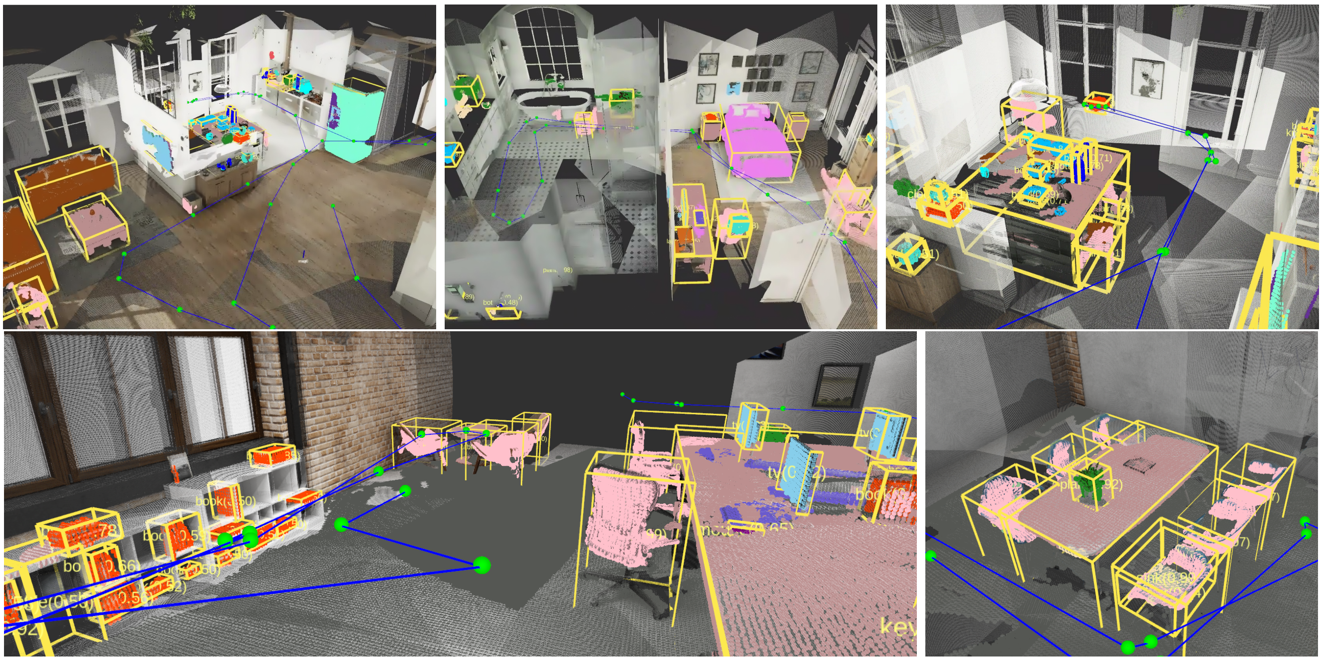
Panoptic hierarchical Semantic SLAM

Submission to the Embodied AI Robotic Vision Scene Understanding Challenge at CVPR 2022

Bosch Corporate Research Semantic SLAM Team



1 Introduction

This document summarizes our method used to participate in the Robotic Vision Scene Understanding Challenge at the Embodied AI Workshop of CVPR 2022 [1]. The team name used on the EvalAI submission page is "Bosch Corporate Research Semantic SLAM".

2 Summary of our Method

Our method utilizes a panoptic / semnatic segmentation frontend together with a custom foreground-background segmentation to create panoptic point clouds per frame. These point clouds are then voxeled in a hierarchical manner, and a label voting for voxel centers is performed, resulting in a global panoptic point cloud. Bounding boxes are extracted from this global panoptic point cloud by either top-down projection or clustering. Finally a filtering step is performed to obtain the final detections.

For the frontend, we use two off-the-shelf pre-trained panoptic / semantic segmentation networks (Detectron2 Panoptic FPN R101 [2] trained on COCO and MMSegmentation UPerNet Swin-B [3] trained on ADE20K) which we apply onto the input images received from the robot. We fuse the output of these networks to make object detections more robust.

Next we create a panoptic point cloud from the panoptic segmentation and the depth-channel of the RGB-D image. To eliminate incorrect associations of panoptic masks with 3D points in the point cloud (e.g. due to inaccurate masks, or due to holey objects like wireframe chairs, plants), we perform basic foreground-background segmentation based on region growing per instance segment on the depth-channel. We then assign the segment-class to the foreground points of each segment. This leaves us with a panoptic point cloud for the current frame, with some false projections already eliminated.

Next we perform a hierarchical voxel-based class voting. To reduce computation time, we operate on a keyframe basis. Keyframes are determined based on rotational and positional distance to the last keyframe. We collect all the per-frame panoptic point clouds since the last keyframe and combine them to a joint panoptic point cloud. This point cloud is then voxelized, and the class of the voxel center for each voxel is determined from all the original points falling into this voxel. In particular we perform a weighted majority vote per voxel using the confidences of the classes from the original segmentation. We retain a confidence distribution of the top classes per voxel for a later global voting. From this step we obtain a voxelized panoptic point cloud for each keyframe.

We then combine all these per-keyframe panoptic point clouds and perform the same voxelization and voting strategy on a global scale. The result is a global voxelized panoptic point cloud of the scene. Objects are finally detected on the basis of this point cloud. For large objects (couches, chairs, refrigerators, etc.), we do a top-down projection of points for each class, and then extract connected components to create bounding boxes for the individual objects. We keep track of extents in Z and overall class distribution during this.

For smaller objects (books, bowls, etc.) we perform per-class Euclidean clustering and determine bounding box extents directly from the individual clusters of points.

Finally, we perform some basic size- and confidence-based filtering and other small adjustments.

3 Closing Remarks

We would like to thank the authors for organizing the challenge, for providing the easy to use simulation environment and for responding quickly to issues that arose along the way. We look forward to participating in future iterations of this challenge.

Bosch Research

Bosch Research is a worldwide corporate division of Bosch, dedicating itself to "Research that really matters". This means that research leads to technologies that impact life with real solutions. Research is conducted in the fields of Artificial Intelligence, IoT, Autonomous Systems and many more. You can find more information on Bosch Research on bosch.com/research.

Contribution

This submission was created by a joint team from the Bosch research locations in Hildesheim and Renningen, Germany. It is part of a larger SLAM-pipeline for predevelopment projects for various robots, ranging from small household robots over industrial robots to autonomous driving applications.

The current members of the Semantic SLAM Team are: Narunas Vaskevicius, Christian Jütte, Reza Sabzevari, Timm Linder, Peter Biber and Stefan Benz. Other colleagues have also contributed to the surrounding system in the past.

References

[1] Niko Sünderhauf. *The Robotic Vision Challenges*. The Robotic Vision Challenges. URL: https://nikosuenderhauf.github.io/roboticvisionchallenges/cvpr2022.html (visited on 06/13/2022).

[2] Yuxin Wu et al. *Detectron2*. https://github.com/facebookresearch/detectron2. 2019.

[3] MMSegmentation Contributors. *MMSegmentation: OpenMMLab Semantic Segmentation Toolbox and Benchmark*. https://github.com/open-mmlab/mmsegmentation. 2020.