An Explainable Robotic Perception Framework for Indoor Scene Understanding

Jihoon Kim

1 Introduction

1.1 Background and Problem Description

The field has witnessed rapid improvements in scene understanding thanks to the introduction of state-of-the-art deep learning methods and ever-improving computational resources. AI with a good scene understanding can be useful in many places. For example, it may help people with visual difficulties by creating an accurate 3D representation of the surrounding environment and providing safe guidance. Also, the recent rise of *"metaverse"* further increases the demand of scene understanding for real-time communication and interaction between its users and the real-world environment [1].

However, 3D scene understanding still remains a very challenging task till this date. This is mainly due to interactions between objects and changes in information across different scenes [2]. Also, there are fundamental differences in the ways humans and machines perceive the environment. Humans are naturally capable of understanding complex relationships between objects and their semantics from an image or a video. However, these are simply a list of numeric values for machines. Therefore, a machine learning agent must be able to extract both geometric and semantic information in order to provide useful knowledge to humans.

The house interior industry may also hugely benefit from the rise of scene understanding. Providing a 3D replica of a real house with editable objects can be hugely attractive to its customers, as they will be able to easily change and customise their houses as wished with low costs in a virtual environment. However, to the best of my knowledge, there has been very little study conducted in this area. Therefore, it will be an interesting and contributable work to develop an entire framework for full indoor perception with robots as a doctorate study.

1.2 Aim, Objectives and Contributions

This research aims to develop a real-time robotics perception framework that constructs a 3D representation of a real-world scene with editable geometric models. The objectives are to:

1. develop a real-time robotic perception and navigation method that captures information from real-world indoor scenes with explainable deep learning methods
2. develop an automated 3D scan to CAD method that uses captured information to reconstruct a scene and convert into editable geometric models with semantic contexts with minimum human interventions

Theoretical and methodological contributions in indoor scene understanding with robotic perception will be made to the relevant field by achieving these objectives.

2 Literature Review

2.1 Perception

2.1.1 Image Classification

2.1.2 Object Detection

2.1.3 Semantic Segmentation

2.1.4 Object Pose Estimation

2.2 Navigation

2.2.1 Reinforcement Learning

A reinforcement learning (RL) agent can be trained to move safely in an unperceived environment while maximising the amount of perceived data.

2.3 3D Reconstruction

3D scan to CAD

2.4 Datasets

A sufficient amount of labelled images is essential to train a machine learning agent to obtain useful information from perception data. Silberman and Fergus [3] and Li et al. [4], McCormac et al. [5] has improved indoor scene understanding by providing large-scale real and synthetic indoor image datasets, respectively.

2.5 Explainable Learning

3 Research Design and Methodology

3.1 Empirical Materials

3.1.1 Robot Platform

TurtleBot with a RGB-D sensor would be sufficient for perception and navigation tasks planned in this project.

3.1.2 Simulation Environment

Gazebo

3.2 Research Tasks and Time Allocation

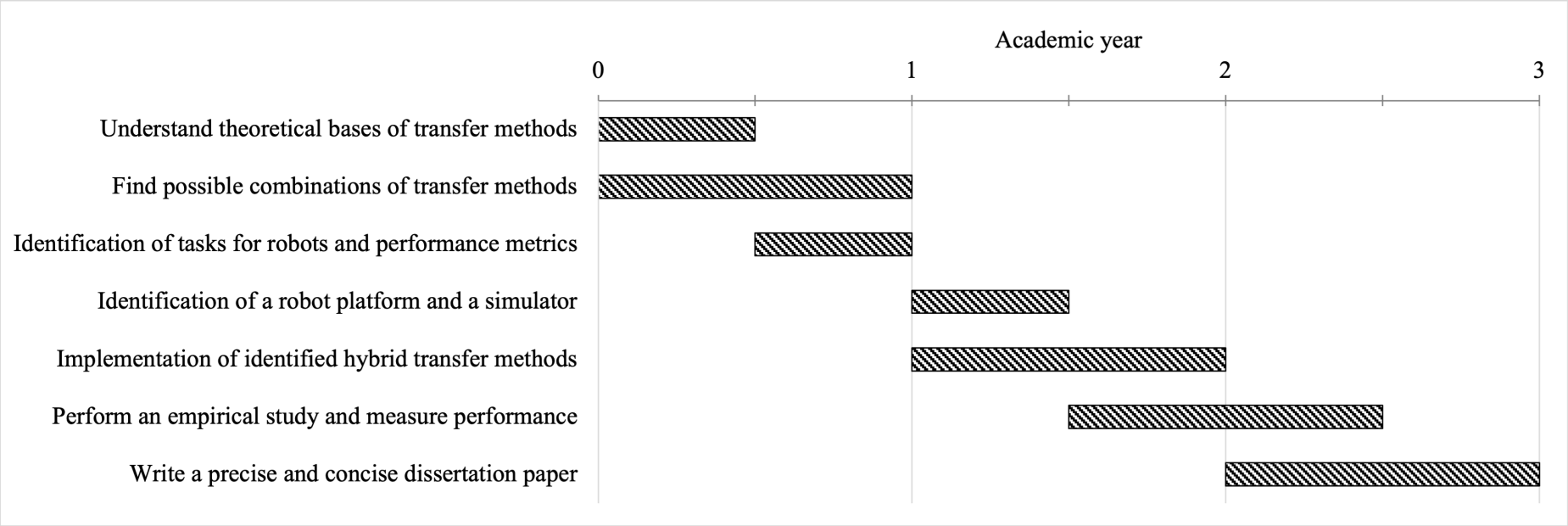

Figure 1: Gantt chart for time allocation of this project.

References

[1] Oculus, "Introducing presence platform: Unleashing mixed reality and natural interaction for oculus developers," 2021, https://developer.oculus.com/blog/introducing-presence-platform-unleashing-mixed-reality-and-natural-interaction-for-oculus-developers/ Last accessed: 06 Feb 2022.

[2] M. Naseer, S. Khan, and F. Porikli, "Indoor scene understanding in 2.5/3d for autonomous agents: A survey," *IEEE access*, vol. 7, pp. 1859–1887, 2018.

[3] N. Silberman and R. Fergus, "Indoor scene segmentation using a structured light sensor," in *2011 IEEE international conference on computer vision workshops (ICCV workshops)*. IEEE, 2011, pp. 601–608.

[4] W. Li, S. Saeedi, J. McCormac, R. Clark, D. Tzoumanikas, Q. Ye, Y. Huang, R. Tang, and S. Leutenegger, "Interiornet: Mega-scale multi-sensor photo-realistic indoor scenes dataset," *arXiv preprint arXiv:1809.00716*, 2018.

[5] J. McCormac, A. Handa, S. Leutenegger, and A. J. Davison, "Scenenet rgb-d: 5m photorealistic images of synthetic indoor trajectories with ground truth," *arXiv preprint arXiv:1612.05079*, 2016.