Robust Autonomous Summoning Capabilities for a Patient Assistive Mobile Robot

Vithanage T. V. R. H.[1], Welangalle P. D.[1], Senaratne N. A. A. N. R.[1], Amarasinghe Y. W. R.[1], Jayathilaka W. A. D. M.[1], and Jayawardane M. A. M. M.[2]

***Abstract*— A novel approach for autonomous summoning of Patient Assistive Mobile Robots (PAMRs) to improve their use in areas with limited infrastructure and healthcare access has been presented. The method has been designed to operate effectively in unstructured environments, eliminating the need for Wi-Fi or other network-dependent Indoor Positioning Systems (IPS). The proposed summoning approach integrates LiDAR-based Simultaneous Localization and Mapping (SLAM), Sound Source Localization, and Stereo Vision technologies. For scenarios where the patient is not within the robot's line of sight, Sound Source Localization guides the robot toward auditory cues. The system utilizes stereo-vision-based gesture recognition and depth estimation for line-of-sight scenarios, allowing patients to summon the robot using natural gestures. LiDAR-based SLAM has been utilized for mapping, localization, path planning, and obstacle avoidance. This multimodal approach ensures the robustness of the summoning capabilities, making it adaptable to diverse environments. Preliminary testing has demonstrated the method's effectiveness, suggesting potential incorporation in PAMRs for healthcare delivery.**

I. INTRODUCTION

Patient Assistive Mobile Robots (PAMRs) have emerged as a promising technology for healthcare delivery, particularly in remote communities with limited facilities and medical personnel. However, for PAMRs to achieve widespread adoption, seamless and reliable patient-robot interaction is critical. A fundamental aspect of this interaction is the ability of patients to effortlessly summon the robot to the patient's location for assistance. Traditional methods often fall short in unstructured environments, where the lack of pre-installed infrastructure creates significant barriers.

The presented approach introduces a summoning technique that leverages the strengths of different sensor modalities. LiDAR-based SLAM provides robust localization and mapping capabilities, sound source localization enables navigation toward the patient in non-line-of-sight scenarios, and stereo vision facilitates natural gesture-based summoning for line-of-sight interaction. Fig. 1 shows the initially developed mobile robotic platform for proof of concept along with the robotic walker platform used for experimental validation of the robustness of the approach.

The subsequent sections provide an in-depth exploration of the proposed solution. Section II, titled 'Related Work', offers a comprehensive review of existing literature and studies in the field. In Section III, 'Proposed System', we detail the design and execution of the multimodal summoning system, elaborating on the hardware components, software algorithms, and the strategy for system integration. Section IV, 'Results', presents an experimental evaluation of the system. This includes the setup for testing, the performance metrics used, and the results obtained. The effectiveness and robustness of the system are analyzed across various scenarios. Finally, in Section V, 'Conclusion', we summarize the key findings and underscore the contributions of the multimodal summoning approach. We also discuss potential avenues for further development and real-world implementation.

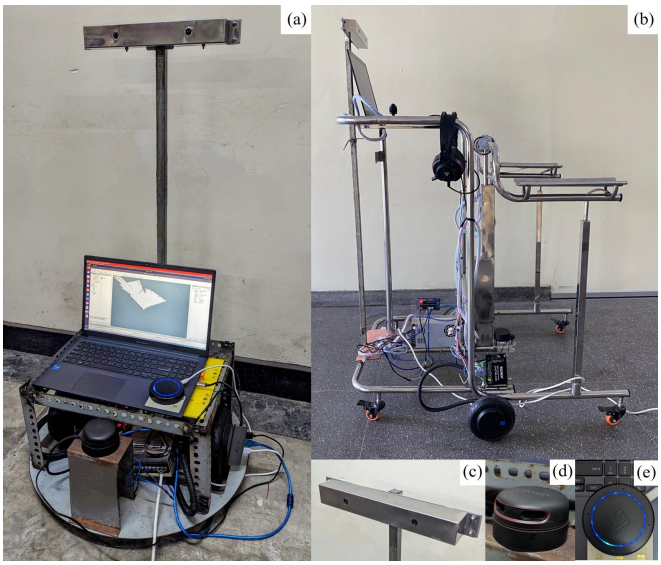


Fig. 1. (a) Mobile Robotic Platform developed for proof of concept (b) Robotic Walker used for Experimental Validation (c) Stereo Camera Setup (d) 2D LiDAR sensor (e) Microphone Array

II. RELATED WORK

In the context of mobile robots, Indoor Positioning Systems are crucial for enabling autonomous navigation and operation within indoor environments where GPS signals are typically unavailable or unreliable. A variety of technologies are applicable to realize IPS. The most prevalent are Computer Vision and LiDAR, followed by IMU (Inertial Measurement Unit), UWB (Ultra-Wideband), VLC (Visible Light Communication), and Wi-Fi [1]. In the radio domain, UWB surpasses Wi-Fi as the most utilized method. However,

*This work was not supported by any organization

[1]Department of Mechanical Engineering, University of Moratuwa, Sri Lanka. (vithanagetvrh.19, welangallepd.19, senaratnenaanr.19, ranama, dumithj)@uom.lk

[2]Department of Obstetrics and Gynaecology, University of Sri Jayawardenapura, Sri Lanka. madurammj@yahoo.com

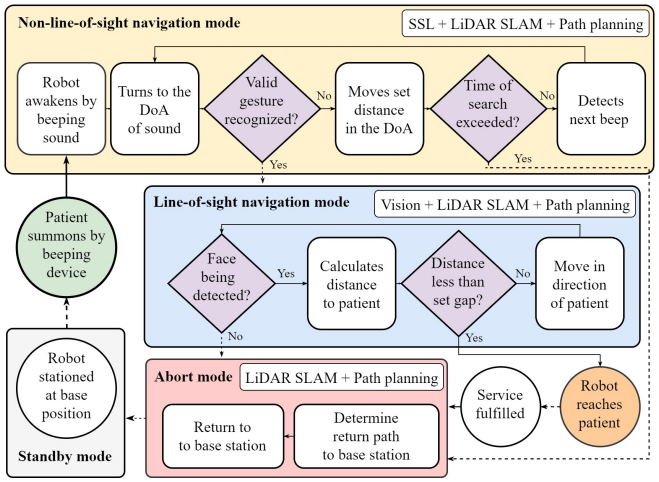


Fig. 2. Navigation Control Architecture

the high cost is a limitation of UWB. While technically simple to implement on a robot, Wi-Fi lacks precision and often needs to be combined with optimization methods like neural networks, deep learning, filters, etc.

Despite the widespread use of radio frequency technology, it's not suitable for areas where radio usage is limited, like hospitals and airports [2]. Non-radio localization methods, particularly Computer Vision and LiDAR, are widely used in indoor robotic localization due to their adaptability to complex environments, and no need for environmental alterations. However, they require high computational resources, which is their main limitation. If computational issues can be addressed, positioning accuracy and speed could be improved. An approach for robot navigation using Visible Light Positioning (VLP) landmarks and SLAM for automatic map calibration is discussed in [3] where the system tracks landmarks and conducts SLAM simultaneously, aligning coordinates on a layout map and a sensor map from SLAM.

Several attempts have also been made to develop affordable localization solutions. The batteryless nurse calling system developed in [4] locates patients and nurses in hospitals by leveraging the existing WLAN (Wireless Local Area Network). It uses distributed RF (Radio Frequency) nodes and WLAN access points for tracking, with positions determined via trilateration and time/angle of arrival. In another work [5], a waypoint-based indoor navigation that uses Bluetooth beacons utilizes directional Bluetooth beacons is presented.

Navigation can also be implemented using Sound source localization which is most useful for non-line-of-sight navigation. Methods for SSL include Energy-Based Localization, Time of Arrival (TOA), and Time Difference of Arrival (TDOA) [6]. Of these methods, widespread adoption is found for Time difference of arrival (TDOA) which works with the time difference between the signals.

A system that uses an array of 8 microphones to accurately localize sounds in three dimensions, even for short-duration sounds, is made possible in [7]. An audio-visual technique for locating hidden sound sources using the TDOA algorithm which is computationally efficient and suitable for dynamic scenarios is presented in [8]. Another audio-visual fusion approach that integrates Sound SSL into visual SLAM is found in [9] which improves visual odometry robustness in SLAM systems.

While existing IPS solutions offer robust navigation in some settings, their reliance on pre-installed infrastructure (e.g., Wi-Fi access points, UWB tags) or user-worn devices significantly limits their applicability in remote areas with limited resources. Advanced methods like LiDAR-based SLAM and computer vision are inherently limited in their application to areas where the user is in direct line-of-sight of the robot. This necessitates a more versatile and adaptable approach to PAMR summoning.

To address these limitations, this research proposes a novel multimodal summoning system that leverages the strengths of LiDAR-based SLAM, sound source localization, and stereo vision. This eliminates the need for pre-existing infrastructure and allows patients to summon the robot using natural gestures or sound cues, making PAMRs more accessible and user-friendly in geographically isolated communities with limited healthcare access.

III. PROPOSED SYSTEM

A. System Overview

The system integrates three key sensor modalities to achieve robust and adaptable robot summoning across diverse settings. LiDAR sensing provides the foundation for autonomous navigation. It enables robot localization and real-time map creation, allowing the robot to understand its surroundings and precisely determine its position. Sound Source Localization tackles scenarios where the patient is out of sight. The system continuously checks for the sound signal of a patient calling for assistance through an audio signal produced using a metronomic beeping device. When a sound is detected, the system pinpoints its direction, guiding the robot toward the patient until visual confirmation is available. Stereo vision facilitates summoning for patients within the robot's field of view. This modality enables accurate depth perception and gesture recognition. Patients can directly summon the robot using pre-defined dynamic hand gestures, fostering a more intuitive interaction.

The system operates in a continuous cycle as indicated in Fig. 2. The method implemented by each subsystem is studied in the following subchapters.

B. LiDAR based SLAM

The execution of four tasks - localization, mapping, path planning, and locomotion - is essential for fully autonomous navigation. [10] A robot requires a map to determine its location within the environment, while to construct a map, the robot needs to understand its own location and the relative positions of its observations. SLAM is a common method used to address the simultaneous needs of localizing robots and mapping their environment.

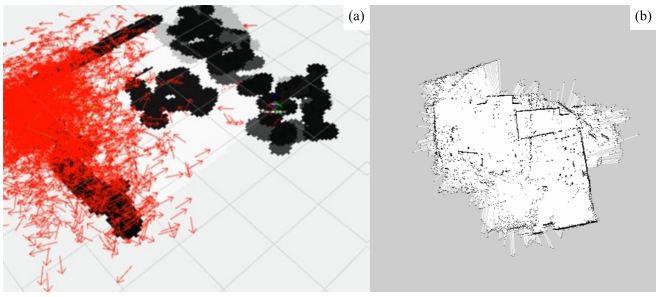


Fig. 3. (a) Cost maps for obstacle avoidance with Dijkstra's algorithm for path planning (b) 2D Map of the area generated by LiDAR-based SLAM

In this study, LiDAR-based SLAM is implemented which is a widely adopted approach suitable for dynamic environments such as hospitals, care homes, rehabilitation centers, etc. A 2D LiDAR sensor was chosen for this application. While 3D LiDAR offers richer data, a 2D sensor provides sufficient information for our purpose while offering advantages in terms of cost and power consumption.

The system employs a Hector-SLAM algorithm for processing LiDAR data. The raw LiDAR data consists of distance measurements to surrounding objects. Hector-SLAM processes this data to identify obstacles and build a map of the environment (Fig. 3), continuously updates the robot's position within the map using odometry data and LiDAR measurements and provides a real-time map representation that can be utilized for robot navigation.

C. Non-line-of-sight navigation

By capturing sound waves from slightly different angles using each microphone, a microphone array can calculate the direction of a sound source using the Time Difference of Arrival technique. It analyzes the time delays or phase differences of the target sound across the microphone array elements. This analysis is used to calculate the direction of the sound source, indicating the patient's location relative to the robot.

If a sound source is identified, the robot uses the Direction of Arrival (DoA) of the sound to create a virtual goal on the real-time map from SLAM. Since the sound source localization provides only direction, the system creates the goal at a pre-determined distance in that direction based on the map and then listens for the sound again. This iterative approach allows the robot to home in on the patient's location even without a direct line of sight.

In place of the user making sounds (claps, voice commands, etc.) a metronomic beeping sound generated from the user's location was used. A mobile phone generated the sounds which is appropriate as a readily available device which eliminated the need for extra instrumentation. Sound at intervals of 5 seconds between beats (12 beats per minute) was used in testing. Sound source separation was performed with onboard firmware of the ReSpeaker Mic Array as shown in Fig. 4.

Two conditions were used for validating non-line-of-sight navigation. Under the first condition, the sound played was

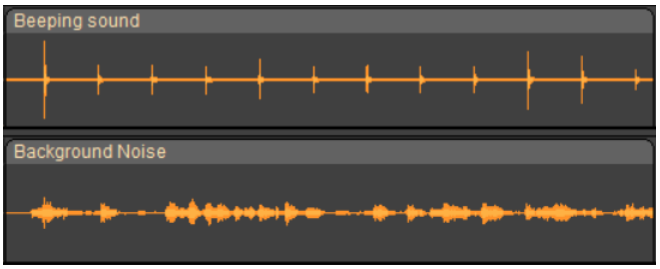


Fig. 4. Separation of metronomic beeping sound from background noise

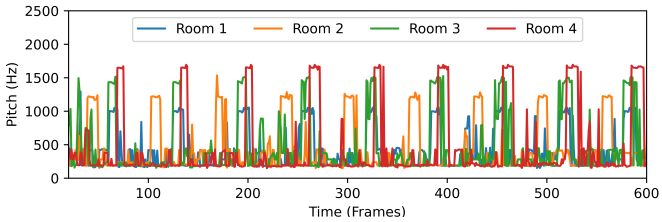


Fig. 5. Unique frequencies for each room under pitch variant condition

unique to each room. This was implemented by playing sounds of different pitch (frequency) for each room (Fig. 5. The robot had information on the specific location of each room on the map and was able to determine the room of the patient by the sound's characteristics. This was implemented on the robot using the librosa [11] library to detect the pitch of the played sound. Under the second condition, the same sound was played at all patient locations, making the pitch invariant of the destination. With this approach, the robustness of the SSL implementation for non-line-of-sight navigation could be elicited.

D. Line-of-sight navigation

Stereo vision utilizes two cameras positioned slightly apart, mimicking human binocular vision. By processing the images from both cameras and analyzing the disparity between the left and right camera images, the system can create a depth map of the scene, allowing the robot to understand the spatial layout of the environment better than a monocular setup. In this study, a custom setup made using two inexpensive webcams and calibrated using the OpenCV library assumed the role of a depth camera.

A dynamic hand gesture needs to be performed by the subject from an upright seated position. To recognize the gestures, mediapipe human pose estimation library [12] is utilized. By locating pose landmarks of either the left or right hand and analyzing their spatial and temporal variation, a dynamic gesture is defined. The choice of gesture is based on its uniqueness to suit an environment with other people present where various hand gestures might be visible to the camera. The easiness of performing the gesture in terms of effort required for a patient was also factored in.

As shown in Fig. 2 if a summoning gesture is detected, the system attempts to locate the face of the user performing the gesture. Although the distance to the user can be computed using hand landmarks, an additional step is performed to

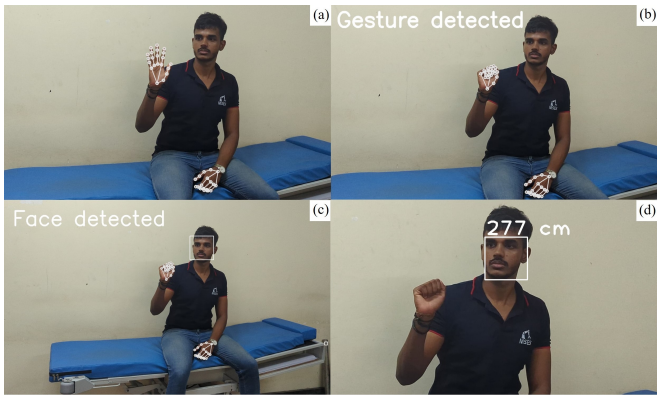


Fig. 6. Gesture recognition and distance estimation procedure (a) Patient in upright sitting position with arm raised (b) Valid gesture sequence detected (c) Face detected (d) Distance calculated with face landmarks

compute distance using face landmarks for two reasons. Firstly, the robot must be approaching from the front of the user to ensure proper human-robot interaction. As the user's gaze can be a good factor for determining body orientation in a seated position, face landmarks were retrieved to compute the distance. Secondly, hands have a greater range of motion with respect to the human body compared to the face, yielding less accurate distance estimations which can be minimized by using face landmarks. A walkthrough of the steps is given in Fig. 6.

After computing the distance to the corresponding patient location, a path is created to gradually minimize the distance. Obstacles recognized are avoided with the aid of the real-time map generated by LiDAR. This allows the robot to navigate toward the patient and locate itself facing the user at the end.

1) Path planning and Movement: This section explores how the fused sensor data is utilized to control the robot's movement and achieve navigation toward the patient. The combined sensor data establishes goal points for the robot to navigate to. The ROS navigation stack leverages the ROS topic *move_base_simple/goal* for the purpose. Within the stack, path planning algorithms utilize cost maps for obstacle avoidance. The dynamic window approach and Dijkstra's Algorithm are used for local planning and global planning, respectively, to determine safe and efficient paths for the robot to reach the identified goal points. Based on the planned path, ROS Navigation sends control commands to the robot's motor controllers, directing its movement.

E. Hardware

The mobile robotic platform measured 50x50x24 cm in dimensions with a turning radius of 24 cm. Differential drive was the optimum choice considering its simplicity and ability to make tight turns with standard (non-holonomic) wheels. Brushless DC hub motors (ZLTECH ZLLG65ASM250-4096, Shenzhen ZhongLing Technology Co., Ltd, China) were a compact and powerful solution that also offered the capability to freewheel, allowing the patient to make minor

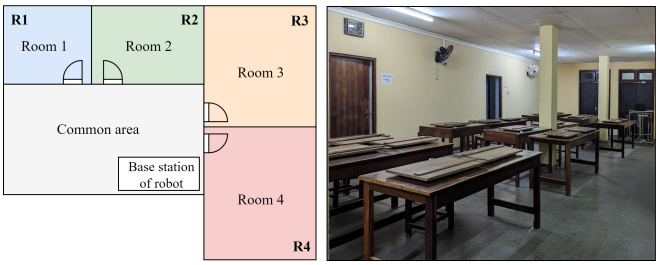


Fig. 7. Floor layout of the experimental area

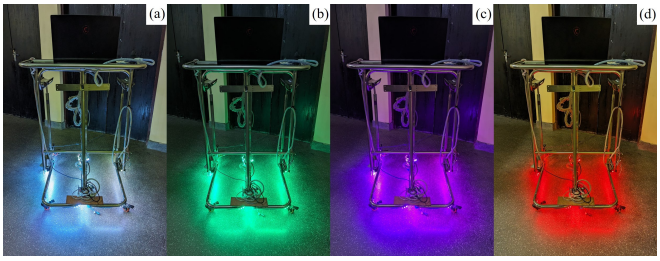


Fig. 8. Visualisation of control modes on robotic walker (a) Standby mode (b) Non-line-of-sight mode (c) Line-of-sight mode (d) Abort mode

adjustments to the robot's position or orientation when the goal location is reached.

A laptop running the Robot Operating System (ROS) performed high-level control while an Arduino Mega microcontroller performed low-level control. The 2D LiDAR sensor used for SLAM was RPLIDAR A1 (SLAMTEC Co., Ltd, China). The stereo vision camera setup was put together with two webcams (Logitech C270) and calibration was performed using the OpenCV library. ReSpeaker USB Mic Array (Seeed Technology Co., Ltd, China) was instrumental in performing SSL and sound source separation. The array's built-in functionalities, such as Automatic Gain Control (AGC) and Noise Suppression (NS), can enhance sound quality and improve source localization accuracy in real-world environments with ambient noise.

F. EXPERIMENTAL SETUP

Experimentation was conducted in a controlled environment; an indoor area with low levels of reverberation. As shown in Fig. 7 the layout features multiple rooms which have doors opening to a common area.

Under non-line-of-sight navigation, the performance of SSL was tested under both conditions; pitch variant by destination and pitch invariant by destination. Under line-of-sight navigation, testing was done for gesture recognition and distance estimation. A subject was instructed to sit upright (elevation of eyes from the ground 145 cm) on a hospital bed (elevation of bed surface from the ground 75 cm).

To validate the system, we employed a robotic walker (Fig. 1 (b)) which implemented the same differential drive control used in the mobile robotic platform. The walker (dimensions 105x65x110 cm) had been designed and developed for research on gait rehabilitation and served as a close example of a Patient Assistive Mobile Robot. The time to reach the

TABLE I

AVERAGE TIME TAKEN (IN MINUTES) BY THE ROBOT TO REACH PATIENTS FROM DIFFERENT ORIGINS

Tele-operated					Pitch variant by destination					Pitch invariant by destination				
patient / robot	R1	R2	R3	R4	patient / robot	R1	R2	R3	R4	patient / robot	R1	R2	R3	R4
Base	1.08	0.80	0.54	0.17	Base	5.62	4.54	2.76	0.87	Base	6.98	5.02	2.94	0.89
R1	-	0.63	0.79	1.10	R1	-	3.96	5.86	6.36	R1	-	5.10	7.75	8.63
R2	0.63	-	0.67	0.84	R2	3.83	-	3.04	5.21	R2	4.80	-	4.22	5.81
R3	0.79	0.67	-	0.51	R3	6.07	3.13	-	2.77	R3	8.56	4.05	-	3.93
R4	1.10	0.84	0.51	-	R4	6.24	5.39	2.83	-	R4	8.90	5.82	3.89	-

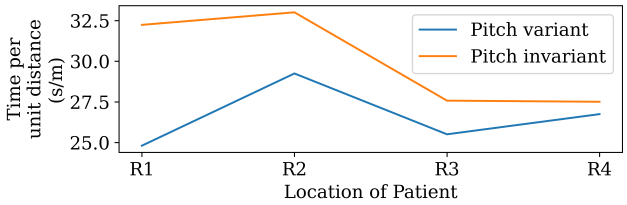
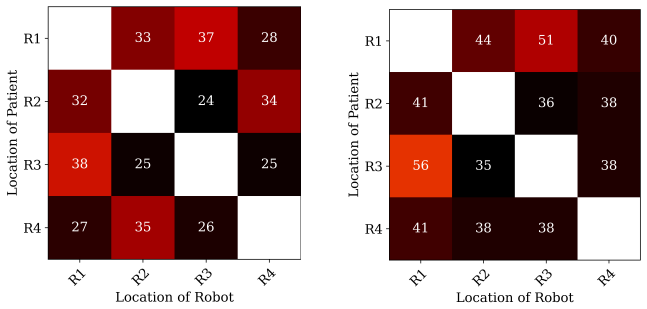


Fig. 9. Deviation of time over unit distance from base to each room compared with the teleoperated condition for trips under different conditions



(a) Pitch variant by destination (b) Pitch invariant by destination

Fig. 10. Deviation of time over unit distance between rooms compared with teleoperated condition (ground truth) for trips under different conditions (units: seconds per meter)

goal location and the frequency of 'abort' mode occurrence was observed for the robotic walker. The mode of operation of the walker was visualized with the help of the light signals given by walker-mounted LED strips as shown in Fig. 8)

IV. RESULTS

A. SSL

Different combinations for the robot's initial location and the patient's location were used. The time for the robot to reach the goal location was recorded for each trip. The time for the robot to reach the goal location under teleoperation was also recorded to serve as a performance benchmark. Each trip was run three times and the average was calculated to minimize random error.

Fig. 9 plots the deviation of times taken for unit distance for the given conditions from the base station to each room. The base station of the robot is closest to Room 4 and gets incrementally farther to Rooms 3, 2, and 1, respectively. Thereby, it can be seen that the pitch variant mode performs slightly better with increasing distance. Fig. 10 shows the variation of times for trips between rooms. By comparison of the respective cells of the two plots, a conclusion can be drawn that the pitch of the sound being a unique identifier for each destination achieved better results in SSL. Significant deviations of both pitch-variant and invariant conditions compared to the teleoperated mode can be attributed to the misinterpretation of the direction of the sound source. The acoustics of each room and the building as a whole take great effect in this regard. However, the SSL algorithm needs further refinement to minimize the observed gap and achieve trip times closer to the teleoperated time.

B. Gesture recognition and distance estimation

Gesture recognition accuracy was evaluated by instructing the subject to perform the summoning gesture along with other gestures that closely resemble it (Fig. 11), to test the system's response to invalid gestures. This was done to assess the system's ability to distinguish between valid and invalid gestures.

The accuracy of the distance estimation was also verified. For this, the test subject and the robot were initially positioned at a known distance apart, with no obstacles in between. The robot then moved in a straight line to reach the patient while the distance values predicted by the computer vision algorithm were recorded concurrently with the wheel odometry data. The actual distance to the patient (decrementing with robot movement) was computed using the odometry data which served as the ground truth for comparing with the predicted distance values. The results indicated that the accuracy of both gesture recognition and distance estimation were within an acceptable range.

C. Validation with Robotic walker

Due to its physical dimensions, navigating through narrow gaps and taking close turns were problematic for the walker and resulted in over 30% more occurrences of the 'abort mode' (due to exceeding the maximum time of search) compared to the mobile robotic platform. The dead reckoning method of sound source proved to be less applicable when

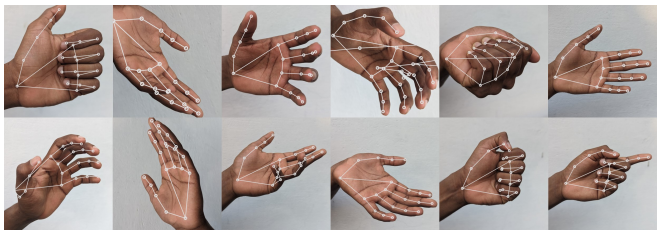


Fig. 11. Invalid hand gestures for testing performance of dynamic hand gesture recognition algorithm

the size of the robot can limit the space available for motion highlighting the requirement of a more precise method for performing SSL.

D. Limitations

In line-of-sight navigation, the range of approach angles for gesture identification and distance estimation was limited; roughly 150 degrees from the patient's perspective. In situations where the robot approaches from the front, there were minimal events of recognition failure. However, when the robot entered the room facing the back of the patient even though the gesture is recognized the face cannot be identified to calculate distance. The practical requirement of the robot to approach the user from the front is to have better interaction. Future versions can work around this problem by instructing the robot to rectify its final pose. The position must be in front of the patient and the orientation opposite to that of the patient.

The accuracy of SSL can be improved with techniques such as Reinforcement Learning [A Smart Robotic Walker with Intelligent Close-Proximity Interaction Capabilities for Elderly Mobility Safety] as dead reckoning proved problematic for longer ranges and complex routes. Testing must be conducted in environments with varying levels of reverberation to ascertain the applicability of SSL in different surrounding conditions. The applicability of SSL for non-line-of-sight navigation diminishes if the surrounding environment is crowded or noisy. To make the locating sound distinguishable, the loudness of the sound would have to be significant. This, inevitably, will make it uncomfortable for the patient and other people in the surroundings.

Intelligent support can further be enhanced by enabling voice interaction, where a wake word can be used to summon the walker to the patient's position. Additionally, with techniques such as Vision and Language Navigation (VLN), the robot can utilize the assistance of the patient to guide itself toward the destination.

V. CONCLUSION

As discussed, a novel, multimodal approach for summoning Patient Assistive Mobile Robots (PAMRs) has been developed to function reliably in diverse and unstructured environments. By integrating LiDAR-based SLAM, Sound source localization, and stereo vision technologies, the proposed method has been able to overcome the limitations of current summoning methods that rely heavily on pre-installed infrastructure. The system's ability to respond to auditory cues and natural gestures ensures robustness and adaptability, making it suitable for real-world deployments. Preliminary testing has demonstrated the effectiveness of this approach, suggesting its potential for incorporation in PAMRs to improve healthcare delivery, particularly in remote areas with limited infrastructure. This work also supports the ongoing efforts to enhance patient-robot interaction, paving the way for the wider adoption of PAMRs in healthcare. Future work will focus on further refining the system and conducting extensive real-world testing to validate its performance and usability.

ACKNOWLEDGMENT

Our sincere thanks go to Mr. Rangika Mark and Ms. Anurisha Dunuwila for providing us with hardware components that were instrumental in our testing procedures.

REFERENCES

[1] J. Huang, S. Junginger, H. Liu, and K. Thurow, "Indoor positioning systems of mobile robots: A review," *Robotics*, vol. 12, no. 2, 2023.

[2] S. r. Zekavat, R. M. Buehrer, G. D. Durgin, L. Lovisolo, Z. Wang, S. T. Goh, and A. Ghasemi, "An overview on position location: Past, present, future," *International Journal of Wireless Information Networks*, vol. 28, pp. 45–76, Mar. 2021.

[3] Y. Wang, B. Hussain, and C. P. Yue, "Vlp landmark and slam-assisted automatic map calibration for robot navigation with semantic information," *Robotics*, vol. 11, no. 4, 2022.

[4] R. Kanan and O. Elhassan, "A combined batteryless radio and wifi indoor positioning system," in *2015 23rd International Conference on Software, Telecommunications and Computer Networks (SoftCOM)*, pp. 101–107, 2015.

[5] E. T.-H. Chu, S. Wang, C. C. Chang, J. W.-S. Liu, J. Hsu, and H.-M. Wu, "Wpin: A waypoint-based indoor navigation system," in *International Conference on Indoor Positioning and Indoor Navigation*, 2019.

[6] C. Rascon and I. Meza, "Localization of sound sources in robotics: A review," *Robotics and Autonomous Systems*, vol. 96, pp. 184–210, 2017.

[7] J.-M. Valin, F. Michaud, J. Rouat, and D. Letourneau, "Robust sound source localization using a microphone array on a mobile robot," in *Proceedings 2003 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2003) (Cat. No.03CH37453)*, vol. 2, pp. 1228–1233 vol.2, 2003.

[8] E. King, A. Tatoglu, D. Iglesias, and A. Matriss, "Audio-visual based non-line-of-sight sound source localization: A feasibility study," *Applied Acoustics*, vol. 171, p. 107674, 2021.

[9] T. Zhang, H. Zhang, X. Li, J. Chen, T. L. Lam, and S. Vijayakumar, "Acousticfusion: Fusing sound source localization to visual slam in dynamic environments," *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 6868–6875, 2021.

[10] Y. D. V. Yasuda, L. E. G. Martins, and F. A. M. Cappabianco, "Autonomous visual navigation for mobile robots: A systematic literature review," *ACM Comput. Surv.*, vol. 53, feb 2020.

[11] B. McFee, C. Raffel, D. Liang, D. P. W. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," in *SciPy*, 2015.

[12] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C.-L. Chang, M. G. Yong, J. Lee, W.-T. Chang, W. Hua, M. Georg, and M. Grundmann, "Mediapipe: A framework for building perception pipelines," *ArXiv*, vol. abs/1906.08172, 2019.