A Visually Context-Aware Robotic System for Object Localization Assistance

Davide Cavicchini
Department of Information Science and Engineering
University of Trento, Italy
davide.cavicchini@studenti.unitn.it

Mario Barbato
Department of Electrical Engineering and Information Technology
University of Naples "Federico II", Italy
Logogramma s.r.l., Italy
mario.barbato@unina.it

***Abstract*—This project presents a modular, visually context-aware robotic system designed to assist users, particularly those with cognitive or physical impairments, in object localization tasks within everyday environments. Leveraging the Retico framework for incremental dialogue processing, the system integrates a vision network that constructs scene graphs from multiple camera feeds and a dialogue network that processes user speech to retrieve relevant information. The vision pipeline uses a transformer-based model (RelTR) for efficient scene graph generation, while the dialogue pipeline employs Whisper for speech recognition and smolAgents for natural language interaction. A pilot evaluation involving interactions with a Misty II robot demonstrated the system's potential, with participants successfully querying object locations across different settings. Results highlight promising accuracy and usability, suggesting avenues for further development in scene representation and socially adaptive interaction.**

***Index Terms*—visual question-answering, scene graphs, social robotics, context-awareness**

I. Introduction

The development of context-aware social robots represents an expanding frontier in contemporary scientific research [1], [2], [3]. A key objective in advancing these systems is to assist individuals, particularly those with cognitive or physical impairments, in performing daily tasks more effectively. Central to this goal is maintaining a dynamic, grounded representation of the surrounding environment, enabling the robot to reason about and communicate pertinent information to the user. A common use case could consist of a user searching for their medication in their home. A robot equipped with a structured, continuously updated understanding of the environment could promptly inform the user of the medication's location, thereby enhancing user independence and reducing cognitive burden.

In this work, our aim is to address one crucial aspect of this challenge by developing a Retico-based modular network capable of interpreting contextual visual data. Specifically, the proposed system processes multiple video feeds from known contexts, such as kitchens, bathrooms, or living rooms, and extracts structured, actionable knowledge. The primary focus is on enabling spatial and semantic reasoning, with the capability to answer queries about the positions and properties of entities within these scenes, enhancing users' awareness on the interested environment and potentially helping them in fulfilling their daily tasks.

II. Background Work

A. Retico Framework

The proposed system is built on top of Retico [4], an open-source framework for building interaction pipelines based on incremental dialogue processing [5]. In this paradigm, each interaction module begins processing incoming data, represented as atomic units called Incremental Units (IUs), before the data is fully received, passing the partially processed information to subsequent modules. This approach contrasts with traditional dialogue processing systems, where data are typically processed only after they are completely available. To keep track of the full stream of information, available only when the input transmission ends, this paradigm introduces different transactional actions. Processed IUs are added to the pipeline to enable real-time processing, but they are committed only when the module is confident in its hypothesis about the data. If the hypothesis changes, the unit is revoked, informing subsequent modules in the pipeline of the correction.

B. Scene Graphs

The visual context-awareness aimed in this project is based on the Scene Graph technology. They are structured representations of visual scenes where objects are depicted as nodes, and their relationships are encoded as directed edges, following a subject-predicate-object structure in order to enable richer semantic understanding compared to flat object detection. Scene graphs provide a formalized way to capture not only the presence of entities but also their interactions and spatial arrangements, facilitating tasks such as image retrieval, question answering, and robotics perception. Recent methods, such as RelTR (Relation Transformer) [6], have advanced scene graph generation by leveraging transformer-based architectures to directly predict relational triplets from images. Unlike traditional two-stage approaches, RelTR eliminates redundant pairwise processing, using parallel attention mechanisms to model object relations more efficiently and accurately in a fully end-to-end manner.

III. System Architecture

We build upon these frameworks and concepts to develop our system for Visual Question Answering. We'll begin by briefly showcasing how our system is structured and the design choices that resulted in the final network.

A. Logical Architecture

To fulfill the set goals, our system needs to be able to process in real time both the visual input from multiple cameras and the speech signal from and to the robot.

Our solution thus involved having two networks:

- **Vision Network**: Receives multiple camera streams from different sources and processes each of them to first produce a scene graph of the view, and then save and index intelligently into memory.
- **Dialogue Network**: Receives microphone input and processes user requests to query the scene graph and retrieve relevant information to assist the user.

The two networks then communicate through tools that will be called by the dialogue manager. A visualization of the proposed network can be seen in Figure 1. Following, the two subnetworks' requirements to achieve their own goals:

Vision Network

V1) *Input Stream Attribution*: Know which video feed comes from which camera, allowing in the next steps the proper storage of the scene data.
V2) *Scene Graph Generation*: Include a component responsible for generating scene graphs from raw image data.
V3) *Scene Processing Time*: Process all incoming input streams with low latency, ensuring a close to real-time information flow.
V4) *Scene Graph Memory*: Independently store and process the produced scene graphs to be later queried.
V5) *Scene Graph Indexing*: Allow easier retrieval of the store data through the use of natural language.

Dialogue Network

D1) *Speech Recognition*: Process incoming data from the microphone to reconstruct the user requests.
D2) *Dialogue Manager*: Understand the user intent and accordingly use the tools at its disposal to gather meaningful information.
D3) *Natural Language Generation*: Generate a relevant and accurate response based on both the user input and retrieved information.
D4) *Speech Generation*: Readily give the information to the user through recognizable and understandable speech.

B. Physical Architecture

To implement the described modular system, we integrated a variety of tools and models across our network. The physical setup combines vision, audio, language models, and a robot interface, all orchestrated using Retico's subscription system. The components were deployed on standard laptop computing hardware and interfaced with the Misty II robot[1] for interaction.

For the scene input streams attribution (req. V1), we take advantage of Retico's *meta_data* field, which can be populated

[1]https://www.mistyrobotics.com/misty-ii

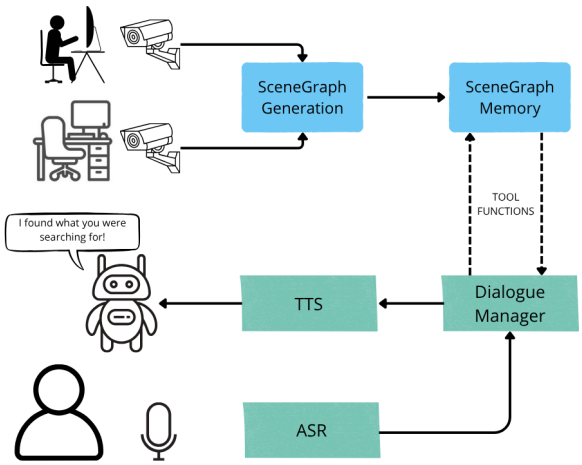

Fig. 1. The system architecture, implemented as a Retico network

when a module is instantiated. Since Retico keeps track of who creates the Incremental Units and the previous processed input IUs that generated them, it's then possible to uncover the origin of the scene graphs and store them appropriately (V4). Concerning the scene graph generation, we adopt the RelTR transformer-based model for scene graph generation. We use this model as it allows for direct triplet extraction from the image data, as already mentioned in II-B, which tackles both requirements V2 and V3. However, due to the way the model outputs its predictions, an entity matching step is required. Fortunately, the model also provides bounding boxes along with the triplets. This allows us to use the simple Intersection over Union (IoU) metric with a threshold to decide when to match two entities.

Finally, we use *Qwen3-Embedding-0.6B* model to store vector representations of the final scene graph triplets. These can later be retrieved using the cosine similarity of the representation computed between them and the vector representation of the query obtained from the same model. By providing tools to define the query at a later stage, we can effectively fulfill the requirement V5.

Regarding the Dialogue Network, we utilize the Whisper model for ASR (D1), and unify the NLU, dialogue management, and NLG (D2-3) using the smolAgents module, which leverages the homonymous HuggingFace library. It employs an LLM to answer queries autonomously while having access to code execution and tool calling. The tools developed are the following:

- *get_camera_names()*: returns all the cameras available by their name (matching the one provided in the *meta_data*)
- *get_scene_graph(camera_name)*: returns the scene graph for a given camera name.
- *query_camera(camera_name, query)*: queries the memory for a given camera name and returns the scene subgraph that matches the query and its neighbors.
- *query_memory(query)*: queries the memory for a given query string and returns for each camera the scene subgraph that matches the query and its neighbors.

Finally, the generated response is given to the Misty II robot, which generates the speech for the user (D4).

IV. Pilot Evaluation

To evaluate the implemented system, a pilot study was conducted. Participants were asked to interact with the Misty II robot, which was placed on a desk, and were instructed to ask questions about three different environments (the experiment setting, an office, and a kitchen), specifically regarding the location of objects and people. As quantitative measures of the system's performance, we computed two metrics: (i) accuracy, defined as the correctness of object identification and spatial localization, and (ii) response time, measured as the number of seconds between the end of the participant's request and the start of Misty's response. Qualitative evaluation involved collecting participants' feedback on system helpfulness and usability, rated on a scale from 1 to 10. Final results are shown in the Table I.

TABLE I
Results from the pilot evaluation process

Participant ID	Accuracy	Time Response (s)	Helpfulness	Usability
1	60%	16.3	7	7
2	40%	14.0	6	7
3	85%	16.7	9	7

V. Discussion

In this study, we proposed a Visual Question Answering system deployed on the Misty II social robot, enabling context-awareness to assist users in everyday tasks, such as locating personal items. Built on the Retico framework, the system architecture relies on Scene Graph generation, which is queried when the user asks the robot, for details about the surrounding environment. A preliminary evaluation was conducted to assess the system's accuracy, response time, usability, and perceived helpfulness, yielding promising results. Future work could focus on integrating more powerful models for scene graph embeddings to improve similarity-based retrieval, and on adding an emotion recognition module to make the interaction socially adaptive. Moreover, a more robust and extensive evaluation is required to scientifically validate the proposed system.

References

[1] C. Recchiuto and A. Sgorbissa, "Diversity-aware social robots meet people: beyond context-aware embodied ai," *arXiv preprint arXiv:2207.05372*, 2022.

[2] X.-T. Truong and T. D. Ngo, "Toward socially aware robot navigation in dynamic and crowded environments: A proactive social motion model," *IEEE Transactions on Automation Science and Engineering*, vol. 14, no. 4, pp. 1743–1760, 2017.

[3] X. Ye, Z. Lin, H. Li, S. Zheng, and Y. Yang, "Active object perceiver: Recognition-guided policy learning for object searching on mobile robots," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 6857–6863, IEEE, 2018.

[4] T. Michael, "Retico: An incremental framework for spoken dialogue systems," in *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pp. 49–52, 2020.

[5] D. Schlangen and G. Skantze, "A general, abstract model of incremental dialogue processing," *Dialogue & Discourse*, vol. 2, no. 1, pp. 83–111, 2011.

[6] Y. Cong, M. Y. Yang, and B. Rosenhahn, "Reltr: Relation transformer for scene graph generation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 9, pp. 11169–11183, 2023.