Cross-Modality Localization for Robot Navigation

Overview

Navigation is pivotal in the development of autonomous systems, including robots and self-driving cars. These systems rely heavily on accurate and robust localization capabilities to operate safely and efficiently in diverse environments. Traditional navigation systems, dependent on Global Navigation Satellite System (GNSS) data, often fail in urban canyons and indoor areas. Mimicking human navigation, which integrates various sensory inputs, autonomous systems can benefit from leveraging multiple sensing modalities. It is therefore essential to develop the capability to perform localization via multiple-modality data from different sensors.

Visual localization using 2D images is a well-established research area in robotics [3, 4]. However, the performance of image-based methods often degrades when facing drastic variations in illumination and appearance caused by weather and seasonal changes. Consequently, researchers have increasingly turned to 3D localization methods [6, 5] that utilize LiDAR point clouds to overcome these limitations. Further advancing this field, recent studies [7, 1, 2] have explored cross-modal approaches that integrate 3D vision with language processing, aiming to enhance localization accuracy and robustness.

In this project, we will design an innovative localization method that integrates multiple sensory modalities (image, text, and 3D point clouds) to function efficiently across diverse and dynamically changing environments. Importantly, we will insert the proposed localization method into a designed navigation system to guide the robot to the target position.

Goals

This project aims to design a localization method capable of fusing visual, text, and 3D point clouds as prompts. In the evaluation, we will build a navigation framework consisting of three modules: mapping, localization, and navigation. We will explore the proposed localization method in the navigation system to guide the robot to a target position.

Tasks

1 Getting familiar with the related literature (e.g., RNR-Map [2], VLMaps [1]) (~ 50h);

2 Setting up dataset and baselines (~ 100h);

3 Setting up the proposed localization pipeline(~ 160h);

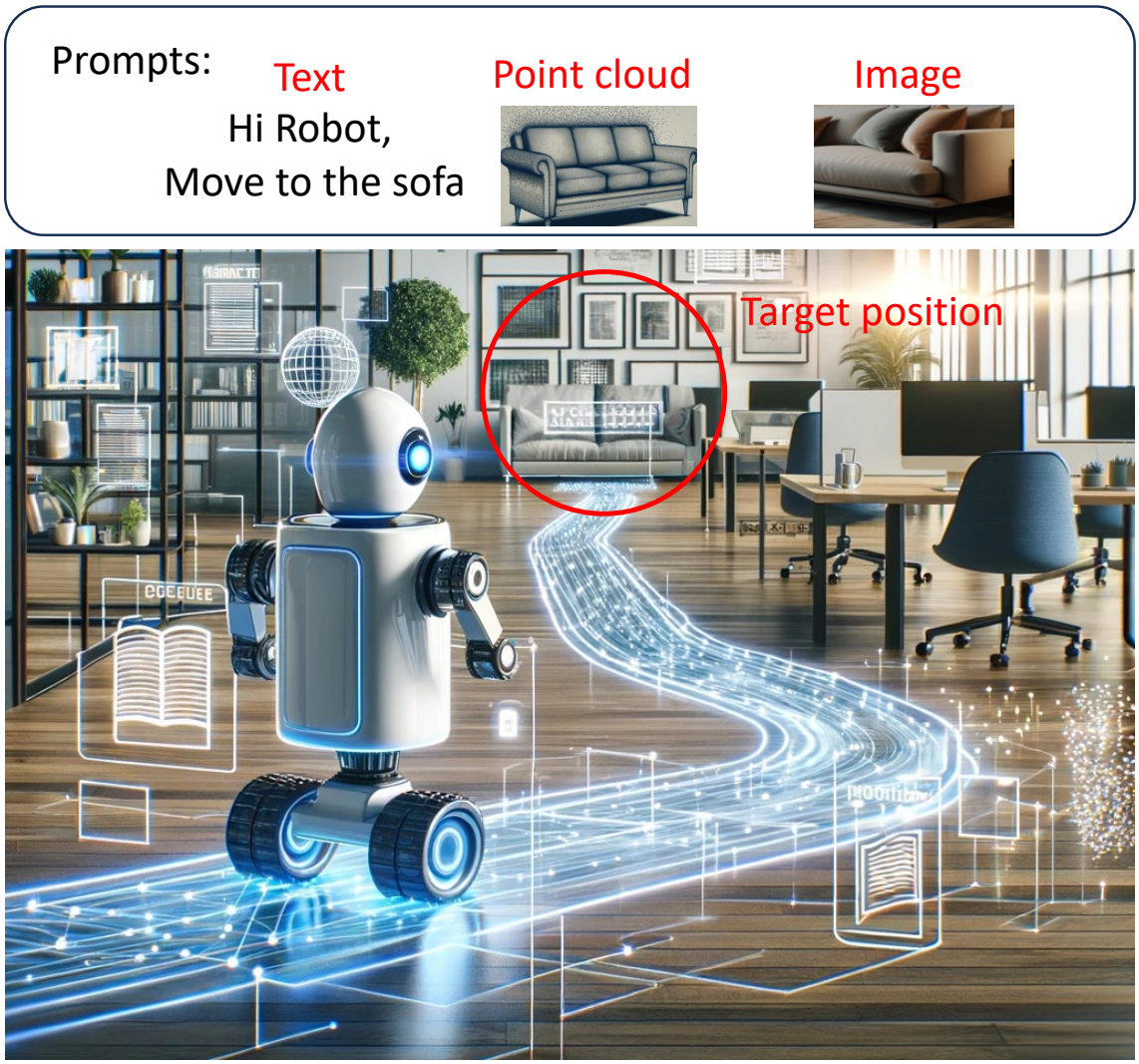

Figure 1: The project aims to design a localization method capable of fusing visual images, texts, and 3D point clouds and navigating the robots to a target location in unseen indoor environments. (The pictures are generated by ChatGPT-4.)

4 Experiments on navigation, e.g. evaluating on unseen scenes (~ 50h);

5 Writing a project report and preparing the presentations (~ 50h).

Contact

Dr. Yan Xia
Email: yan.xia@tum.de

References

[1] Chenguang Huang et al. "Visual Language Maps for Robot Navigation". In: *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*. London, UK, 2023.

[2] Obin Kwon, Jeongho Park, and Songhwai Oh. "Renderable neural radiance map for visual navigation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 9099–9108.

[3] David G Lowe. "Distinctive image features from scale-invariant keypoints". In: *International journal of computer vision* 60 (2004), pp. 91–110.

[4] Paul-Edouard Sarlin et al. "From coarse to fine: Robust hierarchical localization at large scale". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 12716–12725.

[5] Yan Xia et al. "CASSPR: Cross Attention Single Scan Place Recognition". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. Oct. 2023, pp. 8461–8472.

[6] Yan Xia et al. "Soe-net: A self-attention and orientation encoding network for point cloud based place recognition". In: *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*. 2021, pp. 11348–11357.

[7] Yan Xia et al. "Text2Loc: 3D Point Cloud Localization from Natural Language". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024.