DYNAMIC SCENES - SLAM - ACCURATE PHOTOREALISTIC MAPS

[INTRODUCTION TO PROBLEM]

Building robots that operate in households as personal assistants is a longstanding goal of the field of robotics and artificial intelligence. Embodied AI has recently emerged as a research field that emphasizes the usage of AI techniques, such as computer vision and natural language processing within physical entities, to achieve this goal. The survey by Srivastava et al.[24] reports that the top 100 tasks humans want robots to perform in their houses revolve around cleaning, cooking, and rearranging objects. A prerequisite to achieving such composite tasks is the capability to navigate to specified objects present in the environment autonomously.

[PROBLEM DEFINITION]

This is referred to as the “object goal navigation” or “ObjectNav” task in literature [2] and serves as the focal point of this thesis.

[REQUIREMENTS TO SOLVE THE PROBLEM]

Large variations across different households and complexity in terms of their layouts, structures, and objects necessitate a robot to understand the geometric and semantic aspects of its environment for planning.

[PREVIOUS APPROACHES]

Previous works on indoor robot scene understanding have focused on using particle filter-based mapping and localization approaches [8], graph-based metric-semantic maps [11, 3, 9], and prior information like floor plan layouts [29] to map and localize in an environment accurately. For indoor navigation, sampling-based path planners [12], deep reinforcement learning [26], and active sensing using informative path planning [25] have been used.

[SHIFT TOWARDS NEW APPROACHES]

Most approaches rely on deep learning-based perception [22, 21]. Recently, works like Gadre et al. [7, 13] proposed utilizing large-scale transformer-based foundational models for perception in scene exploration, specifically open vocabulary feature detection [19, 17] and segmentation algorithms [14], that provide detection and segmentation of objects with natural language labels in an open world setting unlike previous deep learning methods, which has led to their utilization as a perception backbone for the ObjectNav task.

[MORE IN DETAIL ABOUT THE NEW APPROACHES - INTRODUCTION TO THE BASELINES]

Approaches to tackle the ObjectNav task usually consider the environment to be map-based or map-less. In the map-based scenario, works focus on map construction while navigation is considered a downstream task. Liu et al. [16] demonstrate object search and navigation on an open vocabulary 3D reconstructed static map. Hughes et al. [11] introduce a real-time closed-vocabulary 3D scene graph representation for mapping. Gu et al. [9] improve this by creating an open vocabulary metric-semantic 3D scene graph, but their approach is not real-time. Another notable state-of-the-art approach is VL-Maps by Huang et al. [10] where the authors create a 2D metric semantic map by fusing features from an open vocabulary segmentation network with RGBD camera observations, which can also be used for obstacle avoidance. However, planning is only possible after the environment is mapped and scene exploration is not dealt with in this approach. On the other hand, in the map-less scenario, research focuses on exploration with active perception to locate the object. Recent works in this area employ vision language models (VLMs) as object detectors [13, 7, 23], develop transformer-based architectures [27, 6, 4], and use LLMs for planning [23]. Yokohama et al. [28] develop VLFM to semantically bias frontier-based exploration based on the inductive biases of VLMs, allowing them to explore regions most correlated to the desired object.

[DRAWBACKS OF THE EXISTING METHODS]

However, the drawback of their approach is that they create a new frontier-exploration map for every new object they have to find, which grows with the number of objects, and is memory inefficient. Another major drawback of their approach is the lack of a common map representation that could contain geometric and semantic features of the environment and all objects together.

[INTRODUCTION TO THE PROPOSED APPROACH]

This project will aim to develop an approach for addressing the ObjectNav problem in household environments. The approach will search for a static object by actively exploring a static 3D environment having no prior map. During exploration, it will incrementally build an open-vocabulary map which will be used for active re-planning. As starting points, exploration will be performed similarly to VLFM, and map building will be done akin to VL-MAPS. To be consistent with the assumptions of the above methods, RGB-D images from the forward-facing camera of a mobile robot and its error-free poses are assumed as inputs.

[FACTORS THAT WILL BE ADDRESSED BY THE NEW PROPOSED METHOD]

The project will address the limitations of both baseline methods by developing a hybrid exploration and mapping pipeline, which will actively explore an environment to find an object without an apriori map , index an arbitrary amount of objects into a single consistent map, and support active re-planning.

[EXPERIMENTS DESIGNED TO EVALUATE THE CLAIMS - DATASETS, METRICS]

Experiments will aim to show that the method could actively explore the environment to find a number of objects and build a map during exploration. We plan to evaluate our approach against the baselines VL-MAPS [10] and VLFM [28] representing state-of-the-art methods for open vocabulary mapping and exploration, respectively. We will be using metrics like the success rate of finding the object, distance traveled [1], and time taken to reach the object, which will highlight the robustness and efficiency of using a hybrid mapping-exploration approach.

[DISCUSSION ON THE EXPECTED FINAL RESULTS OF THE THESIS]

As VL-MAPS requires a manual exploration of the environment for map creation before it can be used to perform navigation, our incremental approach would be faster. Also compared to VLFM, which only relies on exploration for object search, our hybrid approach would be quicker and more memory efficient. Experiments on household scenes in a high-fidelity simulator like Habitat [18] or AI2 Thor [15] and using real-world datasets such as the HM3D and the MP3D Dataset [20, 5] will aim to show that the method works on realistic household scenes and can provide a basis for ObjectNav in household environments

Active Perception and Mapping for Open Vocabulary Object Goal Navigation

[INTRODUCTION TO PROBLEM]

Building robots that operate in households as personal assistants is a longstanding goal of the field of robotics and artificial intelligence. Embodied AI has recently emerged as a research field that emphasizes the usage of AI techniques, such as computer vision and natural language processing within physical entities, to achieve this goal. The survey by Srivastava et al.[24] reports that the top 100 tasks humans want robots to perform in their houses revolve around cleaning, cooking, and rearranging objects. A prerequisite to achieving such composite tasks is the capability to navigate to specified objects present in the environment autonomously.

[PROBLEM DEFINITION]
This is referred to as the “object goal navigation” or “ObjectNav” task in literature [2] and serves as the focal point of this thesis.

[REQUIREMENTS TO SOLVE THE PROBLEM]
Large variations across different households and complexity in terms of their layouts, structures, and objects necessitate a robot to understand the geometric and semantic aspects of its environment for planning.

[PREVIOUS APPROACHES]
Previous works on indoor robot scene understanding have focused on using particle filter-based mapping and localization approaches [8], graph-based metric-semantic maps [11, 3, 9], and prior information like floor plan layouts [29] to map and localize in an environment accurately. For indoor navigation, sampling-based path planners [12], deep reinforcement learning [26], and active sensing using informative path planning [25] have been used.

[SHIFT TOWARDS NEW APPROACHES]
Most approaches rely on deep learning-based perception [22, 21]. Recently, works like Gadre et al. [7, 13] proposed utilizing large-scale transformer-based foundational models for perception in scene exploration, specifically open vocabulary feature detection [19, 17] and segmentation algorithms [14], that provide detection and segmentation of objects with natural language labels in an open world setting unlike previous deep learning methods, which has led to their utilization as a perception backbone for the ObjectNav task.

[MORE IN DETAIL ABOUT THE NEW APPROACHES - INTRODUCTION TO THE BASELINES]

Approaches to tackle the ObjectNav task usually consider the environment to be map-based or map-less. In the map-based scenario, works focus on map construction while navigation is considered a downstream task. Liu et al. [16] demonstrate object search and navigation on an open vocabulary 3D reconstructed static map. Hughes et al. [11] introduce a real-time closed-vocabulary 3D scene graph representation for mapping. Gu et al. [9] improve this by creating an open vocabulary metric-semantic 3D scene graph, but their approach is not real-time. Another notable state-of-the-art approach is VL-Maps by Huang et al. [10] where the authors create a 2D metric semantic map by fusing features from an open vocabulary segmentation network with RGBD camera observations, which can also be used for obstacle avoidance. However, planning is only possible after the environment is mapped and scene exploration is not dealt with in this approach. On the other hand, in the map-less scenario, research focuses on exploration with active perception to locate the object. Recent works in this area employ vision language models (VLMs) as object detectors [13, 7, 23], develop transformer-based architectures [27, 6, 4], and use LLMs for planning [23]. Yokohama et al. [28] develop VLFM to semantically bias frontier-based exploration based on the inductive biases of VLMs, allowing them to explore regions most correlated to the desired object.

[DRAWBACKS OF THE EXISTING METHODS]

However, the drawback of their approach is that they create a new frontier-exploration map for every new object they have to find, which grows with the number of objects, and is memory inefficient. Another major drawback of their approach is the lack of a common map representation that could contain geometric and semantic features of the environment and all objects together.

[INTRODUCTION TO THE PROPOSED APPROACH]

This project will aim to develop an approach for addressing the ObjectNav problem in household environments. The approach will search for a static object by actively exploring a static 3D environment having no prior map. During exploration, it will incrementally build an open-vocabulary map which will be used for active re-planning. As starting points, exploration will be performed similarly to VLFM, and map building will be done akin to VL-MAPS. To be consistent with the assumptions of the above methods, RGB-D images from the forward-facing camera of a mobile robot and its error-free poses are assumed as inputs.

[FACTORS THAT WILL BE ADDRESSED BY THE NEW PROPOSED METHOD]

The project will address the limitations of both baseline methods by developing a hybrid exploration and mapping pipeline, which will actively explore an environment to find an object without an apriori map , index an arbitrary amount of objects into a single consistent map, and support active re-planning.

[EXPERIMENTS DESIGNED TO EVALUATE THE CLAIMS - DATASETS, METRICS]

Experiments will aim to show that the method could actively explore the environment to find a number of objects and build a map during exploration. We plan to evaluate our approach against the baselines VL-MAPS [10] and VLFM [28] representing state-of-the-art methods for open vocabulary mapping and exploration, respectively. We will be using metrics like the success rate of finding the object, distance traveled [1], and time taken to reach the object, which will highlight the robustness and efficiency of using a hybrid mapping-exploration approach.

[DISCUSSION ON THE EXPECTED FINAL RESULTS OF THE THESIS]

As VL-MAPS requires a manual exploration of the environment for map creation before it can be used to perform navigation, our incremental approach would be faster. Also compared to VLFM, which only relies on exploration for object search, our hybrid approach would be quicker and more memory efficient. Experiments on household scenes in a high-fidelity simulator like Habitat [18] or AI2 Thor [15] and using real-world datasets such as the HM3D and the MP3D Dataset [20, 5] will aim to show that the method works on realistic household scenes and can provide a basis for ObjectNav in household environments