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. An artificial intelligence (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 object reconstruction 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

A perception framework should perform scene image classification to give the object detection and semantic segmentation models more general and useful information about a scene. Image classification is a basic, yet very important and challenging task for scene understanding. For example, perceived data must be represented with the best method (e.g., point cloud, 3D mesh) to maximise the useful information while minimising the computational complexity. Further, the proposed framework must be able to work well with limited amount of data, 3D deformations and background clutter that occur in the real-world conditions.

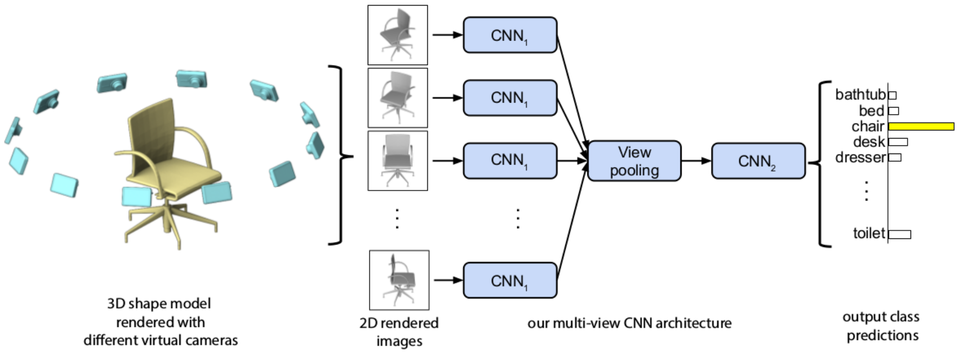

Figure 1: Multi-View CNN proposed by Su et al. [5]

Socher et al. [3] introduced the use of recurrent neural network (RNN) with convolutional layers to extract low and high-level features from the perceived RGB-D data and improve the classification performance. Further, Wu et al. [4] proposed the use of convolutional Deep Belief Network (DBN) that detects patterns from unlabelled data and showed that DBN can be used to learn joint distributions from 3D voxels. The use of convolutional layers reduces the computational costs compared to fully-connected (FC) layers thanks to lower number of parameters from weight sharing.

Su et al. [5] proposed multi-view convolutional neural network (MVCNN) that uses multiple 2D render views of 3D objects for classification, shown in Figure 1. It would be possible to capture multiple images of a 3D object in different views as it is planned to use video-recording robots for perception. Qi et al. [6] argued that MVCNNs perform better than volumetric CNNs due to input resolutions. The 3D representation of an object must be reduced to train a deep learning model in a reasonable timeframe.

2.1.2 Object Detection

The object detection algorithm in this project must be able to perform *"amodal object detection"*, i.e., to find the location, shape and orientation of an object in a scene. Similar to other detection techniques in the field, the initial object detection approaches were based on handcrafted features [7]. However, the most recent approaches utilise deep neural network based algorithms (e.g., region proposal generation [8]). The main challenges of object detection in this project will be the highly cluttered real-world scenes, understanding scene contexts (explained in Section 2.1.1) and an imbalanced distribution of objects.

2.1.3 Semantic Segmentation

In order to reconstruct a human-interpretable scene, it is important for this project to obtain semantically meaningful information from a given scene. However, it remains one of the most challenging tasks as obtaining dense pixel-level predictions is difficult due to occulusions and cluttering backgrounds in the real world. Also, it is heavily affected by the appearance and scale changes of objects as a robot obtains data while moving in an environment.

Conditional Random Field (CRF) has been the most popular choice for semantic segmentation with RGB-D data [9, 10] due to its flexiblity on model contextualisation. Due to the recent rise of deep learning based methods, CNNs have been increasingly used [11, 12] in the field, providing high-resolution segmentation. Couprie et al. [13] labelled the RGB-D images and videos of indoor scenes using multiscale CNN to learn features directly with remarkable accuracies. A similar technique that exploits temporal consistency in a video could be used for scene segmentation in this project.

2.1.4 Object Pose Estimation

Pose estimation is crucial for 3D CAD reconstruction task of objects in this project. It is often addressed jointly with object detection explained in Section 2.1.2 due to their similarities [14]. The most challenging tasks for pose estimation are to detect objects and estimate their orientations concurrently and to build an algorithm that works well when a robot is changing positions in an environment. The CNN-based models have shown promising results for pose estimation as well [15]. Krull et al. [16] further proved that the CNN can be used with objects in differenet shapes and appearances, and it does not solely optimise to the geometry or appearance. Therefore, a CNN-based object pose detection model similar to [15, 16] will be suitable for the detection of positions and poses of objects for this project.

2.2 3D Object Reconstruction

3D reconstruction is the most important task for this project to build a representation of a real-world scene. It is challenging to reconstruct a scene with incomplete information. It is advantageous that a robot platform will be capturing information as it moves in a scene, giving data from many different views with continuity.

KinectFusion [17] is a great reference for this project where RGB-D indoor dense information was used for surface mapping and tracking. A simple RGB-D sensor called *"Kinect"* and a GPU hardware were used to model real-world scenes to dense surfaces in real-time, showing reasonable computational costs. For this project, however, the objects in a scene must be modelled individually in order create editable models. Choy et al. [18] used long short-term memory (LSTM) reconstruct objects from synthetic data with arbitrary views. A technique similar to this may be used to perform 3D object reconstruction in this project.

2.3 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 [19] and Li et al. [20], McCormac et al. [21] has improved indoor scene understanding by providing labelled large-scale real and synthetic indoor image datasets, respectively. These datasets will be suitable to train and test the learning models for many different tasks stated above.

2.4 Explainable Learning

A reinforcement learning (RL) agent must be able to move safely in an unperceived environment while maximising the amount of useful information in perceived data. A simple deep RL method can be used to train an agent in a virtual indoor environment. Samek et al. [22] surveyed the methods to explain deep learning models that are usually applied in a *"black-box"* manner. Computing sensitivity of the prediction according to changes in input and decomposing the decision in terms of input variables were found two approaches to explaining predictions of deep learning models.

3 Research Design and Methodology

3.1 Empirical Materials

3.1.1 Robot Platform

TurtleBot with a RGB-D sensor (e.g., Kinect of Microsoft) would be sufficient for perception and navigation tasks planned in this project. The robot platform should be tall enough to perceive as much as possible; it will not be able to perceive objects on a table if it's too small.

3.1.2 Simulation Environment

A simulation environment must be capable of fully utilising the RGB-D image and video datasets to be used for learning models. Gazebo may be used with robot operating system (ROS) used as a primary data processing and control platform.

3.2 Research Tasks and Time Allocation

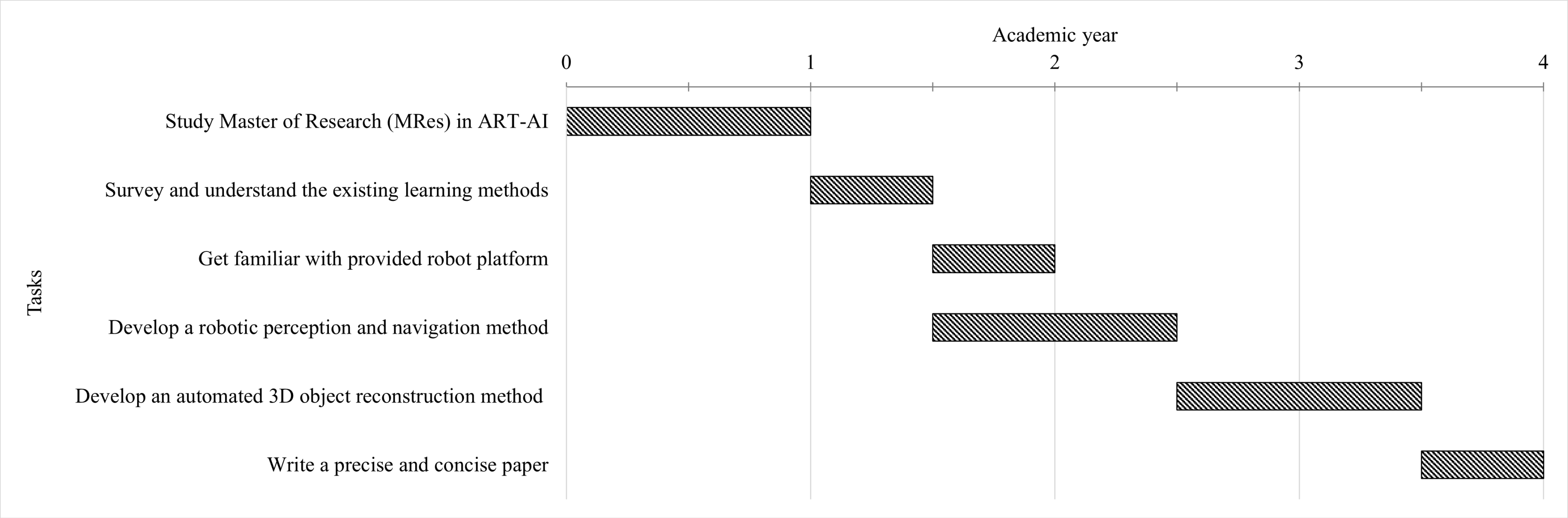

Figure 2: Gantt chart for time allocation of this project.

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] R. Socher, B. Huval, B. Bath, C. D. Manning, and A. Ng, "Convolutional-recursive deep learning for 3d object classification," *Advances in neural information processing systems*, vol. 25, 2012.

[4] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, and J. Xiao, "3d shapenets: A deep representation for volumetric shapes," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1912–1920.

[5] H. Su, S. Maji, E. Kalogerakis, and E. Learned-Miller, "Multi-view convolutional neural networks for 3d shape recognition," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 945–953.

[6] C. R. Qi, H. Su, M. Nießner, A. Dai, M. Yan, and L. J. Guibas, "Volumetric and multi-view cnns for object classification on 3d data," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 5648–5656.

[7] S. H. Khan, X. He, M. Bennamoun, F. Sohel, and R. Togneri, "Separating objects and clutter in indoor scenes," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 4603–4611.

[8] X. Chen, K. Kundu, Y. Zhu, A. G. Berneshawi, H. Ma, S. Fidler, and R. Urtasun, "3d object proposals for accurate object class detection," *Advances in neural information processing systems*, vol. 28, 2015.

[9] A. Hermans, G. Floros, and B. Leibe, "Dense 3d semantic mapping of indoor scenes from rgb-d images," in *2014 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2014, pp. 2631–2638.

[10] Z. Deng, S. Todorovic, and L. Jan Latecki, "Semantic segmentation of rgbd images with mutex constraints," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1733–1741.

[11] S. Gupta, R. Girshick, P. Arbeláez, and J. Malik, "Learning rich features from rgb-d images for object detection and segmentation," in *European conference on computer vision*. Springer, 2014, pp. 345–360.

[12] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[13] C. Couprie, C. Farabet, L. Najman, and Y. LeCun, "Indoor semantic segmentation using depth information," *arXiv preprint arXiv:1301.3572*, 2013.

[14] N. Sedaghat, M. Zolfaghari, E. Amiri, and T. Brox, "Orientation-boosted voxel nets for 3d object recognition," *arXiv preprint arXiv:1604.03351*, 2016.

[15] P. Wohlhart and V. Lepetit, "Learning descriptors for object recognition and 3d pose estimation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3109–3118.

[16] A. Krull, E. Brachmann, F. Michel, M. Y. Yang, S. Gumhold, and C. Rother, "Learning analysis-by-synthesis for 6d pose estimation in rgb-d images," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 954–962.

[17] R. A. Newcombe, S. Izadi, O. Hilliges, D. Molyneaux, D. Kim, A. J. Davison, P. Kohi, J. Shotton, S. Hodges, and A. Fitzgibbon, "Kinectfusion: Real-time dense surface mapping and tracking," in *2011 10th IEEE international symposium on mixed and augmented reality*. IEEE, 2011, pp. 127–136.

[18] C. B. Choy, D. Xu, J. Gwak, K. Chen, and S. Savarese, "3d-r2n2: A unified approach for single and multi-view 3d object reconstruction," in *European conference on computer vision*. Springer, 2016, pp. 628–644.

[19] 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.

[20] 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.

[21] 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.

[22] W. Samek, T. Wiegand, and K.-R. Müller, “Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models,” *arXiv preprint arXiv:1708.08296*, 2017.