3D Map Generation with Shape and Appearance Information

Taro Yamada* and Shuichi Enokida
Kyushu Institute of Technology, Iizuka-shi, Fukuoka, Japan

Keywords: Robotics, Visual SLAM, Semantic Segmentation.

Abstract: It is clear from the numerous reports of recent years that interest in the research and development of autonomous mobile robots is growing and that a key requirement for the successful development of such self-directed machines is effective estimations of the navigable domain. Furthermore, in view of the differing characteristics of their physical performance capabilities relative to specific applications, specific estimations must be made for each robot. The effective assessment of a domain that permits successful robot navigation of a densely occupied indoor space requires the generation of a fine-grained three-dimensional (3D) map to facilitate its safe movements. This, in turn, requires the provision of appearance information as well as space shape ascertainment. To addresFs these issues, we herein propose a practical Semantic Simultaneous Localization and Mapping (Semantic SLAM) method capable of yielding labeled 3D maps. This method generates maps by class-labeling images obtained via semantic segmentation of 3D point groups obtained with Real-Time Appearance-Based Mapping (RTAB-Map).

1 INTRODUCTION

Interest in the research and development of autonomous mobile robots has rapidly grown in recent years because such machines promise to reduce the human burdens needed to perform routine functions such as providing cleaning and security services at train stations, shopping malls, and other public facilities. However, for a robot to move autonomously in an actual environment, a key requirement is its ability to effectively estimate its navigable domain, and this ability will vary depending on the robot's size, suspension capability, and other factors. Accordingly, such estimations must be determined for each robot.

It is also necessary to consider the robot body structure in order to facilitate safe navigation and to determine efficient routes. This is particularly true when it is necessary for a robot to navigate a densely occupied indoor space, in which case its ability to make a detailed determination of the most proximate environment is essential. Such environmental recognition information may involve using appearance information to distinguish objects, states, and material properties based on appearances.

One means of obtaining such appearance information is semantic segmentation, which is a method that may also facilitate the recognition of objects and materials at the pixel level from images of the robot's environment. However, because appearance information is two-dimensional (2D), even if a step is present in front of a door, as shown in Fig. 1, it will be identified as a single domain.

In contrast, shape information can be more immediately useful because it yields fine-grained three-dimensional (3D) information, thus permitting recognition of steps and slopes. As a result, the obtained shape information can ensure the recognition of steps and other 3D objects, and ascertainments based on that capacity are thus applicable to navigable domains.

However, shape information cannot, by itself, identify domain states or materials. Accordingly, in the present study, as a process for estimating navigable domains of robots that considers their physical structure, we herein propose a Semantic Simultaneous Localization and Mapping (Semantic SLAM) method capable of generating labeled 3D maps and obtaining both shape and appearance information simultaneously.

* https://www.kyutech.ac.jp

Yamada, T. and Enokida, S.
3D Map Generation with Shape and Appearance Information.
DOI: 10.5220/0010812900003124
In *Proceedings of the 17th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2022) - Volume 4: VISAPP*, pages 743-748
ISBN: 978-989-758-555-5; ISSN: 2184-4321

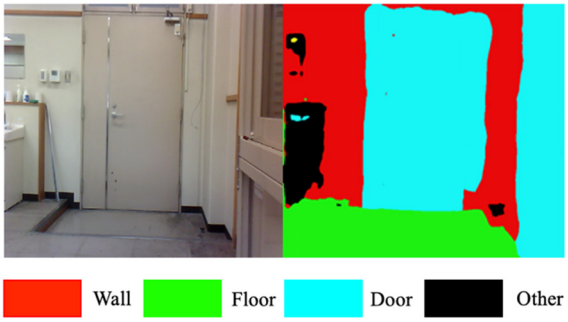


Figure 1: Identifying a floor with a step as a continuous single-level domain.

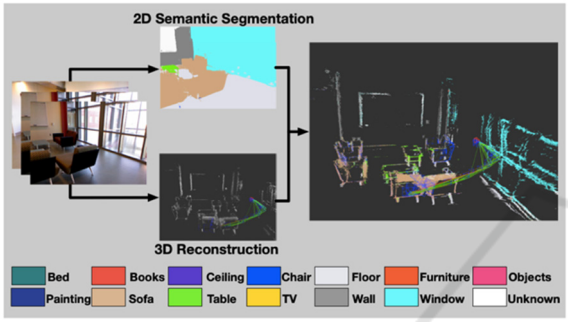


Figure 2: Result shown in Li et al.

2 RELEVANT STUDIES

SLAM studies with labeled 3D point group class information have been performed by numerous researchers in recent years, and the resulting methods can be separated into two main categories.

In one category, the objects in the image are first distinguished, after which image-based mapping is then added to generate the 3D point group. One method belonging to this category, which was proposed by Li et al, produces semantic segmentation output using a convolutional neural network (CNN) and Large-Scale Direct Monocular-SLAM (LSD-SLAM), thereby producing results such as those shown in Fig. 2.

Another method in this category, also proposed by Liu et al, incorporates object detection by Single Shot MultiBox Detector (SSD) into ORB(Oriented FAST and Rotated BRIEF)- SLAM2 to generate 3D point groups. However, since both of these methods employ a camera as a sensor, they are strongly affected by camera parameter divergence, light reflection, halation, and other such effects, and neither can generate dense point groups. As a result, they are unsuitable for autonomous mobile robot obstacle avoidance or navigation.

To resolve these problems, the method proposed in Ref, which uses an RGB-D camera as the sensor, combines the use of Real-Time Appearance-Based Mapping (RTAB-Map) and the YOLOv2 single-stage real-time object detection model to perform highly accurate real-time detection. More specifically, the method uses RTAB-Map to generate high-density broad-ranging labeled 3D maps. Meanwhile, YOLOv2 object detection enables labeling of bags, mugs, cars, and other such objects – but not walls, floors, or other such domains.

The second category comprises methods in which directly labeled 3D point groups are generated with PointNet or some other deep learning process. However, it must be noted that identification methods based on 3D point groups are far more difficult than image-based recognition methods and generally lead to network construction complexity. Additionally, the necessity of using complete 3D point groups as network input makes those methods inappropriate for online mapping with SLAM, while the inability of this method type to consider appearance information also makes it difficult to distinguish between doors and walls, earthen and flooring surfaces, and other planar surface states.

In contrast, in conjunction with a method for labeling of high-density expansive 3D point groups generated via an RGB-D camera, our proposed SLAM method enables the generation of 3D maps capable of fine-grained discernment of the robot's environment through the use of close real-time semantic segmentation capable of recognizing both domain and object content.

3 3D MAP GENERATION WITH SHAPE AND APPEARANCE INFORMATION

3.1 Shape Information Acquisition with RTAB-Map

RTAB-Map is an open-source library available since 2013. Its current expanded range of functions enables the generation of high-density, high-resolution maps in correspondence with arrays of stereo and RGB-D cameras and light detection and ranging (LiDAR) sensors to provide practical SLAM. Figure 3 shows the main nodes of RTAB-Map, which consists essentially of a basic structure with mounted memory processing referred to as Graph-base SLAM.

Inputs, which are synchronized and transferred to the graph-based SLAM, consist of camera input

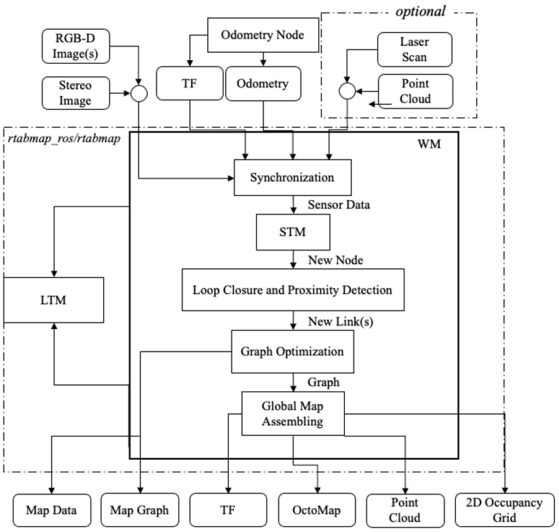


Figure 3: RTAB-Map main node schematic.

including the calibration message, transform(TF) defining the sensor positioning relative to the robot, odometry from a freely selected source, and (optionally) 2D LiDAR laser scans or 3D point clouds.

The map structure of the graph-based SLAM algorithm is graphic with linked nodes. An adjacent link with a node is generated by the Short-Term Memory(STM) from the synchronized input, loop closure input produced by an image-based bag-of-words, and/or a new link added by laser scan-based proximity detection.

The constructed graph is optimized and output as a 3D point cloud or an occupancy grid map by the global map assembly module. Additionally, a memory management function is mounted on the RTAB-Map, and the number of nodes used for loop closure detection is limited by working memory (WM) and long-term memory (LTM) memory management, thus maintaining real-time performance even if the map grows to a large-scale size.

3.2 Appearance Information Acquisition by Semantic Segmentation

For appearance information, a Bilateral Segmentation Network (BiSeNet) was used to provide real-time semantic segmentation. BiSeNet is designed to resolve the tradeoff between real-time and output image precision.

As shown by its network schematic in Fig. 4, BiSeNet is characterized by its use of two paths. In the spatial path, a features map holding spatial information is generated by three convolution layers (stride=2), while in the context path, multiple reduced-size feature maps are output by the Xception module. A features map bearing contextual information is output by their passage through an attention refinement module (Fig. 5) holding global average pooling.

Lower- and higher-level feature maps obtained in the spatial and context paths, respectively, are efficiently combined in the feature fusion module (Fig. 6) to facilitate high-precision semantic segmentation. This method yielded a 68.4% mean intersection-over-union (IOU) at a rate of 105 frames per second (FPS) for Cityscapes test data.

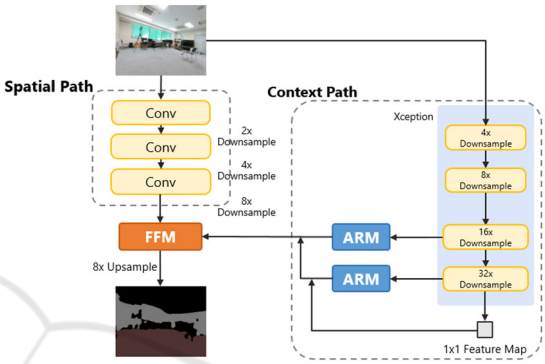


Figure 4: BiSeNet schematic.

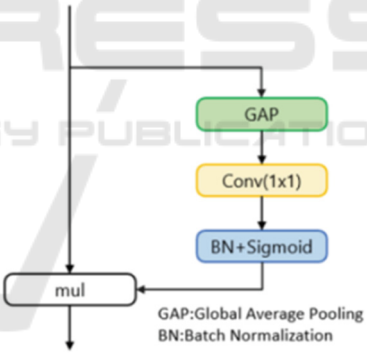


Figure 5: ARM schematic.

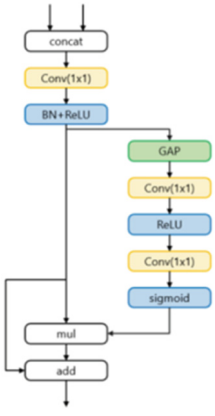


Figure 6: FFM schematic.

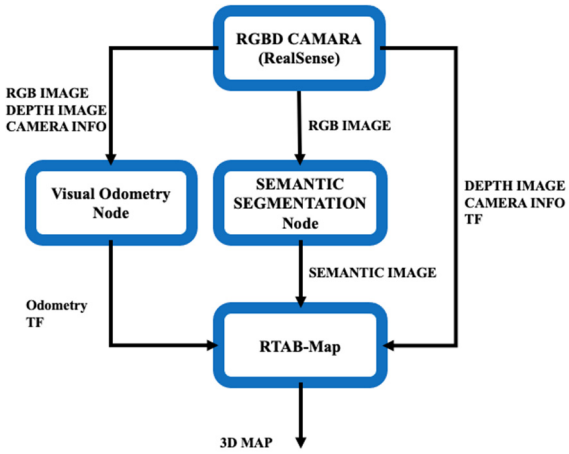


Figure 7: ROS composition used in the present study.

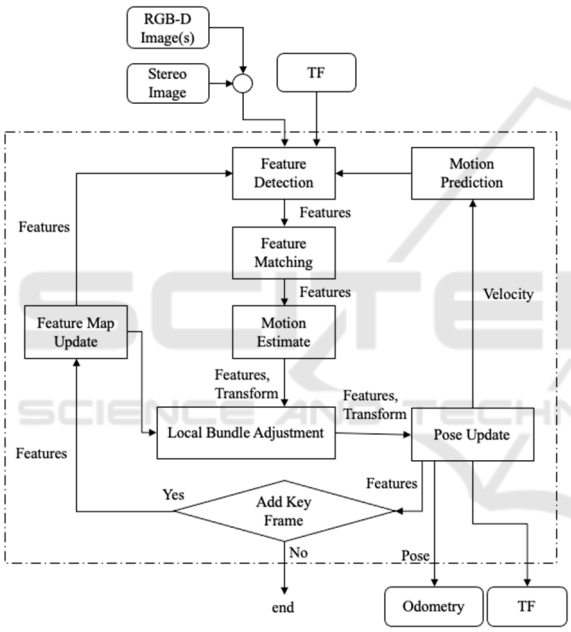


Figure 8: Visual odometry node schematic.

3.3 Shape and Appearance Information Integration

In the present study, we constructed an internodal communication system using the Robot Operating System (ROS) composition shown schematically in Fig. 7. For odometry input to RTAB-Map, we use the visual odometry provided by the RTAB-Map library, as shown schematically in Fig. 8.

The feature point is first extracted from the input image (feature detection), after which the Binary Robust Independent Elementary Features (BRIEF) descriptor of the detected feature is then compared with the feature map, and matched (feature matching) by nearest-neighbor search.

When a corresponding point is computed, the current rate of frame conversion on the feature map is calculated (motion estimate) using Perspective-n-Point (PnP) packaged with OpenCV. Finally, the total number of key frame features on the feature map is optimized (local bundle adjustment) with g^{2}o, which is an open-source C++ framework for optimizing graph-based nonlinear error functions, and the odometry is updated (pose update).

The feature points extracted after input by the visual odometry node are reused to generate the image vector for loop closure detection at the RTAB-Map main node. Next, a visual-based odometry estimation and an RGB image-based loop closure are performed. Here, it should be noted that the semantic segmentation output image is only used for 3D point group coloring and that all inputs to the RTAB-Map nodes in the present study are unified at 30 FPS and synchronized completely using ROS time stamps.

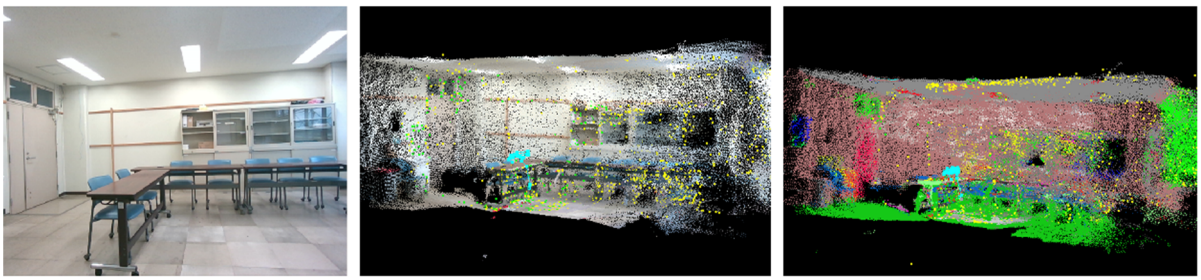
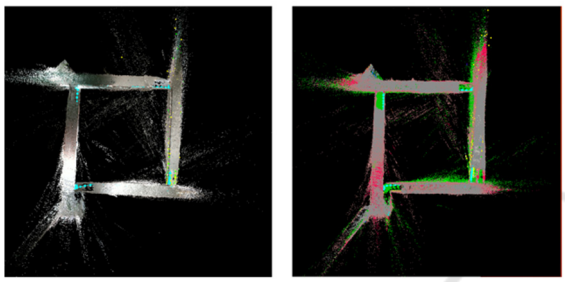


Figure 9: (a) Movement ascertainment environment, (b) pre-existent RTAB-Map output, (c) output of RTAB-Map labeled with a semantic segmentation output image.

4 MOVEMENT ASCERTAINMENT

The proposed method was implemented with an Intel RealSense D455 as the RGB-D camera. ADE20K was used as the training data for semantic segmentation. ADE20K provides images and pixel-level annotations of various scenes and objects, and classifies 150 classes. odometry of RTAB-Map. F2M provided by the library was used. Figure 9 shows a room mapping result produced via this method, and Fig. 10 shows the mapping result produced by walking one lap around a corridor (97.0m) and returning to the starting position.



(a) (b)

Figure 10: Output with loop closure detection: (a) pre-existent RTAB-Map, (b) RTAB-Map output labeled with a semantic segmentation output image.

Table 1: Processing speed in the odometry approach

RTAB-Map Odometry	o_{avg} (msec)
F2F[7]	37
F2M[7]	70
DVO[17]	37
Fovis[18]	21
ORB2-RTAB[19]	54
Ours(F2M+Segmentation)	107

The processing speeds of other odometry approaches are shown in Table 1. Table 1 shows that although the processing time of the proposed method is longer (107msc) than that of the conventional F2M method (70msc), the proposed method retains a certain degree of real-time performance and odometry is functional. Correct loop closure detection by operation ascertainment of the proposed method shown in Fig. 10 was also confirmed.

Comparison with Relevant Studies. The map point group completed by the method proposed in Li et al. was only semi-dense because its generation was based on LSD-SLAM. In contrast, it was possible to complete a high-density map in this study because it was generated based on RTAB-Map formation. Additionally, despite the large scale of the map, the real-time function was well maintained as a result of the mounted memory management function.

The method proposed by Mao et al. labels the object detected by YOLOv2 on the 3D map generated by RTAB-Map. In contrast, the present method utilizes semantic segmentation rather than object detection, and thus permits labeling of floor, ceiling, wall, and other domains as well as content objects. However, a problem remains, as the present study confirmed the presence of 3D points with semantic segmentation for objects yielding indistinct object boundary demarcation, which can prevent correct labeling.

When using the direct segmentation of 3D point groups described in PointNet, as proposed by Qi et al, it is difficult to distinguish between doors and walls or other regions on a given plane. In contrast, the present study confirmed that doors and walls could be correctly distinguished using our proposed method, as shown in Fig. 9(c).

5 CONCLUSION

Herein, we proposed a SLAM method that uses labeled 3D point group data obtained using an RGB-D camera and verified its operation. For generating a labeled 3D map, we implemented BiSeNet real-time semantic segmentation to perform classification in 2D images and used corresponding depth imaging to label 3D point groups in RTAB-Map.

The use of RTAB-Map enabled long-term high-precision Semantic SLAM, but when performing object labeling, the inter-object boundaries were sometimes indistinct, which could prevent correct labeling. Accordingly, our future studies will focus on enhancing semantic labeling while giving consideration to edge information and 3D point group clustering. It is also considered likely that higher resolution could be gained by inputting semantic segmentation output images in visual odometry and by canceling feature points representing noise extracted from mobile objects in odometric estimations and loop closure detections. Furthermore, we may also consider improvements to robot navigable domain estimation and travel route planning with the labeled 3D maps.

ACKNOWLEDGEMENTS

This work was supported by JSPS KAKENHI Grant Number JP18H01463.

REFERENCES

Junji EGUCHI and Koichi OZAKI, “Extraction Method of Travelable Area by Using of 3D-laser Scanner - Development of Autonomous Mobile Robot for Urban Area”, Transactions of the Society of Instrument and Control Engineers, Vol52, No3, 152/159,2016.

Hideaki Suzuki, Akihisa Oya, Shinichi Yuda, “Obstacle Avoidance of Mobile Robot Considering 3D Shape of Environment”, Robomec , 1998.

X. Li and R. Belaroussi, “Semi-dense 3d semantic. mapping from monocular slam”, Computer Vision and Pattern Recognition, 2016.

W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed and C. Y. Fu, et al, SSD: Single Shot MultiBox Detector. European Conference on Computer Vision, Springer International Publishing, 21–37, 2016.

Mingyuan Mao, Hewei Zhang, Simeng Li, and Baochang Zhang, “SEMANTIC-RTAB-MAP (SRM): A semantic SLAM system with CNNs on depth images”, Mathematical Foundations of Computing, 2019.

R. Q. Charles, H. Su, K. Mo and L. J. Guibas, “Pointnet: Deep learning on point sets for 3d classification and segmentation”, IEEE Conference on Computer Vision and Pattern Recognition, 2017.

M. Labbé and F. Michaud, “RTAB-Map as an Open-Source Lidar and Visual SLAM Library for Large-Scale and Long-Term Online Operation,” in Journal of Field Robotics, vol. 36, no. 2, pp. 416–446, 2019.

M. Labbé and F. Michaud, “Long-term online multi-session graph-based SPLAM with memory management,” in Autonomous Robots, vol. 42, no. 6, pp. 1133-1150, 2018.

M. Labbé and F. Michaud, “Online Global Loop Closure Detection for Large-Scale Multi-Session Graph-Based SLAM,” in Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, 2014.

M. Labbé and F. Michaud, “Appearance-Based Loop Closure Detection for Online Large-Scale and Long-Term Operation,” in IEEE Transactions on Robotics, vol. 29, no. 3, pp. 734-745, 2013.

M. Labbé and F. Michaud, “Memory management for real-time appearance-based loop closure detection,” in Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 1271–1276, 2011.

Changqian Yu et al, “BiSeNet: Bilateral segmentation network for real-time semantic segmentation”, European Conference on Computer Vision, pp. 325-341, 2018.

Chollet, F. “Xception: Deep Learning with Depthwise Separable Convolutions”, IEEE Conference on Computer Vision and Pattern Recognition, 2017.

Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. “The cityscapes dataset for semantic urban scene understanding”. In Proceed- ings of the IEEE conference on computer vision and pattern recognition, pages 3213–3223, 2016.

Kummerle, R., Grisetti, G., Strasdat, H., Konolige, K., and Burgard, W. “g2o: A general framework for graph optimization”. In Proceedings IEEE International Conference on Robotics and Automation, pages 3607–3613, 2011.

Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao , Sanja Fidler, Adela Barriuso and Antonio Torralba, ”Semantic Understanding of Scenes through the ADE20K Dataset”, International Journal on Computer Vision , 2018.

Kerl, C., Sturm, J., and Cremers, D., “Dense visual SLAM for RGB-D cameras”, In Proceedings IEEE/RSJ International Conference on Intelligent Robots and Systems, pages 2100–2106, 2013.

Huang, A. S., Bachrach, A., Henry, P., Krainin, M., Maturana, D., Fox, D., and Roy, N., “Visual odometry and mapping for autonomous flight using an RGB-D camera”, In Proceedings International Symposium on Robotics Research, 2011.

Mur-Artal, R. and Tard os, J. D., “ORB-SLAM2: An open-source SLAM system for monocular, stereo and RGB-D cameras”, IEEE Transactions on Robotics, 33(5):1255–1262, 2017.