

Teledrive: An Embodied AI based Telepresence System

Snehasis Banerjee¹, Sayan Paul¹, Ruddra Roychoudhury¹, Abhijan Bhattacharya¹, Chayan Sarkar¹, Ashis Sau¹, Pradip Pramanick¹, Brojeshwar Bhowmick¹

¹Visual Computing and Embodied AI, TCS Research, Kolkata, India.

Contributing authors: snehasis.banerjee@tcs.com; p.sayan@tcs.com; ruddra.roychoudhury@tcs.com; abhijan.bhattacharyya@tcs.com; sarkar.chayan@tcs.com; ashis.sau@tcs.com; pradip.pramanick@tcs.com; b.bhowmick@tcs.com;

Abstract

This article presents ‘Teledrive’, a telepresence robotic system with embodied AI features that empowers an operator to navigate the telerobot in any unknown remote place with minimal human intervention. We conceive Teledrive in the context of democratizing remote ‘care-giving’ for elderly citizens as well as for isolated patients, affected by contagious diseases. In particular, this paper focuses on the problem of navigating to a rough target area (like ‘bedroom’ or ‘kitchen’) rather than pre-specified point destinations. This ushers in a unique ‘AreaGoal’ based navigation feature, which has not been explored in depth in the contemporary solutions. Further, we describe an edge computing-based software system built on a WebRTC-based communication framework to realize the aforementioned scheme through an easy-to-use speech-based human-robot interaction. Moreover, to enhance the ease of operation for the remote caregiver, we incorporate a ‘person following’ feature, whereby a robot follows a person on the move in its premises as directed by the operator. Moreover, the system presented is loosely coupled with specific robot hardware, unlike the existing solutions. We have evaluated the efficacy of the proposed system through baseline experiments, user study, and real-life deployment.

Keywords: Telepresence, Cognitive robotics, Embodied AI, AreaGoal, Person Following

1 Introduction

This article¹ primarily focuses on a telepresence robotic system in the context of remote caregiving. The pandemic situation demanded ‘social distancing’ as the new normal. Yet careful monitoring of patients in isolation must be taken care of without risking the lives of ‘caregivers’. Even without the pandemic, there is a shortage of caregivers in different parts of the world, which is expected to be acute during the next pandemic outbreak [1]. The availability of caregiver service must be done in a democratized manner

such that individual care is possible for geographically distant individuals. A telepresence robot can address part of this issue. However, a major hindrance to the wider deployment of telepresence systems is the ease of use, particularly for a non-expert user. Existing telepresence systems, like Double², provide a manual navigation capability, which is often cumbersome for a user in a non-familiar or semi-familiar environment. Moreover, manual navigation requires continuous user intervention to move the remote robot from place to place within the remote environment. Furthermore, existing telepresence systems are tightly

¹Data Availability Statement:
The data that support the findings of this study is available from
<https://github.com/facebookresearch/habitat-lab#data>

²<https://www.doublerobotics.com/>

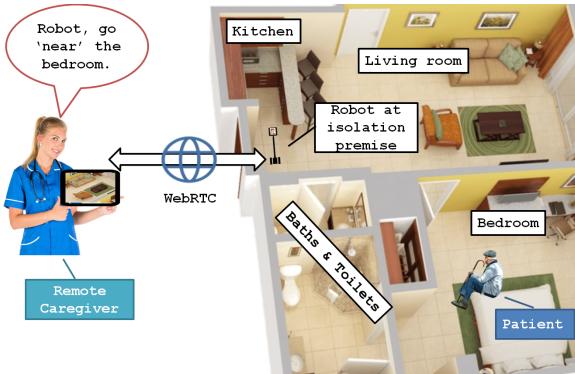


Fig. 1: ‘Teledrive’ system is given AreaGoal-based navigation instruction in remote caregiving use case.

coupled with the robot hardware (like Amy³), which makes it difficult for enhancement, particularly by third-party developers. Thus, hardware-independent robot software development is needed for ease of feature addition, which has been exactly followed in this work through loosely coupled software modules.

A telepresence system typically maintains a real-time connection with an application at the caregiver’s end and acts as an ‘avatar’ of the caregiver at the patient’s premise. The caregiver must navigate the patient’s premise in real-time through the Avatar interface based on audio-visual feedback as part of the ongoing real-time multimedia communication (Fig. 1). In most systems, the robot Avatar is maneuvered by the remote caregiver via manual instructions using on-screen navigation buttons, a joystick, a keyboard, etc. However, in an unknown premise (in the case of a tele-doctor), it would be too tedious for the caregiver to manually navigate the robot to the patient’s location. Hence, we conceive a system in which a caregiver can provide remote verbal instruction to the robot to navigate near a desired location inside the room (e.g., ‘bedroom’). The scope of the application can be further extended to isolation wards in health centers. Speech-based human-robot interaction (HRI) increases the usability and acceptability of the robot [2] in this context. Hence, a need is felt for a software architecture that can facilitate various embodied AI algorithms and machine learning models (and technology enablers) to work in harmony in the aforementioned scenario.

1.1 Motivation

Navigation is the key task of a telepresence robot. To establish the motivation for this work, let us take the help of the remote caregiver scenario depicted in Fig. 1. Based on scene analysis on the ego view (front camera view), the robot is able to move to a position near the intended goal location. Once the robot reaches ‘near’ that position, the caregiver can take manual control and perform finer control through the on-screen navigation buttons. In this example, the robot is connected with the caregiver’s PDA over the Internet through a WebRTC-based communication protocol. The robot is at the entrance of the patient’s premises. The old aged patient is in the bedroom whose vitals can be recorded by the robot like heart rate [3], emotions [4] etc. The caregiver verbally instructs the robot to navigate to the bedroom. Once the robot is able to locate itself around the bedroom, the caregiver can manually lead the robot to the bed where the old patient is waiting. This motivates us to develop a navigation capability for the following tasks - (a) PointGoal: going to a point location (b) ObjectGoal: searching and finding an object category (c) AreaGoal: navigating to a region category. Also, while navigating the environment, the robot may get confused about the next step or get stuck at a spot – in such a scenario, it is essential to leverage dialogue exchange with the operator to clarify the robot’s next move. Usually in a caregiver scenario or in a meeting, different users from varied remote locations need to converse and take control of the robot in turns. This requires a remote telecommunication protocol that can handle multiple users, also called multi-party communication. In this work, we have built a system catering to these requirements. To illustrate further, let us assume that a user gives a speech instruction ‘check if the cup is in the bedroom’ which gets processed into a task instruction. In this case, the robot needs to invoke the AreaGoal task to find the bedroom region. Next, the robot needs to search for the object ‘cup’ invoking the ObjectGoal task. It is to be noted that PointGoal is invoked in intermediate points for local transitions – or intermediate points between the start and end goal. If there are two or more cups in the bedroom, the robot can ask a question back (dialogue exchange) to verify if indeed target goal is achieved.

Before deploying the algorithms in a real robot, the standard way is to train and test them on an actuation-enabled virtual robot in a simulator on some photo-realistic datasets of indoor scenes or in a created

³<https://www.amyrobotics.com/>

scene using SfM [5]. Matterport3D [6], Replica [7] and Gibson [8] are some popular 3D datasets mapped from real world house layouts. Among the choice of simulators for the tasks, we have used AI Habitat [9] (a simulation platform dedicated to Embodied AI research) as state-of-the-art benchmarks have also used the same framework. AI Habitat enables developers to test their algorithms in a 3D engine, where input is the robot's actuation commands and output is the actuation in the scenes loaded from the dataset. AI Habitat provides access to the ego view camera of the virtual robot in the form of an RGB frame, depth frame, and optionally semantic annotation of the scene. The odometry information is also available at each time step of the robot alongside the ego camera view.

The concept of Embodied AI is as follows - artificial agents or virtual robots, operating in 3D environments take as input egocentric perception and output actions based on a given goal task. This also includes training of embodied AI agents in realistic 3D simulators, before transferring the learned skills to reality. The presented telepresence system conforms to the embodied AI philosophy in the following manner. The user instruction is received using a human-robot interaction module where robot grounds the instruction [10] and ambiguity is resolved with respect to the observed scene using principles presented in [11, 12]. The other three embodied AI downstream tasks, namely PointGoal, ObjectGoal, and AreaGoal, as defined in [13], are taken care of – initially in 3D simulators and then transferred to real-world deployment. Additionally, the person-following module is trained and tested in a 3D simulator before transferring to a real-world robot.

1.2 Contributions

The primary contributions of this paper are enlisted:

- We have concretely defined the problem specification of a set of ‘Areagoal’ tasks and specified the metrics, benchmark, and approach with promising results. This is tied with individual contributions in Human-Robot Dialogue Exchange, Pointgoal, and Objectgoal tasks as enablers. To the best of our knowledge, this is the first work on the ‘AreaGoal’ downstream task tested on benchmark, backed by a real-life deployment.
- We have carried out a user study about the usability and functionality of the features presented in the

Teledrive system and revealed interesting insights based on real human experience.

- We have presented a software framework containing embodied AI features (including person following) that runs on ROS 1 or ROS 2 compliant robotic hardware in contrast to prior art where software features and hardware are tightly coupled [14].
- We have developed a custom communication framework to enable real-time exchange of multi-media and control signals between the robot and the remote operator. This is achieved through custom exchange semantics built on standard WebRTC APIs. This enables an end-to-end web browser-based system. We have also built interfaces between the WebRTC APIs and the low-level robot control APIs. In lines of the operator-to-robot channel, we have created custom semantics on WebSocket for exchange between the Robot and the Edge in real-time for offloading of computation from the robot and fetching the computation outcome into the robot.
- We have developed a device-agnostic responsive web browser-based frontend, that makes the solution platform-independent and easily accessible over URL by remote users having internet connectivity across mobile device variants.

The overall organization of the paper is as follows. Section 2 provides a comparison table of the state-of-the-art telepresence systems, highlighting relevant related works. In section 3, we present the software system architecture. Section 4 discusses task understanding and dialogue exchange with robots. Section 5 describes various aspects of cognitive navigation with a focus on AreaGoal and ‘person following’ tasks. ‘PointGoal’ and ‘ObjectGoal’ tasks are also discussed in short. Section 6 describes the browser-based remote operator interface and user studies. Finally, we conclude the paper in section 7 along with future work.

2 Related Work

There has been a significant amount of work done in the robotic telepresence catering to a multitude of solutions [27] [28] [29] [30] [31] [32] [33] [34] [35] [36] [37]. The majority of telepresence systems are intended for use in an office setting. Some are suited for elderly care, healthcare, and robotic research. Most of the prior art has focused on the hardware capabilities and allied features. However, we focus on the

Table 1: Comparison of features in existing telepresence systems.

software platform itself that can run on any ROS-compliant hardware with some adaptation at the hardware level. In terms of software, we have included several components to perform embodied AI tasks to give a flavor of cognitive intelligence to the telepresence scenario. The closest work is ENRICHME [15] which supports some simple software-level cognitive features. However, we offer a significant shift towards complex and sophisticated embodied AI features that have been tested to work on a real robot.

Standalone efforts on embodied AI tasks have recently gained the focus of the research community - like cognitive navigation [38] and human-robot dialogue. A number of works on Speech based HRI in robots have focused on accessibility [39] and cloud infrastructure for HRI using speech [40], however, the area of speech HRI for general purpose telepresence scenario has limited peer-reviewed work. In cognitive navigation, for the ‘PointGoal’ problem, there exists some work that uses frontier-based planners [41, 42]. There has been some work on the person following by robot [43, 44], however, our work gives control to the remote user in order to follow a person, making it suitable for the telepresence scenario; and introduces engineering contributions to execute the task in the real world. In recent times, the ‘ObjectNav’ task has grabbed the attention of the robotics research community. The initial work on Semantic Visual Navigation [45] uses scene priors learnt from the standard image dataset. Their work uses Graph Convolutional Networks (GCNs) to embed prior knowledge into a Deep Reinforcement Learning framework and is based on an Actor-Critic model. However, that work lacked a concrete decision model when two or more objects are in the same scene. The proposed method makes an improvement on GCN-based ObjectNav [46] by using trajectory data priors along with a learnt region-object joint embedding learned from a GCN training to perform the ObjectNav task. We have presented comparative results with the end-to-end RL-based ObjectNav – SemExp [47] to support the efficiency of the approach. In the context of the ‘AreaGoal’ problem, there is no dedicated work, apart from some work on scene graph analysis [48] to decide the region. Another work on robot navigation learning is based on sub-routines from egocentric perception [49] that shows an example run for the AreaGoal task in the ‘washroom’ region. However, the work neither gives away details of the AreaGoal task specifically nor performed any benchmark studies on the AreaGoal task. Their work was more focused on the PointGoal task.

In contrast, our paper tackles the AreaGoal problem in great detail backed by results. Hence, the combination of multiple software features within the telepresence framework is a unique proposition of this work, which is further evident from the comparison of features in prior art as presented in table 1.

3 System architecture

In this section, we provide a high-level overview of Teledrive (Fig. 2 and Fig. 3). The edge-cloud hybrid architecture is based on the principles laid out in [50, 51, 52]. It comprises four major subsystems – Communication, Embodied Dialogue Exchange, Embodied Navigation, and User Interface.

Robotic Telepresence enables virtual presence of a distant human (aka Operator) through an in-situ Robot (aka Avatar) which is maneuvered in real-time by the Operator, while having multimedia conferencing with the Avatar side participants over the Internet. The recent pandemic has underscored the importance of such systems. The Edge-enabled architecture helps offload computation required for the cognitive tasks by the robot. The Edge may be an in-situ computer connected over local WiFi or it may be inside the network provider’s infrastructure. The communication mechanism coupling the entire system has two major aspects: (1) to realize collaborative multi-presence session (CMS) and (2) to coordinate amongst computation modules, distributed between the Edge and the CMS on the Avatar. The CMS is maintained by WebRTC compatible browsers on every peer. The first part is ensured by a unique application layer on WebRTC. It supports a unique hybrid topology to address diverging Quality of Service (QoS) requirements of different types of data. The A/V is exchanged between the Avatar and the remote users through a cloud centric star topology over the SRTP based WebRTC media channel. But the delay-sensitive control signals from the active Operator to the Avatar is exchanged over a P2P data-channel on SCTP established directly between the Operator and Avatar on demand. This is unlike usual WebRTC topologies which support either mesh or star topologies using its inherent P2P (peer-to-peer) mechanism.

Navigation, though natural to humans is non-trivial for robots. The ‘PointGoal’ problem is solved using the egocentric views obtained from the robot, which are then combined to make a global view of the environment using a spatial transform. The global top-view map is then used to plan a trajectory using

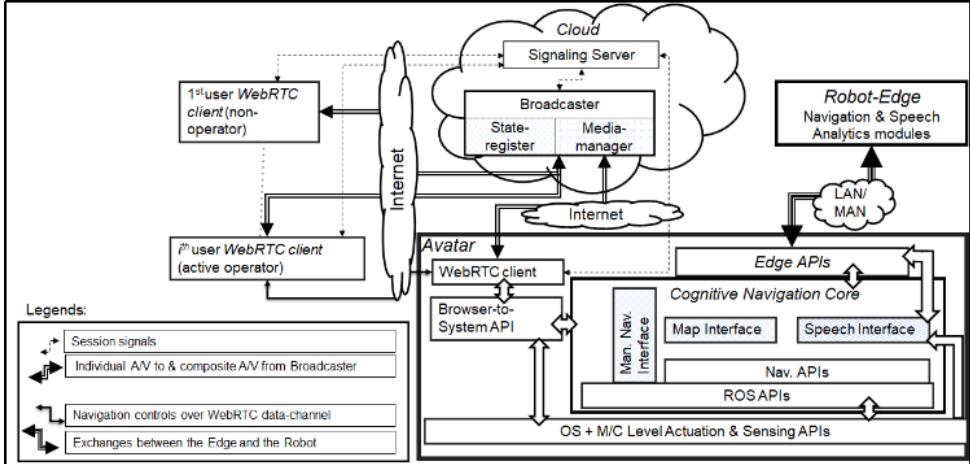


Fig. 2: Edge-Cloud topology of the distributed networked Embodied AI of Teledrive.

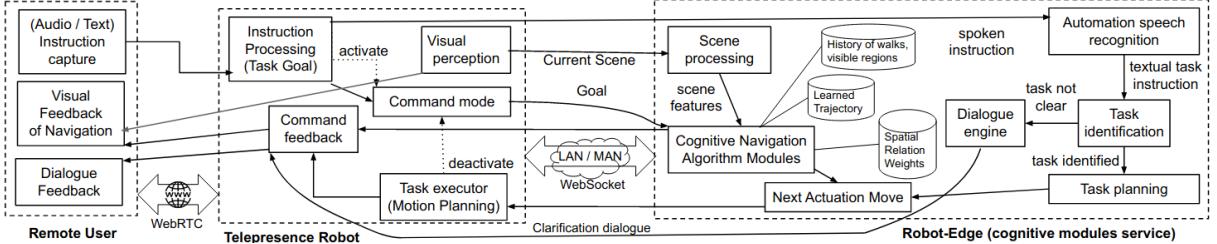


Fig. 3: System architecture of Teledrive showing distributed set of building blocks spread across four entities – master device, cloud server, robot (avatar), and robot-edge.

the A* planner. The Dialogue engine helps in disambiguation of tasks instructions specified verbally using a pre-trained model ‘wav2vec’, which is passed on to task identifier [53] to understand the goal and invoke the corresponding task module as a service.

In the context of the caregiver example, the telepresence system maintains a real-time connection with an application at the caregiver’s end and acts as an Avatar of the care giver at the patient’s premise. The caregiver must navigate the patient’s premise via robot avatar in real-time, based on audio-visual feedback. The caregiver can provide a remote verbal instruction to the robot to navigate near a desired location inside the room (e.g., ‘bedroom’). The speech based human-robot interface understands the desired goal from the voice. Based on the in-situ analysis of the semantic map derived from the live captured frames, the robot is able to move to a position near to the intended location. Once the robot reaches near that goal, the caregiver can take manual control and perform finer control through on-screen navigation buttons. The

robot is connected with the caregiver’s mobile device over the Internet through the WebRTC based communication protocol. The following sections discuss the individual building blocks in detail, which are needed to support the caregiver telepresence scenario, amongst other possible use cases.

4 Human Robot Interaction

In a telepresence system, the robotic platform is primarily used to interact with human beings. Instead of manually controlling the robot through a continuous set of control commands, autonomous action by the robot based on high-level command is quite desirable [54, 2]. This requires a robust human-robot interaction (HRI) module.

Fig. 4 depicts the HRI module embedded with our Teledrive system. Since the speech interface is the most intuitive and user-friendly way to interact with a robot, Teledrive also supports speech-based task instruction to the robot. However, most of the time

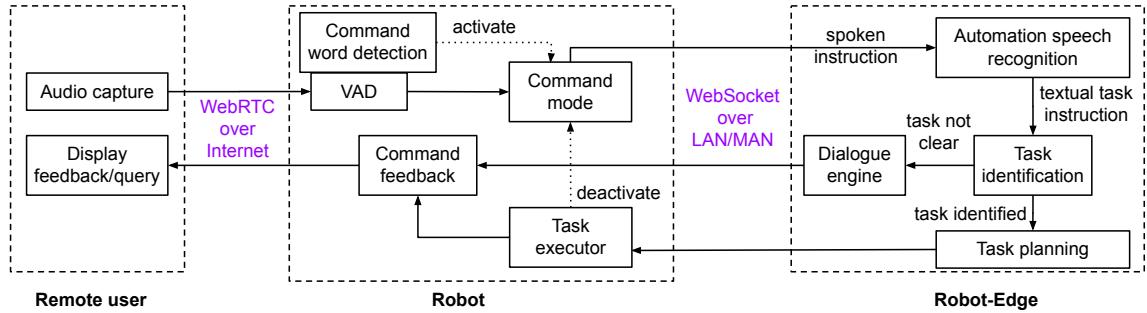


Fig. 4: Embedding of human-robot interaction mechanism into Teledrive.

what the master says is meant for the remote audience. To distinguish the control instruction for the robot and ordinary speech, we devised a protocol of prefixing the control instruction with a command word. This is common for many personal assistant systems where a wake word is spoken first, before instructing the assistant. Now whenever the robot receives an audio signal, the voice activity detection (VAD) module processes to find the command word. If there is no command word, no further action is taken. On the other hand, if the command word is detected, the command mode is activated. The command word is detected using a CNN-based binary classifier. The model takes any audio chunk of a specified duration (upper limit to pronounce the command word) and outputs whether the audio chunk contains the command word or not.

In the command mode, the received audio is recorded locally on the robot until a silence of significant duration is detected by the VAD, which signifies the end of instruction. Then, the recorded audio is transmitted to the robot edge for further processing. At the robot edge, the audio is translated to text first. For this, we use our own automated speech recognition (ASR) model for embodied agent [55, 56]. This is an adaptation of the ‘wav2vec’ ASR model, which is trained using a transformer-based deep neural network [57]. Then, the textual instruction is passed to a task identification module. This module identifies the task type and the parameter(s) from the natural language instruction. An example is ‘navigate to the bedroom’ where the task is navigation and the parameter is ‘bedroom’ mapped to an area lookup table. If the task understanding is successful, the execution plan is communicated to the robot, where the plan is executed through API calls. At this point, the robot exits from the command mode. In case the robot edge fails to decode the task instruction, it generates a query for

the operator (Master) that is passed onto the operator’s devices through the robot communication channel.

Another aspect is handling ambiguity in HRI. Following the principles set in the work [11], the ambiguity in goal of the task instruction can be categorized, and accordingly the type of question back can be formulated to aid the robotic agent in selecting the next course of action to complete the task.

5 Embodied AI Navigation Tasks

One of the principal tasks that an embodied AI agent needs to perform very well is navigation. In this regard, [13] has classified the navigation tasks into *PointGoal* (go to a point in space), *ObjectGoal* (go to a semantically distinct object instance), and *AreaGoal* (go to a semantically distinct area). The *AreaGoal* problem (also called *AreaNav*) has been elaborated here, while other tasks are described in brief. We evaluate the ‘ObjectGoal’ navigation task results on 4 evaluation metrics as specified in [13] and [47] and refer the metrics in Appendix (section 9).

For experimentation and benchmark perspective, a fixed discrete action space is the norm followed by the peer community for embodied AI tasks. However, in real life deployment in robots, the step length and the rotation angle can be specified, provided training is done on a wide action space. For the person following task, the robot is free to move and rotate by values supported by the software running on robot hardware. So we could specify precise angle of rotation and step length of movement in that case. For a robot with a camera only in the front, a backward motion is not allowed to prevent (a) collision (b) uncertainty as very little information comes into view due to reverse motion. So alternatively, a backward motion is done as 180^0 turn in left or right, and then going forward.

The baselines for Cognitive Navigation are: (a) Random exploration: the robot does random movements, until the goal is reached or there is a timeout. However, this includes collision avoidance intelligence. (b) Frontier-based exploration: The robot tries to explore the entire layout of an indoor house, and while doing so will encounter the target goal at some point of time. This in some way is a frontier search method, avoiding visiting the same place again, and the success mostly depends on how close the target area is to the randomly initialized robot pose. Usually an upper limit of number of steps is kept depending on the dataset indoor space layout, and target needs to be reached within those steps. ‘PointNav’ task is leveraged by both ‘ObjectGoal’ and ‘AreaGoal’ modules in order to move from one point to another pre-specified point location (by performing local motion planning). To ensure that during navigation, the robot maintains the trajectory with best possible network connectivity, we have used a zero learning, source-agnostic approach as presented in [58].

5.1 PointNav Module

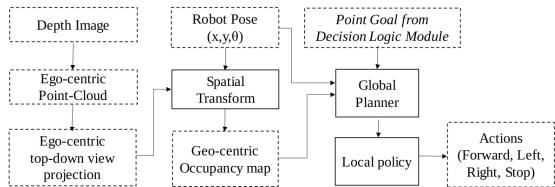


Fig. 5: Method overview of PointNav Module.

Given a start location (x_1, y_1) and an end goal location (x_2, y_2) for a ground robot, the task of this module is to help the robot navigate to the target within a certain predefined threshold without colliding with any static obstacles present in the environment. The robot is provided with an RGB-D camera sensor and wheel encoders that gives at each timestep – the RGB image, depth image (raw or refined [59]) and absolute pose of the robot with respect to the start location. The depth image is converted to an egocentric point cloud using known camera intrinsics, and then is further flattened along the vertical axis (z-axis) to obtain an egocentric top-down projection occupancy map, containing the obstacles that robot has perceived from its ego view.

As this problem is solved using a differential drive robot executing a motion on flat surface, the absolute pose of the robot with respect to the start can be resolved into 3 variables: (x, y, θ) – x being the forward displacement, y being the displacement to the left perpendicular to x and θ being the anti-clockwise rotation of the robot with respect to it’s center. Given the wheel encoder differences at each timestep, the values of (x, y, θ) can be defined using [60]. Our experiments with the Double3 Robot show, that on a flat surface (coherent with office, or home environments with hard carpet), the wheel odometry of Double3 has an average drift of 1 cm per metre of motion; and rotation error is negligible. In absence of a motion capture system or LiDAR, this was done by running the robot manually over loops of 3 metres to 10 metres so that the start and end positions of the trajectory is the same; and then measuring the drift accumulated by the robot wheel odometer. We reason that since each subtask of the Area-goal problem (or any similar problem) is independent, and the absolute position of the robot is not needed across the whole trajectory, if we can accumulate the pose from trajectory’s start position, till the trajectory ends, the purpose is served.

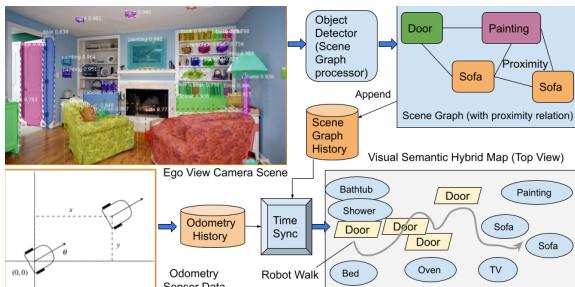
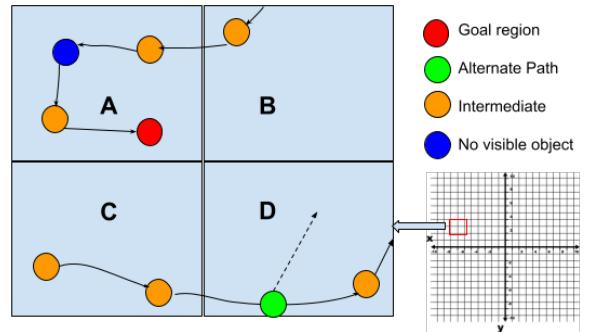
This egocentric map is in camera co-ordinate space, and is transformed to the world co-ordinate space using the absolute pose at that timestep by the spatial transform block (ST) to update the geocentric map. The Spatial transform block [61] transform each point in the egocentric point cloud into it’s position with respect to the start location. This updated geocentric map, current location and the goal location is fed to the Global Planner at every timestep t , which generates a feasible short but safe trajectory or path from the current location to the goal location. The global planner used here is a modified implementation of the A* Algorithm [62], where the free-space cost map is weighted more near the obstacles. As a result, the generated trajectories or paths are planned at a safe distance from the obstacles. It tries to find a good balance between finding the shortest path to the goal and avoiding collisions with obstacles. This planned path and the current pose is then fed to the Local Policy which is a heuristic planner that finds the next best action to be performed by the robot to align and follow the path generated by the global planner. The heuristic planner [61] follows the global plan waypoints by deciding one of the actions out of ‘forward, turn left, turn right’, by understanding it’s orientation with respect to the nearest global planner waypoint.

Table 2: Results for ‘PointNav’ task.

Number of Trials	Successful trials	Success (%) ↑
994	905	91.05

We use a modular approach to solve this ‘Pointnav’ task where mapper, global planner and local policy is stacked one after the other. This decoupled approach helps our method to generalize to different environments without the need for re-training or fine-tuning as compared to end-to-end learned approaches. It also helps in sim-2-real transfers. We use AI Habitat [9] as our evaluation framework. The habitat Pointgoal challenge 2020 [63] uses a standard split for the Gibson Dataset. We have assumed ground truth pose for our evaluations. We use the validation split of the Gibson Dataset [64] which consists of 994 episodes where the robot is spawned at a random environment at a random start location and then given a random goal location that is reachable from the start point. An episode is deemed successful if the robot reaches the goal within 500 steps from the start location. It is assumed to have reached the goal if it gets within 0.36 metres (twice the robot radius which is 0.18 metres) of the goal location. We also tested this module on our real robot – Double3 [65] which is equipped with Intel Realsense RGB-D cameras, however due to the lack of motion capture systems, we have not provided results to our evaluation on the real robot for this subtask here. However, we plan to provide an evaluation in future. We kept the above action space same in both simulation evaluation and real-world testing for smoother sim-2-real transfer. The results on AI Habitat with Gibson dataset are presented in the Table 2.

5.2 Visual Semantic Map Module

**Fig. 6:** Semantic map generation from scene graph sequences learned from random walks in simulator.**Fig. 7:** Grid Map for back tracking to the next unexplored landmark point for ‘AreaGoal’ task.

When the robot is moving in an unknown environment, it is imperative to create a map of its surroundings while it is moving. While a traditional Visual SLAM-based approach [66] helps in ‘PointNav’ problem, however, for tasks needing higher level semantic understanding of the scene, a metric map or voxel map is not sufficient. Hence, an alternative Visual Semantic Map is introduced that uses odometry data of the robot combined with perception (via RGB camera) to create a map that is both metric (having relative distance level granularity) and topological (retaining connection between scene graphs). The map is further enhanced with external semantic level features to link regions and objects in the robot’s ego view. As seen in Fig. 6, the history of scene graphs extracted from a sequence of scenes along with their time synced odometry data helps in generating the map incrementally. It is to be noted that raw image data from camera is processed to build the scene graph, and not semantic factual information as in the case of stream reasoning [67] [68]. Scene graph processor takes in the RGB camera image as input and applies object categorization using YOLO [69]. It populates a graph with nodes as objects and edges as proximity relation between objects. The initial position of the robot is taken as (0,0,0) in 3-D co-ordinate space facing east. In case the robot is navigating in the same floor or elevation, the z-axis (height) will be always positive, whereas translation of robot will happen in (x,y) plane. The ‘z’ axis data points observed from ego view helps in aiding the robot to look in specific portions of 3D space while searching an object. As an example, ‘paintings’ should be hanging on a wall at some height above floor, whereas ‘sofa’ will be grounded on the floor.

The specific data-structure used for this is a rectangle grid as shown in Fig 7. It is computationally

inefficient to search each and every point when the robot needs to look for the next unexplored landmark. Hence the map is initialized with a grid structure which is a large enough area in comparison to a regular house layout. Each grid cell has a set of Boolean tags – explored or unexplored; having alternative branch path or not. Each grid cell can consist of multiple points from where the robot has taken an observation – this also include the objects identified in the view. Each of the points has a view angle associated with it. In a navigation run of moving forward, the side view objects and openings which were kept out of view will be ignored. So later, if backtracking is needed, the information of those unexplored view angles can be used for further inspection in those areas. If a robot has reached a dead end, say $(x, y, \theta : 2, 3, 30^0)$, and need to backtrack, it can find the nearest cell to backtrack from the list of alternative path points (shown in green color in Fig. 7) kept in a list. Also, when the robot is exploring, it can look ahead a few grid cells (if previously visited) to find the best grid cell to move to maximize the objective of the specific long term navigation task. The green circles are alternate odometry points that the robot has visited in past. The orange points are the transition points in each step, assuming discrete steps of a fixed distance. Blue circle denotes a position, where the robot could not see any visible object in that view (may be free space like hall or a failure in object detection algorithm to identify any object). In such situation, it needs to take the next move randomly by invoking the PointNav module to come out of that situation of indecision. The red circle is the location where the robot needs to reach currently – it can be the final goal or an intermediate goal given by the PointNav module to make it come out of a stuck region. Hence, this module helps in taking decisions where past navigation history and observations can be leveraged, by querying this representation.

5.3 ObjectNav Module

It is imperative that out-of-view object finding is a needed feature of the telepresence system. As an example, suppose a user wants to find where an object (say medicinal supplies) is at a remote location. Instead of manually driving and searching the entire indoor area, this task can help overcome the absence of physical doctors in safe homes in contagion scenarios. A single robot used across multiple patients can carry instruction of different doctors as well as patients. As seen in Fig. 8, the user gives instruction

from the operator (Master) end to find an object. The visual perception (camera feed of scene) of the robot is continuously processed to identify current objects and regions in view.

A 3-layered Graph Convolution Network (GCN) is trained with spatial relation weights and object finding trajectories that lead to a success. The number of connection layers in GCN (Fig. 9) is selected as three, as more number of edge chains (paths) does not yield enough distinct node level embeddings, as otherwise most nodes will get connected with some other distant node if graph path length is not restricted. The input to the GCN is the spatial relational weights of each object and region as a node along with their adjacency matrix, and output is 128 dimensional node embedding. The relational weights among objects with objects and objects with regions is learned by a combination of Visual Genome [71] extracted weights and extraction from random agent exploration on large number of AI Habitat scenes. This structure is called Spatial Relational Graph (SRG). The GCN takes two inputs during training: (i) input features for every node i , represented as a $N \times D$ matrix (N : number of nodes, D : number of input features); and (ii) graph structure in the form of an adjacency matrix A of size $N \times N$ [72]. It produces an output of dimension $N \times E$ where E is the dimension of the embedding. The *region* and *object* categorical values are mapped to integer values using the *one-hot encoding vector* to avoid bias, i.e., the index of the node has value 1 and other values are zeros. At evaluation time, based on a policy of object finding, given visible regions, history of walks and target object, the aforementioned trained model is used to calculate similarity score of current walk with trained walk data of same category of object. During runtime, the actuation move command is sent to the ‘Robot’ for execution using AI Habitat’s Geodesic Shortest Path Follower [73] navigation planner. The success rate of this methodology has been found to be around 94% for realistic indoor scenarios tested for 19 common indoor objects in 6 scene groups of Matterport3D dataset. The results are enlisted in Table 3.

5.4 AreaNav Module

The research community has mostly focused on the *ObjectGoal* problem, where as there is no dedicated work for the *AreaGoal* problem. Hence there is no crisp definition of the problem statement and how to

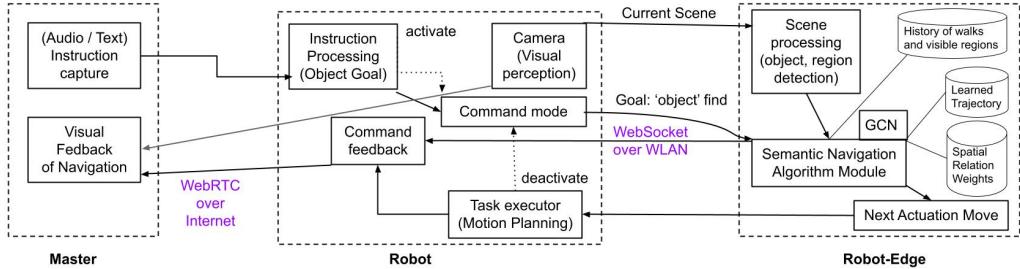


Fig. 8: Semantic navigation for ‘ObjectGoal’ task in telepresence scenario.

Table 3: Comparison of *ObjectGoal* method with baselines.

Method	Success ↑	SPL ↑	SoftSPL ↑	DTS ↓
Random	0.006	0.0049	0.0363	6.6547
Frontier Based Exploration [70]	0.598	0.3703	0.3891	4.2478
Our Method	0.94133	0.658	0.67931	0.3047

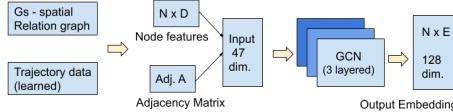


Fig. 9: Training GCN to encode embedding of information in SRG based on ‘valid’ trajectories.

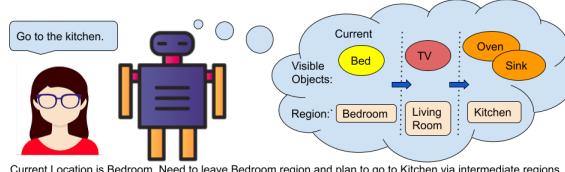


Fig. 10: Example of ‘AreaGoal’ task of navigation.

decide task completion. Hence, we introduce the task definitions as below.

- Definition of Area (A): an area in the context of robot navigation is a region or zone in space where the robot can navigate. Unreachable areas or obstacle-blocked areas are not considered.
- Definition of Concrete Boundary (P): The boundary of an area is the region outlining the polygonal sides of it. However, there may be cases like passage, door, openings where there is an overlap between two or more areas. Concrete Boundary marks the line after crossing which there is zero membership of that sub-area to other areas / regions.

Next, we concertize the problem statement and divide it into three problems as follows:

- Problem 1: A robot R needs to traverse to enter the *concrete* boundary P of an area A , given a goal task to navigate there. Hence, just getting a view of the area is not sufficient, the robot needs to be within the area for the task to be complete. This we denote as *AreaGoal* task.
- Problem 2. The above task completes when the area / region comes into robot perception view. This is a softer *AreaGoal* problem. This can come handy when just the outside view serves the purpose. This we denote as *ViewAreaGoal* task.
- Problem 3. The robot needs to navigate to the centroid point or centroid sub-region within a radius of 1.0 m of the goal area’s mathematical centroid. However, for this problem, the layout of the area needs to be known beforehand by fusion of external knowledge or learned apriori by exploration. There can be blockage in the center of an area – meaning no navigable point to go to. In that case the point closest to the navigable centroid is considered. This we denote as *CenterAreaGoal* task. While the earlier problem definitions are for unknown mapless environment, the latter either requires a metric map or run-time map generation based on the approximate navigable center point by taking into view the depth of surrounding structure (like walls) and free space.

We present results on the first two problems in a realistic simulation environment of AI Habitat and later

test it on real world housing apartments to establish the efficacy of the approach.

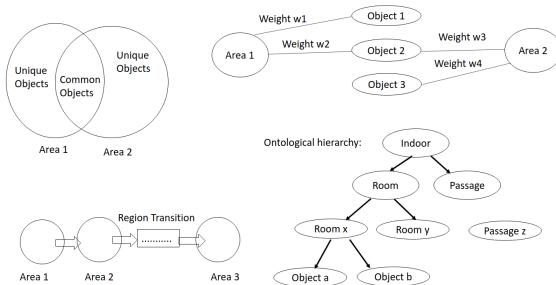


Fig. 11: Various ways to represent the concept of area.

As shown in Fig. 11, two or more areas can have common objects while some areas will have tight coupled unique objects. The unique objects aid in area identification with high accuracy. Also, region transition paths plays a complementary role in the area search. If a robot is in current location within area A, and target area is T, then based on whether A and T are adjacent or not, or what intermediate areas need to be traversed through – a navigation planning decision can be taken. Another aspect is related to edge weights between objects and regions as nodes. This spatial relational graph (SRG) will aid in statistically deciding the current area among the set of regions. Finally, a good way to represent indoor environment is by means of ontology, where regions can be divided into passages and room enclosures; and rooms can be subdivided into specific areas like bedroom, toilet. Each area can be represented as a composition of objects. In this paper, to solve the ‘AreaGoal’ task, we have leveraged this concepts.

5.4.1 AreaNav Methodology

The main task of ‘AreaGoal’ class of problems can be broken into two subtasks: identifying the area; and navigation from one area to another. Once the robot starts in a random location, first, it needs to identify the area it is currently in. Next, it needs to go out of the current area if it is not the target area. If there are multiple openings from current area, it needs to select the most statistically close one to the target area, and go there. If after taking that choice of path, the area is not reached, it needs to backtrack to an earlier viable branch position to continue the area search.

As shown in Fig. 12, the proposed system comprises of an input space comprising of sensor observations of RGB-D image and odometry readings, while the output is a member of the action space (left, right, forward) with goal of moving to the target area. The target area is specified in human instruction, processed by HRI module as shown in Fig. 13. The HRI module will help in identifying the task type as well as handling cases where robot is stuck at a place and needs human intervention to move towards the target. Based on object sets detected over a stream of scenes, the robot predicts the region based on object-region relations, region-region transitions and learnt GCN embeddings. The GCN was trained using random walks over large number of AI Habitat scenes to extract embedding representation for each node (object and regions). Then, based on aforementioned inputs a decision is taken to move towards target. As navigating from one area to another is a long term goal, it is broken into local goals that are handled by the ‘PointGoal’ module (Fig. 14), discussed in Section 5.1. Based on the robot’s current pose and the given goal location, the ‘PointGoal’ module plans the shortest navigable path. Also, when the robot is stuck in an area, an external fair point is given as a goal to the ‘PointGoal’ module for it to explore out of that stuck area. An external fair point is an unvisited point lying within the initialized grid, but at some pre-specified distance away from current robot location. This fair point is sampled from points close to the least sampled corner (based on earlier choices) among the four corners of the grid.

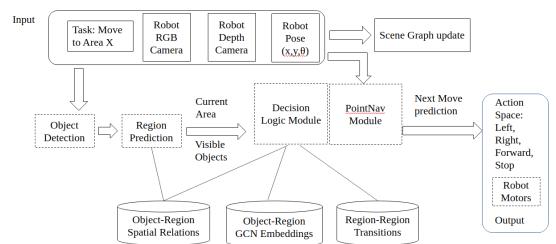


Fig. 12: System architecture of AreaNav module comprising submodules.

The above task is dependent on several software entities. They are enlisted below.

(a) Region Relation Graph: An indoor space is comprised of objects and areas (regions). There are some specific objects like cup, chair; and there are

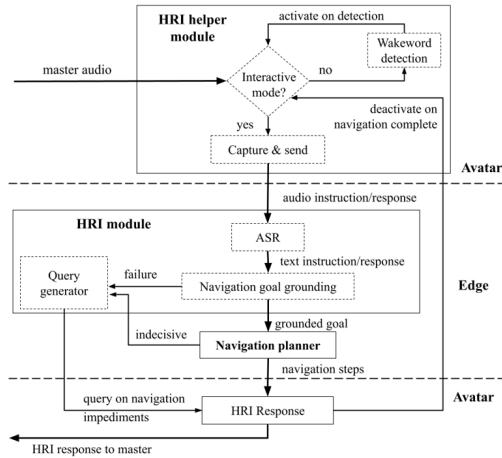


Fig. 13: Invocation of HRI module for task understanding and disambiguation.

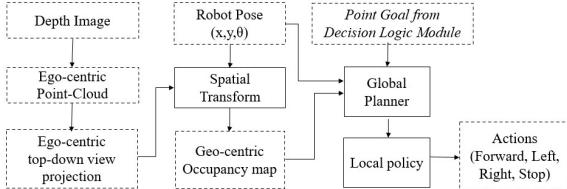


Fig. 14: Invocation of ‘PointGoal’ module for navigation towards objects or grid map corners.

some generic objects like wall, ceiling, floor which are continuous. There are three types of generic spatial relations: (a) object near another object, like table near chair (b) objects situated within a region, like bed as an object in region ‘bedroom’ (c) regions closely connected with other regions, like regions ‘dining room’ and ‘kitchen’. The list of indoor objects (as a subset of MS COCO [74]) considered are – bench, bottle, wine glass, cup, fork, knife, spoon, bowl, banana, apple, sandwich, orange, broccoli, carrot, hot dog, pizza, donut, cake, chair, sofa, potted plant, bed, dining table, toilet, TV monitor, laptop, mouse, remote (TV), keyboard, cell phone (mobile phone), microwave, oven, toaster, sink, refrigerator, book, clock, vase, scissors, teddy bear, hair drier, and toothbrush. The regions (areas or zones) considered are: bathroom, bedroom, dining room, study room, kitchen, living room, toilet, balcony and passage.

Two separate approaches were used to create a weighted relation graph. An entry in that relation

graph is ‘bed’ and ‘bedroom’ having very high relationship weight close to 1.0. In the first approach, these relations were extracted and normalized through a user survey comprising questions and responses such as how close two objects and regions are on an interval scale of 0 to 1. In approach B, the weights were learnt via observations registered in random walks in AI Habitat environment on a large number of indoor realistic scenes. Through ablation studies in various indoor layouts and scenes, it was found that manual survey based relation weights, provided better results for the ‘AreaGoal’ task.

(b) Region Transition Graph: The ‘AreaGoal’ problem specifically deals with navigating from region ‘A’ to target region ‘Z’ via intermediate regions as per layout of the indoor environments. In this regard, the past history of traversal through regions can guide whether a robot is moving towards a target region. As an example, when navigating from ‘kitchen’ to ‘bedroom’ the robot will have to generally navigate via intermediate region of ‘dining room’.

(c) Object Category Recognition: Identification of an area is generally determined by the type of objects the area contains. As an example, an area ‘bedroom’ will contain the object ‘bed’ and optionally ‘chair’, ‘clock’, ‘cell phone’, etc. In this regard, MaskRCNN [75] and YOLO [69] based approaches trained on MS Coco dataset has been used extensively in literature. However, contrary to simulation environments like AI Habitat, where object category identification can have ground-truth known via semantic annotation, the current framework was made in a way to work in real world, without using any ground-truth. In our case, robot observations (RGB-D image frames and odometer readings) are taken as input and action is given as output - this is the only communication between the simulation environment (treated as black box) and our modules. Through studies in both black box simulation settings and real world indoor environments, YOLO was found to perform better than Mask-RCNN for real-time robot observation processing in indoors, and hence was adopted here.

In order for the robot to navigate out of an area, it is important to detect openings. The object detection algorithm (YOLO) was pre-trained for additional 3 object classes: ‘open door’, ‘closed door’ and ‘free area’ (passage). These classes do not exist in off-the-shelf YOLO models. The training was done by manually annotating 100 images per class taken from simulation environment and real world (Google Images)

using the Visual Object Tagging Tool (MS VoTT⁴). Another alternative way that was employed was looking for rectangular contours in depth image as shown in Fig. 15.



Fig. 15: Detection of opening based on rectangular contours by processing depth observation.

(d) Obstacle Avoidance: Although many robots now a days has inbuilt obstacle avoidance mechanism based on depth sensor, however, it works in a passive way, i.e. when a ‘move forward’ instruction is given and there is an obstacle in front, the robot will try to move, but its sensors will prohibit it from executing the task. Hence, alternatively, a better way is to analyse depth images and post diving the image into ‘free’ and ‘occupied spaces’ based on gray-scale threshold value, the robot is prevented from issuing non-navigable commands, thereby saving step counts.

(e) Reflective Surface Avoidance: A typical indoor house may consist of a large number of mirrors and reflective surfaces. If a robot relies on its depth sensor for finding free spaces, a mirror will give wrong depth information resulting in robot collision. Hence, two complementary strategies are taken. Firstly, a check is kept if current and past RGB image observations across robot steps are very similar using a threshold parameter. Secondly, the approach of Mirror3D [76] is adopted to identify reflective surfaces and correct depth estimates to aid robot in successful navigation.

(f) Object to PointGoal Mapper: When an object say ‘bed’ related to target region ‘bedroom’ comes in view of the robot, it needs to move towards ‘bed’ to maximize its chance of being within the target region. This is done by a mapper module, that takes the RGB-D bounding box of identified object and maps it to a point on the 2D navigable surface. This point can be

passed to a ‘PointGoal’ local planner (like ROS Nav2 Planner to plan the route avoiding obstacles.

(g) PointGoal Planner: As discussed in earlier subsection 5.1, the point to point navigation will be done by a ‘PointGoal’ planner. This module also aids in planning a path to a distant grid point, when the need is to go out of a region, and no openings are detected by aforementioned object recognizer methods.

(h) Visible Area Identifier: In the ‘ViewAreaGoal’ task, the task is said to be complete when the target area comes in view or objects related to that area comes in view. This is accomplished by seeing related objects to target region in subsequent image frames with high confidence, provided the objects are at a distance from current robot position. Sometimes, it may happen that due to occlusion or failure in object detection algorithm, the robot needs to enter the area in order for related objects to be detected. Then this problem reduces to standard ‘AreaGoal’ task.

(i) Area Identifier: In the ‘AreaGoal’ task, the robot does a full 360^0 rotation at intervals. The rotation is triggered either by (a) rotation interval counter or (b) when it is detected that robot has entered a new region (based on passing detected openings) or (c) objects dissimilar to last region is detected. The robot invokes the edge values of the Region Relation Graph for each of the objects O in the rotation view to predict the most probable region R as shown below:

$$R = \max(\sum(p(O_x)*p(R_y)*c(O_i)*p(R_y | O_x)/N)) \quad (1)$$

Here, $p(O_x)$ denotes probability of an object x based on frequency of occurrence among all objects. Similarly $p(R_y)$ denotes the probability of a region among all regions based on the frequency of occurrence in observations. $c(O_i)$ denotes confidence score of the object detector for that object class. $p(R_y | O_x)$ denotes conditional probability that robot is in one of ‘y’ regions based on the view of a single object ‘x’. While doing a 360^0 rotation, if an object with high affinity to a region comes in view with good detection confidence score, the rotation is stopped and area inferred. This saves step counts and unnecessary full rotations.

(j) Area Center Identifier: Where as the ‘AreaGoal’ task is completed when the robot enters the target area, in case of ‘AreaCenterGoal’ task, the robot needs to be close to the center of target area or its closest navigable point. Once target area is identified using aforementioned methodology (i), next the robot needs

⁴<https://github.com/microsoft/VoTT>

to navigate to a center point. This can be achieved by invoking the mapper module to identify the coarse boundary from depth observation and estimating a center point that is navigable. A simple implementation will be a heuristic based estimation to calculate a point and pass it on to ‘PointNav’ module. The point in 2D plane can be calculated as follows:

$$C = (\text{centroid}(\text{visible objects' average depth points}) + \text{centroid}(\text{coarse region map perimeters})) / 2$$

In future work, we will tackle this problem separately as it involves advances in relative pose estimation and depth sensor noise correction. We have kept this task out of scope of this paper.

(k) Backtracking to Landmark: The Visual Semantic Map as discussed in Section 5.2, helps in keeping track of unexplored places in a grid. It also marks certain places as landmarks having more than one branches (openings) and objects having high affinity with target region. In case, the robot meets a dead end or has followed a branch path, instead of exploring, it can directly invoke the ‘PointNav’ module to go to the nearest unexplored landmark grid cell.

The high-level algorithmic workflow for the Area Goal task is shown in Algorithm 1. Initially a square grid map with empty cells is initialized with 4 corner points. At first, the robot performs a rotation to identify current area, and do area prediction at intervals or when robot observes objects closely tied with target area. If there are mirrors or blockages, the robot turns around for new views. Once robot identifies that current area is not target area, it searches for openings and free space to enter newer areas. Intermittently, it moves towards objects having highest relation to target area. In case of dead end, robot backtracks to last stored landmarks (unexplored openings or unexplored object directions in the map). Finally, if landmarks are exhausted, robot tries to plan path towards the four corners of the grid map one by one, expecting that target area will lie in one of the paths. It is to be noted that grid map gets updated with new information at each observation.

5.4.2 AreaNav Experimental Results

20 scene groups of Matterport 3D (MP3D) dataset were used to devise tasks for different target regions (areas). Ground truth information of the regions were used as boundaries of areas. Visual inspection was also carried out in scene groups giving poor results to identify the underlying cause, and adapt the algorithm

Algorithm 1 Pseudocode of AreaGoal task

1: Parameters:

```

2: img ← RGB-D camera egocentric image stream;
3: actuation ← commands to robot wheels;
4: sf ← link to various software modules;
5: ta ← target area as per instruction
6: map ← generated map for storing landmarks
7: odom ← location of the robot in 2D space
8: Initialization:
9: map ← empty N*N size grid map, say N = 30m
10: corners ← four corner co-ordinates of grid map
11: steps ← 0 // step count of the actuation moves
    rotateCount ← 0; rotateFlag ← 1; taflag ← 0
12: while ta NOT reached AND steps < 500 do
13:   Wait for actuation completion;
14:   if rotateFlag == 1 AND rotateCount == 0 then
15:     do a 360° rotation to scan area
16:     area ← sf.areaPredict(img sequences)
17:     if area == ta then taflag ← 1; break;
18:   end if
19: end if
20: if sf.estimate(img) finds mirrors / block then
21:   do a 90° rotation towards {left or right}
22: end if
23: if sf.estimate(img) finds openings X then
24:   sf.PointGoal(X1 bounding box's center)
25: end if
26: if software.YOLO(img) contains objects then
27:   if probability(area | object) > 0.9 then
28:     rotateFlag ← 1; rotateCount ← 0;
29:     return (to do rotation);
30:   else sf.PointGoal(object most related to ta)
31:   end if
32: else
33:   if map.landmark is NOT exhausted then
34:     sf.PointGoal(nearest landmark)
35:   else
36:     sf.PointGoal(corners[least accessed])
37:   end if
38: end if
39: if rotateCount > 20 then rotateCount ← 0
40: end if
41: update map(odom, img); rotateCount += 1;
42: end while
43: if taflag == 1 then
44:   print 'Area $ta reached in $steps steps';
45: else
46:   print 'Task Incomplete after 500 steps';
47: end if

```

heuristics. Through ablation studies, it was found contrary to ‘ObjectGoal’ task, learnt GCN embeddings do not enhance ‘AreaGoal’ task - hence it is not used in the baseline experiments. Considering MP3D house layouts being quite large, the upper limit of step count was kept at 500, by which, if the robot is unable to reach the target area, the search terminates as failure.

As can be seen from the results Table 4, we take note of both success and step count. Success as a metric is important to tell whether a random movement can ever reach a target area in allowed steps. Next metric step count tells how quickly the specified method based navigation reached the target area. Hence, for random movement based navigation we also get success many times, but the step count is large compared to other methods. This gives us an insight, that there are certain areas which are difficult to navigate to (if at all can be navigable autonomously for a specified layout), whereas there are certain areas that are simpler to traverse to.

The proposed method was also successfully tested in real life settings of indoor home. For use in real world with noise, depth sensor distance up to 2.5 meters were considered for map building or object localization. Fig. 18 shows an example snapshot of the robot navigating to the ‘bedroom’ area after getting initialized at ‘dining room’ region.

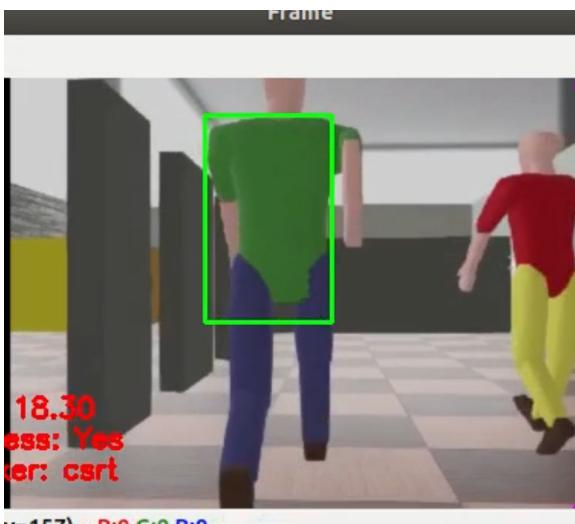


Fig. 16: Person following in Webots with 2 persons.



Fig. 17: Person following in real world office settings.

5.5 Person Following

One of the important features of an embodied AI based Telepresence robot is to follow a target person when instructed to do so. This feature becomes useful when an elderly person needs to be monitored while doing some activity. Another use case is when a caregiver needs to be followed by a robot between different wards, the doctor being remotely connected over internet. Fig. 16 shows how a person is followed based on first person ego view in Webots, a ROS compliant simulation environment. The choice of Webots was made as a test ground, as AI Habitat 2 did not allow addition of person objects with physics based trajectories. Fig. 17 depicts person following algorithm working in real world settings. The algorithm 2 for person following is enlisted below. Initially the scene is processed using YOLO object detector to detect persons; and the user is asked to tag a person who needs to be followed. The features of the person selected (bounding box) is passed on to (a) Discriminative Correlation Filter Tracker (CSRT) [77] algorithm (b) deep learning based GOTURN tracker [78] (c) one class Support Vector Machine (SVM) [79], trained by data augmentation techniques of the single snapshot of the user tagged person pixels of bounding box. The training continues in online fashion while the tracker identifies the person in subsequent frames. The training stops when tracker confidence is low or person goes out-of-view. The former two trackers help in tracking a person object from a stream of sequential image frames. However, when the person gets out of view due to speed, or occlusion and reappears, this SVM based person re-identification method helps in detecting the specific person, where the former two trackers fail. The final target bounding box is decided based on a weighted voting amongst the tracker and the SVM model. In case of person becomes out of view, the SVM model weight matter more as shown in the algorithm. Generally, for such scenarios the robot follows

Table 4: Comparison of above *AreaGoal* method with baselines for different target areas (Goal).

Target goal	Method	AreaViewGoal Task:		AreaGoal Task:	
		Success ↑	Step ↓	Success ↑	Step ↓
Bathroom or Toilet	Random	0.9	230	0.9	212
	Frontier-based [70]	0.9	128	0.9	147
	Our Method	0.95	110	0.95	122
Bedroom	Random	0.85	354	0.8	423
	Frontier-based	0.95	178	0.95	182
	Our Method	0.95	125	0.95	136
Dining room	Random	0.9	290	0.9	244
	Frontier-based	0.95	240	0.95	246
	Our Method	1.0	204	1.0	220
Study room	Random	0.5	442	0.3	489
	Frontier-based	0.7	390	0.65	430
	Our Method	0.9	280	0.85	343
Kitchen	Random	0.9	290	0.9	301
	Frontier-based	0.95	157	0.95	173
	Our Method	1.0	122	1.0	147
Living room	Random	0.6	482	0.55	497
	Frontier-based	0.85	137	0.85	143
	Our Method	0.95	110	0.95	119
Average across areas	Random	0.775	332	0.725	361
	Frontier-based	0.83	205	0.875	221
	Our Method	0.958	159	0.942	182

**Fig. 18:** Real life example of ‘AreaGoal’ task. Starting at dining room, the robot needs to navigate to bedroom.

the last trajectory move. Finally, if even then person is not found, then a 360° rotation followed by random moves is the way out. At each frame, the scene area is divided into three parts (Left, Forward, Right) and based on maximum overlap of bounding box of target person with scene area, the decision to move in which direction is taken. To avoid collision with target person, and maintaining a safe distance while

following, the bounding box dimensions of target person is tracked to see if it lies within a threshold. If the threshold is crossed, the robot stops till the person moves away from close proximity. Simultaneously, map data is also learned to avoid static obstacles in future traversals.

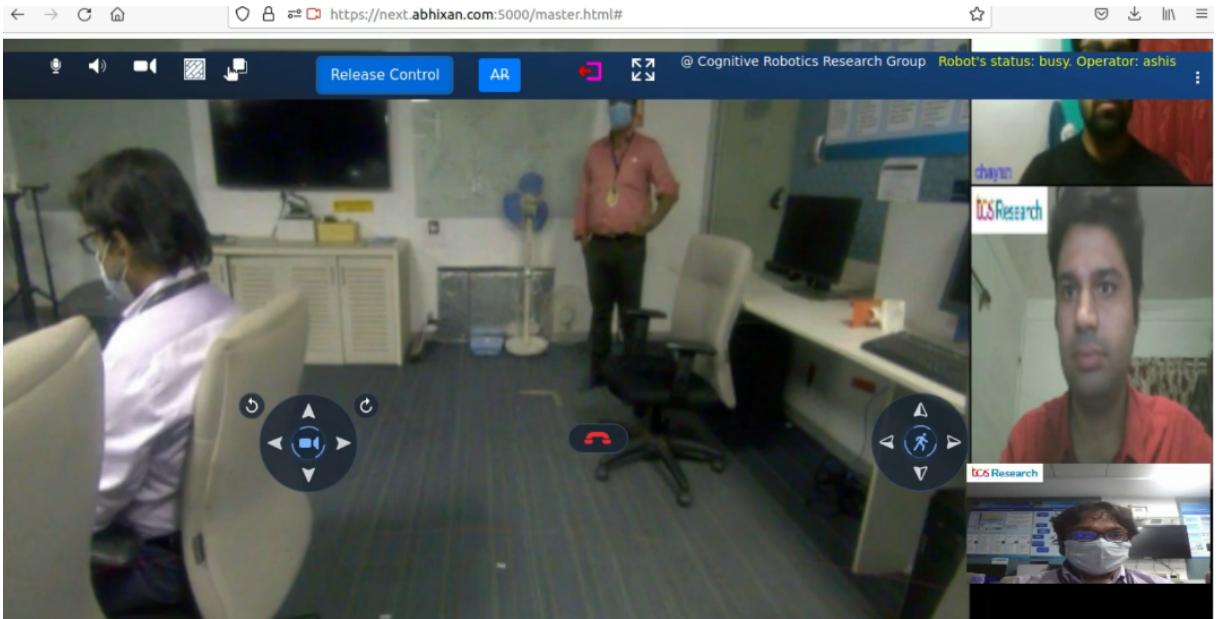


Fig. 19: Browser based user interface for remote user.

6 User Interface and User Study

To support ease of use and portability, a browser based user interface (as shown in Fig. 19) is developed. The user joins over the internet using a secured authentication process. The background video is a teleconference setup with the main video stream being of the robot's front camera.

The controls, visible as icons in the figure, from left to right are: (1) Microphone - to control the audio of remote user, (2) Speaker - to control the hardware speaker of device on which browser interface is loaded, (3) Video - to turn on or off the self camera view, (4) Map - to display pre-created map for this location (if any). This in turn will help in map based point goal navigation. (5) Map Size control - to zoom in and zoom out from the map (6) A toggle button named 'Take Control' or 'Release Control' - for the remote user to take control of the robot located in a different location. In the screenshot, user named 'Ashis' can be seen in control of the robot. (7) 'AR' toggle button to enable a user to click on the screen to create an augmented reality icon marker, where the robot should move to (8) Exit icon - to sign off from current session (9) Fullscreen - to hide controls and get full view of robot's camera feed. The central 3 controls are as follows: (1) Camera tilt icon set - to tilt or rotate the camera (2) Hangup button - to end the session of the teleconferencing call (3) Robot control - to

move the robot in all four directions, with center being the 'Stop' command. Similar to existing real-time conferencing software, the right hand side is dedicated to display other remote users logged into the session.

On a study conducted over 30 users across different demographics and geographic locations, the system was found to be responsive and easy to use. The study results are listed in Fig. 20. The 30 remote users were located at different cities across the world and they were asked to login from their device browsers to test the features of the Teledrive solution and rate their experience in survey questionnaire on a scale of 1 to 5 (5 being excellent). The response time for each feature was calculated based on a user clicking a button or giving a command on the remote user side and the corresponding trigger in the robot and is listed in the above figure – this was calculated without user intervention. The user study tested the following features: (a) Audio-Video Conferencing over WebRTC (b) Manual Navigation (c) Map based Navigation (d) Pin Goal Navigation (dropping a pin) (e) Speech based HRI to move to some location like object for Object-Goal and area for AreaGoal. The user was given a option to select the relative importance of a feature. The feature-wise collective Mean Opinion Score

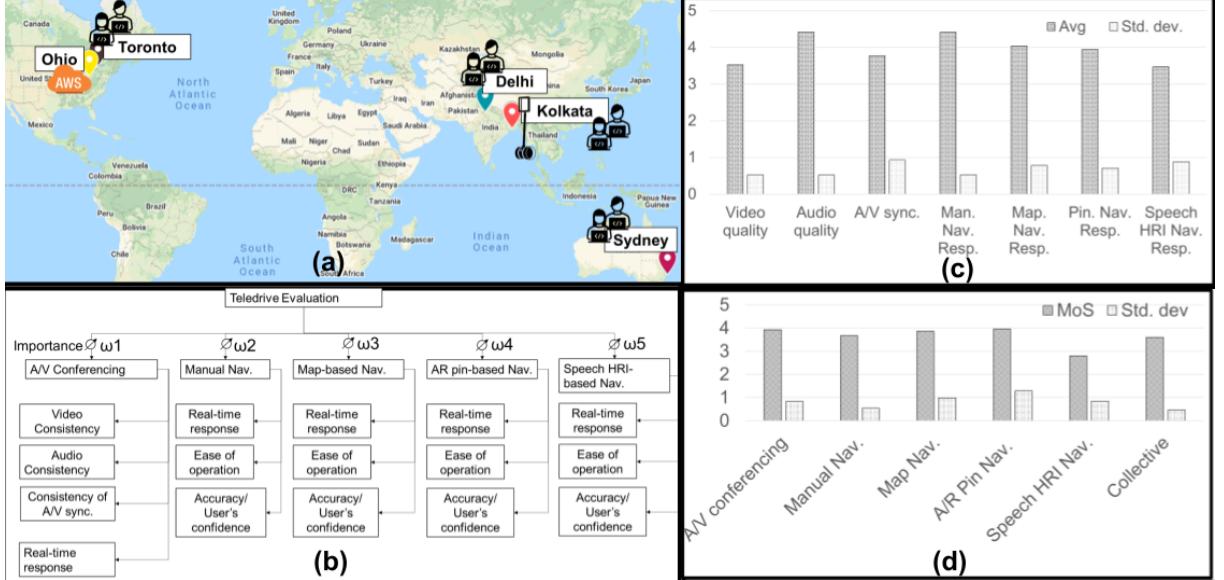


Fig. 20: (a) Varied remote location of users (b) Survey questionnaire on which ratings were given (c) Response time feedback for the surveyed features (d) Quality of user experience for the surveyed features.

(MOS) computed as:

$$M_k = \frac{\sum_{j=1}^N \frac{w_{kj} \sum_{i=1}^{A_k} a_{ij}}{A_k}}{\sum_{j=1}^N w_{kj}} \quad (2)$$

Here, M_k = MOS for the kth feature; a_i = rating of the i-th attribute of the k-th feature by j-th user; A_k = total number of attributes in k-th feature; w_{kj} = importance of k-th feature for the j-th user; N = total number of users.

The robot side interface is similar to Master (remote) side except lack of controls which are unnecessary. The robot side interface is needed in order for the robot side user (say a patient or elderly person) to see the remote users (say a team of doctors) to perform some basic operations like audio-video control on the robot side. As a security feature, the Master operator has no direct control on the edge server where ROS 2 is running with software modules as ROS nodes. It is through the frontend communication channeling that Master can invoke the software feature executions.

7 Conclusion

In this article, we presented an architecture for embodied AI tasks targeted toward a telepresence setup, especially in patient care scenarios. We have described approaches of individual embodied AI modules with a

specific focus on the ‘AreaGoal’ and ‘person following’ problem. We also back our claims with relevant results. In future work, we expect to adjoin additional embodied AI modules for the tasks of language-grounded navigation, contextual scene understanding, gesture-based navigation, etc. Apart from presented use cases, this work will be expanded for deployment in office conference scenarios, retail stores, and factory workshops requiring supervision tasks.

References

- [1] Lal, A. *et al.* Pandemic preparedness and response: exploring the role of universal health coverage within the global health security architecture. *The Lancet Global Health* **10** (11), e1675–e1683 (2022).
- [2] Pramanick, P. *et al.* Enabling human-like task identification from natural conversation, 6196–6203 (IEEE, 2019).
- [3] Gupta, P., Bhowmick, B. & Pal, A. Accurate heart-rate estimation from face videos using quality-based fusion, 4132–4136 (2017).
- [4] Gupta, P., Bhowmick, B. & Pal, A. Exploring the feasibility of face video based instantaneous heart-rate for micro-expression spotting (2018).
- [5] Bhowmick, B., Patra, S., Chatterjee, A., Madhav Govindu, V. & Banerjee, S. Divide and

Algorithm 2 Pseudocode of Person Following Robot

```
1: Parameters:
2: image ← RGB camera egocentric image stream;
3: actuation ← commands to robot wheels;
4: software ← link to software modules;
5: target ← target person bounding box to follow;
6: sa ← area of scene (eg. Left, Middle, Right);
7: Initialization:
8: persons ← bounding boxes from
   software.YOLO(image);
9: target ← select a person's bbox from view;
10: features ← software.extract(target);
11: tracker1 ← software.GOTURN(features);
12: tracker2 ← software.CSRT(features);
13: model ← software.OneClassSVM(features);
14: lastMove ← Stop; weight ← 1/3;
15: while No command to stop Person Follow task do
16:   Wait for actuation completion;
17:   if software.YOLO(image) contains persons
    then
18:     a ← software.detect(persons, tracker1);
19:     b ← software.detect(persons, tracker2);
20:     c ← software.detect(persons, model);
21:     target ← voting( a/3, b/3, weight * c );
22:     weight ← 1/3;
23:     bbox ← bounding box dimension of target
24:     if bbox ≥ threshold t then
25:       actuation ← STOP;
26:     else
27:       m ← max. overlap of target with sa
28:       actuation ← sa[m];
29:       lastMove ← actuation;
30:       model ← software.update(target);
31:     end if
32:   else
33:     actuation ← lastMove; weight ← 1;
34:   end if
35: end while
```

- conquer: A hierarchical approach to large-scale structure-from-motion. *Computer Vision and Image Understanding* **157**, 190–205 (2017).
- [6] Chang, A. *et al.* Matterport3d: Learning from rgb-d data in indoor environments. *arXiv preprint arXiv:1709.06158* (2017).
 - [7] Straub, J. *et al.* The replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797* (2019).
 - [8] Xia, F. *et al.* Gibson env: Real-world perception for embodied agents, 9068–9079 (2018).

- [9] Savva, M. *et al.* Habitat: A platform for embodied ai research, 9339–9347 (2019).
- [10] Sriram, N. N. *et al.* Talk to the vehicle: Language conditioned autonomous navigation of self driving cars, 5284–5290 (2019).
- [11] Pramanick, P., Sarkar, C., Banerjee, S. & Bhowmick, B. Talk-to-resolve: Combining scene understanding and spatial dialogue to resolve granular task ambiguity for a collocated robot. *Robotics and Autonomous Systems* **155**, 104183 (2022).
- [12] Pramanick, P., Sarkar, C., Paul, S., dev Roychoudhury, R. & Bhowmick, B. Doro: Disambiguation of referred object for embodied agents. *IEEE Robotics and Automation Letters* **7** (4), 10826–10833 (2022).
- [13] Anderson, P. *et al.* On evaluation of embodied navigation agents. *arXiv preprint arXiv:1807.07567* (2018).
- [14] Macenski, S., Foote, T., Gerkey, B., Lalancette, C. & Woodall, W. Robot operating system 2: Design, architecture, and uses in the wild. *Science Robotics* **7** (66) (2022).
- [15] Coşar, S. *et al.* Enrichme: Perception and interaction of an assistive robot for the elderly at home. *International Journal of Social Robotics* **12** (3), 779–805 (2020).
- [16] Amy Robot. <https://www.amyrobotics.com/indexproducten>. [Online; accessed 12-Sep-2022].
- [17] Wu, X., Thomas, R., Drobina, E., Mitzner, T. & Beer, J. An evaluation of a telepresence robot: User testing among older adults with mobility impairment (2017).
- [18] Lewis, T., Drury, J. & Beltz, B. Evaluating mobile remote presence (mrp) robots, 302–305 (2014).
- [19] PadBot Robot. <https://www.padbot.com/>. [Online; accessed 12-Sep-2022].
- [20] Tsui, K. M., Desai, M., Yanco, H. A. & Uhlik, C. Exploring use cases for telepresence robots, 11–18 (IEEE, 2011).
- [21] Ohmni Robot. <https://ohmnilabs.com/products/ohmni-telepresence-robot/>. [Online; accessed 12-Sep-2022].
- [22] Boteyes Robot. <https://boteyes.com/>. [Online; accessed 12-Sep-2022].
- [23] Orlandini, A. *et al.* Excite project: A review of forty-two months of robotic telepresence technology evolution. *Presence: Teleoperators and Virtual Environments* **25** (3), 204–221 (2016).

- [24] Beam Pro Robot. <https://telepresencerobots.com/robots/suitable-technologies-beam-pro/>. [Online; accessed 12-Sep-2022].
- [25] Lutz, C. & Tamò, A. Privacy and healthcare robots—an ant analysis. *We Robot* (2016) .
- [26] Hung, C.-F., Lin, Y., Ciou, H.-J., Wang, W.-Y. & Chiang, H.-H. *Foodtemi: The ai-oriented catering service robot*, 1–2 (IEEE, 2021).
- [27] Melendez-Fernandez, F., Galindo, C. & Gonzalez-Jimenez, J. A web-based solution for robotic telepresence. *International Journal of Advanced Robotic Systems* **14** (6), 1729881417743738 (2017) .
- [28] Tuli, T. B., Terefe, T. O. & Rashid, M. M. U. Telepresence mobile robots design and control for social interaction. *International Journal of Social Robotics* 1–10 (2020) .
- [29] Soares, N., Kay, J. C. & Craven, G. Mobile robotic telepresence solutions for the education of hospitalized children. *Perspectives in health information management* **14** (Fall) (2017) .
- [30] Herring, S. C. Telepresence robots for academics. *Proceedings of the American Society for Information Science and Technology* **50** (1), 1–4 (2013) .
- [31] Ng, M. K. *et al.* A cloud robotics system for telepresence enabling mobility impaired people to enjoy the whole museum experience **2015** (10th), 1–6 (2015) .
- [32] Michaud, F. *et al.* Telepresence robot for home care assistance. 50–55 (2007) .
- [33] Tan, Q. *et al.* Toward a telepresence robot empowered smart lab. *Smart Learning Environments* **6** (1), 1–19 (2019) .
- [34] Cesta, A., Cortellessa, G., Orlandini, A. & Tiberio, L. Long-term evaluation of a telepresence robot for the elderly: methodology and ecological case study. *International Journal of Social Robotics* **8** (3), 421–441 (2016) .
- [35] Monroy, J., Melendez-Fernandez, F., Gongora, A. & Gonzalez-Jimenez, J. *Integrating olfaction in a robotic telepresence loop*, 1012–1017 (IEEE, 2017) .
- [36] Beno, M. *Work flexibility, telepresence in the office for remote workers: A case study from austria*, 19–31 (Springer, 2018) .
- [37] Kristoffersson, A., Coradeschi, S. & Loutfi, A. A review of mobile robotic telepresence. *Advances in Human-Computer Interaction* **2013** (2013) .
- [38] Duan, J., Yu, S., Tan, H. L., Zhu, H. & Tan, C. A survey of embodied ai: From simulators to research tasks. *IEEE Transactions on Emerging Topics in Computational Intelligence* (2022) .
- [39] Tsui, K. M. *et al.* Accessible human-robot interaction for telepresence robots: A case study. *Paladyn, Journal of Behavioral Robotics* **6** (1) (2015) .
- [40] Deuerlein, C., Langer, M., Seßner, J., Heß, P. & Franke, J. Human-robot-interaction using cloud-based speech recognition systems. *Procedia CIRP* **97**, 130–135 (2021) .
- [41] Batinovic, A., Petrovic, T., Ivanovic, A., Petric, F. & Bogdan, S. A multi-resolution frontier-based planner for autonomous 3d exploration. *IEEE Robotics and Automation Letters* **6** (3), 4528–4535 (2021) .
- [42] Chattopadhyay, P., Hoffman, J., Mottaghi, R. & Kembhavi, A. *Robustnav: Towards benchmarking robustness in embodied navigation*, 15691–15700 (2021) .
- [43] Cosgun, A., Florencio, D. A. & Christensen, H. I. *Autonomous person following for telepresence robots*, 4335–4342 (2013) .
- [44] Cheng, X., Jia, Y., Su, J. & Wu, Y. *Person-following for telepresence robots using web cameras*, 2096–2101 (IEEE, 2019) .
- [45] Yang, W., Wang, X., Farhadi, A., Gupta, A. & Mottaghi, R. Visual semantic navigation using scene priors. *arXiv preprint arXiv:1810.06543* (2018) .
- [46] Tatiya, G. *et al.* Knowledge-driven scene priors for semantic audio-visual embodied navigation (2021) .
- [47] Chaplot, D. S., Gandhi, D. P., Gupta, A. & Salakhutdinov, R. R. Object goal navigation using goal-oriented semantic exploration. *Advances in Neural Information Processing Systems* **33** (2020) .
- [48] Liu, A.-A. *et al.* Toward region-aware attention learning for scene graph generation. *IEEE Transactions on Neural Networks and Learning Systems* (2021) .
- [49] Kumar, A., Gupta, S. & Malik, J. *Learning navigation subroutines from egocentric videos*, 617–626 (PMLR, 2020) .
- [50] Bhattacharyya, A. *et al.* *Teledrive: An intelligent telepresence solution for “collaborative multi-presence” through a telerobot*, 433–435 (IEEE, 2022) .
- [51] Sau, A., Bhattacharyya, A. & Ganguly, M. *Teledrive: a multi-master hybrid mobile telerobotics system with federated avatar control*,

- 102–114 (Springer, 2021).
- [52] Sau, A., Bhattacharyya, A., Ganguly, M. & Mahato, S. K. *An edge-inclusive webrtc-based framework to enable embodied visual analytics in telerobot*, 228–230 (IEEE, 2023).
- [53] Sarkar, C., Mitra, A., Pramanick, P. & Nayak, T. *tagE: Enabling an embodied agent to understand human instructions*, 8846–8857 (Association for Computational Linguistics, Singapore, 2023).
- [54] Pramanick, P., Sarkar, C. & Bhattacharya, I. *Your instruction may be crisp, but not clear to me!*, 1–8 (IEEE, 2019).
- [55] Pramanick, P. & Sarkar, C. *Can visual context improve automatic speech recognition for an embodied agent?*, 1946–1957 (Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022).
- [56] Pramanick, P. & Sarkar, C. *Utilizing prior knowledge to improve automatic speech recognition in human-robot interactive scenarios*, HRI ’23, 471–475 (Association for Computing Machinery, New York, NY, USA, 2023).
- [57] Schneider, S., Baevski, A., Collobert, R. & Auli, M. wav2vec: Unsupervised pre-training for speech recognition. *arXiv preprint arXiv:1904.05862* (2019).
- [58] Ganguly, M., Mahato, S., Sau, A. & Bhattacharyya, A. *Sensing best-connected future path for mobile telerobot: A radio-source location agnostic approach*, 524–532 (IEEE, 2024).
- [59] Patra, S., Bhowmick, B., Banerjee, S. & Kalra, P. *High resolution point cloud generation from kinect and HD cameras using graph cut*, 311–316 (SciTePress, 2012).
- [60] Clark, C. Boteyes Robot. <https://www.hmc.edu/lair/ARW/ARW-Lecture01-Odometry.pdf>. [Online; accessed 12-Sep-2022].
- [61] Chaplot, D. S., Gandhi, D., Gupta, S., Gupta, A. & Salakhutdinov, R. Learning to explore using active neural slam. *arXiv preprint arXiv:2004.05155* (2020).
- [62] Russell, S. J. & Norvig, P. *Artificial Intelligence: a modern approach* 3 edn (Pearson, 2009).
- [63] Abhishek Kadian* *et al.* *Sim2Real Predictivity: Does Evaluation in Simulation Predict Real-World Performance?*, Vol. 5, 6670–6677 (2020).
- [64] Xia, F. *et al.* *Gibson env: real-world perception for embodied agents* (IEEE, 2018).
- [65] <https://www.doublerobotics.com/>.
- [66] Taketomi, T., Uchiyama, H. & Ikeda, S. Visual slam algorithms: a survey from 2010 to 2016. *IPSJ Transactions on Computer Vision and Applications* **9** (1), 1–11 (2017).
- [67] Mukherjee, D., Banerjee, S. & Misra, P. *Towards efficient stream reasoning*, 735–738 (Springer, 2013).
- [68] Banerjee, S. & Mukherjee, D. System and method for executing a sparql query (2018). US Patent 9,898,502.
- [69] Jiao, L. *et al.* A survey of deep learning-based object detection. *IEEE access* **7**, 128837–128868 (2019).
- [70] Yamauchi, B. *A frontier-based approach for autonomous exploration*, 146–151 (IEEE, 1997).
- [71] Krishna, R. *et al.* Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision* **123** (1), 32–73 (2017).
- [72] Kipf, T. N. & Welling, M. Semi-supervised classification with graph convolutional networks. *preprint arXiv:1609.02907* (2016).
- [73] AI Habitat’s Shorest Path Follower. https://aihabitat.org/docs/habitat-sim/habitat_sim.nav.GreedyGeodesicFollower.html. [Online; accessed 12-Sep-2022].
- [74] Lin, T.-Y. *et al.* *Microsoft coco: Common objects in context*, 740–755 (Springer, 2014).
- [75] Bharati, P. & Pramanik, A. in *Deep learning techniques—r-cnn to mask r-cnn: a survey* 657–668 (Springer, 2020).
- [76] Tan, J., Lin, W., Chang, A. X. & Savva, M. *Mirror3d: Depth refinement for mirror surfaces*, 15990–15999 (2021).
- [77] Farkhodov, K., Lee, S.-H. & Kwon, K.-R. *Object tracking using csrt tracker and rcnn.*, 209–212 (2020).
- [78] Held, D., Thrun, S. & Savarese, S. *Learning to track at 100 fps with deep regression networks*, 749–765 (Springer, 2016).
- [79] Cyganek, B. *Framework for object tracking with support vector machines, structural tensor and the mean shift method*, 399–408 (Springer, 2009).

8 Statements and Declarations

8.1 Acknowledgements

Thanks to Dr. Balamuralidhar Purushothaman of TCS Research for guidance and motivation for this work.

8.2 Funding

The authors have received research support from Tata Consultancy Services Limited, India.

8.3 Code and Data Availability

Code is IP protected, and data on which experiments are done is available for download from the link: <https://github.com/facebookresearch/habitat-lab#data>.

8.4 Author Contributions

Conceptualization: [Snehasis Banerjee, Abhijan Bhattacharya]; Methodology: [Snehasis Banerjee, Sayan Paul, Ashis Sau, Pradip Pramanick]; Writing and Content: [Snehasis Banerjee, Abhijan Bhattacharya, Rudradaev Roychowdhury, Chayan Sarkar]; Supervision: Brojeshwar Bhowmick.

8.5 Ethics Approval

Not Applicable

8.6 Consent to Participate

Informed consent was obtained from all individual participants included in the user study.

8.7 Consent to Publish

User study data is anonymized and not published in this work.

9 Appendix

The embodied navigation metrics are presented below.

1. Success : It is the ratio of the successful episode to total number of episodes. An episode is successful if the agent is at a distance ≤ 1.0 m from the target object at the end of episode.

2. SPL (Success weighted by path length) : It measures the efficiency of path taken by agent as compared with the optimal path. This is computed as follows:

$$SPL = \frac{1}{N} \sum_{i=1}^N S_i \cdot \frac{l_i}{\max(p_i, l_i)} \quad (3)$$

where N is the number of test episodes, S_i is a binary indicator of success, l_i is the length of the shortest path to the closest instance of the goal from the agent's starting position and p_i is the length of the actual

path traversed by the agent. SPL ranges from 0 to 1. Higher SPL indicates better model performance for the following reasons,

- High SPL signifies that the agent trajectory length is comparable to the shortest path from the agent's starting position to the goal.
- SPL also suggests that the agent has reached the closest instance of the target goal category 't' from its starting position.

3. SoftSPL: One of the shortcomings of SPL is that it treats all failure episodes equally. This is addressed in the SoftSPL metric by replacing S_i , the binary indicator of success (whether goal reached or not as 0 or 1), by $episode_progress$, a continuous indicator. This $episode_progress$ ranges from 0 to 1 depending on how close the agent was to the goal object at episode termination.

$$SoftSPL = \frac{1}{N} \sum_{i=1}^N \underbrace{\left(1 - \frac{d_i}{\max(l_i, d_i)}\right)}_{episode_progress} \cdot \left(\frac{l_i}{\max(p_i, l_i)}\right) \quad (4)$$

where N is the number of test episodes, l_i is the length of the shortest path to the closest instance of the goal from the agent's starting position, p_i is the length of the actual path traversed by agent and d_i is the length of the shortest path to the goal from the agent's position at episode termination. Similar to SPL, higher SPL indicates better model performance.

4. Distance to Success (DTS) : It signifies the distance between the agent and the permissible distance to target for success at the end of a search episode.

$$DTS = \max((\|x_T - G\|_2 - d), 0) \quad (5)$$

where x_T is the L2 distance of the agent from the Goal at the end of the episode, d is the success threshold. A lower DTS value at episode termination indicates that the agent is closer to the goal category. Therefore, a lower DTS value indicates better performance.

5. Step count: In the AI Habitat environment, the robot action space consists of three actions - turn right by 30 degrees, turn left by 30 degrees, and move forward by 0.25 metres. Each such action is counted as one step. A single rotation is also considered as a step. A full 360^0 rotation equates to 12 total steps. The less the number of steps taken, the better is the method. Hence, step count is considered as a metric.