbretière, F. a Zhang, C. (2024). EyeEcho: continuous and low-power facial expression tracking on glasses. *CHI Conference on Human Factors in Computing Systems*.

May-Britt Moser, D. C. R. a Moser, E. I. (2024). Place cells, grid cells, and memory. *CHI Conference on Human Factors in Computing Systems*.

Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J. a Zhao, H. (2024). Depth anything v2. *arXiv:2406.09414*.