ACSO/BahiaRT@Home 2025 Team Description Paper

Gabrielle F. S. Carvalho[1] Elias R. da Silva[1] Kaique W. S. da Silva[1] Luis J. S. Junior[1] Pedro H. O. dos Santos[1] Pedro L. J. Santos[1] Reinaldo Júnior[1] Samuel J. Cesar[1] Vitória F. N. Matos[1] José Grimaldo da Silva Filho[1] Marco A. C. Simões[1] Jorge A. Campos[1] Josemar R. de Souza[1] Ana Patrícia F. M. Mascarenhas[1]

***Abstract*—This paper presents the BahiaRT team. It describes an autonomous service robot named BILL and its capabilities, such as navigation, manipulation, people and object recognition, human-robot interaction and decision-making as well as It's hardware and software systems. Furthermore, the paper highlights BahiaRT research interests and scientific contributions. Complementary information about our robot BILL, its codes and research are available at the team's GitLab repository: https://gitlab.com/bahiart/athome**

***Index Terms*—RoboCup Brazil, BILL, service robotics, Object Manipulation, XTTS-V2, RoboCup@Home.**

I. Overview

The Center of Computer Architecture, Intelligent Systems and Robotics (ACSO) at the University of the State of Bahia (UNEB) has been participating in RoboCup with the BahiaRT team since 2009 in leagues such as 2D Soccer Simulation, Mixed Reality, 3D Soccer Simulation, and @Home.

The service robot proposal of BahiaRT for the RoboCup@Home league is called BILL (Bot Intelligent Large capacity Low cost). It was created in 2014 due to research projects in assistive robotics. Its main goal is to assist humans in everyday tasks. It has some capabilities to do this, such as communicating with humans through natural and sign language processing, recognizing people and objects and navigating through unknown environments.

Over the years, BILL has participated in both the RoboCup@Home and Brazilian RoboCup@Home competitions, securing notable achievements. In the RoboCup@Home league, BILL achieved 13th place in 2015, 21st place in 2016 and 13th place and 2025. In the Brazilian RoboCup@Home league, it earned second place in 2015, 3rd place in 2016 and 2017, 6th place in 2023, and seventh place in 2024.

This paper presents the third generation of BILL, named **BILL Estranho**, and its main improvements: re-engineering the hardware with new components, redefining the architecture, changing speech recognition model and transitioning the operating system to ROS2. This year's updates focused on other functionalities like human-robot interaction, face and object recognition, navigation, and manipulation.

University of the State of Bahia

∗ This work is partially funded by CNPq/PIBIC, FAPESB/IC and UNEB/PICIN

[1]Centro de Pesquisa e Arquitetura de Computadores, Sistemas Inteligentes e Robótica (ACSO), Universidade do Estado da Bahia (UNEB), Salvador, BA, Brazil. teambahiart@gmail.com

The remainder of this paper is organized as follows: Section II introduces our research group's interests and achievements. Section III details the team's contributions to the league. Section IV describes a task performed by BILL, and Section V presents the conclusions and future work.

II. Advances in innovative technology and research interests

The main research interests of the BahiaRT team involve robotics and artificial intelligence, specifically focusing on human-robot interaction, object recognition, and the fault tolerance of BILL's hardware.

Regarding human-robot interaction, in recent years, the team has focused on research to improve BILL's communication using Large Language Models. Additionally, the team has been working on enabling interaction with hearing-impaired individuals through sign language.

For object detection, we have been working on improving recognition accuracy through the integration of Optical Character Recognition (OCR) [3] and image recognition to enhance the identification of objects manipulated by BILL.

Concerning fault tolerance, BILL's current release adopts a method that we call Fox-Dog. This concept, created by the ACSO research group, consists of a nature analogy: a dog chases the fox, and the fox chases the dog until one of them fails. Utilizing this, the model provides more security and infallibility for BILL to control the front and rear left/right motors. The method ensures that if one microcontroller fails, the other takes over.

A. Navigation

Navigation is the cornerstone for efficient execution and interaction with the robot environment. BILL's navigation components include encoder output, odometry, Slam-toolbox, a Behavior-Tree-based navigation node called bt-navigator (ROS2), Adaptive Monte Carlo Localization - AMCL (ROS2), map-server (ROS2), and a 360° laser scanner.

The odometry module utilizes data from the encoders and LiDAR to estimate BILL's movements in space. Additionally, the Behavior-Tree Navigator leverages odometry data to calculate the trajectory to a desired target. Once all data are published, simultaneous mapping and localization using AMCL are activated, integrating 360-degree laser scan data.

The Simultaneous Localization and Mapping (SLAM) is responsible for mapping the environment and providing self-localization within this map. First, the incremental mapping

package constructs the map using Slam_toolbox [4] and a LIDAR sensor generates the grid map. Then, the path plan is created based on the occupancy grid map, which is updated using the navfn_planner. The shortest path to the goal is then computed using the D*Lite [8] algorithm to ensure obstacle avoidance during incremental mapping. Motion planning, which is responsible for processing the path plan and determining the required linear and angular motions, is triggered. It applies the kinematic control law and sends commands to the low-level control system.

B. Vision

This module handles the reception, processing, and response to external stimuli through image capture. It consists of three main sub-modules: (i) object detection, integrating image recognition and OCR; (ii) facial recognition; and (iii) communication using sign language.

Concerning object detection, BILL uses YOLO [6] version 8, a technology capable of real-time detection across a wide range of objects. The training process involved photos of distinct objects captured in various locations and orientations. The team used img with varying numbers of objects per frame to assess the impact of different parameters and the quantity of training data on recognition accuracy. The RoboFlow [7] software supports the labelling process. To enhance the accuracy of our object detection system, we integrate YOLO with Tesseract [9], an OCR tool. This approach has already proven effective in industries such as manufacturing and traffic management, and our team believes that it could also make valuable contributions to the service robot domain.

Inspired by BILL's focus on high-capacity, low-cost solutions, this research presents a two-stage system for efficient facial recognition. The first stage employs the Haar-cascade [10] algorithm from OpenCV [11] to rapidly detect and isolate faces within an image. The second stage utilizes Dlib's [12] landmark detection, mapping 68 key facial coordinates to capture unique features. This extracted data is then used to train the facial recognition model. To gauge the system's effectiveness, we conducted recognition tests with 55 students from the University of the State of Bahia. After a brief training period, the algorithm demonstrated its ability to identify specific individuals within a group setting.

Sign language communication aims to facilitate robot-human interaction for hearing impairments people. This first release uses Brazilian Sign Language, known as LIBRAS [13]. Figure 1 illustrates the communication process. The person performs the sign, which is captured by BILL's camera and processed by our solution, RoboSign [2]. It converts the sign into a text instruction and uses it as input for BILL's natural language processing, similar to voice interaction, improving BILL's communication accessibility. We are currently working on a new sign dataset in American Sign Language (ASL) using the same process.

The RoboSign solution uses the MediaPipe [14] landmark detection software to capture signs for dataset creation. Then, a recurrent neural network (LSTM) [15] uses this dataset

Fig. 1. Communication process using sign language.

for training and validation. To assess RoboSign's accuracy, a LIBRAS professor from the University of the State of Bahia interacted with BILL in real time. The results achieved 85% accuracy in recognizing the signs, demonstrating its potential to make BILL more accessible in the human-robot interaction. More information about this research can be found in our GitLab repository, which is linked in the abstract section.

C. Manipulation

Object manipulation plays a crucial role in the interaction with a domestic environment. BILL is equipped with a 5 DoF TurtleBot Robotic Arm, with clamp, built from a set of 5 Dynamixel AX-12+ model servo motors, 4 of which are for joints and 1 for the claw. The system has a joint with a vertical prismatic degree of freedom, driven by a stepper motor, to level the arm when handling objects. This system was controlled by an Arduino Arbotix-M, but for this year, it was modified for a new controller based on Arduino MKR Zero board with Robotis Dynamixel Shield designed for this purpose. This improvement was motivated by the fact that the new board is compatible with the Micro-Ros system.

D. Speech Recognition and Voice

Voice is the most widely used form of human-robot interaction, and we have adopted it as the primary communication method with BILL. Following BILL's upgrade to ROS2, we integrate the XTTS-V2 software to enable voice interaction. Our team's speech recognition system employs advanced machine learning techniques, specifically the GPT-2 [16] model for understanding and generating responses in natural language and DistilBERT [17] for question-answering in specific contexts.

The integration of XTTS-v2 for speech synthesis and ChatGPT (version 2) for natural language understanding and generation enables smooth and intuitive voice interaction with the service robot. This allows users to communicate through spoken commands and receive natural, context-aware responses, making the robot more effective in domestic and social settings. The voice-based interface enhances accessibility, especially for individuals who may have difficulty using traditional input methods. Considering some problems when using the past models, the team has decided to change from Google Speech Recognition (GTTS) to XTTS-V2, for two main reasons.

- **Offline solution:** A significant advancement in this system is its ability to operate entirely offline. Both speech processing and language understanding run locally on the robot, eliminating the need for an internet connection. This increases the robot's autonomy and reliability, making it suitable for environments with limited or no connectivity, and improving privacy by keeping all data processing on-device.
- **Using GPU acceleration.:** To support the computational demands of real-time voice interaction, the system leverages GPU acceleration. Running inference tasks on a GPU dramatically reduces response time, allowing the robot to process speech and generate answers quickly and efficiently. This ensures a seamless user experience, even when handling complex tasks or multiple interactions in rapid succession.

Furthermore, GPT-2 has been fine-tuned with a custom dataset of command phrases and natural language interactions relevant to BILL. The training was conducted using the TensorFlow [18] and PyTorch [19] libraries. Additionally, we use the Speech Recognition library to interface with XTTS-V2.

To train the DistilBERT [17] algorithm, we adopted the SQuAD [20] dataset from Stanford University, along with a dataframe divided into three columns: question, answer, and context. During the inference phase, the system captures voice commands through BILL's array of microphones and preprocesses it to enhance clarity and reduce background noise. The audio is then transcribed into text using XTTS-V2. For general conversation, the transcribed text is fed into the fine-tuned GPT-2 model, which interprets it and generates an appropriate response. Similarly, the transcribed text is fed into the DistilBERT [17] algorithm for quiz tasks. For more information about the GPT-2 model, access our GitLab repository, which is linked in the abstract.

Audio capture is a critical component of our speech recognition system. To optimize sound quality, BILL is equipped with a high-quality microphone strategically positioned in the operator clothing for optimal performance. The audio processing pipeline includes noise reduction to filter out background sounds and enhance the signal-to-noise ratio. Additionally, real-time processing ensures simultaneous audio capture and analysis, enabling rapid responses to voice commands.

This comprehensive approach to speech recognition and voice interaction aims to create a robust and reliable system that enhances BILL's human-robot interaction, making it more intuitive and efficient.

III. Contributions

This section summarizes BILL's recent contributions.

BILL Estranho's electronic structure was improved with better protection against short circuits and higher power. The system utilizes ROS2, micro-ROS (which integrates ROS2 onto microcontrollers) [21], and ESP32. In an unprecedented way, Fault Tolerance is applied using the Fox-Dog method in Service Robots. The fault-tolerant Fox-Dog method is

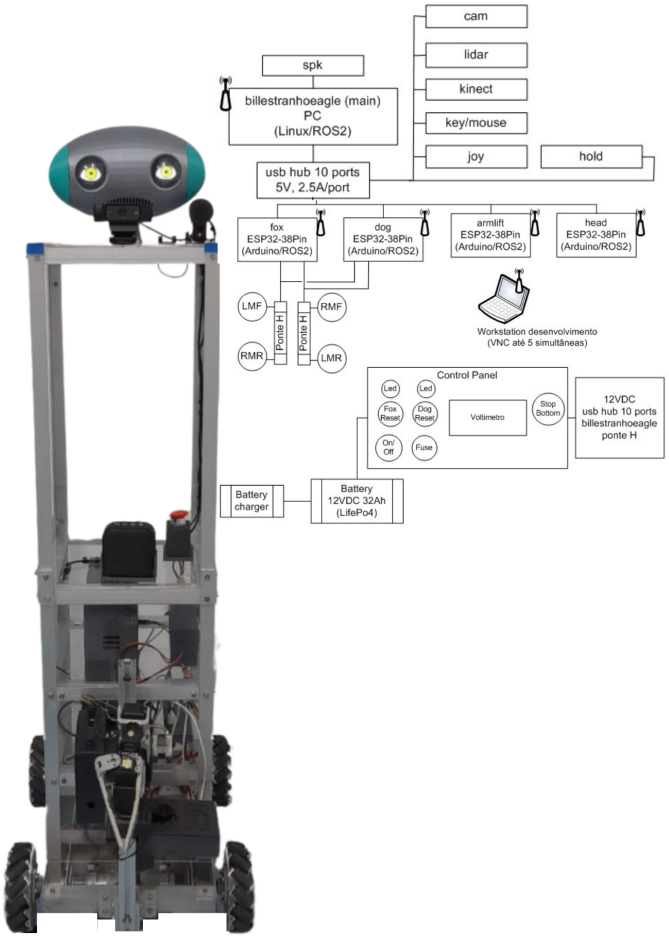

Fig. 2. BILL Estranho.

implemented using two ESP32 microcontrollers replicated with ROS2. These microcontrollers control the left and right front motors (LMF/RMF) and the left and right rear motors (LMR/RMR). If one microcontroller fails, the other takes over (2).

For better accessibility in human-robot communication, the RoboSign solution enables BILL to recognize Brazilian Sign Language (LIBRAS), allowing communication with people with hearing impairments. RoboSign is available at our GitLab repository, and other teams can use, study, and increase it.

Finally, concerning object detection, we are currently working on the integration of YOLO and Tesseract for Optical Character Recognition and we have some results that indicate an improvement in detecting objects that have labels, e.g. supermarket products.

IV. Example of Task Performed by BILL

This section describes the "Receptionist" task, which integrates some of BILL's capacities. The task starts with introducing a person to BILL, i.e., the person interact with BILL, telling him their name, something they're is interested in, and their favorite drink, while having his face trained.

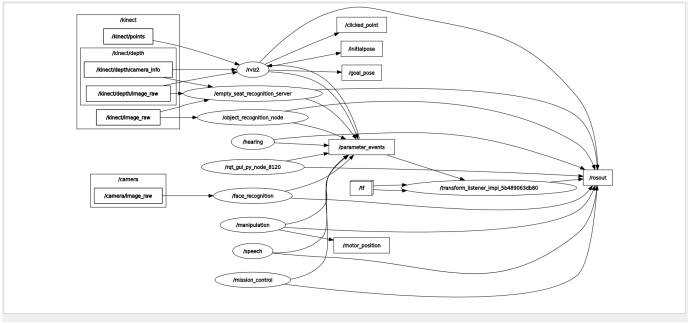

Fig. 3. People Recognition task.

After the introduction, the person moves to the drinks area, where BILL recognize and say if there is their favorite drink in the table. Then, BILL guides the person to the living room, where it recognize and say if there's an empty seat for the person or if there isn't. As soon as the person seats or keep standing (in case of not having an empty seat), BILL start to introduce the new guest to the others in the living room, telling them their interests and saying if they have similar interests. In order to implement it, we developed a main control that orchestrates vision, speech, and navigation resources. It comprises four nodes: *imageCaptureNode*, to take pictures of the person; *peopleRecognitionNode*, to train and recognize a person; *turnAroundNode*, to move BILL to the crowd; and *voiceNode*, responsible to the whole communication between BILL and the person. Figure 3 presents the RQT Graph of this task.

Similarly, BILL can perform a common everyday task at the demonstration level: picking up an object and placing it on a designated surface. To accomplish this, BILL positions himself in front of the object's original location, then tilts his manipulator toward the object, grasps it, and lifts it. Next, he moves to the destination, stops before it, positions his claw over the designated surface, and releases the object.

V. Conclusion

This paper introduced BILL, the BahiaRT Team's service robot for the RoboCup@Home league. BILL has the necessary capabilities to perform most of the competition's tasks. Among the latest improvements, we highlight the adoption of LLMs for speech interaction and the integration of object detection using YOLO and Tesseract. Both have significantly increased BILL's accuracy. Additionally, the fault-tolerance system has enhanced reliability, providing extra security for BILL's actions and task performance.

Our current work focuses on motion detection and sign recognition. We have achieved promising results in motion detection. However, this area is broad and multidisciplinary, requiring extensive work to achieve advanced capabilities. Regarding sign recognition, BILL is an inclusive robot capable of communicating through sign language. We are currently upgrading its dataset to recognize signs in multiple languages.

References

[1] Gomes, Vanessa Martins: Reft: sistema integrado de reconhecimento de emoções na fala e no texto para o idioma brasileiro. Departamento de Ciências Exatas e da Terra, UNEB, Campus I.

[2] Brito, Aron Caiuá Viana de: Robosign: IA aplicada para interação entre deficientes auditivos e robôs de serviço. Departamento de Ciências Exatas e da Terra, UNEB, Campus I.

[3] Ye, Q., Doermann, D.: Text detection and recognition in imagery: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence **37**(7), 1480–1500 (2015)

[4] Macenski, S., Jambrecic, I.: SLAM Toolbox: SLAM for the dynamic world. Journal of Open Source Software **6**(61), 2783 (2021)

[5] Koenig, S., Likhachev, M.: D* Lite. In: Eighteenth National Conference on Artificial Intelligence (AAAI), pp. 476–483 (2002)

[6] Jocher, G. et al.: YOLO by Ultralytics. GitHub repository. https://github.com/ultralytics/yolov5, last accessed 2025/07/31

[7] Roboflow. https://roboflow.com, last accessed 2025/07/31

[8] D*Lite. https://github.com/Sollimann/Dstar-lite-pathplanner, last accessed 2025/07/31

[9] Smith, R.: An overview of the Tesseract OCR engine. In: Ninth International Conference on Document Analysis and Recognition (ICDAR), pp. 629–633 (2007)

[10] Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 511–518 (2001)

[11] Bradski, G.: The OpenCV Library. Dr. Dobb's Journal of Software Tools (2000)

[12] Kazemi, V., Sullivan, J.: One millisecond face alignment with an ensemble of regression trees. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1867–1874 (2014)

[13] Gameiro, E. et al.: A Brazilian Sign Language Video Database for Automatic Recognition. ResearchGate Preprint. https://www.researchgate.net/publication/348336044, last accessed 2025/07/31

[14] Lugaresi, C. et al.: MediaPipe: A Framework for Building Perception Pipelines. arXiv preprint arXiv:1906.08172 (2019)

[15] Hochreiter, S., Schmidhuber, J.: Long Short-Term Memory. Neural Computation **9**(8), 1735–1780 (1997)

[16] Radford, A. et al.: Language Models are Unsupervised Multitask Learners. OpenAI Technical Report (2019)

[17] Sanh, V., Debut, L., Chaumond, J., Wolf, T.: DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108 (2019)

[18] Abadi, M. et al.: TensorFlow: Large-scale machine learning on heterogeneous systems. https://www.tensorflow.org, last accessed 2025/07/31

[19] Paszke, A. et al.: PyTorch: An Imperative Style, High-Performance Deep Learning Library. Advances in Neural Information Processing Systems (NeurIPS) **32** (2019)

[20] Rajpurkar, P., Zhang, J., Lopyrev, K., Liang, P.: SQuAD: 100,000+ Questions for Machine Comprehension of Text. arXiv preprint arXiv:1606.05250 (2016)

[21] micro-ROS Project. https://micro.ros.org, last accessed 2025/07/31