Who Was Where: Natural Language Verbalization of Localized Persons from a Humanoid Robot's Episodic Memory

Joana Plewnia and Tamim Asfour

***Abstract*— The ability to verbalize past locations of people is essential for natural human-robot interaction in multi-person environments. We present a pipeline enabling the humanoid robot ARMAR-7 to answer person-related queries about its past, such as "Where did you last see Joana?". Our approach combines *OpenPose* for pose detection, *InsightFace* for face recognition, and a memory-based cognitive architecture to create person instances with symbolic spatial representations integrated with an existing verbalization framework. Evaluation in a laboratory environment demonstrates successful verbalization. Analysis reveals that verbalization quality depends critically on face recognition accuracy. We discuss challenges and future directions for person-aware spatial verbalization.**

I. INTRODUCTION

The ability to verbalize spatial and temporal information is crucial for effective human-robot interaction, enabling robots to communicate their intentions, explain past actions, and share their understanding of the environment. Recent advances in natural language generation and embodied AI have led to significant progress in robot verbalization capabilities, particularly for object- or action-related queries (e. g. [1], [2], [3], [4]). Several such works have addressed the challenge of verbalizing object locations and their spatiotemporal history, including systems being able to answer questions regarding where objects were last seen ([3], [4]).

However, as robots are increasingly deployed in real-world environments involving multiple humans, such as nursing homes, offices, or domestic settings, the ability to track and verbalize information about people becomes equally essential. Unlike objects, which can often be identified through prior knowledge databases or modern vision-language models (VLMs), tracking humans introduces distinct challenges. The robot system must integrate face recognition and human localization with symbolic spatial representations, such as "at the kitchen counter", while maintaining consistency with the representations in the robot's memory.

In this work, we present a pipeline that enables the humanoid robot ARMAR-7 to answer person-related spatial questions such as *"Where did you last see Joana?"* by addressing these challenges. Our approach combines face recognition, spatial reasoning, and natural language generation within the robot's cognitive architecture [5], and extends our episodic memory and verbalization framework [4] to handle verbalization of person-related experiences. While the robot could previously verbalize objects and actions, it can now also verbalize information about its experience with people. We demonstrate the system's functionality, discuss deployment results, and identify open challenges and future research directions in person-aware spatial verbalization. Our main contribution is demonstrating the extensibility of our verbalization framework [4] through the use of appropriate symbolic memory aggregation with minimal modifications.

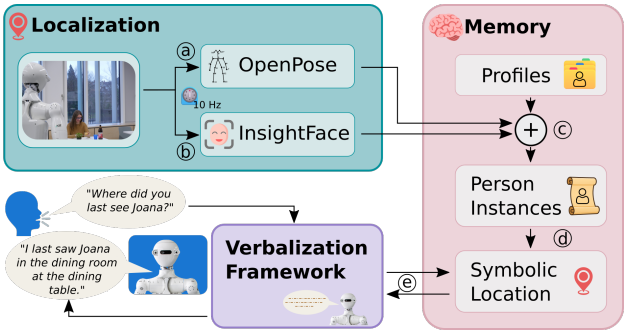

Fig. 1. **Person-aware spatial verbalization pipeline.** The system processes visual input through ⓐ *OpenPose* and ⓑ *InsightFace* for localization and identification, then ⓒ aggregates this with memory profiles to create Person Instances. These are enriched with symbolic spatial representations of the environment ⓓ to create symbolic localizations of a person. When queried (e), the verbalization framework retrieves relevant instances and generates natural language responses combining symbolic and metric spatial information.

II. BACKGROUND

Our approach builds upon a previously developed memory-centric cognitive architecture [5] and verbalization framework [4]. The memory system [5] serves as the central *active* component mediating between high-level symbolic reasoning and low-level sensorimotor representations. Its hierarchical organization—spanning memory servers, entity types, and individual instances—combined with episodic timestamping, supports both spatial and temporal queries about the robot's past experiences, and enables queries about both what the robot perceived and when it perceived it. The verbalization framework [4] converts these internal memory representations into natural language responses through question definitions. Each question definition specifies three strategies: Selection (identifies relevant entities via memory IDs), Search (determines temporal traversal of snapshots), and Evaluation (assesses snapshot relevance and extracts information). Question Templates with placeholders match user queries, while Answer Templates construct responses from retrieved memory data. This framework allows the

This work has been supported by the European Union's Horizon Europe Widening Program through the HERON project, the Carl Zeiss Foundation through the JuBot project, and the German Federal Ministry of Research, Technology and Space (BMFTR) under the Robotics Institute Germany (RIG).

The authors are with the High Performance Humanoid Technologies Lab, Institute for Anthropomatics and Robotics, Karlsruhe Institute of Technology (KIT), Germany E-mails: {joana.plewnia,asfour}@kit.edu.

robot to answer questions about objects and actions by systematically querying its episodic memory and generating contextually appropriate natural language responses.

III. APPROACH

A. Method

As illustrated in Figure 1, our pipeline for person-aware spatial verbalization integrates perception, memory, and natural language generation within ARMAR-7's cognitive architecture. The system begins by detecting humans in the robot's field of view and estimating their 3D positions relative to the robot's coordinate frame, providing the spatial foundation for tracking person locations throughout the environment ⓐ. Once humans are detected, the robot identifies them through face recognition ⓑ, associating detected individuals with known person identities stored in its memory system. Building on this perceptual information, the system aggregates data from different memory segments to create unified person instances that combine identity, spatial location, and temporal information ⓒ. These person instances are then integrated with the robot's broader understanding of its environment through symbolic spatial representations that combine information about rooms, objects in the scene, and person instances ⓓ, enabling higher-level descriptions beyond raw coordinates, such as "near the table" or "in the kitchen." When a question about a person's location is posed (e. g., "Where did you last see Joana?"), the system retrieves the most recent information about that person from memory and generates a natural language response that combines symbolic information (room, proximity to furniture) with subsymbolic details (metric distances, relative positions) to produce human-understandable spatial descriptions ⓔ.

B. Implementation

For human pose detection and localization ⓐ, we employ *OpenPose* [6] to extract human poses from the robot's RGB-D camera stream, calculating 3D positions in the global coordinate frame. For face recognition ⓑ, we utilize the *buffalo_l* model from *InsightFace* [7] , which employs RetinaFace-10GF [8] for face detection and a ResNet50 model [8] trained on the WebFace600K dataset [9] for recognition. Face recognition runs continuously at 10Hz, with 3D face positions determined by computing the median depth value across face region pixels. The robot maintains person profiles in its memory, each containing at least one reference headshot used for recognition. The aggregation of person instances ⓒ is triggered whenever new pose or face recognition data becomes available. Person instances combine a person's identity from memory profiles, 3D position from *OpenPose*/depth data, and time information into a unified symbolic entity representation. The system maintains unique tracking IDs, provided by *OpenPose*, to handle pose updates even when face recognition is temporarily unavailable. Symbolic spatial information ⓓ is automatically updated each time new person location data is received, calculating the room containing the person, the nearest object-centric location, and the distance to that location. For verbalization ⓔ, we extend the framework proposed in [4] by defining a new *QuestionDefinition* for questions about past or current locations of persons. The system constructs responses based on available information granularity: "I have last seen Joana in the living room." when only room information is available, or "I have last seen Joana in the kitchen, close to the fridge." when object-centric location data enriches the description.

IV. EARLY RESULTS

We evaluated our pipeline in a controlled laboratory environment with three spatial regions (kitchen, living room, dining area) containing typical household objects and appliances (dishwasher, kitchen counter, dining table, etc.). The evaluation consisted of 40 queries about 8 different persons, with episodic memory consolidated over a 2-week period. Questions were posed up to 2 hours after each person was last observed.

The general pipeline successfully integrates all components and produces verbalizations for person location queries. To assess verbalization quality, we manually compared the system's generated responses against ground truth observations, categorizing results into five classes: correct (both room and object-centric information accurate) 47.5%, correct but only room information (room accurate, no object-centric detail provided) 45%, incorrect object information (room correct but object reference wrong) 2.5%, and incorrect (fundamental location error) 5.0%. Our analysis reveals that the verbalization framework's accuracy is heavily dependent on the underlying perceptual components. Face recognition presents the most significant challenge, occasionally producing false positives where people are detected despite no one being present. Such errors propagate through the pipeline and result in incorrect verbalizations. In contrast, human pose-based localization proved relatively stable and robust across different scenarios. From a computational perspective, the integrated components operate efficiently enough to support human detection at approximately 10Hz, providing sufficiently frequent updates for real-time person tracking and verbalization in typical interaction scenarios.

V. CONCLUSION AND FUTURE WORK

We presented a pipeline that extends robot verbalization capabilities from object-centric to person-aware spatial descriptions, enabling humanoid robots to answer questions such as "Where did you last see Joana?". Our approach demonstrates that existing verbalization frameworks, such as the one presented in [4], can be effectively adapted to handle person location queries by integrating face recognition and human tracking with symbolic spatial reasoning. Early results confirm the viability of this approach, though verbalization quality remains fundamentally tied to the reliability of underlying perception systems, particularly face recognition. Future work should focus on improving person detection robustness, introduce strategies for handling data privacy issues, and enrich the verbalization by conveying information about uncertainties in the recognition and localization process.

References

[1] C. DeChant, I. Akinola, and D. Bauer, "Learning to summarize and answer questions about a virtual robot's past actions," *Autonomous robots*, vol. 47, no. 8, pp. 1103–1118, 2023.

[2] A. Anwar, J. Welsh, J. Biswas, S. Pouya, and Y. Chang, "Remembr: Building and reasoning over long-horizon spatio-temporal memory for robot navigation," *arXiv preprint arXiv:2409.13682*, 2024.

[3] L. Bärmann, C. DeChant, J. Plewnia, F. Peller-Konrad, D. Bauer, T. Asfour, and A. Waibel, "Episodic memory verbalization using hierarchical representations of life-long robot experience," in *IEEE/RAS International Conference on Humanoid Robots (Humanoids)*, 2025, pp. 783–790.

[4] J. Plewnia and T. Asfour, "Combining episodic memory and llms for the verbalization of robot experiences," in *IEEE/RAS International Conference on Humanoid Robots (Humanoids)*, 2025, pp. 531–538.

[5] F. Peller-Konrad, R. Kartmann, C. R. G. Dreher, A. Meixner, F. Reister, M. Grotz, and T. Asfour, "A memory system of a robot cognitive architecture and its implementation in armarx," *Robotics and Autonomous Systems*, vol. 164, pp. 1–20, 2023.

[6] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh, "Openpose: Realtime multi-person 2d pose estimation using part affinity fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[7] J. Deng, J. Guo *et al.*, "InsightFace: 2d and 3d face analysis project," https://github.com/deepinsight/insightface, 2022, accessed: 2025-01-09.

[8] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[9] Z. Zhu, G. Huang, J. Deng, Y. Ye, J. Huang, X. Chen, J. Zhu, T. Yang, D. Du, J. Lu *et al.*, "Webface260m: A benchmark for million-scale deep face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 2, pp. 2627–2644, 2022.