A Modular Robotic Platform for Rehabilitation of Hemispatial Inattention

Roberto Soldi[1], Bruna Guerra[1], Stefania Sozzi[1], Francesco Lunghi[1], Lorenzo Vecchi[1], Leo Russo[1], Micaela Schmid[1] and Stefano Ramat[1,*]

[1] *Bioengineering Laboratory, University of Pavia, Pavia 27100, Italy.*

Abstract

Hemispatial inattention (HI) is a disabling and often persistent consequence of right-hemisphere stroke, typically resistant to conventional rehabilitation approaches. This study reports the results of preliminary tests on a modular robotic platform developed to support HI rehabilitation by integrating autonomous navigation, patient identification, speech-based interaction, patient response assessment, and adaptive therapy delivery. The system leverages multimodal sensing, including RGB-D cameras, skeletal tracking via the MediaPipe framework, and automatic speech recognition, to deliver personalized sensorimotor and cognitive exercises targeting the neglected hemispace. The navigation protocol allows the correct positioning of the robot in front of the patient. The monitoring of the session leverages markerless head pose estimation to study the patient response to the delivered stimuli in terms of spatial exploration. Verbal input is interpreted through a rule-based intent recognition module, enabling context-aware dialogue management and real-time adaptation of the therapeutic session. A preliminary feasibility study was conducted in the Living Lab of the CoE REDI (Centre of Excellence on Rehabilitation Devices and Digital Instruments) at the University of Pavia. The results demonstrated the platform's ability to accurately identify multiple individuals simultaneously, correctly assigning their names and roles. It also achieved a 94% accuracy rate in interpreting spoken commands. Additionally, the system was able to assess patient engagement by leveraging head pose estimation, with a root mean square error (RMSE) of less than 10°. These preliminary results support the potential of the system to deliver structured, responsive, and data-driven rehabilitation experiences. The ongoing work includes full-system integration, deployment of a centralized patient database, and clinical trials to evaluate efficacy and usability in real-world settings.

Keywords

Hemispatial Inattention, Human-Robot Interaction, Assistive Robotics

1. Introduction

Hemispatial inattention (HI), also known as unilateral spatial neglect, is a frequent, debilitating consequence of right-hemisphere stroke. It is characterized by the inability to attend to stimuli on the contralesional side of space, often leading to impaired mobility, delayed recovery, and increased dependence in carrying out activities of daily living [1], [2]. Despite rehabilitation efforts, HI remains resistant to conventional therapies in many cases, with persistent deficits observed even in the chronic phase [3]. In parallel, the last two decades have witnessed a growing integration of robotic systems into neurorehabilitation practice. Robots are capable of delivering structured, intensive, and repeatable interventions while simultaneously capturing quantitative data on patient performance [4], [5]. These systems, equipped with multimodal sensors such as RGB-D cameras, motion trackers, and physiological monitors, offer kinematic and behavioral monitoring abilities, allowing them to analyze movement trajectories, posture, and timing based on sensor data, to enable adaptive therapy protocols grounded on objective, quantifiable metrics [6], [7].

To address the complex needs of patients with HI, rehabilitation systems should go beyond simple task delivery and incorporate perceptive and social interaction capabilities. Advances in socially assistive

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

[*] Corresponding author.

roberto.soldi01@universitadipavia.it (R. Soldi); brunamariavittoria.guerra@unipv.it (B. Guerra); stefania.sozzi@unipv.it (S. Sozzi); cesco.lunghi@gmail.com (F. Lunghi); lorenzo.vecchi01@universitadipavia.it (L. Vecchi); leo.russo01@universitadipavia.it (L. Russo); micaela.schmid@unipv.it (M. Schmid); stefano.ramat@unipv.it (S. Ramat)

0009-0004-9671-2199 (R. Soldi); 0000-0002-5169-4431 (B. Guerra); 0000-0002-6837-651X (S. Sozzi); 0009-0001-9800-2228 (L. Russo); 0000-0003-2537-8723 (M. Schmid); 0000-0001-5932-186X (S. Ramat)

robotics and human–robot interaction (HRI) suggest that embodied agents can foster engagement, motivation, and adherence to therapy by leveraging natural communication modalities, including speech and facial expression [8], [9]. In this context, robots should act not only as motor assistants but as intelligent, context-aware companions that understand spoken input, detect emotional states, and tailor interaction accordingly [10].

This work proposes a modular robotic platform aimed at supporting the rehabilitation of patients with HI through a combination of multisensory cueing, autonomous navigation, speech-based interaction, and adaptive therapy delivery. The robot will navigate autonomously and safely within the clinical environment, using SLAM (Simultaneous Localization and Mapping) and obstacle avoidance techniques to approach the patient at an appropriate distance for interaction. MediaPipe-based head tracking will be used in conjunction with speech and facial recognition to support an interactive and personalized rehabilitation session. The system will then propose dynamic visual stimuli, head-orienting cues, upper-limb exercises, and cognitive tasks delivered via a touchscreen interface to redirect attention toward the neglected hemispace. The verbal interaction module, based on speech and text recognition [11], will be designed to support context-aware and adaptive communication capable of fully understanding and responding to the patient's intentions throughout the rehabilitation process. Real-time data, including head orientation (to monitor head direction and spatial attention), reaction time (how quickly the patient responds to stimuli), task accuracy (correctness of responses), and error rates (frequency of mistakes), will be continuously analyzed to estimate the patient's engagement and attentional focus. These dynamic indicators will drive the robot's adaptive behavior, enabling adjustments in exercise difficulty, pace, or modality, which will be proposed to the therapist for validation, to maintain an optimal therapeutic challenge tailored to the patient's current ability.

For example, if the system detects signs of inattention (e.g., reduced head movement toward targets) or cognitive overload (e.g., high error rates or slow responses), it will reduce task complexity or suggest a break. Conversely, successful and efficient task completion will trigger a progressive increase in difficulty, in accordance with principles of challenge-based neuroplasticity [12] [13]. The goal is to maintain an optimal therapeutic balance, ensuring that exercises are neither too difficult nor too simple, thereby sustaining patient motivation and maximizing engagement throughout the session. A therapist will nonetheless supervise the session and intervene as needed, ensuring safety, personalization, and clinical oversight. In this model, the robot functions as a consistent, responsive therapeutic.

A preliminary feasibility study has been conducted in the Living Lab of the CoE REDI (Centre of Excellence on Rehabilitation Devices and Digital Instruments) at the University of Pavia. The study utilized the TIAGo robot developed by PAL Robotics, on which we developed the proposed modular platform [14]. This robot originates as an open source, multipurpose social and assistive robot with a large set of sensors and actuators to be controlled for a thorough integration with the environment in which it operates. Differently from other systems like ReoGo [15] or Armeo [16] which physically support the patient, we focused on the assistive features of TIAGo to develop it as a rehabilitation guide and evaluation instrument.

Validation experiments focused on key capabilities required for safe and adaptive operation, including autonomous navigation, patient identification and interaction, and head orientation tracking. The results demonstrate the system's potential to support personalized, data-driven rehabilitation experiences in a clinical context, confirming the suitability of the TIAGo platform for real-world deployment in neurorehabilitation settings.

1.1. System Overview

The proposed rehabilitation platform is organized into five core modules, each responsible for a critical subset of functionalities: Navigation, Patient Identification, Interaction, Therapy, and Evaluation. All modules are implemented within a unified software architecture based on the Robot Operating System (ROS), which provides flexible middleware for integrating sensors, control algorithms, and user interface components. ROS nodes handle the modular functions and communicate in real time, ensuring that perception and action are tightly coupled throughout the therapy session. The current version of the TIAGo robot operates on ROS1; however, given the planned deprecation of ROS1, a custom communication bridge has been developed to enable interoperability with components running on ROS2, ensuring future-proof integration and compatibility with evolving robotic software ecosystems.

- **Navigation Module**: this module manages autonomous movement and positioning of the robot in the clinical environment. It utilizes SLAM techniques to construct a map of the surroundings and localize the robot within it. A combination of 2D lidar and depth camera data feeds into the SLAM algorithm, allowing the robot to detect obstacles and plan safe paths toward the patient's location or designated waypoints. During a session, the navigation module drives the robot approaching the patient at an appropriate distance for interaction, and it can also reposition the robot as needed (for example, to ensure the patient's face remains in view of the camera). The navigation system runs under ROS's navigation stack, leveraging standard packages for path planning and obstacle avoidance while enabling real-time adjustments if a person or object enters the robot's path.
- **Patient Identification Module**: once the robot reaches the general vicinity of the patient, the identification module confirms the patient's identity. We implement a face recognition system using the RGB camera data: the robot's vision software detects human faces in its field of view and matches them to a pre-recorded profile of the target patient.
- **Patient Interaction Module**: the interaction module handles all engagement with the patient using automatic speech recognition (ASR) for input and text-to-speech (TTS) to produce its output. This allows patients to respond to the robot's questions or instructions verbally, for example, saying "I'm ready" to begin, or answering simple queries during the session. The ASR system is designed around a limited domain vocabulary (related to therapy exercises and basic needs) to maximize accuracy in the clinical context.
- **Therapy Module**: it delivers a series of sensorimotor and cognitive exercises to the patient, specifically targeting the neglected hemispace. Exercises include head-orienting responses to the robot's visual stimuli, upper limb reaching actions, and touchscreen-based cognitive games. Head orientation and motor behavior are tracked in real time using the robot's RGB-D camera in combination with a MediaPipe-based skeletal tracking algorithm [17]. These data are continuously analyzed to assess patient compliance and attentional engagement and fed to the Patient Interaction Module to motivate the patient throughout the session.
- **Evaluation Module**: it is responsible for aggregating data from the session and providing analytical feedback. All events and performance data acquired during the session are timestamped and logged in a secure database: this includes quantitative metrics (scores, times, error counts) and sensor readings (head orientation over time, etc.). The evaluation component will analyze these data to produce a session assessment by computing performance indices such as overall accuracy, average reaction time, number of omission errors (e.g., stimuli on the neglected side not responded to), and improvement compared to previous sessions. These figures will be fed back into the system's adaptive algorithms so that the following session's difficulty settings can be adjusted according to the patient's progress. Over longitudinal use, this module supports trend analysis, allowing therapists to track recovery of attention functions over weeks of therapy.

The system's modules organization is summarized as a block diagram in Figure 1.

2. Materials and Methods

2.1. Navigation Module: Setup and Testing

The custom-designed navigation module enables the TIAGo robot to approach a person in its sight and align in front of it when the correct distance is reached. The target distance is set to 1.75 meters from the patient for safety while allowing TIAGo to monitor most of the body of the patient.

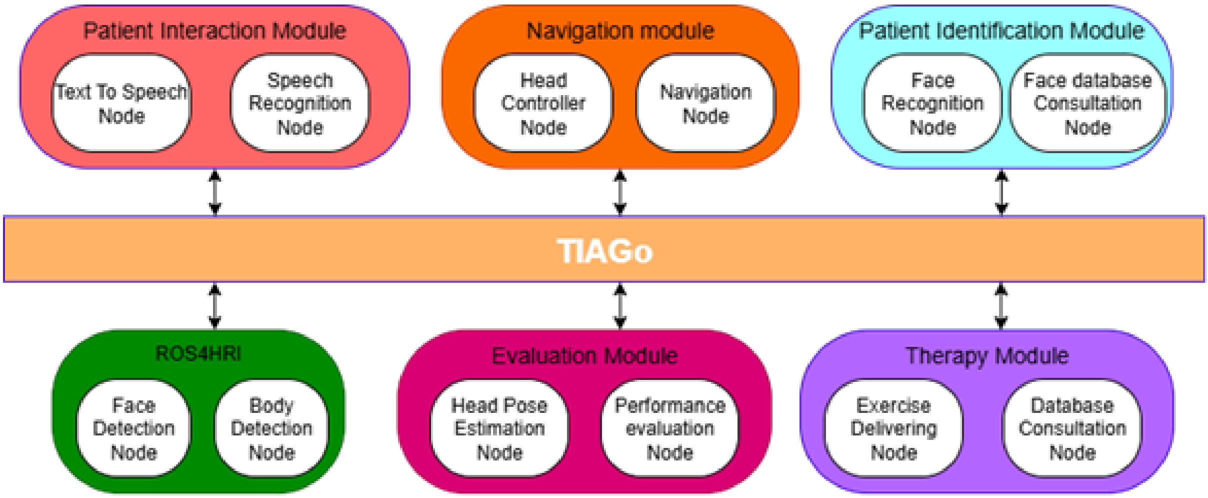

Figure 1: Block diagram of the system that summarizes the module's organization.

The first node involved in this module controls the pan and tilt motion of the robot head to search for a person around the room and even follow its movement. The person detection is enabled by ROS for Human-Robot Interaction (ROS4HRI) package that implements two main nodes for body and face detection. Both nodes rely on MediaPipe machine learning models that reconstruct a stick figure of the body pose and the 3D face mesh.

Two navigation modalities are implemented:

- **Autonomous mode**: the robot operates in this mode whenever the detected person is farther than 2.75 meters. The robot uses its on-board sensors to navigate through the environment relying on a pre-built map of the surroundings. The target in this operating mode is 1.3m from the final target position which in turn is set to 1.75m in front of the person's frame.
- **Manual mode**: activated when the robot is closer than 1.3m from the final target. The navigation behavior of the robot is structured into four sequential phases: **Rotate to target, Move to target, Rotate to face the person** and **Done**.

To verify the robustness and the accuracy of the system we performed 12 experimental trials testing the developed navigation algorithms involving three healthy subjects from our laboratory personnel. Ground-truth measurements were gathered using a 6-cameras stereophotogrammetric system (SMART-DX EVO, BTS Bioengineering, Italy) made available by the REDI center of excellence of Pavia. This instrument is considered the gold standard tool for motion tracking and kinematic analysis. A total of ten reflective markers were affixed on both the robot (four on the corners of the laptop tray and three on the head) and the subject (two on the inferior angles of the scapulae and one on the 7th cervical vertebra) to record their movement at high resolution (100Hz, 1mm precision).

Four experimental trials were conducted for each of the three participants. At the beginning of each trial, the participant was positioned approximately 4 meters from the robot. Throughout the trials, the robot navigated toward a dynamically computed target located 1.75 meters from the participant, utilizing a hybrid approach that combined autonomous and manual control. Following this initial interaction, the participant moved diagonally to a secondary position approximately 1.5 meters from the first, prompting the robot to re-engage its positioning behavior towards the newly defined target.

To compare the robot measurement capabilities with the ground truth, six parameters were computed:

- **Navigation time** (s): time taken by the robot to reach the computed target.
- **Path length** (m): total length of the trajectory executed by the robot during navigation.
- **Distance to first subject position** (m): euclidean distance between the robot and the subject after the first approach, i.e. at the first "Done" message.

- **Alignment angle at first position** (°): angular difference between the robot's frontal axis and the subject's orientation measured after the first approach.
- **Distance to second subject position** (m): Euclidean distance between the robot and the subject after the second approach, i.e. at the second "Done" message.
- **Alignment angle at second position** (°): angular difference between the robot's frontal axis and the subject's orientation measured after the second approach.

2.2. Patient Identification Module: Setup and Testing

The face recognition module enables the robot to identify individuals, such as patients and clinical staff within its visual field in real time. The module operates in two distinct phases: an offline learning phase and an online identification phase. During the learning phase, the system captures multiple facial images per subject using the robot's RGB camera. These samples are processed using the MediaPipe Face Recognition framework [14], which extracts compact facial embeddings. The embeddings are then linked to metadata such as name and role (e.g., patient, therapist) and stored in a structured database. In the identification phase, the robot continuously processes the live video stream to detect and identify faces. MediaPipe compares each detected face against the stored representations using cosine similarity. A modifiable similarity threshold, typically set at 0.8, determines whether a face is confidently matched to a known identity. This ensures a balance between sensitivity and robustness to false positives. The resulting identity and role information is used to personalize the session, i.e., select the therapy and behavior based on interlocutor roles, and maintain logs for patient-specific tracking. The system can also distinguish known individuals from unknown ones, supporting safety and selective interaction logic.

To validate the face recognition module, a dataset was created including seven individuals: two labeled as "doctors", five as "patients", and one as a "nurse." Each participant stood facing the robot at a distance of one meter while facial data were recorded for 30 seconds. For each individual, a unique five-letter identifier was assigned and stored along with their role. To test the system's performance all the seven individuals were asked again, in a random order, to stand in front of the robot for 30 seconds to verify the correct recognition of the person's identity. Two new subjects were introduced to verify that the system would correctly categorize them as "unknown". During the identification phase, the system detects faces in real time and matches them against the stored profiles, providing as output the identifier, the assigned role, and a confidence score indicating the reliability of the match. A second experiment was performed to assess the system's ability to recognize multiple individuals simultaneously, achieved by instructing two subjects at a time to stand in front of the robot.

2.3. Patient Interaction Module: Setup and Testing

The speech recognition and interaction module enables the robot to engage patients in meaningful verbal exchanges throughout the rehabilitation session. This functionality is essential not only for delivering task instructions and monitoring progress, but also for assessing the patient's willingness to exercise, emotional state and to motivate the patient while performing the intended tasks. The module follows a structured pipeline composed of three main components: automatic speech recognition (ASR), intent detection through keyword-based mapping, and a dialogue state manager.

The ASR component transcribes the patient's spoken input using the cloud-based Google Speech-to-Text service, configured for the Italian language. The resulting text is processed by a rule-based intent recognizer, which matches the utterance against predefined phrases associated with specific intent categories, such as *start_confirmation* (e.g., "yes", "okay", "I'm ready"), refusal ("no", "I don't want to"), discomfort ("I feel unwell", "I'm tired"), *exercise_completed*, and *out_of_context*. This deterministic approach avoids the need for model training while remaining effective in the constrained clinical domain, where patient responses are expected to follow known patterns.

Internally, the intent recognizer uses a set of handcrafted regular expressions (regex) applied to normalized text. Each input is first converted to lowercase, stripped of diacritics (e.g., "è" ⟶ "e"), and trimmed of whitespaces. The regex engine sequentially evaluates the utterance against the expressions associated with each intent category. The first matching pattern determines the classification, ensuring deterministic behavior consistent with the clinical context.

Once the intent is classified, a finite-state dialogue manager governs the interaction flow by transitioning across a predefined set of dialog states and selecting the corresponding robot response, which is then output through the text-to-speech node. The main states include:

- **Introduction**: the robot prompts the patient for its name and identifies the patient through both speech and facial recognition.
- **Start_Rehabilitation**: the system checks if the patient is ready to begin. If the patient refuses or expresses hesitation, the robot responds empathetically and may propose a delay, a simpler task, or a motivational message. The transition to the next state occurs only when readiness is confirmed.
- **Exercise**: the robot describes the current task and initiates its execution. During this phase, the system continuously monitors verbal input for signs of discomfort (e.g., pain, frustration). If discomfort is detected, the robot temporarily halts the exercise, offers supportive dialogue, and decides, based on the context and predefined thresholds, to resume, adjust the task, or skip to the next one.
- **Next_Exercise_Proposal**: after each exercise or set of repetitions, the robot proposes the next activity. If the patient shows signs of fatigue or refuses to continue, the robot attempts to re-engage the patient through encouragement or to simplify the upcoming task. Persistent refusal may trigger early termination of the training session and intervention of the therapist.
- **End_Therapy**: this state is reached when the planned sequence of exercises is completed or if the robot detects repeated expressions of discomfort or refusal, signaling that continuing might be counterproductive. The robot then informs the therapist and ends the session with a closing message.

The interaction system architecture relies on structured JSON files to manage patient profiles, therapy plans, and dialogue content. Each patient follows a personalized exercise program, while verbal responses are drawn from a varied response bank, enabling natural, non-repetitive interaction based on dialog state and intent.

Fallback strategies are implemented to ensure robustness. If the ASR fails to capture input or the recognized text does not match any known pattern, the robot prompts the patient to repeat or reformulate (e.g., “Sorry, can you repeat?”). The rule-based intent recognition module was evaluated using a balanced dataset comprising 175 utterances, evenly split across five intent categories: *start_confirmation*, *exercise_completed*, *out_of_context*, discomfort, and refusal. Each utterance was manually labeled with the corresponding ground-truth intent to allow for performance analysis. The module processes normalized input text through a sequence of handcrafted regular expressions, applying fixed priority matching to determine the most likely intent.

2.4. Evaluation Module: Setup and Testing

The core feature of the evaluation module is currently the head tracking, used to assess the patient’s behavior during rehabilitation. To validate the accuracy of this function, 10 healthy subjects were enrolled to perform head movements aimed at validating the measurements of head rotation around the three reference axes.

Two synchronized data acquisition systems were employed:

- **Ground truth** head rotation was computed on data coming from a 6-cameras stereophotogrammetric system (SMART-DX EVO, BTS Bioengineering, Italy). Four reflective markers were positioned on a bike helmet worn by the subject to be used as landmarks for angles computation. Two markers were positioned on the left and right sides, one on the top and the last one on the front of the helmet. Yaw, pitch and roll angles referred to the resting position were computed.

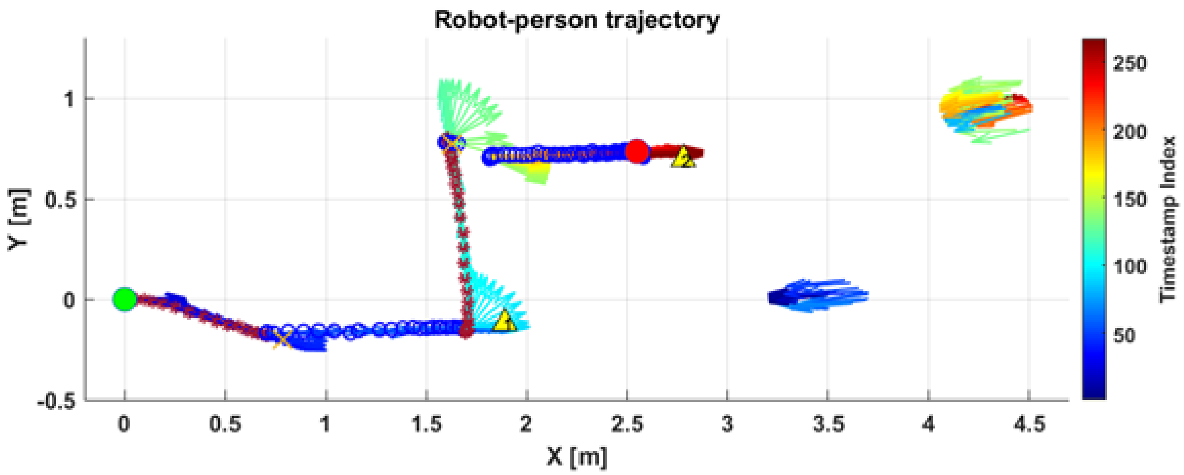

Figure 2: An example of the robot's navigation path, starting at the green dot and ending at the red one. It combines autonomous navigation (red asterisks) with manual navigation (blue circles).

The subjects sat on a chair and were asked to initially maintain the head in the resting position looking straight in to the camera in order to define the initial position, then they were asked to perform 6 head rotations around the 3 axes, one for each direction, in the following order: Yaw left, Yaw right, Pitch up, Pitch down, Roll left, Roll right.

The comparison of the two systems was conducted using the three orientation angles, each calculated independently. To assess the accuracy of the MediaPipe measurements around each axis, the following evaluation metrics were employed:

- Root Mean Square Error (RMSE), which quantifies the average angular deviation between the recordings of two systems;
- Pearson's correlation coefficient (r), which assesses the degree of linear association between the time series of head angle measurements performed by the two systems;
- Maximum Absolute Error, which identifies the largest instantaneous angular difference observed across the trials.

3. Results

3.1. Navigation Module

Figure 2 presents the path followed by the robot during a representative experimental trial incorporating both autonomous and manual navigation phases. The yellow triangles are the two desired targets at 1.75 meters in front of the subject. On the right, arrows indicate the subject location and its orientation as measured by the robot. The trajectory originates from the initial robot location (green dot) and traverses a series of waypoints leading toward target destinations derived from the two positions that the subject occupied during each trial. Segments executed via autonomous control are indicated by red asterisks, while those under manual operation are shown as blue circles. The robot's final position is marked with a red dot. Figure 3 summarizes the comparison results.

3.2. Patient Identification Module

The performance of the face recognition module was evaluated by computing the mean confidence score for each individual over the 30 seconds recognition acquisition. Results showed high and consistent identification accuracy across roles. The mean confidence scores were as follows: Subject1 (nurse) –

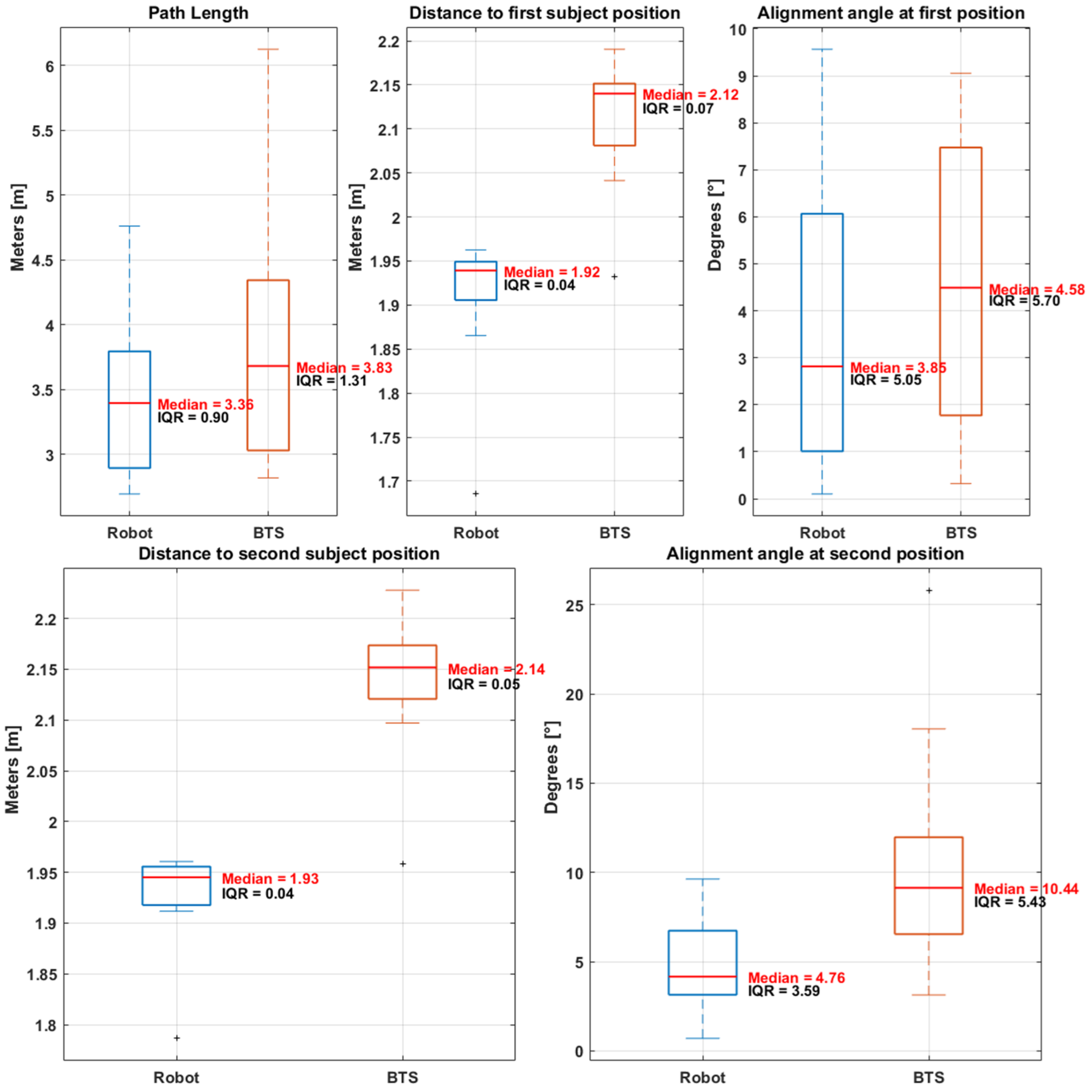

Figure 3: Boxplot report of the comparison between stereophotogrammetry and robot measures.

0.93, Subject2 (patient) – 0.91, Subject3 (doctor) – 0.90, Subject4 (doctor) – 0.92, Subject5 (patient) – 0.91, Subject6 (patient) – 0.92, and Subject7 (patient) – 0.90. These values indicate that the system was able to reliably recognize all registered individuals with high confidence, regardless of role. The two unknown subjects were correctly recognized as not being part of the database of recorded people. Stable confidence levels across categories highlight the robustness of the identification algorithm and its ability to detect unknown people under controlled conditions. The system is also capable of recognizing and identifying multiple individuals (two) simultaneously present within the same scene, enabling contextualized and role-aware interaction. This functionality is essential for supporting flexible, multi-user scenarios in clinical environments.

3.3. Patient Interaction Module

To evaluate the performance of the rule-based intent classification module, we tested it on a balanced test set of 175 utterances, evenly distributed across five intent classes: *start_confirmation*, *exercise_completed*,

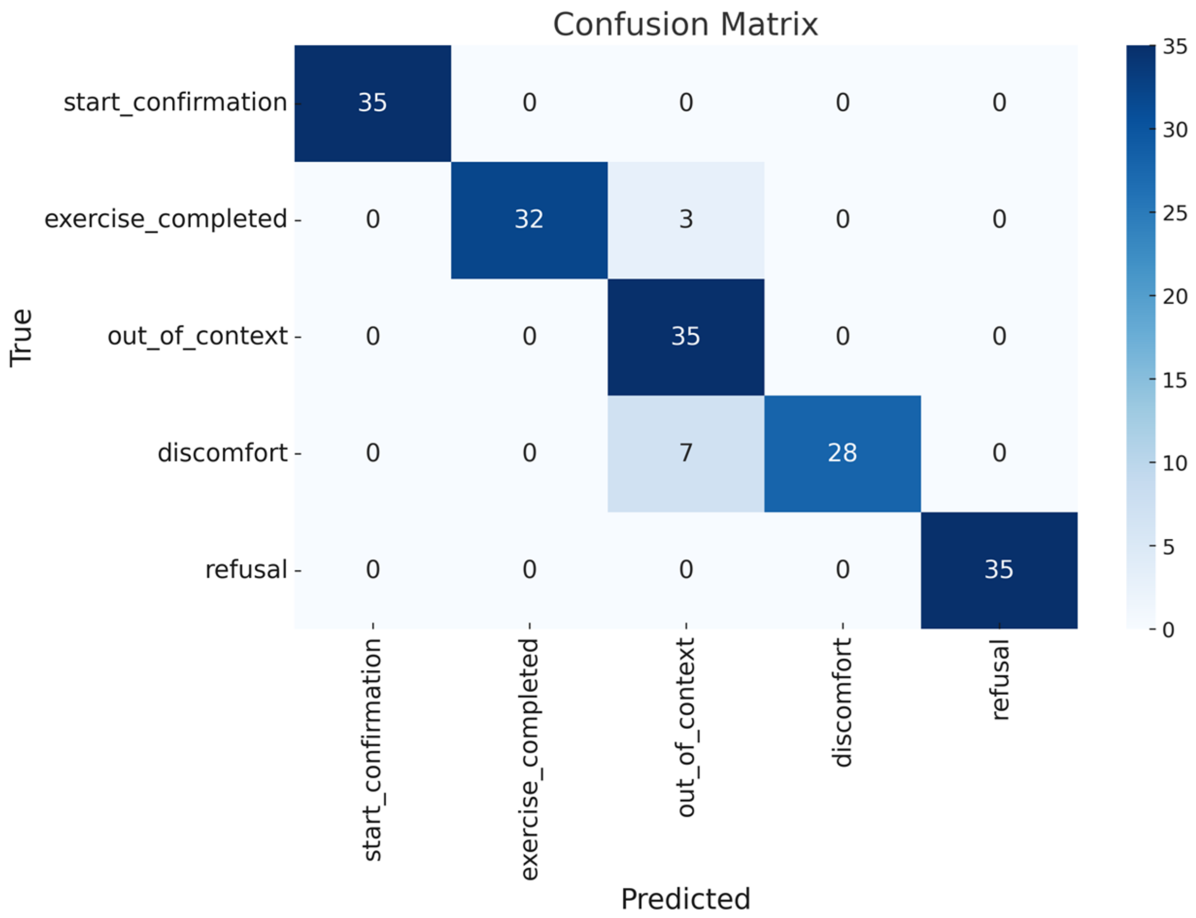

Figure 4: Confusion matrix of the rule-based intent classification module. The only errors occurring are recorded for the class discomfort which is misclassified as *out_of_context* 7 times and the class exercise$_c$*ompletedthatwasclassifiedasout_of_context3times*.

out_of_context, discomfort, and refusal, with 35 samples per class. Each utterance in the test set was manually annotated with a ground-truth intent label. The classifier was evaluated using standard performance metrics, including accuracy, precision, recall, and F1-score for each class. Input sentences were normalized (lowercased, stripped of accents, and whitespace-trimmed) before applying a series of handcrafted regular expressions (regex) designed to detect intent. These regex patterns were applied in a fixed priority order across intent categories. The rule-based classifier achieved an overall accuracy of 94%. The system achieved perfect precision and recall (1.00) for the *start_confirmation* and refusal intents. The *exercise_completed* and discomfort classes had slightly lower recall values (0.91 and 0.80, respectively), due to some utterances being misclassified as *out_of_context*. Conversely, the *out_of_context* class achieved perfect recall, but with a lower precision (0.78), consistent with its fallback role in the rule-matching hierarchy. A confusion matrix of the classification results is shown in Figure 4.

3.4. Evaluation Module

We assessed the viability of a non-invasive head-pose estimation method by comparing angles of yaw, pitch, and roll derived from MediaPipe Face Mesh landmarks with those measured by a stereophotogrammetric system. In a representative trial (Figure 5),the temporal profiles of each angle are overlaid: MediaPipe estimates are shown in blue and BTS reference data in red. Throughout the sequence, the two traces remain closely aligned, yaw and roll show some minor discrepancies, while pitch exhibits

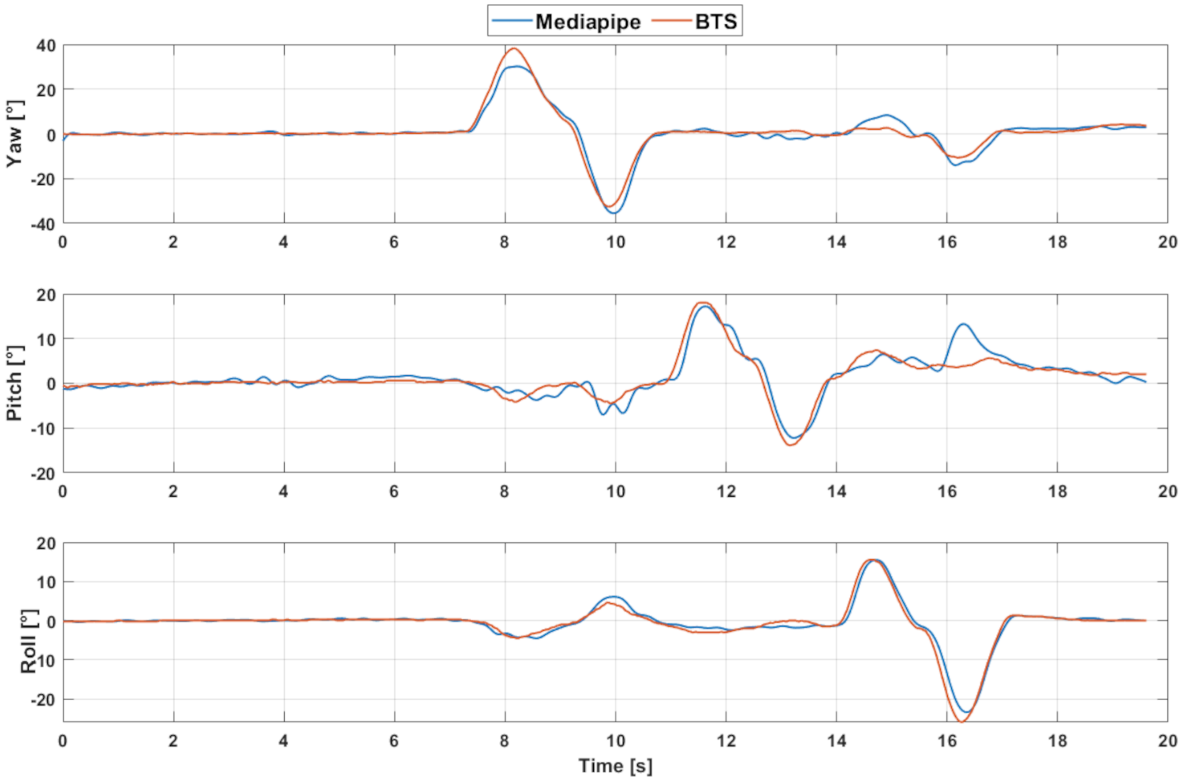

Figure 5: Head angles (Yaw, Pitch and Roll) calculated by MediaPipe (blue traces) and from the data recorded with the stereophotogrammetric system (red traces).

a larger offset. This increased divergence in pitch likely arises because upward or downward tilts move facial features away from the camera's central axis, degrading the accuracy of 2D landmark detection and consequently the 3D orientation reconstruction by MediaPipe. Statistical analysis using Pearson's correlation coefficient, root-mean-square error (RMSE), and maximum deviation reveals excellent concordance across all three axes (with yaw and roll both having r > 0.98) and RMSE values within the thresholds for real-time clinical monitoring (less than 10° - Figure 6).

4. Discussion and Future Directions

The preliminary results on the operation of the proposed robotic platform are encouraging, demonstrating the feasibility of integrating robotic navigation with speech, vision, and interaction capabilities to support personalized rehabilitation for patients with HI. Its modular architecture, built on the ROS and structured around JSON-based configurations, facilitates flexible development, targeted evaluation, and rapid adaptation of each subsystem to specific therapeutic goals and user needs. As a future development, the full migration of the TIAGo robot to ROS 2 is planned, enabling deployment of all software modules directly on the robot and ensuring long-term maintainability and compatibility with evolving robotic middleware standards. This will imply the suppression of the bridge container, which we expect to impact positively on the robot's processing times.

Preliminary testing of key modules yielded promising results. The navigation framework allows a fast and safe navigation towards the patient, correctly following its movement in space. The rule-based intent classification module achieved high accuracy across essential interaction categories, including confirmation and refusal of the treatment session, hinting its suitability for structured clinical dialogues. The face recognition module also performed reliably, successfully identifying known individuals while rejecting unknown ones based on a compact but realistic dataset consisting of seven participants with simulated clinical roles (e.g., patient, doctor, nurse). Notably, the system demonstrated the capability to

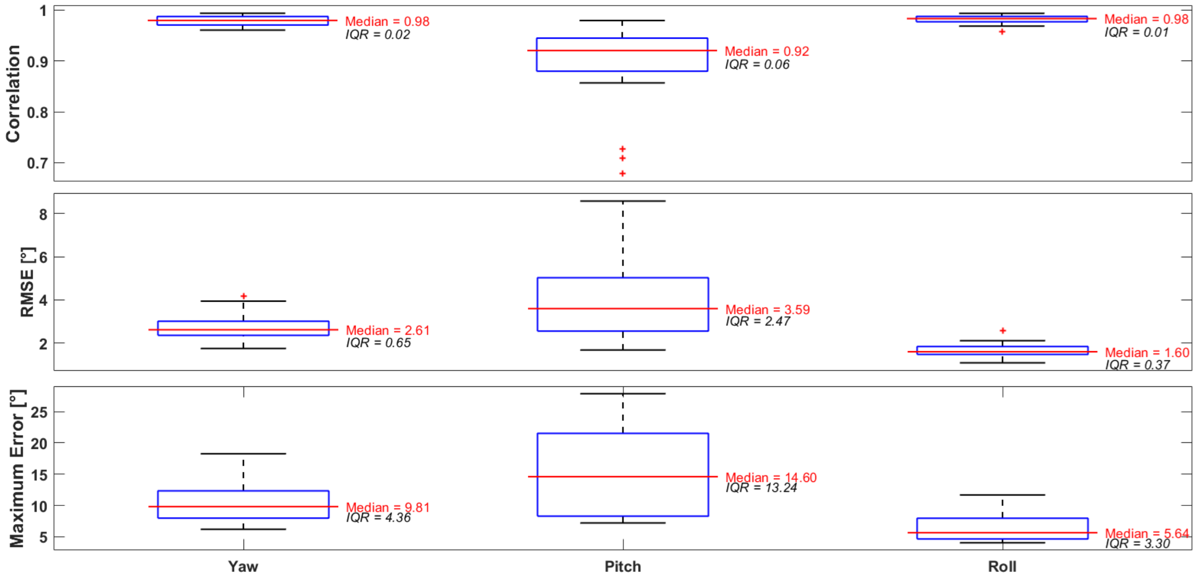

Figure 6: Boxplot representing the distribution around its central tendency of the metrics computed for the evaluation module. Median and Interquartile Range (IQR) are reported.

detect and distinguish multiple individuals simultaneously within the same scene, supporting contextualized and role-aware interaction. The evaluation module showed that markerless pose estimation accuracy is suitable for clinical applications with and RMSE much lower than 10° for yaw and roll rotations. These findings collectively highlight the system's potential to deliver structured, personalized rehabilitation sessions with real-time feedback and adaptive behavior.

Nevertheless, several limitations must be addressed. The intent recognition system currently relies on handcrafted regular expressions, which—while effective in constrained domains—lack the flexibility to handle diverse or paraphrased patient responses. This could limit generalizability in real-world deployments, but system computational power is limited, and state of the art Large Language Models might be too much demanding for getting fast responses; although the system can access the web for software updates, we do not want to completely depend on internet access for remote language processing services. Similarly, although the face recognition module performed well in initial tests, its evaluation was limited to a small group of participants under controlled conditions. The system's robustness under more challenging and variable clinical settings remains to be validated. Furthermore, while individual modules have been tested in isolation, full-system integration and continuous operation across multi-session rehabilitation programs are still under development. Finally, head pose estimation is limited to differential measurements, since the position of the RGB camera relative to the stereophotogrammetric system was unknown. As a result, the comparison relies on relative rotations from the resting head orientation, which itself is not precisely defined.

Future work will focus on overcoming these limitations to advance the platform toward clinical readiness. Specifically, the integration of machine learning–based natural language understanding (NLU) models will improve the system's ability to process unconstrained speech and adapt to patient variability. The face recognition pipeline will be expanded to include larger and more diverse datasets, enabling better generalization to real-world conditions. The evaluation module will be extended to take into consideration the tracking of the upper body of the patient and specifically that of the upper limbs for evaluating the response to pick-and-place exercises proposed by the robot. In the next iteration of this work, the head pose tracking will be implemented directly on the images collected by the robot onboard camera. Additionally, work is ongoing to refine the therapy and evaluation modules for automated adaptation of task difficulty based on real-time patient performance and engagement metrics.

Before the deployment for use in clinical rehabilitation, pilot studies in clinical environments will assess the system's usability, therapeutic effectiveness, and acceptance by both patients and healthcare professionals. These evaluations will provide essential feedback for refining the platform and ensuring its suitability for real-world deployment. In parallel, the current JSON-based configuration system will be replaced by a secure, centralized database to manage patient demographics, therapy schedules, exercise parameters, and performance. This upgrade will enhance scalability, support integration with hospital information systems, and enable long-term monitoring of patient progress across rehabilitation sessions.

Acknowledgments

This work was supported by the Italian Ministry of Research, under the complementary actions to the NRRP "Fit4MedRob - Fit for Medical Robotics" Grant (PNC0000007).

Declaration on Generative AI

During the preparation of this work, the authors used generative AI tools (specifically, OpenAI's GPT-4) to assist with grammar and spelling checks. Furthermore, the custom dataset for the evaluation of the patient interaction module was created with generative AI tools (Google Gemini and OpenAI's GPT-4).

References

[1] H. B. Coslett, Neglect syndromes, in: Encyclopedia of the Human Brain, 2002, pp. 719–732.

[2] G. Kerkhoff, Neurovisual rehabilitation: Recent developments and future directions, Journal of Neurology, Neurosurgery & Psychiatry 74 (2003) 710–712.

[3] M. Jehkonen, et al., Visual neglect as a predictor of functional outcome one year after stroke, Acta Neurologica Scandinavica 104 (2001) 408–415.

[4] A. M. Okamura, et al., Medical robotics for training and rehabilitation, IEEE Transactions on Biomedical Engineering 61 (2014) 1611–1624.

[5] D. J. Reinkensmeyer, L. E. Kahn, Understanding and treating arm movement impairment after chronic brain injury: progress with the arm guide, Journal of Rehabilitation Research and Development 37 (2000) 653–662.

[6] T. D. Feasel, et al., Kinematic evaluation of upper-limb robotic rehabilitation devices: A review, Journal of NeuroEngineering and Rehabilitation 10 (2013) 119.

[7] G. M. Dissanayake, et al., Multimodal sensing and feedback in assistive robotic systems, IEEE Access 8 (2020) 102847–102865.

[8] M. Tapus, C. Tapus, M. J. Mataric, Socially assistive robots for individuals with cognitive impairments, Interaction Studies 11 (2010) 479–502.

[9] T. Belpaeme, et al., Social robots for education: A review, Science Robotics 3 (2018) eaat5954.

[10] F. P. Romano, et al., Towards affective hri: Emotion recognition in real-time for socially assistive robotics, in: Proceedings of IEEE RO-MAN, 2021.

[11] D. Griol, et al., A conversational agent for personalized rehabilitation therapy, Expert Systems with Applications 145 (2020) 113120.

[12] J. Kim, et al., Affective computing in robot-assisted therapy: Current challenges and future directions, IEEE Reviews in Biomedical Engineering 13 (2020) 349–364.

[13] M. A. Maier, et al., Principles of neurorehabilitation after stroke based on motor learning and brain plasticity mechanisms, Frontiers in Human Neuroscience 13 (2019) 344.

[14] PAL Robotics, Tiago – mobile manipulator robot, https://pal-robotics.com/robot/tiago/, 2025. [Accessed: Jun. 23, 2025].

[15] T. Takebayashi, Y. Uchiyama, K. Domen, Automatic setting optimization for robotic upper-extremity rehabilitation in patients with stroke using reogo-j: a cross-sectional clinical trial, Scientific Reports 14 (2024) 25710. doi:10.1038/s41598-024-74672-2.

[16] G. Galeoto, A. Berardi, M. Mangone, L. Tufo, M. Silvani, J. González-Bernal, J. Seco-Calvo, Retracted: Galeoto et al. assessment capacity of the armeo® power: Cross-sectional study, Technologies 11 (2023) 125. doi:10.3390/technologies12110213, also published in Technologies 2024, 12, 213.

[17] C. Lugaresi, et al., Mediapipe: A framework for perceiving and processing reality, in: Third Workshop on Computer Vision for AR/VR at IEEE Computer Vision and Pattern Recognition (CVPR), volume 2019, 2019.