Small Object Navigation with Context Information

Jiaming Wang[1], **Harold Soh**[1,2]

[1] Dept. of Computer Science, National University of Singapore
[2] Smart Systems Institute, National University of Singapore
jamie-w@nus.edu.sg, harold@comp.nus.edu.sg

Abstract

We introduce a novel framework designed to effectively address the object goal navigation task, specifically focusing on smaller daily household objects such as bowls or mugs. These diminutive objects present challenges to existing SLAM-based semantic mapping methods, because the object detection module employed in the mapping pipeline struggles to accurately detect them. To address this limitation, we propose to use a probabilistic semantic map. It is updated by a trained mapping module that incorporates contextual information to estimate the likelihood of object presence within the agent's current field of view. Subsequently, this probabilistic map guides the agent towards more promising areas for object search. Our experimental results demonstrate that the proposed method outperforms a strong baseline by 26% in Small Object Navigation tasks.

Introduction

Efficient navigation and object search in unfamiliar environments are crucial objectives for the development of intelligent service robots. State-of-the-art methods for addressing this challenge commonly employ SLAM-like pipelines, in which a semantic map is constructed and updated as the agent explores the surroundings. To update the semantic map, the agent first constructs a 3D point cloud from the RGBD observation, where each point is labeled by a pre-trained object detection module such as MaskRCNN (He et al. 2017). The point cloud is subsequently projected onto a 2D top-down map by pooling along the height axis. Based on the current semantic map, a high-level policy, either analytic or learned, is responsible for selecting the next target location. Simultaneously, a low-level policy governs the agent's actions towards achieving the high-level objectives (Chaplot et al. 2020a,b; Blukis et al. 2022; Ramakrishnan et al. 2022).

However, the success of this process heavily depends on the accurate detection of object classes by the pre-trained object detection module. If the object detection fails to identify the object, the entire process can fail. Searching for small objects presents a particular challenge due to the presence of noisy sensors and limited resolution, which can hinder the object detection module's ability to accurately detect or recognize the object. Consequently, prior approaches have been observed to perform inadequately on smaller objects as demonstrated in our experiment and other prior works (Min et al. 2021).

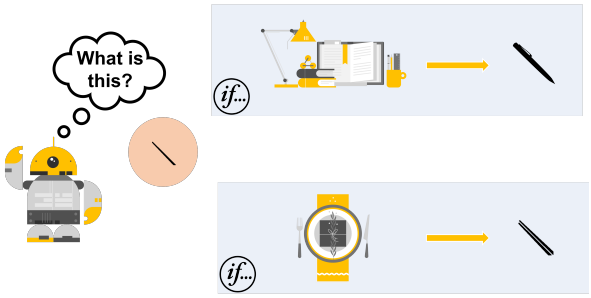

Figure 1: This work explores using such contextual information during object goal navigation. Humans can effectively use context information for object recognition. For example, a small elongated object is more likely a pen if it is on an office desk. Alternatively, if it was on a kitchen table, it is more likely a utensil, e.g., a pair of chopsticks.

In contrast, humans and animals have demonstrated remarkable proficiency in detecting objects even with limited visual cues. We typically utilize contextual information to aid the detection of a small object: the elongated object is more likely to be a pen if it is on a study desk, or a chopstick if it is on a dining table (Figure 1). Prior studies in the field of neuroscience (Park et al. 2011; Bar 2004) highlight the importance of incorporating contextual information in the design of algorithms for object recognition and detection, especially in cases where the objects in question are small and difficult to detect. This idea has previously been explored in the computer vision community (Liu et al. 2021; Tong, Wu, and Zhou 2020). However, there have been few research works addressing the small object navigation task as a whole, where successful execution of the task not only requires a better object detection model but also a unified system that allows the agent to efficiently reason about uncertainty in the environment.

Our framework employs an end-to-end trained neural network to learn object associations - for example, mugs

are typically found on tables - directly from visual inputs (RGBD). Our experiment demonstrates that the learned system can infer the presence of small objects by utilizing contextual information, i.e. surrounding large objects, even when the agent is far from the object that a pre-trained object detection module failed to detect. The acquired mapping module can be used to guide the agent towards areas that are more likely to contain the specified object. Our proposed method substantially outperforms a strong baseline in the small object navigation (SONav) task, as evidenced by a **26% absolute increase** in success rate.

Small Object Navigation (SONav) Task

We adopt a similar definition as Object Goal Task (Chaplot et al. 2020b; Wani et al. 2020). In this task, the agent is required to explore an environment that is initially unknown and locate one or more objects based on their specified object class names, such as TV. At each step, the agent receives egocentric RGB and depth images, as well as GPS and compass readings. The action space comprises four discrete actions: MOVE FORWARD, TURN LEFT, TURN RIGHT, and STOP. An episode is considered successful when the agent declares the object as found using the STOP action when it is within a certain threshold distance d_s from the goal object. The episode terminates after a fixed number of timesteps.

A significant distinction between the task we consider in this work and prior studies lies in the fact that we focus on *small* daily objects commonly found in households, namely 'MUG', 'BOWL', 'SNACK BOX', 'BOOK', 'FRYPAN', and 'EARPHONE'. In contrast, previous works predominantly focus on larger goal objects. For instance, in (Chaplot et al. 2020b), the objective is to find objects such as 'CHAIR', 'COUCH', 'PLANT', 'BED', 'TOILET', and 'TV'. Due to imperfect perception and the small size of these daily household objects, our task presents a greater challenge for the agent.

Small Object Navigation with Context Information

System Overview

Figure 2 provides an overview of our proposed system. The system operates in the following manner: starting with the current RGBD observation, the probabilistic mapping module predicts the probability of different objects' existence, by levering both visual information and contextual information as described in the following section. This module produces a top-down ego-centric map, a $C \times M \times M$ tensor, where C represents the total number of object classes in the environment, and M denotes the agent's vision range. Each entry in this tensor reflects the agent's belief regarding the presence of a specific object in the corresponding cell of the physical world (5cm $\times$5cm). The local ego-centric map is then utilized to update the global probabilistic map m_p, which is represented by a $C \times N \times N$ tensor, using differentiable spatial transformations. It is worth noting that we update the probabilistic map in the logit space, allowing for more efficient Bayesian updates, as discussed in the next section.

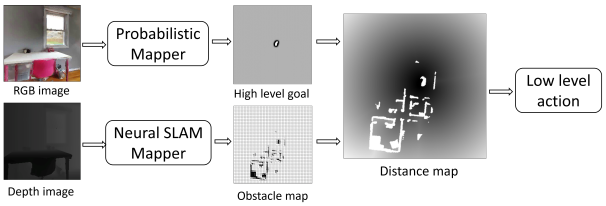

Figure 2: Overview of the proposed system.

Concurrently, we incorporate the neural SLAM module (Chaplot et al. 2020a) to predict the local obstacle map based on the depth image. This prediction is used to update the global obstacle map m_o. Subsequently, the updated probabilistic map m_p is utilized by the agent to select the high-level goal, while the obstacle map m_o is employed to calculate the goal distance map. At the beginning of each episode when the probabilistic map is empty, we use a frontier-based (Yamauchi 1997) exploration policy to explore the environment. Finally, the next low-level discrete action can be determined using the distance map.

Updating the Probabilistic Map

The probabilistic map incorporates sequences of observations by performing approximate Bayesian updates (Persson et al. 2007):

$$p(e_{i,c}|o^T, o^{T-1:1}) = \frac{p(e_{i,c}|o^{T-1:1})}{1-p(e_{i,c}|o^{T-1:1})} \times \frac{p(e_{i,c}|o^T)}{1-p(e_{i,c}|o^T)} \times \frac{1-p(e_{i,c})}{p(e_{i,c})} \times (1-p(e_{i,c}|o^T, o^{T-1:1})) \quad (1)$$

where $p(e_{i,c}|o^T)$ is the probability that there exists a specific object c in the cell i given the observation o at timestep T, and $p(e)$ is the prior probability. The equation can be simplified in the logit space as:

$$L(e|o^T, o^{T-1:1}) = L(e|o^T) + L(e|o^{T-1:1}) - L(e) \quad (2)$$

where $L(\cdot)$ denotes the logit of $p(\cdot)$

Probabilistic Mapping Module

The architecture of the probabilistic mapping module is depicted in Figure 3. We utilize the initial 8 layers of the pre-trained Resnet50 model (He et al. 2016) as the backbone for our visual encoder. Since the original Resnet50 model was not trained with the depth channel, we further train a convolutional neural network specifically for processing the depth image. Subsequently, we concatenate the depth features with the processed RGB features along the channel dimension.

To obtain the latent context vector, we crop the current semantic map to generate a fixed-size (128 $\times$ 128) local map with the agent positioned at the center. This local map was then processed by a convolutional neural network to produce a latent context vector z. We combined it with the output of the visual encoder as input for the decoder. The decoder first mix the information from both the visual encoder and the context encoder using an MLP. Then it decodes the processed information using a transposed convolutional neural network to generate the predicted ego-centric map.

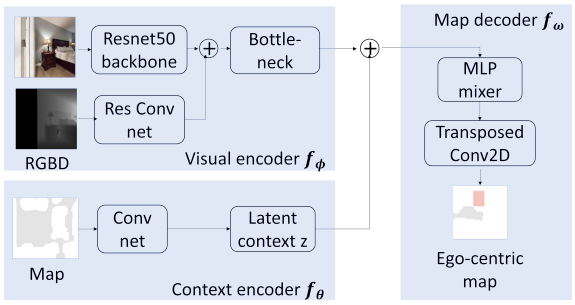

Figure 3: The probabilistic mapping module."+" means concatenation.

Method	Success ↑	SPL ↑	DTS ↓
BIG OBJECT GOAL			
Random	0.01	0.01	5.93
Neural SLAM	0.72	0.479	3.998
Ours	**0.73**	**0.492**	**3.954**
SMALL OBJECT GOAL			
Random	0	0	5.23
Neural SLAM	0.43	0.186	1.957
Ours	**0.69**	**0.382**	**0.294**

Table 1: Experiment results.

We train the proposed probabilistic mapping module by maximizing the observed data log-likelihood: $\theta, \phi, \omega = \arg\max_{\theta,\phi,\omega} \log p(s|o,z)$, where θ, ϕ, ω are the parameters of the context encoder, visual encoder and map decoder respectively; s is the ground truth semantic map that is only observed during training, o is the RGBD observation and z is the encoded latent context.

We employ the binary cross entropy (BCE) loss function to optimize the aforementioned objective. However, naively using the BCE loss leads to the neural network disregarding smaller objects. This occurs because the final loss is computed by averaging across all C class objects. Consequently, smaller objects that occupy fewer cells receive fewer training signals, gradually becoming overshadowed by larger objects. To mitigate this issue, we propose normalizing the BCE loss across the channel dimension (C) based on the size of the object depicted in the local map.

The normalized BCE loss for object class C, denoted as $\mathcal{L}^C$, is defined as follows:

$$\mathcal{L}^C = \left(\sum_i y_i^C\right)^{-1} \sum_i y_i^C \log \hat{y}_i^C + \beta \sum_i (1 - y_i^C) \log(1 - \hat{y}_i^C)$$

where $y_i^C \in 0, 1$ represents the ground truth label for object class C in cell i, $\hat{y}_i^C$ denotes the predicted value, and β is a hyperparameter that regulates the degree of penalty imposed on false positive predictions. By adopting this approach, each object class receives training signals within comparable ranges, thereby enabling the neural network to equally learn how to predict each object.

Priority Sampling

To achieve better sample efficiency during training, we use a priority-sampling-like technique to encourage the agent to learn more from its past errors. To do that, we discretize the agent's configuration space, which consists of the agent's pose(x,y) and the agent's facing direction(θ), into grids. Then we use a reply buffer to store the BCE loss calculated at each grid.

To train the probabilistic mapper, we first sample some agent's configurations according to a distribution define below, then we render the observations at the sampled configurations by the simulator. The sampling distribution is defined as: $p(x, y, \theta) = \frac{1}{Z}\mathcal{L}_{x,y,\theta}^{BCE} \times \gamma^{N_{x,y,\theta}}$, where $p(x, y, \theta)$ is the probability of selecting that specific configuration (x, y, θ), Z is the normalization constant, $\mathcal{L}_{x,y,\theta}^{BCE}$ is the BCE loss at that configuration, γ is some scaling constant that is less than 1, and N is the number of visits at that configuration.

Experimental Setup

We evaluate the proposed method by conducting experiments on the iGibson scene dataset (Xia et al. 2018) using the Habitat simulator (Szot et al. 2021). The original iGibson scene dataset lacked diversity in terms of small objects and their variations. To address this limitation, we enhance the 3D scenes in the iGibson dataset by introducing randomly selected common household objects, such as books, bowls, mugs, etc. Moreover, we follow commonsense knowledge to determine appropriate receptacles for each object. For instance, bowls are commonly placed on kitchen countertops or dining tables in dining rooms. For the purpose of training, we generated 8000 unique episodes using eight distinct house layouts. Additionally, for testing, we generated 200 episodes using two novel layouts.

Results

We compare our approach with a baseline method using neural SLAM (Chaplot et al. 2020b) with frontier-based exploration (FBE) (Yamauchi 1997) as the high-level exploration policy. It is worth noting that the winner of CVPR 2022 Object Goal Navigation Challenge used a combination of the neural SLAM approach with a trained exploration policy. Learning a more efficient exploration policy is beyond the scope of this research, as we focus on the probabilistic semantic mapping that leverages contextual information. We test both methods using the same FBE exploration policy.

We evaluate the performance of our method and the baseline using three metrics:

- **Success**: the number of successful episodes divided by the total number of evaluated episodes;
- **SPL**:success weighted by path length. This metric measures the efficiency of the agent to locate the goal object;
- **DTS**: distance to success. This is the distance between the agent and the goal object minus the success threshold when the episode ends.

The experimental results (Table 1) suggest that searching for small objects presents significant challenges, as indicated

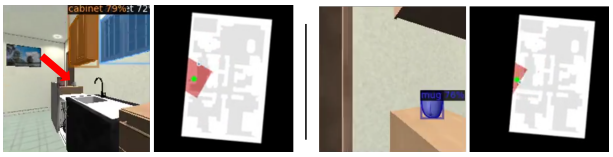

Figure 4: Qualitative result of the proposed method. Left: the object detection module was unable to detect the goal object (the mug, as indicated by the red arrow) when the agent is far from the object. Right: our proposed probabilistic map learns that the mug can be usually found on a cabinet, thus leads the agent to take a closer look at the cabinet and successfully find it.

by the decline in performance for both methods. Our proposed method exhibits a strong improvement in both success rate and SPL compared to the neural SLAM baseline, effectively narrowing the gap in small object search tasks. Notably, the substantial increase in SPL (+105%) demonstrates the efficiency of our method in locating the goal object, attributed to the guidance provided by the proposed probabilistic semantic map. Furthermore, we observe that the trained probabilistic mapper can learn to predict the presence of small objects based on contextual cues when the object detection module fails to detect them (Figure 4).

Conclusion

In this paper, we presented a probabilistic method for the SONav task where the key idea is to leverage contextual information. Our experiment in the Habitat simulation environment suggests our proposed method substantially outperforms a recent SLAM-based method on the SONav task.

Acknowledgements

This research is supported by the National Research Foundation, Singapore under its Medium Sized Center for Advanced Robotics Technology Innovation.

References

Bar, M. 2004. Visual objects in context. *Nature Reviews Neuroscience*, 5(8): 617–629.

Blukis, V.; Paxton, C.; Fox, D.; Garg, A.; and Artzi, Y. 2022. A Persistent Spatial Semantic Representation for High-level Natural Language Instruction Execution. In *Proceedings of the 5th Conference on Robot Learning*, 706–717. PMLR. ISSN: 2640-3498.

Chaplot, D. S.; Gandhi, D.; Gupta, S.; Gupta, A.; and Salakhutdinov, R. 2020a. Learning To Explore Using Active Neural SLAM. In *International Conference on Learning Representations (ICLR)*.

Chaplot, D. S.; Gandhi, D. P.; Gupta, A.; and Salakhutdinov, R. R. 2020b. Object Goal Navigation using Goal-Oriented Semantic Exploration. In *Advances in Neural Information Processing Systems*, volume 33, 4247–4258. Curran Associates, Inc.

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. 2017. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, 2961–2969.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778. Las Vegas, NV, USA: IEEE. ISBN 978-1-4673-8851-1.

Liu, Y.; Sun, P.; Wergeles, N.; and Shang, Y. 2021. A survey and performance evaluation of deep learning methods for small object detection. *Expert Systems with Applications*, 172: 114602.

Min, S. Y.; Chaplot, D. S.; Ravikumar, P. K.; Bisk, Y.; and Salakhutdinov, R. 2021. FILM: Following Instructions in Language with Modular Methods. In *International Conference on Learning Representations*.

Park, S.; Brady, T. F.; Greene, M. R.; and Oliva, A. 2011. Disentangling scene content from spatial boundary: complementary roles for the parahippocampal place area and lateral occipital complex in representing real-world scenes. *Journal of Neuroscience*, 31(4): 1333–1340.

Persson, M.; Duckett, T.; Valgren, C.; and Lilienthal, A. 2007. Probabilistic Semantic Mapping with a Virtual Sensor for Building/Nature detection. In *2007 International Symposium on Computational Intelligence in Robotics and Automation*, 236–242.

Ramakrishnan, S. K.; Chaplot, D. S.; Al-Halah, Z.; Malik, J.; and Grauman, K. 2022. PONI: Potential Functions for ObjectGoal Navigation with Interaction-free Learning. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 18868–18878. New Orleans, LA, USA: IEEE. ISBN 978-1-66546-946-3.

Szot, A.; Clegg, A.; Undersander, E.; Wijmans, E.; Zhao, Y.; Turner, J.; Maestre, N.; Mukadam, M.; Chaplot, D. S.; Maksymets, O.; Gokaslan, A.; Vondruš, V.; Dharur, S.; Meier, F.; Galuba, W.; Chang, A.; Kira, Z.; Koltun, V.; Malik, J.; Savva, M.; and Batra, D. 2021. Habitat 2.0: Training Home Assistants to Rearrange their Habitat. In *Advances in Neural Information Processing Systems*, volume 34, 251–266. Curran Associates, Inc.

Tong, K.; Wu, Y.; and Zhou, F. 2020. Recent advances in small object detection based on deep learning: A review. *Image and Vision Computing*, 97: 103910.

Wani, S.; Patel, S.; Jain, U.; Chang, A. X.; and Savva, M. 2020. MultiON: Benchmarking Semantic Map Memory using Multi-Object Navigation. In *NeurIPS*.

Xia, F.; Zamir, A. R.; He, Z.; Sax, A.; Malik, J.; and Savarese, S. 2018. Gibson env: Real-world perception for embodied agents. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 9068–9079.

Yamauchi, B. 1997. A frontier-based approach for autonomous exploration. In *Proceedings 1997 IEEE International Symposium on Computational Intelligence in Robotics and Automation CIRA'97.'Towards New Computational Principles for Robotics and Automation'*, 146–151. IEEE.