Group Technical Report

Team Homo-SAP-iens
Nicholas Keatley, Doran Moison, Hussein Naddaf,
Dhruv Pandit, Raj Tagore, Vaibhav Tewani

***Abstract*—With museum attendance on the decline and museums are struggling to stay afloat with a lack of funding [1]. This is a downward spiral with a lack of funding leading to a lack of tour guides who can show people around and answer questions which leads to a decline in interest and attendance [2]. To solve these issues the use of robot guides is proposed. The novelty of a robot guide would entice visitors while simultaneously allowing them to obtain information about the artwork and exhibits.**

***Index Terms*—robotics, LLM, object detection, TTS**

I. Introduction

The aim of our project is to use an existing robot as a museum curator and guide. The robot uses object recognition to detect paintings and then uses an LLM to provide further information. It can also recognize if a person is trying to ask it questions and then answer said question. The robot used in the project is the TIAGo robot. It was used due to its availability and height that allows it to easily perceive artwork and museum visitors.

The report is consists of section II which is a summary of the reviewed literature, III, which describes the methods used in the implementation of the project, IV which gives a summary of the results obtained, and V which provides a conclusion and describes any potential future work.

II. Literature Review

A. Robots in Galleries

Stefano Rosa (2024) discusses experimental trials and lessons learned from employing a tour guide robot in which she outlined the research history of robotics in galleries. For this application both of mobile base and humanoid robots were used such as Pepper, Rhino Burgard, Asimo, and Robovie. However, none of these papers utilized TIAGo for this task [3].

The basic requirements to have a robotics tour guidance are building navigation system, object recognition model, and linguistic communication system. However, in some applications such as [4], they add more subsystems to make their robot - which was Robovie-R Ver.2 - more socially accessible. In this research the robot read the humans expression to detect how the person is willing to interact with the robot and ask them a question based on that.

B. Object Detection

Object detection is pivotal for robots to interact effectively with their surroundings, especially in scenarios like guiding visitors through galleries. Such environments pose challenges such as varying lighting and diverse objects. YOLO V8, a cutting-edge algorithm, outshines others like RCNN, SPPNet, and Mask R-CNN for this task due to its real-time performance, high accuracy, and ease of training with custom databases [5]. YOLO V8's single neural network predicts object positions and classes swiftly, making it ideal for real-time applications. Its balance of speed and accuracy suits humanoid robots in dynamic gallery settings, ensuring accurate object recognition and seamless navigation. Leveraging YOLO V8, robots enhance visitor experiences by efficiently navigating through diverse gallery spaces.

III. Methodology

A. Robot Selection and ROS Overview

Of the various robots available to us (NAO, Miro, Pepper and TIAGo) only Pepper and TIAGo were observed to be tall enough to view the museum exhibits and interact with museum visitors. The TIAGo Robot was finally selected due to the presence of an RGB-D camera that could be used for object detection and recognition. It also has a laser that could be used for mapping and obstacle avoidance. Once the robot was selected, various tasks were split into ROS nodes that could be integrated for the final implementation.

Mapping out the system as shown in Fig.1 was feasible due to the architecture of ROS, which allows us to communicate between different Nodes using Topics efficiently in real-time. We also ensured we were using the appropriate ROS packages and message formats so that we would not have to rebuild any wheels, we could leverage the vast database of pre-written ROS code, and our system would integrate the with real TIAGo Robot well.

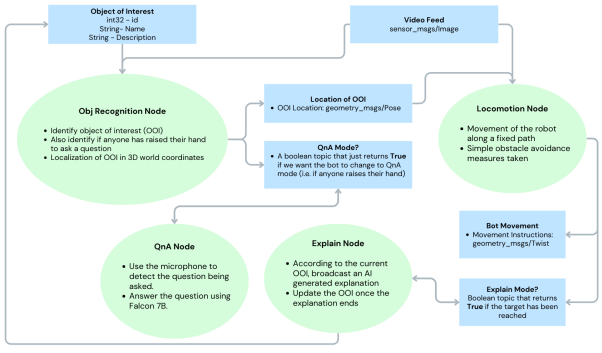

Fig. 1. ROS Nodes

B. Object Detection and Localization

YOLO V8 algorithm was used for object detection. It takes image frames as input and outputs boundary boxes around detected objects, labelling them accordingly . The model was specifically trained to identify three paintings: "The Mona Lisa" "The Starry Night" and "The Last Supper", a bag representing a masterpiece in the gallery, and a human hand in a raised position. We have collected a dataset of 500 images. These images were captured under different lighting conditions and The orientations to challenge the model's ability to recognize the objects in various scenarios. The dataset was split into a training set of 400 images and a validation set of 100 images. TIAGo's RGB camera was calibrated using OpenCV chessboard camera calibration functions to achieve object localization. This process allowed us to determine the focal lengths (f_x, f_y) and the centre of the image (C_x, C_y). By making use of the intrinsic matrix of the RGB camera and the depth information (z) from the integrated depth camera, we were able to calculate the world coordinates of the centres of the objects relative to TIAGo's RGB camera coordinate system (X, Y, Z) using equations 1 to 3 [6].

$$X = \frac{(x - C_x) \times z}{f_x} \tag{1}$$

$$Y = \frac{(y - C_y) \times z}{f_y} \tag{2}$$

$$Z = z \tag{3}$$

Where x,y are the coordinates of the center of the object in the pixel's coordinates.

C. Locomotion and Obstacle Avoidance

The locomotion node takes an input from the location node i.e the location of it's next objective (museum exhibit/visitor asking question) using geometry_msgs and mobile_base_controller that provide linear and angular vectors for locomotion.

Navigating in a museum can be a challenging task for a robot because of the large crowds, especially if these gather around the robot guide during explanations. In order to safely move from one artwork to another, the robot must make use of its multiple sensors to detect any obstacles that could be in its way. Therefore, the obstacle detection node subscribes to the laser and depth camera topics to receive data from both sensors. Using the information provided by the sensors, the robot can know if there is an obstacle in front of it, in which case it indicates it must stop. At this point, the robot may be in one of two cases, each requiring a different action from the guide. Either the obstacle is a person, and the robot waits for them to move, or the obstacle is not and TIAGo must find a new path. This decision was made because as visitors gather around the guide during the tour, finding a path through the crowd would be much more complicated and time-consuming than simply asking people to make way rather than risking the freezing robot problem [7]

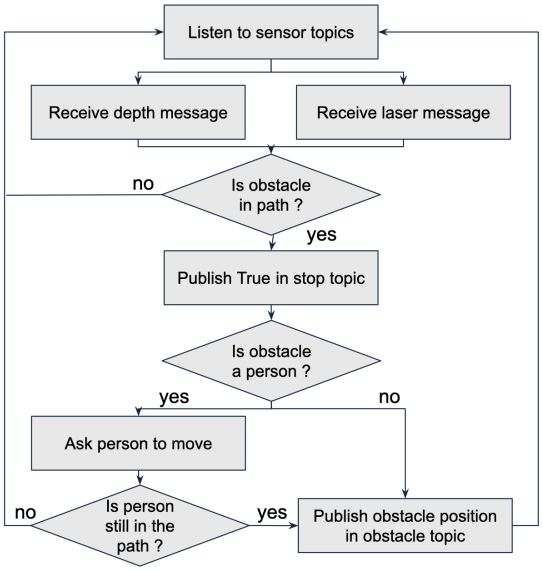

Fig. 2. **Obstacle Avoidance Flowchart**

D. LLM integration

1) Explanation of Museum Pieces: The ChatGPT API was integrated with the TIAGo robot to provide contextual information about the art pieces encountered during the tour. This integration allowed the robot to dynamically generate descriptions and answer questions related to the artworks.

The integration process involved two key components: the YOLO model for artwork detection and the ChatGPT API for natural language processing. When the TIAGo robot encounters an artwork, it utilizes the YOLO model to identify the specific art piece within its field of view. The YOLO model outputs the relevant information about the artwork, including its unique identifier.

Subsequently, the TIAGo robot leverages the ChatGPT API to generate contextual descriptions for the identified artwork. By providing the unique identifier as input to the ChatGPT API, the robot retrieves relevant information from its knowledge base and formulates a comprehensive description of the artwork. This description includes details such as the artist, historical background, artistic style, and any notable features or symbolism.

Furthermore, the integration extends to the question-answering capability of the TIAGo robot. When a visitor poses a question without explicitly specifying a particular painting, the TIAGo robot utilizes the YOLO model's output to infer that the question pertains to the currently observed artwork. By incorporating the YOLO model's output as additional input to the ChatGPT API, the robot generates a response that is contextually relevant to the artwork in question.

2) Answering questions from visitors: For answering questions from audiences, a number of machine learning models were investigated, all involving transformers with self-attention mechanisms. While this task could be accomplished with a single end-to-end model [8], it was decided that

breaking this into separate modules would allow greater ease of component testing and debugging. These sub-components are:

- Audio transcription - all audio surrounding the robot is transcribed as human speech into text using Automatic Speech Recognition (ASR) for more effective processing.
- Question Filtering - the text is classified as either a question or non-question.
- Answer Generation - for each question, an answer is generated as text.
- Voice Synthesis - the answer is rendered as synthetic human speech.

OpenAI developed transformers for multilingual audio transcription, Whisper , which analyses overlapping image patches of audio spectrograms as input to the transformer. While the ASR of Whisper's training data includes 96 languages, approximately two-thirds of this is in English and transfer learning has poor results across non-Indo-European languages [8].

Multiple versions are available, with Whisper Tiny (37 million parameters) being very effective at transcribing audio to text. Other versions such as Base to LargeV3 are capable of audio-to-audio question-answering, however this end-to-end approach was rejected for aforementioned reasons and so Whisper Tiny was found to be sufficient for our purposes While inputs have a maximum limit of the input vector size, continuous audio can be 'chunked' into 30-second segments for processing.

Question Filtering was tested with both BERT Large [9] and RoBERTa [10] (Robustly Optimized BERT Approach). While both are based on BERT (Bidirectional Encoder Representations from Transformers), BERT Large uses a much larger dataset and more parameters (340M). RoBERTa (10M parameters) uses optimised pre-training for improved performance across multiple tasks including text-classification, and uses a larger and more diverse text corpus for training. In the SQuAD benchmark (Stanford Question-Answering Dataset), RoBERTa outperforms BERT at 94.6 to 90.9 in F1-score and so was chosen for its improved performance and more compact size (10M versus 340M parameters).

For generating answers in text, a number of models were investigated. MambaGPT-3B [11] is based on a selective state-space model which approximates the method of transformers over a smaller range for more efficient computation and smaller model size. However the model is not finetuned using RLHF (Reinforcement Learning with Human Feedback) for dialogue-based interaction, and so it made factual errors for simple questions and give repeating statements. Hermes Mixtral 8x7B [12] uses a mixture-of-experts architecture (combining muiltiple transformers for inferece) for greater versatility. It is fine-tuned with RLHF and performs well on benchmarks such as GPT4All and BigBench, however it requires substantial resources for text-generation that is unsuitable for this project. requires large dependencies for sub-component Falcon7B-Instruct [13] the model was found to give accurate, unbiased and relevant answers and so was chosen for this component.

For generating answers, it was found that sampled questions often depend on contextual information from the environment without explicitly mentioning them, especially questions that reference nearby objects such as "Can you tell me about this painting?". In order for the robot to generate answers including this context, the last recognised museum object is prepended to the prompt, e.g. "You are standing next to the painting 'Starry Night'. The robot's social role is also inserted into the prompt, describing that the transformer should generate the output of a museum tour guide.

Voice Synthesis, converting the generated answer to speech, was accomplished using a text-to-speech model. This was carried out with fine-tuned with VITS (Variational Inference with adversarial learning for end-to-end Text-to-Speech) [14], based on Coqui TTS. This allows specific individual voices to be generated, by taking existing audio of the individual that is paired with text, generated by transcribing the audio with OpenAI Whisper. For demonstration purposes, we chose the nature documentary narrator Sir David Attenborough, however future work should use voices of consenting individuals or new voices that do not correspond to existing individuals.

IV. Results

The detection model achieved outstanding results, with a zero-validation loss in the labelling accuracy and a validation loss of 0.5 in the boundary boxing accuracy after 400 iterations of training. These results indicate the model's high reliability in both recognizing and accurately determining the positions of objects within an image frame. It also provides a high response rate making it perfect for real-time applications such as ours.

For localization part, our model was able to combine the information from the TIAGo's depth and RGB cameras. It first takes frame from RGB camera as shown in Fig.3 and then label the Object Of Interest (OOI) within the frame and add the boundary box, then it finds the centre of the object in pixels and find its depth using the depth image shown in Fig.4. And finally, it calculates the 3D world coordinate of the centre of OOI and add them to the RGB frame as shown in Fig.3.

The robot successfully detects questions, generates sufficient responses that are relevant and accurate, and converts these responses into synthetic speech that is clear and comprehensible.

V. Conclusion and Future Work

For the question and answering node, future work can involve audio pre-processing to remove noise (e.g. amplifying higher frequencies), and biased content can be avoided using NeMo guardrails [15]. The dialogue can also be improved with contextual memory using vector databases [16] - which save embeddings of all previous statements in a conversation, which can be searched by the transformer using nearest-neighbour to refer to previous discussions. User experience can also be improved by making the robot proactive i.e. asking museum

Fig. 3. Processed TIAGo's RGB camera frame

Fig. 4. TIAGo's Depth Image

visitors for questions rather than wait for the visitor to ask a question.

In terms of ethics, many museums lack critical funding. Technological investments can lead to direct competition for resources between labour costs, collection pieces, and essential maintenance [2]. Especially when UK grant funding has fallen by 20% over the last decade [1].

Museums have had to make short-term repairs or delay essential maintenance work and working conditions impact on staff morale - as an example, at The Wallace Collection, London, a section of masonry fell from the portico in 2018 due to deterioration in the supporting beams [2].

Technology investment can lead to direct competition in funding for labour, collection pieces, and essential maintenance Therefore our work aims to supplement staff and resources for cultural institutions, rather than replace them.

REFERENCES

[1] E. Mills, "Museum collections at risk through lack of maintenance funding," 2020.

[2] S. M. Sleeter, "Technology and its impact on the future of museums." https://stmupublichistory.org/publiclyhistorians/technology-and-its-impact-on-the-future-of-museums/, 2018. Accessed: 2024-04-03.

[3] S. Rosa, M. Randazzo, E. Landini, S. Bernagozzi, G. Sacco, M. Piccinino, and L. Natale, "Tour guide robot: a 5g-enabled robot museum guide," *Frontiers in Robotics and AI*, vol. 10, p. 1323675, 2024.

[4] Y. Kobayashi, T. Shibata, Y. Hoshi, Y. Kuno, M. Okada, and K. Yamazaki, ""i will ask you" choosing answerers by observing gaze responses using integrated sensors for museum guide robots," in *19th International Symposium in Robot and Human Interactive Communication*, pp. 652–657, 2010.

[5] G. Boesch, "Object Detection in 2024: The Definitive Guide - viso.ai — viso.ai." https://viso.ai/deep-learning/object-detection/. [Accessed 02-04-2024].

[6] Y. Mori, N. Fukushima, T. Yendo, T. Fujii, and M. Tanimoto, "View generation with 3d warping using depth information for ftv," *Sig. Proc.: Image Comm.*, vol. 24, pp. 65–72, 01 2009.

[7] P. Trautman and A. Krause, "Unfreezing the robot: Navigation in dense, interacting crowds," in *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 797–803, 2010.

[8] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International Conference on Machine Learning*, pp. 28492–28518, PMLR, 2023.

[9] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[10] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[11] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," *arXiv preprint arXiv:2312.00752*, 2023.

[12] A. Q. Jiang, A. Sablayrolles, A. Roux, A. Mensch, B. Savary, C. Bamford, D. S. Chaplot, D. d. l. Casas, E. B. Hanna, F. Bressand, *et al.*, "Mixtral of experts," *arXiv preprint arXiv:2401.04088*, 2024.

[13] E. Almazrouei, H. Alobeidli, A. Alshamsi, A. Cappelli, R. Cojocaru, M. Debbah, É. Goffinet, D. Hesslow, J. Launay, Q. Malartic, *et al.*, "The falcon series of open language models," *arXiv preprint arXiv:2311.16867*, 2023.

[14] E. Casanova, J. Weber, C. D. Shulby, A. C. Junior, E. Gölge, and M. A. Ponti, "Yourtts: Towards zero-shot multi-speaker tts and zero-shot voice conversion for everyone," in *International Conference on Machine Learning*, pp. 2709–2720, PMLR, 2022.

[15] T. Rebedea, R. Dinu, M. Sreedhar, C. Parisien, and J. Cohen, "Nemo guardrails: A toolkit for controllable and safe llm applications with programmable rails," *arXiv preprint arXiv:2310.10501*, 2023.

[16] X. Xie, H. Liu, W. Hou, and H. Huang, "A brief survey of vector databases," in *2023 9th International Conference on Big Data and Information Analytics (BigDIA)*, pp. 364–371, IEEE, 2023.