MARTI: Market Assistant Robot for Transit & Interaction

Daniel Pawlak
daniel.pawlak@kcl.ac.uk
Ismail Hendryx
ismail.hendryx@kcl.ac.uk
Kautilya Chappidi
kautilya.r.chappidi@kcl.ac.uk
Nii Tettey
nii.tettey@kcl.ac.uk
Dev Atreya
dev.atreya@kcl.ac.uk
Mohamed Mohamed
mohamed.m.mohamed@kcl.ac.uk

***Abstract*—This paper presents MARTI (Market Assistant Robot for Transit & Interaction), a Pepper-based humaoid robot developed for customer assistance in retail environments. MARTI leverages its sensors like colour and depth cameras, microphones and lasers for human and gesture detection, emergency response and product identification. The robot attempts SLAM for navigation for product retrieval and guidance, responding to voice commands when interacting with customers. Challenges using legacy systems and processing limitations are discussed alongside suggestions for future improvements for performance in real-world scenarios.**

***Index Terms*—computer vision, human-pose, human detection, SLAM, human-robot interaction**

I. Introduction

Humanoid robots are becoming more and more integrated into dynamic environments with human presence [1], hence creating the need for safe and responsive human interaction [2] to provide aid and information. There are many examples of these robots. This paper explores the use of these in mall and supermarket scenarios with a strict focus of sensing the surroundings and developing a robot perception that creates a foundation the chosen Pepper robot can use to base its behaviour around. This robot was chosen in particular due to its vast range of sensors (including both colour and depth cameras, sonar, microphones, accelerometers and ultrasonic) and the vast range of actuators to inhibit human interaction with ease. The robot has been worked with in person, however, it has also been explored in a simulation environment to showcase its capabilities and designed systems. This paper will discuss previous works and state-of-the-art in the field, the objectives with hypothetical questions to be answered by reaching established aims, the developed methodology to reach those aims, the corresponding results, and finally conclude the project with ideas for future improvements and challenges faced.

II. Literature Review

A. Motivations

There is a vast amount of evidence of growing interest and potential capabilities with studies [3]–[5] looking into consumers' perception of the usefulness, social capabilities and human collaboration in retail environments to enhance service and affect purchasing behaviours. These show there is a clear drive in this sector to implement humanoid robots, leading to the idea of developing MARTI using Pepper. It is important to keep in mind the ethics of this to create a robot that is an honest and reliable assistant rather than a manipulative salesperson [6], [7] for trustworthy integration into human life.

B. Previous Works

One of the most important aspects that should always be addressed first is effectively and safely sharing space [8] during interactions and movement addressed by SLAM (simultaneous localization and mapping), which enables mapping out the environment to avoid collisions seen in many motion planning research [9]–[11]. Past this foundation, the project can focus on specialist tasks such as human detection and behaviour analysis, which involves acknowledging humans in the sensing space and picking up gestures. This can lead to shoplifting detection systems [12], [13] and human pose landmarking [14], [15] that the robot can react to accordingly by calling for help in case of a safety or health emergency. Another staple of retail assistance is the detection and segmentation of products in these stores to assist in retrieval and identification seen in various studies [16]–[18], leading to the identification and proposal of these store products. To communicate with potential customers the robot should be able to listen, understand and respond to any spoken queries achieved by audio recognition algorithms such as [19], [20].

C. State-of-Art

Ameca [21] is a recent humanoid robot purposefully built for human-robot interaction, equipped with cutting-edge technology allowing it to exhibit facial expressions. Combined with natural language processing capabilities and lifelike motions, it makes it capable of highly engaging interactions [22]. Apart from this conversational capability, it is also equipped with a range of perception systems such as high-resolution cameras for vision, microphones for audio recognition and depth sensors for spatial awareness. Ameca can understand human gestures and emotions whilst integrating complex motion planning algorithms for natural navigation through dynamic environments. This combination of perceptual and motion skills shows the inspiration this project can take to develop a shopping assistant.

III. Objectives and Aims

Overall, this project intends to create a shopping assistant robot that can help efficiently locate and guide consumers to various products and locations within a locally mapped store environment. Human interaction is aimed to be achieved via human detection, gesture tracking and conversation to respond and react to queries and potential emergencies. To answer whether this is possible on the Pepper platform there are certain objectives this project aims to complete:

- **Human and Shoplifting Detection:** Detect the presence of customers whilst identifying potential shoplifting incidents.
- **Voice Command Recognition:** Allowing the robot to recognise specific phrases to trigger the correct responses.
- **SLAM:** Mapping out the environment to enable efficient path planning and obstacle avoidance.
- **Object Recognition:** Enables the assistant to identify and point out items.
- **Pose Landmarking:** Identifies customer movement to detect gestures and identify potential emergencies.

IV. Methodology

A. Experimental Setup

For this experiment, the Pepper platform was used seen in Figure 1 however, this turned out to show some limitations that had to be accounted for. This includes the constraint to using Python 2.7, no internet access for any API requests, as the local network had to be used and some problems with stable, prolonged camera access. Due to time constraints, most work developed was done using our own camera testing and eventually a simulation environment using QiBullet that allowed virtual connections to the sensors offered on Pepper.

B. Experimental Procedure

*1) **Human and Shoplifting Detection**:* This script employs a multi-step approach to detect faces, eyes, and track the pupil using a combination of YOLO (You Only Look Once) and Haar Cascade classifiers. Initially, YOLO is used to detect faces in a video stream, where the developed model outputs bounding boxes around detected faces. Once a face is detected, the region of interest (ROI) is extracted and converted to grayscale for further analysis. Next, the Haar Cascade classifier is applied to detect eyes within the detected face bounding box. For each eye detected, the script applies thresholding to segment the pupil region, which is then further processed to identify the largest contour, assumed to be the pupil. Finally, the detected face and pupil are highlighted by drawing bounding boxes and circles, respectively, on the video feed, which is displayed in real-time. This approach leverages both deep learning (YOLO) and classical computer vision (Haar cascades and thresholding) to provide accurate face, eye, and pupil detection. For theft detection, a Roboflow model was trained to run alongside these scripts, triggering an alert upon detecting suspicious activity and notifying security via a theoretical dummy application.

*2) **Voice Command Recognition**:* The purpose of the speech recognition section was to convert verbal commands from the user into discrete actions. This process consisted of two main components: the speech-to-text algorithm and the matching algorithm. The Speech-to-text (STT) algorithm classifies audio signals from the user and appropriately categorises them under words and phrases from the vocabulary list. The matching algorithm takes as input the words and phrases from the STT algorithm and interprets them to determine the appropriate response. For the STT program to classify sound samples, a machine learning classifier AI model is required, along with a vocabulary list of all possible words recognised by MARTY. However, this comes at the cost of increased processing requirements, as well as additional storage for model parameters and the vocabulary. An alternative approach was to use a cloud-based service such as Google STT. However, this required an internet connection, which was incompatible with the network IP-based connection used to control the robot.

Ultimately, it was decided to use Naoqi's integrated speech-to-text algorithm due to its superior optimisation for the Pepper robot. This simplified the model and eliminated the need for intermediaries. Given MARTI's highly specialised function, a comprehensive vocabulary list was deemed unnecessary. The algorithm produced a string containing the identified word as its output.

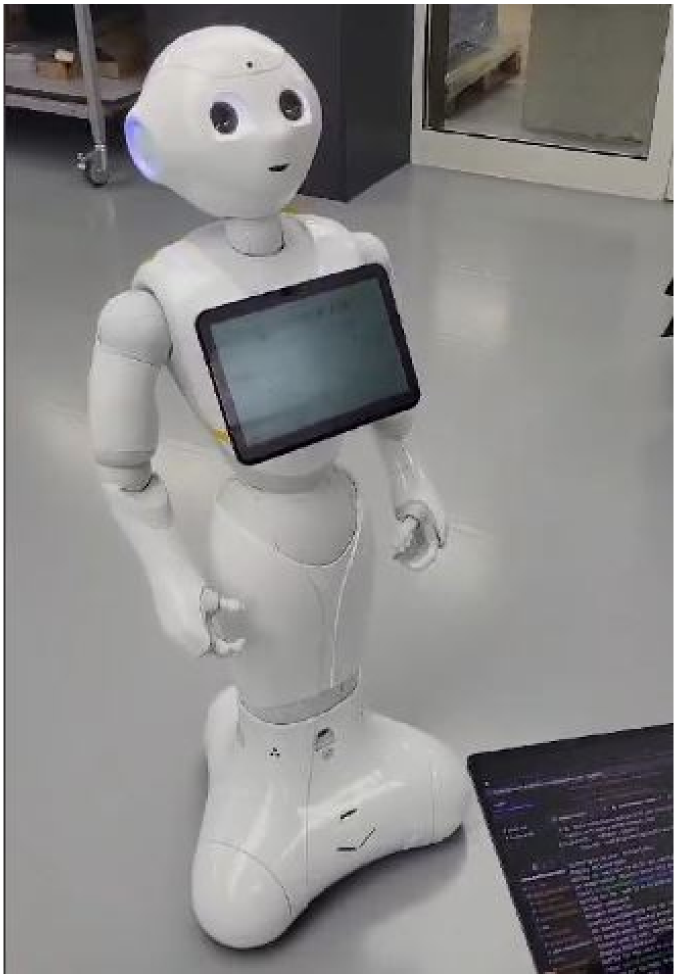


Fig. 1. Interfacing with Pepper in the lab

*3) **SLAM**:* Initially, the plan focused on implementing the ORB-SLAM2 or PySLAM for real-time localisation and mapping, which never materialised due to hardware constraints; Mainly a lack of a GPU to execute these approaches. To compensate, a logic-based motion planning algorithm was developed to simulate Pepper's movement within a virtual environment. This QiBullet framework was used to develop trajectory mapping and obstacle avoidance based on start and goal coordinates, mimicking locations of items and spaces within a retail space. This was built on geometric reasoning and conditional checks to navigate around virtual obstacles.

*4) **Object Recognition**:* The purpose of this section is to enable object recognition. The Pepper robot is designed to navigate through aisles and identify objects as they are encountered. The object recognition system consists of two main components: feature extraction and classification. The algorithm design was to have:

- **Input:** The model accepts an image as input.
- **Feature Extraction:** A sequential model extracts visual information from the image and encodes it into a compact, multi-dimensional feature vector.
- **Classification:** This feature vector is fed into a network (referred to as sequential-15), which classifies the image into one of seven grocery categories.
- **Output:** The final layer produces a probability distribution over the seven classes, indicating the likelihood that the input image belongs to each category.

The initial plan was to use Naoqi's ready-made functions to access the top camera, which would provide a live stream. Images were intended to be captured at 10 frames per second and processed by the object recognition algorithm. However, due to the Pepper robot's limited processing capabilities, this process could not be executed on the robot itself. Instead, the image processing was offloaded to an external laptop connected via Wi-Fi.

During testing, several issues arose, including compatibility problems. To overcome these challenges, the project shifted to a simulation environment using PyBullet. In this virtual setting, the Pepper robot was tasked with navigating an aisle and recognizing three different objects. A dedicated Python pipeline was developed to support this simulation.

*5) **Pose Landmarking**:* This area was addressed using a popular top-down landmarking approach [23] by bounding boxes around the human areas and placing estimated key points within these boxes. Knowing the locations of these key points on the frames of the image received the proposed method to detect waving and individual lying will be explained. For the waving gesture, the position of the wrist is compared to the corresponding shoulder, and when it is above the shoulder, this is recorded and changes in the x-coordinate frame are analysed looking at direction changes imitated a side to side motion seen in waving; If this is done for long enough (defined by a threshold) it is acknowledged by the system. For lying down, the angle between the mid-points of the torso and hip is calculated using the vector gathered from the difference in positions on the frame; If this is closer to being horizontal (defined by an angular threshold such as 60°) then this is identified as lying down that can be used as a trigger to call for emergency assistance as this could mimic someone unconscious or slipping in the store.

V. Results and Discussion

A. Human and Shoplifting Detection

Despite Pepper's limitations, a real-time tacking was achieved using deep learning models. These models were optimised with various thresholding and hyper-parameter scaling, leading to the accuracies seen in Table I.

Model	mAP	Precision	Recall
Human Detection	95.2%	95%	90.2%
Theft Detection	90%	87.9%	83.5%

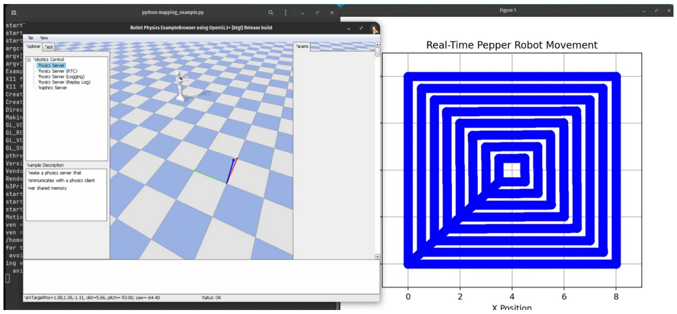
TABLE I
PERFORMANCE METRICS FOR HUMAN AND THEFT DETECTION MODELS

B. Voice Command Recognition

Ultimately, we decided to use Naoqi's integrated speech-to-text algorithm due to its superior optimisation for the Pepper robot. This simplified the model and eliminated the need for intermediaries. Given MARTI's highly specialised function, a comprehensive vocabulary list was deemed unnecessary. The algorithm produced a string containing the identified word as its output.

C. SLAM

In the virtual environment, real-time mapping of Pepper's movement was achieved and can be seen in Figure 2. This behaviour is extremely useful during motion execution to check whether the robot is still on track to its goals and if any deviations that are extremely likely in a dynamic environment have occurred.



Create a physics server that communicates with a physics client over shared memory

camTargetPos=1.08,1.28,-1.31, dist=5.66, pitch=-53.00, yaw=-64.40 Status: OK

Fig. 2. Real-time mapping of Pepper in the simulation environment on Matplotlib

For the detection and avoidance algorithm, a hard-coded approach was used wherein Pepper used its laser sensors to check for obstacles. When an obstacle is detected, pepper starts moving right while facing the obstacle. It keeps going till the front laser reading is greater than the threshold set. However, the laser has a range of 600 so when the outermost laser beam toward the right side is clear, the robot assumes that the obstacle has been avoided successfully and starts moving straight, colliding into the obstacle. Therefore, an extra strafe right command is given so it fully clears the obstacle and moves onwards. An example of reaching an obstacle can be seen in Figure 3.

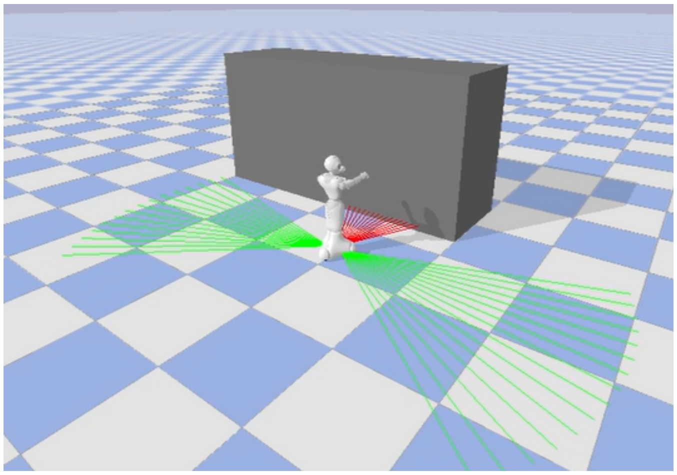


Fig. 3. Obstacle Detection using lasers on the simulation environment

D. Object Recognition

Testing on the physical Pepper robot revealed that obtaining a live camera stream was not feasible. The robot's GUI encountered issues and displayed various errors, and its operating system ran an outdated version of Python that is no longer supported. In contrast, the virtual environment did not experience these problems. However, a new challenge emerged: the simulation had to be optimized to run efficiently on a standard CPU, as the team did not have access to a GPU. The robot was able to move and recognize objects when these functions were executed separately in the virtual environment, but attempting to run all functions concurrently proved challenging. Therefore, optimizing and executing the code under these constraints was a significant challenge, and as a result, the simulation was only partially successful.

E. Pose Landmarking

The human landmarking has been finalized to categorise seen customers as either standing, waving or lying on the floor seen in Figure 4. This approach was lightweight enough to operate without introducing delay and had great accuracy within a wide range of angles. The most problematic behaviour to pick up is lying on the floor directly and facing the front of the robot, where the angle difference can be difficult to pick up; This problem can be more reliably fixed by incorporating depth cameras to detect this case more efficiently.

VI. Conclusion and Future Work

Overall, MARTI shows the importance of accurate and responsive perception systems to create a reliable shopping assistant that is built up through many different blocks seen throughout this paper. Most methods resulted in fairly accurate and responsive systems with the bigesest let down being not enough time left to deploy all of these systems in person on Pepper.

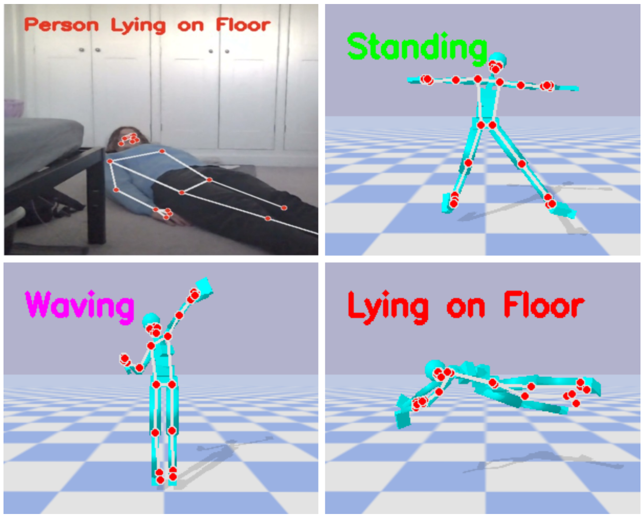


Fig. 4. Examples of results given from human landmarking on both real-life subjects and in the simulation environment

This project also highlighted the challenges of working with legacy robotic systems like Pepper, particularly its reliance on Python 2.7, leading to significant compatibility issues with modern libraries like TensorFlow. Persistent connectivity failures prevented script execution for nearly seven weeks, severely limiting progress. Alternative approaches such as simulation should have been explored earlier. This experience underscored the importance of adaptability, proactive problem-solving, and contingency planning in robotics projects.In the future, most technological constrictions discussed previously could be addressed with hardware, software and perception enhancements similar to work seen in a similar project [24].

References

[1] Y. Tong, H. Liu, and Z. Zhang, "Advancements in humanoid robots: A comprehensive review and future prospects," *IEEE/CAA Journal of Automatica Sinica*, vol. 11, no. 2, pp. 301–328, 2024.

[2] P. Samarathunga, M. Valori, G. Legnani, and I. Fassi, "Assessing safety in physical human–robot interaction in industrial settings: A systematic review of contact modelling and impact measuring methods," *Robotics*, vol. 14, 02 2025.

[3] C. S. Song and Y.-K. Kim, "The role of the human-robot interaction in consumers' acceptance of humanoid retail service robots," *Journal of Business Research*, vol. 146, pp. 489–503, 07 2022.

[4] Y. Okafuji, S. Song, J. Baba, Y. Yoshikawa, and H. Ishiguro, "Influence of collaborative customer service by service robots and clerks in bakery stores," 2022. [Online]. Available: https://arxiv.org/abs/2212.10687

[5] B. Jo and C. S. Song, "Applying human-robot interaction technology in retail industries," p. 839, 11 2019.

[6] O. Bendel and L. M. D. S. Alves, "Should social robots in retail manipulate customers?" 2022. [Online]. Available: https://arxiv.org/abs/2206.14571

[7] A. De Santis, B. Siciliano, A. Luca, and A. Bicchi, "An atlas of physical human-robot interaction," *Mechanism and Machine Theory*, vol. 43, pp. 253–270, 03 2008.

[8] Y. Chen, Y. Luo, C. Yang, M. O. Yerebakan, S. Hao, N. Grimaldi, S. Li, R. Hayes, and B. Hu, "Human mobile robot interaction in the retail environment," *Scientific Data*, vol. 9, no. 1, p. 673, Nov. 2022.

[9] M. F. Ahmed, K. Masood, V. Fremont, and I. Fantoni, "Active SLAM: A review on last decade," *Sensors (Basel)*, vol. 23, no. 19, Sep. 2023.

[10] J. Qiao, J. Guo, and Y. Li, “Simultaneous localization and mapping (SLAM)-based robot localization and navigation algorithm,” *Applied Water Science*, vol. 14, no. 7, p. 151, Jun. 2024.

[11] C. Gómez, M. Mattamala, T. Resink, and J. Ruiz-del Solar, *Visual SLAM-Based Localization and Navigation for Service Robots: The Pepper Case*. Springer International Publishing, 2019, p. 32–44. [Online]. Available: http://dx.doi.org/10.1007/978-3-030-27544-0$_3$

[12] Y. Yamato, Y. Fukumoto, and H. Kumazaki, “Security camera movie and erp data matching system to prevent theft,” in *2017 14th IEEE Annual Consumer Communications amp; Networking Conference (CCNC)*. IEEE, Jan. 2017, p. 1014–1015. [Online]. Available: http://dx.doi.org/10.1109/CCNC.2017.7983275

[13] G. A. Martínez-Mascorro, J. R. Abreu-Pederzini, J. C. Ortiz-Bayliss, A. Garcia-Collantes, and H. Terashima-Marín, “Criminal intention detection at early stages of shoplifting cases by using 3d convolutional neural networks,” *Computation*, vol. 9, no. 2, p. 24, Feb. 2021. [Online]. Available: http://dx.doi.org/10.3390/computation9020024

[14] P. Schneider, R. Memmesheimer, I. Kramer, and D. Paulus, “Gesture recognition in rgb videos usinghuman body keypoints and dynamic time warping,” 2019. [Online]. Available: https://arxiv.org/abs/1906.12171

[15] N. Rashvand, G. A. Noghre, A. D. Pazho, S. Yao, and H. Tabkhi, “Exploring pose-based anomaly detection for retail security: A real-world shoplifting dataset and benchmark,” 2025. [Online]. Available: https://arxiv.org/abs/2501.06591

[16] M. G. S. Murshed, E. Verenich, J. J. Carroll, N. Khan, and F. Hussain, “Hazard detection in supermarkets using deep learning on the edge,” 2020. [Online]. Available: https://arxiv.org/abs/2003.04116

[17] P. Follmann, B. Drost, and T. Böttger, “Acquire, augment, segment enjoy: Weakly supervised instance segmentation of supermarket products,” 2018. [Online]. Available: https://arxiv.org/abs/1807.02001

[18] S. Qiao, W. Shen, W. Qiu, C. Liu, and A. Yuille, “Scalenet: Guiding object proposal generation in supermarkets and beyond,” 2017. [Online]. Available: https://arxiv.org/abs/1704.06752

[19] Y. Tang, J. Pino, C. Wang, X. Ma, and D. Genzel, “A general multi-task learning framework to leverage text data for speech to text tasks,” 2021. [Online]. Available: https://arxiv.org/abs/2010.11338

[20] K. Lakhotia, E. Kharitonov, W.-N. Hsu, Y. Adi, A. Polyak, B. Bolte, T.-A. Nguyen, J. Copet, A. Baevski, A. Mohamed, and E. Dupoux, “Generative spoken language modeling from raw audio,” 2021. [Online]. Available: https://arxiv.org/abs/2102.01192

[21] E. Ackerman, “Video friday: Ameca humanoid,” *IEEE Spectrum*, 2021. [Online]. Available: https://spectrum.ieee.org/video-friday-ameca-humanoid

[22] K. Berns and A. Ashok, ““you scare me”: The effects of humanoid robot appearance, emotion, and interaction skills on uncanny valley phenomenon,” *Actuators*, vol. 13, no. 10, p. 419, 2024. [Online]. Available: https://www.mdpi.com/2076-0825/13/10/419

[23] T. D. Nguyen and M. Kresovic, “A survey of top-down approaches for human pose estimation,” 2022. [Online]. Available: https://arxiv.org/abs/2202.02656

[24] P. Magri, J. Amirian, and M. Chetouani, “Upgrading pepper robot s social interaction with advanced hardware and perception enhancements,” 2024. [Online]. Available: https://arxiv.org/abs/2409.01036