Towards Adaptive Assistance: A Preliminary Architecture for Dynamic User Profiling in Social Robots

Ritesh Sharma[1,*], Allen Marshall[1], David Montgomery[1] and Rose Gamble[1,2]

[1] *Institute for Robotics and Autonomy, The University of Tulsa, Tulsa, Oklahoma, United States*
[2] *Tandy School of Computer Science, The University of Tulsa, Tulsa, Oklahoma, United States*

Abstract

As assistive robots become more common in human-shared environments, there is a growing need for a robust interaction framework that supports personalization, adaptability, and context-awareness. This work-in-progress paper presents a preliminary framework using the Hello Robot Stretch 3 platform to explore dynamic user profiling for personalized, context-aware assistance. The approach integrates four core modules: real-time scene analysis using deep neural networks for object detection and localization; persistent user profiling through facial recognition and emotion analysis using the DeepFace framework; navigation control; and a Large Language Model (LLM)-based conversational interface. The main purpose of these modules is to enable the robot to recognize individuals, learn their preferences for verbal interaction, and provide contextual assistance through intelligent navigation and object location services. Initial implementation demonstrates promising results in structured indoor environments, although challenges such as processing latency and environmental complexity remain. An initial evaluation was performed for object mapping, face detection, and emotion recognition to test the system, with experimentation of the system capabilities currently limited to research staff. Nevertheless, this early-stage work lays the foundation for future development in adaptive assistive robotics.

Keywords

Social Robotics, User Profiling, LLM Integration, Personalized HRI, Assistive Navigation

1. Introduction

The integration of robots with artificial intelligence (AI) is an exciting and important area of research, especially when it comes to designing systems that can assist people in personalized and meaningful ways. In recent years, researchers have emphasized the importance of making human-robot interaction (HRI) more adaptive by allowing robots to recognize and respond to individual user needs. For example, surveys on user profiling and behavioral adaptation in HRI have shown that people expect robots to detect with whom they are interacting and adjust their behavior to match user preferences, communication styles, or emotional states [1]. These adaptations are essential for building trust and maintaining engagement over longer periods [2].

Despite these advances, many existing robotic systems remain limited in their ability to offer truly integrated support. In particular, tasks like scene understanding (e.g., identifying objects and locations), user modeling (e.g., tracking user identity and preferences), and conversational interaction are often handled in separate modules. For example, vision-language models for social navigation [3], proxemic-aware navigation systems [4], person following behaviors [5], large language models for robotics applications [6], emotion detections [7], and active learning based user profiling systems [8] have all shown effectiveness in their respective domains. However, integrating these diverse components into cohesive and well-coordinated systems that can operate effectively in dynamic real-world environments remains a significant challenge [9]. Current methods usually treat these capabilities in isolation: spatial mapping focuses only on understanding the environment, user profiling works without considering

Workshop on Social Robotics for Human-Centered Assistive and Rehabilitation AI (a Fit4MedRob event) - ICSR 2025

[*] Corresponding author.

ris7188@utulsa.edu (R. Sharma); allen-marshall@utulsa.edu (A. Marshall); dmm3950@utulsa.edu (D. Montgomery); gamble@utulsa.edu (R. Gamble)
https://sharmrit.github.io/Homepage (R. Sharma); https://utulsa.edu/people/rose-gamble/ (R. Gamble)
0000-0003-1160-3918 (R. Sharma); 0000-0002-8423-3172 (A. Marshall); 0009-0002-6040-3090 (D. Montgomery); 0000-0003-3206-5557 (R. Gamble)

spatial context, and navigation often lacks personalization. As a result, the robot may struggle to provide consistent and context-aware assistance, especially in situations where understanding both the environment and the user at the same time is essential.

To address this gap, we propose a unified framework that bridges recent advances in large language models (LLMs), deep learning-based object detection, user profiling, and adaptive social navigation into a single integrated system. Using the Hello Robot Stretch 3 as an experimentation and demonstration platform, our approach builds on the capabilities of existing deep learning models and OpenAI's LLMs to perform real-time scene analysis and user profiling, respectively. The proposed framework allows the robot to automatically create spatial maps of static objects in the environment, such as chairs, tables, or medical equipment, while simultaneously detecting and profiling human users. The system considers conversational norms like appropriate interpersonal distance, drawing on research in proxemics [4]. The system collects user history to personalize interactions and target support during tasks, such as object search or location assistance.

The main contribution of this early work is the design and testing of a unified framework to enable robots to offer intelligent, adaptive, and personalized assistance by integrating three core capabilities. First, it combines deep learning-based scene analysis with LLM-driven user profiling, allowing for continuous updates to the robot's understanding of the environment and individual preferences. Second, it includes a personalized dialogue system that adapts speech content and tone based on prior or current user interactions, improving engagement and accessibility. Third, the framework incorporates a context-aware navigation module with spatial memory, enabling the robot to assist users in locating and reaching objects or destinations while avoiding static and moving obstacles. These components together form a framework for a cohesive, human-centered system that supports more natural and effective human-robot interaction.

2. System Architecture

Our proposed system architecture consists of four integrated modules that work together to provide personalized context-aware assistance: (1) Scene Analysis Module, (2) User Profiling System, (3) Adaptive Navigation Controller and (4) Personalized Conversation Engine. Figure 1 illustrates the overall system architecture.

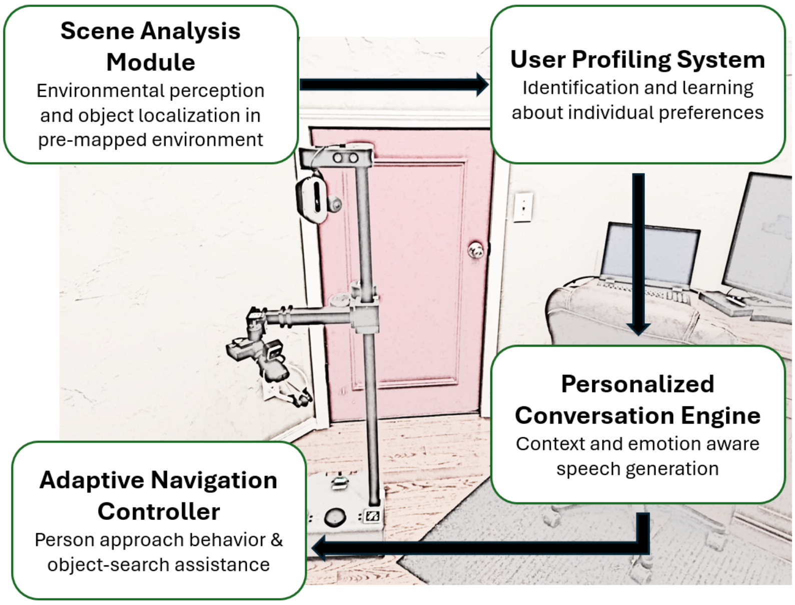

Scene Analysis Module
Environmental perception and object localization in pre-mapped environment
User Profiling System
Identification and learning about individual preferences
Personalized Conversation Engine
Context and emotion aware speech generation
Adaptive Navigation Controller
Person approach behavior & object-search assistance

Figure 1: System Architecture showing four integrated modules for assistive robots.

2.1. Scene Analysis Module

The Scene Analysis Module operates within a pre-mapped environment and uses deep learning models to enable continuous perception. As the robot autonomously explores its surroundings, it captures visual input and processes it through Hello Robot's existing Stretch Deep Perception module [10] to detect both human occupants and environmental features. Stretch Deep Perception is a deep learning module using YOLOv5 and OpenVINO models for object and face detection, respectively. The Scene Analysis Module performs two core functions. The first core function scans the environment to identify approachable individuals in the scene.

The second core function simultaneously identifies and catalogs static elements in the scene, such as furniture (e.g., "table", "chair"), architectural features (e.g., "door", "kitchen"), and semantic waypoints (e.g., "exit"). All identified elements are recorded into a persistent spatial memory map (in JSON format) representing observations from the past to enable short-term spatial reasoning. This map serves as a foundation for object search, spatial reasoning, and navigation assistance. Importantly, the system remembers what it has recently seen, allowing it to answer user queries like "Have you seen my medicine recently?" or guide users by navigating to the observed objects when asked "Can you take me to the nearest chair?".

2.2. User Profiling System

The User Profiling System manages individual user identification and preference learning through a dynamic persistent storage mechanism. Each user profile contains identification parameters, interaction history, and personalized interaction preferences. When the system encounters a new individual, it initiates profile creation, gathering identification information and establishing baseline interaction preferences using approaches similar to those developed for robot-human personality matching in rehabilitation contexts [11]. For known users, the system retrieves existing profiles and updates them based on new interactions. It analyzes user communication patterns and adjusts its own speech style and interaction pace to match individual preferences, following established principles for affective-sensitive companion systems [12]. This personalization can be extended to content selection, with the system learning which types of information and assistance each user finds most valuable.

2.3. Adaptive Navigation Controller

The Adaptive Navigation Controller manages the robot's physical movement while integrating collision avoidance for static and dynamic obstacles using 2D LiDAR mounted on the Hello Robot Stretch 3 platform. Whenever an obstacle is detected, a replanning request is sent to the path planner, which responds with a new path to help the robot avoid obstacles.

One of the important functions of this module is to approach humans detected by the Scene Analysis Module while respecting social proxemics norms [4]. When approaching a human, the robot moves to an appropriate conversational distance before activating the interaction protocol.

The robot relies on its spatial memory and real-time spatial information to navigate to the user's desired location. When a user requests help locating an object or reaching a previously seen destination, the controller queries the environmental map built by the Scene Analysis Module. The robot can either provide verbal directions or physically lead the user to the target, asking them to follow. This navigation support is informed by the robot's memory of recent visual input and the precomputed map, enabling it to respond intelligently to commands such as "Can you take me to the chair you saw earlier?".

2.4. Personalized Conversation Engine

The Personalized Conversation Engine acts as the primary interface for human-robot communication and is built on top of customized LLMs derived from OpenAI's foundation model [13]. It maintains continuity across sessions by referencing past interactions and adapting responses to the user's current context and preferences. The engine draws from the User Profiling System to align its tone and conversational structure to the individual, as indicated in the research surveyed in [1].

The system improves its responsiveness through real-time emotion detection using DeepFace analysis [14, 15], which processes facial expressions to identify emotional states based on confidence scores for detected faces. To demonstrate the usability of our framework, we targeted the emotional states of “Happy”, “Anger”, “Sad” or “Fear” for their relevance in social interaction. Studies have shown that happiness, sadness, and anger are the most reliably recognized emotional states, while fear is often misinterpreted as anxiety or surprise—highlighting the practical trade-offs involved in emotion selection [16]. Additionally, limiting the number of emotional categories has been found to improve detection accuracy, supporting the use of a small, well-separated set in real-time HRI systems [17]. Once emotions are detected, the conversation engine employs a two-step modification process that first adjusts the response of the base language model through emotion-specific prompt engineering, modifying word choices and sentence structures to match the detected emotional state. It then modifies text-to-speech parameters, such as speaking rate and vocal emphasis, accordingly.

In addition to handling everyday queries and task instructions, the system incorporates domain-specific protocols, such as those found in [18], to deliver more specialized assistance. Its flexibility allows the robot to converse naturally, providing not just general interaction but targeted, useful responses aligned with the user’s needs and the environment’s current state.

3. Integrated Functionality and Implementation

Our assistive robotic framework is designed to provide context-aware, personalized support by integrating user recognition, environmental awareness, adaptive conversation, and real-time navigation. It continuously scans the pre-mapped environment to detect human faces and approaches individuals to initiate interaction. For known individuals, it offers personalized interactions based on stored profiles, addressing them by name and referencing prior engagements. For new users, the system begins by collecting information and building a user profile, setting the foundation for future personalized exchanges.

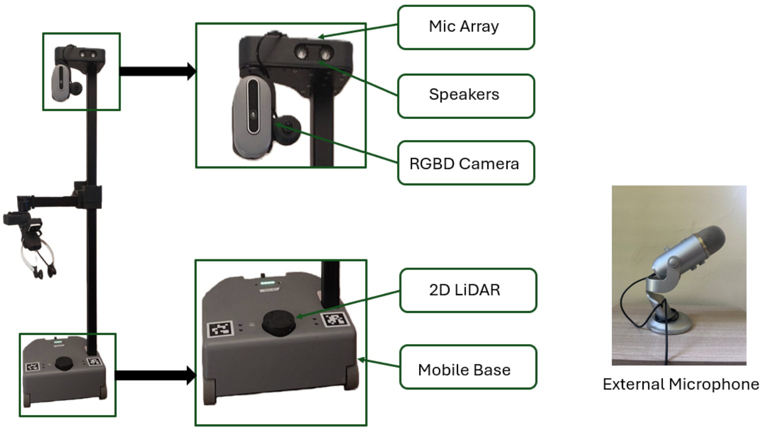


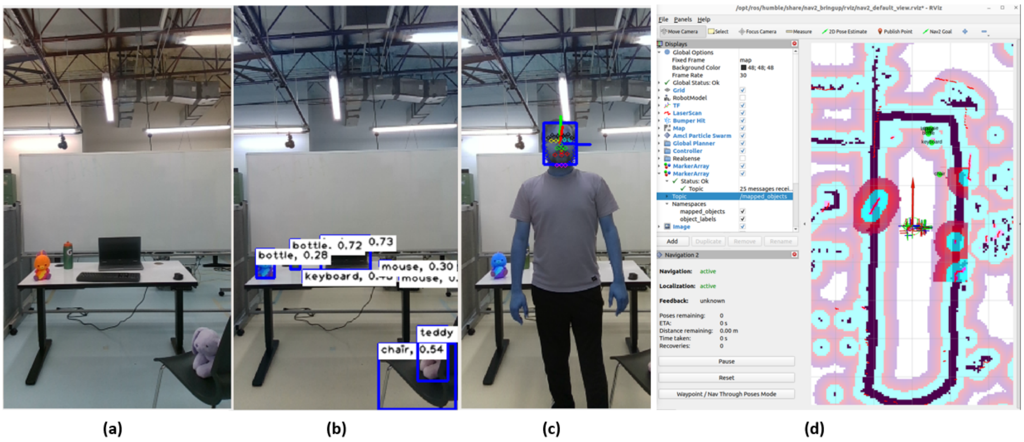
Figure 2: Hello Robot Stretch 3 platform and external studio grade microphone.

Our system is built on the commercially available Hello Robot Stretch 3 platform (Figure 2), a compact and lightweight mobile manipulator. It features a differential-drive mobile base equipped with a 2D LiDAR for localization and obstacle avoidance, a head-mounted Intel RealSense D435if RGBD camera, a microphone array, and speakers to support multimodal perception and user interaction [19]. To enhance speech recognition, a studio-grade external microphone is integrated. These sensing and interaction modalities, combined with the onboard computer (Intel NUC 12), make the system ideal for implementing the initial architecture, as they allow the system to maintain situational awareness; store spatial memory of objects and environmental features; and engage in context-aware, speech-based navigation and assistance. Our proposed framework allows the robot to respond to spatial queries (for example, “Where did you see the coffee mug?”) and supports personalized conversations grounded in prior interactions and user-specific patterns.

4. Evaluations

To assess the architecture, we conducted capability evaluations in three main categories: (1) object detection and mapping, (2) face detection, and (3) emotion recognition for personalized interaction.

Object Detection and Mapping: Figure 3 shows our initial assessment of the object detection and mapping pipeline, starting from the raw frame (Figure 3(a)). When the raw frame passes through the deep perception module, bounding boxes and confidence scores are generated for the identified objects as shown in Figure 3(b). The system produces more detections than the actual number of objects. To address this, confidence-based filtering is applied, where each object category is assigned a threshold below which detections are excluded from the mapping process.



Global Options
Fixed Frame map
Background Color 48; 48; 48
Frame Rate 30
Global Status: Ok
Grid ☑
RobotModel ☐
TF ☑
LaserScan ☑
Bumper Hit ☑
Map ☑
Amcl Particle Swarm ☑
Global Planner ☑
Controller ☑
Realsense ☐
MarkerArray ☑
MarkerArray ☑
Status: Ok
Topic 25 messages recei...
Topic /mapped_objects
Namespaces
mapped_objects ☑
object_labels ☑
Image ☑

Navigation: active
Localization: active
Feedback: unknown
Poses remaining: 0
ETA: 0 s
Distance remaining: 0.00 m
Time taken: 0 s
Recoveries: 0

Figure 3: Shows the visualization of object detection and mapping pipeline: (a) the raw camera input frame, (b) object detection with class labels and confidence scores, (c) face detection with bounding box and orientation, and (d) spatial localization of detected objects (except humans) based on custom confidence-based filtering in RViz2, with the arrow indicating the robot's position and viewing direction.

Figure 3(c) demonstrates successful human detection within the scene, while Figure 3(d) shows the RViz2 interface, displaying the spatial location of the detected objects on the map. While the system detects most objects effectively, certain limitations remain as some objects are occasionally undetected or poorly mapped. For example, the table and the orange toy as seen in Figure 3(b) are not detected correctly and are not mapped in Figure 3(d). These limitations are attributed to the constraints of the YOLOv5 architecture used in the deep perception module, highlighting the need for improved detection models or complementary sensing modalities to achieve comprehensive scene understanding in complex indoor environments. Figure 4 shows an enlarged view of the map showing detected objects with green blobs and a robot with an arrow showing the robot's orientation.

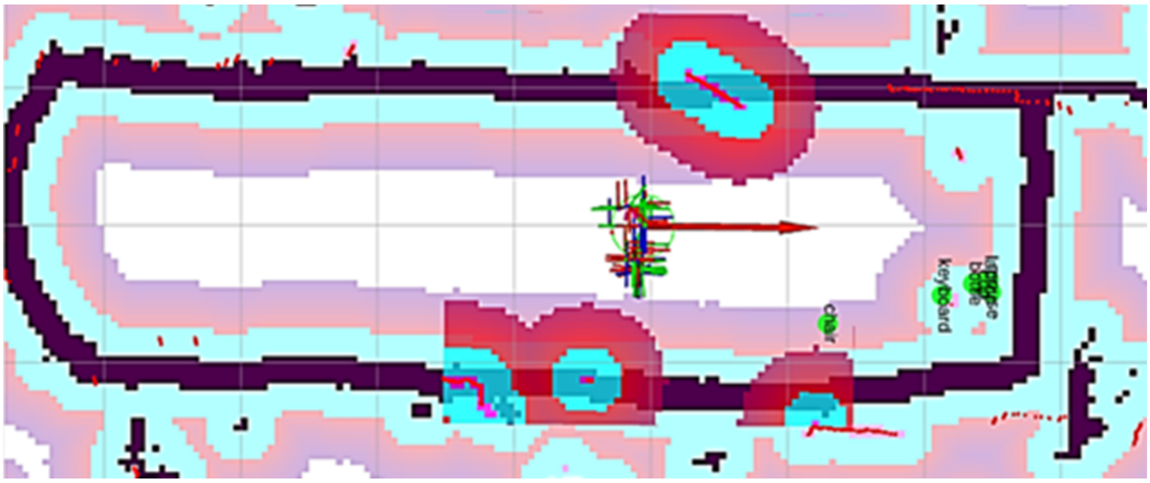


Figure 4: Enhanced view of the object detection results (shown in Figure 3(d)), highlighting object mapping.

Face Detection: Building on scene perception, we evaluate face detection capabilities, which serve as a prerequisite for user profiling and personalized navigation. In a small-scale trial with three staff members using approximately 30 face samples, the system achieves a 70% face detection success rate using the OpenVINO model. Figure 5 illustrates various successful and failed cases. Detection errors are generally associated with challenging conditions such as low-light environments, excessive subject distance, and reflective surfaces. However, the system demonstrated robustness to common appearance variations, maintaining reliable detection when participants wore caps (Figure 5(d)) or were observed from side-profile view (Figure 5(e)). Failures in user localization occurred when the bounding box appeared at an incorrect location (Figure 5(a)), when no detection was made despite a visible face in the scene (Figure 5(b)), or when multiple bounding boxes (Figure 5(c)) were detected within a frame, resulting in navigation errors which in turn affected robot's navigation ability to reach the person for interaction.

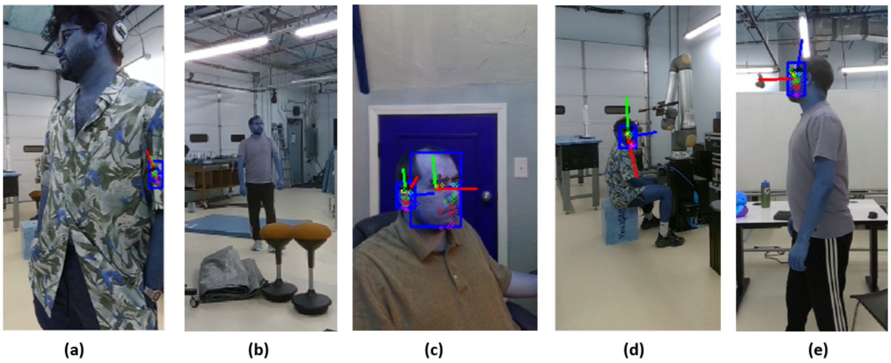


Figure 5: Face detection successes and failures. (a), (b) and (c) show failed detections, including false positives and false negatives. (d) and (e) demonstrate successful detections, including a case where the person is wearing a cap and a side-profile view, respectively.

Emotion Recognition: To evaluate the system's capability for emotion recognition, we instruct four individuals to make faces representing specific emotions in front of the Stretch camera and record the output of the emotion recognition module. A data point is collected for each emotion report produced by the module during the recorded interval for each emotion. The confusion matrix presented in Figure 6 shows that there is considerable variability in the classification performance in different emotional categories. The "Other" category in the confusion matrix represents cases where the DeepFace model reported no face, multiple faces with different emotions, or an emotion other than Happy, Sad, Anger, or Fear.

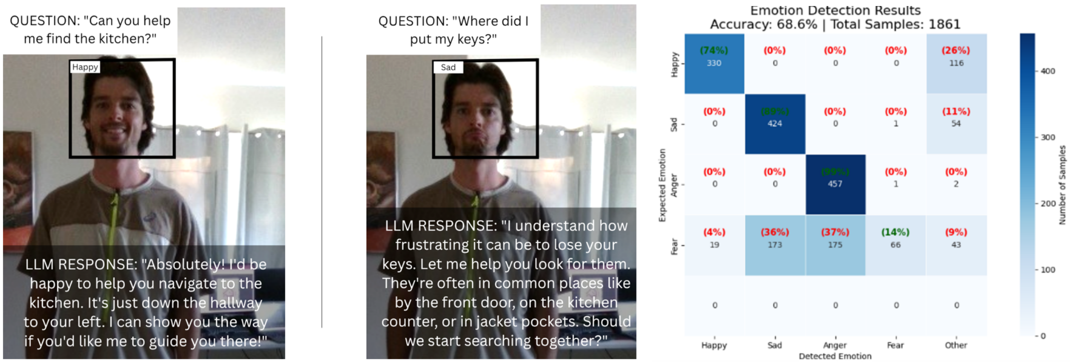
The model achieves an overall classification accuracy of 68.6% across 1,861 samples using only 4 faces. The performance metrics were computed using standard formulas:

$$\text{Precision} = \frac{TP}{TP + FP} \quad \text{Recall} = \frac{TP}{TP + FN}$$

where TP is true positives, FP is false positives, and FN is false negatives.

The confusion matrix demonstrates varying performance across different emotion classes. Anger recognition achieves the highest recall at 99.3% (457/460), indicating successful identification of nearly all anger instances, though precision is moderate at 72.3% (457/632) due to false positives from other emotion categories. Sad emotion recognition shows balanced performance with 88.5% recall (424/479) and 71.0% precision (424/597). Happy emotion recognition exhibits asymmetric performance characteristics, with moderate recall of 74.0% (330/446) but high precision of 94.6% (330/349), suggesting conservative but accurate prediction behavior. Fear recognition presents the most significant challenge, demonstrating a clear trade-off between sensitivity and specificity. The fear class achieves an extremely low recall of 13.9% (66/476), indicating that 410 out of 476 fear instances are misclassified, yet exhibits remarkably high precision of 97.1% (66/68). This pattern suggests frequent misclassification of fear, with the majority of fear instances being redistributed to other negative emotion categories, particularly sad (173 instances) and anger (175 instances). These findings indicate that while the model demonstrates conservative accuracy when predicting fear, it fails to capture the majority of true fear expressions, likely due to overlapping feature representations among negative emotional states.

Figure 6 shows the emotion detected as “Happy” and “Sad” which is then used by the LLM to adjust its tone when responding to the person.



QUESTION: "Can you help me find the kitchen?"

LLM RESPONSE: "Absolutely! I'd be happy to help you navigate to the kitchen. It's just down the hallway to your left. I can show you the way if you'd like me to guide you there!"

QUESTION: "Where did I put my keys?"

LLM RESPONSE: "I understand how frustrating it can be to lose your keys. Let me help you look for them. They're often in common places like by the front door, on the kitchen counter, or in jacket pockets. Should we start searching together?"

Figure 6: Illustrates the LLM’s personalized response based on the emotion detected using DeepFace and the confusion matrix for detected emotions.

5. Conclusion and Future Work

This work introduces a preliminary framework for assistive social robots that integrates deep learning based environmental perception, spatial memory guided navigation, and conversational adaptability using large language models. The framework allows personalized, context-aware interactions through dynamic user profiling and real-time scene understanding. Initial testing on the Hello Robot Stretch 3 platform demonstrates the feasibility of the approach in structured indoor settings, highlighting its potential for human-centered assistance.

Future work will focus on three main directions. First, we plan to enhance object and face detection by incorporating state-of-the-art methods, including visual language models, and to develop a multimodal navigation system that functions effectively in previously unseen environments while preserving user privacy. Second, we plan to explore proactive assistance features such as fall detection using pose estimation and behavioral modeling to identify potential health risks, with careful attention to privacy and reliability. Third, we intend to conduct small-scale clinical trials in simulated care settings and investigate integration with existing healthcare infrastructures, including electronic health records and caregiver support systems.

Overall, this preliminary work aims to advance the system toward real-world deployment, contributing to the development of intelligent, adaptive, and trustworthy assistive robotic platforms.

6. Acknowledgements

This work was performed under the following financial assistance award 60NANB24D221 from the U.S. Department of Commerce, National Institute of Standards and Technology.

Declaration on Generative AI

During the preparation of this work, the author(s) used ChatGPT for grammar and spelling checks, as well as paraphrasing. The author(s) reviewed and edited all content and take full responsibility for the final version.

References

[1] S. Rossi, F. Ferland, A. Tapus, User profiling and behavioral adaptation for hri: A survey, Pattern Recognition Letters 99 (2017) 3–12. doi:10.1016/j.patrec.2017.06.002.

[2] M. K. Lee, J. Forlizzi, S. Kiesler, P. Rybski, J. Antanitis, S. Savetsila, Personalization in hri: a longitudinal field experiment, in: Proceedings of the Seventh Annual ACM/IEEE International Conference on Human-Robot Interaction, HRI '12, Association for Computing Machinery, New York, NY, USA, 2012, p. 319–326. doi:10.1145/2157689.2157804.

[3] D. Song, J. Liang, A. Payandeh, A. H. Raj, X. Xiao, D. Manocha, Vlm-social-nav: Socially aware robot navigation through scoring using vision-language models, IEEE Robotics and Automation Letters 10 (2025) 508–515. doi:10.1109/LRA.2024.3511409.

[4] J. Rios-Martinez, A. Spalanzani, C. Laugier, From proxemics theory to socially-aware navigation: A survey, International Journal of Social Robotics 7 (2015) 137–153. doi:10.1007/s12369-014-0251-1.

[5] C. Granata, P. Bidaud, A framework for the design of person following behaviors for social mobile robots, in: 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems, 2012, pp. 4652–4659. doi:10.1109/IROS.2012.6385976.

[6] J. Wang, E. Shi, H. Hu, C. Ma, Y. Liu, X. Wang, Y. Yao, X. Liu, B. Ge, S. Zhang, Large language models for robotics: Opportunities, challenges, and perspectives, Journal of Automation and Intelligence 4 (2025) 52–64. doi:10.1016/j.jai.2024.12.003.

[7] D. Mccoll, A. Hong, N. Hatakeyama, G. Nejat, B. Benhabib, A survey of autonomous human affect detection methods for social robots engaged in natural hri, Journal of Intelligent & Robotic Systems 82 (2016) 101–133. doi:10.1007/s10846-015-0259-2.

[8] M. Maroto-Gómez, S. Marqués-Villaroya, J. C. Castillo, Álvaro Castro-González, M. Malfaz, Active learning based on computer vision and human–robot interaction for the user profiling and behavior personalization of an autonomous social robot, Engineering Applications of Artificial Intelligence 117 (2023) 105631. doi:10.1016/j.engappai.2022.105631.

[9] J. Atuhurra, Leveraging large language models in human-robot interaction: A critical analysis of potential and pitfalls, 2024. doi:10.48550/arXiv.2405.00693. arXiv:2405.00693.

[10] Hello Robot, Deep perception, https://docs.hello-robot.com/latest/ros2/deep_perception/, 2024. Accessed: July 3, 2025.

[11] A. Tapus, C. Ţăpuş, M. J. Matarić, User—robot personality matching and assistive robot behavior adaptation for post-stroke rehabilitation therapy, Intelligent Service Robotics 1 (2008) 169–183. doi:10.1007/s11370-008-0017-4.

[12] G. Castellano, P. Mcowan, Towards affect sensitive and socially perceptive companions, 2013, pp. 42–53. doi:10.1007/978-3-642-37346-6_5.

[13] J. Achiam, et al., Gpt-4 technical report, arXiv preprint arXiv:2303.08774 (2023). doi:10.48550/arXiv.2303.08774.

[14] S. I. Serengil, A. Ozpinar, Lightface: A hybrid deep face recognition framework, in: 2020 Innovations in Intelligent Systems and Applications Conference (ASYU), 2020, pp. 1–5. doi:10.1109/ASYU50717.2020.9259802.

[15] S. I. Serengil, A. Ozpinar, Hyperextended lightface: A facial attribute analysis framework, in: 2021 International Conference on Engineering and Emerging Technologies (ICEET), 2021, pp. 1–4. doi:10.1109/ICEET53442.2021.9659697.

[16] C. Mishra, R. Verdonschot, P. Hagoort, G. Skantze, Real-time emotion generation in human-robot dialogue using large language models, Frontiers in Robotics and AI 10 (2023). doi:10.3389/frobt.2023.1271610.

[17] F. Alonso-Martín, M. Malfaz, J. Sequeira, J. F. Gorostiza, M. A. Salichs, A multimodal emotion detection system during human–robot interaction, Sensors 13 (2013) 15549–15581. doi:10.3390/s131115549.

[18] B. Zitkovich, et al., RT-2: vision-language-action models transfer web knowledge to robotic control, in: Proceedings of The 7th Conference on Robot Learning, volume 229 of *Proceedings of Machine Learning Research*, PMLR, 2023, pp. 2165–2183. URL: https://proceedings.mlr.press/v229/zitkovich23a.html.

[19] Hello Robot, Stretch 3 – A fully integrated mobile manipulator, https://hello-robot.com/stretch-3-product, 2025. Accessed: August 18, 2025.