International Journal of Advanced Culture Technology Vol.9 No.4 302-306 (2021)
DOI https://doi.org/10.17703/IJACT.2021.9.4.302

IJACT 21-12-36

Visual Positioning System based on Voxel Labeling using Object Simultaneous Localization And Mapping

[1]Tae-Won Jung, [2]In-Seon Kim, [3]Kye-Dong Jung

[1]Department of Immersive Content Convergence, Kwangwoon University,
20 Kwangwoon-ro, Nowon-gu, Seoul 01897, Korea
[2]Department of Smart System, Kwangwoon University,
20 Kwangwoon-ro, Nowon-gu, Seoul 01897, Korea
[3]Ingenium College of Liberal Arts, Kwangwoon University,
20 Kwangwoon-ro, Nowon-gu, Seoul 01897, Korea
{onom, kisidid, gdchung}@kw.ac.kr

Abstract

Indoor localization is one of the basic elements of Location-Based Service, such as indoor navigation, location-based precision marketing, spatial recognition of robotics, augmented reality, and mixed reality. We propose a Voxel Labeling-based visual positioning system using object simultaneous localization and mapping (SLAM). Our method is a method of determining a location through single image 3D cuboid object detection and object SLAM for indoor navigation, then mapping to create an indoor map, addressing it with voxels, and matching with a defined space. First, high-quality cuboids are created from sampling 2D bounding boxes and vanishing points for single image object detection. And after jointly optimizing the poses of cameras, objects, and points, it is a Visual Positioning System (VPS) through matching with the pose information of the object in the voxel database. Our method provided the spatial information needed to the user with improved location accuracy and direction estimation.

Keywords: *Visual Positioning System, Simultaneous Localization and Mapping, Augmented Reality, Deep Learning*

1. INTRODUCTION

Object detection and SLAM are two important techniques in computer vision and augmented reality. In augmented reality, 3D objects also need localization for more physical interaction, and various sensors that can directly provide depth measurement such as laser-range finders and stereo or RGBD cameras can be used. Most of the state of art monocular approaches solve object detection and SLAM separately, relying on prior object models in general circumstances. Our method is a VPS through matching with Voxel Labeling database in indoor space using object detection and object SLAM, which can greatly improve 3D indoor environment understanding and augmented reality and virtual reality technology.

2. RELATED RESEARCH

Manuscript received: October 26, 2021 / revised: November 10, 2021 / accepted: December 7, 2021
Corresponding Author: gdchung@kw.ac.kr
Tel: +82-2-940-5288
Professor, Ingenium College of liberal arts, Kwangwoon University, Seoul, Korea

2.1 Cuboid using Single Image 3D Object Detection

Detecting 3D objects from a single image is much more difficult than 2D because more object pose variables and camera projection geometry need to be considered. Existing 3D detection approaches can be divided into two categories: with prior models based or without prior models.

With prior models: object poses that best suit RGB images can be found through key-point Perspective n-Point (PnP) matching, manually created texture functions, or the latest deep network [1, 2].

Without prior models: objects are generally displayed as cuboids. A typical approach is to combine geometric modeling and learning. Objects are generated through the vanishing point by a combination of Manhattan edges or rays, and many 3D boxes are thoroughly sampled and then selected based on various situational features [3, 4]. CubeSLAM has expanded to operate without predicting the size and direction of the object using projection features to find a cube that fits perfectly into the 2D boundary box and estimating the vanishing point and one corner. As shown in Figure 1 also allows the remaining seven corners to be calculated analytically [5].

2.2 Object SLAM

ObjectSLAM is a point-based visual SLAM algorithm, such as ORB SLAM and DSO [6, 7], and achieved impressive results in general environments. Recently, object-augmented mapping combining objects and augmentation proposed. ObjectSLAM generally has a method in which object and SLAM are decoupled or coupled. The decoupled approach first builds a SLAM point cloud map, and then further detects and optimizes 3D object poses based on point cloud clustering and semantic information. Although it showed improved results compared to 2D object detection, the separation approach may fail if SLAM cannot build high-quality maps because the SLAM part has not been changed. The coupled approach is usually called object-level SLAM. Semantic Structure from Motion (SfM) jointly optimizes camera poses, objects, points and planes and SLAM++ is a practical SLAM system that uses RGB-D cameras and prior object models [8]. Recently, real-time monocular object SLAM used a prior object model and there is also some end-to-end deep learning-based SLAM without object representations, such as DVSO [9, 10].

3. VPS BASED ON VOXEL LABELING USING OBJECT DETECTION

3.1 System Structure

Our method combines 3D object detection with SLAM pose estimation by generating cuboids from 2D image detection. As shown in Figure 1, after estimating position of camera and pose of object, it is a VPS through matching with our proposed voxel labeling database. First, a 2D object is detected from a 2D image acquired by a camera with a deep learning network such as YoLo convolutional neural networks. The cube of the object is estimated through the recognition of the 2D bounding box and sampling of the vanishing point. The selected cuboid is then further optimized with point and camera tracking via multi-view Bundle Adjustment (BA). Objects provide geometry and scale constraints in BA and depth initialization for difficult-to-triangulate points. The object's cuboid is then matched with a database of known object's cuboid positions and poses based on voxel labeling. After matching each cuboid, the estimated camera position in the camera coordinate system is matched to a voxel-based database, and its position in the indoor 3D space can be determined.

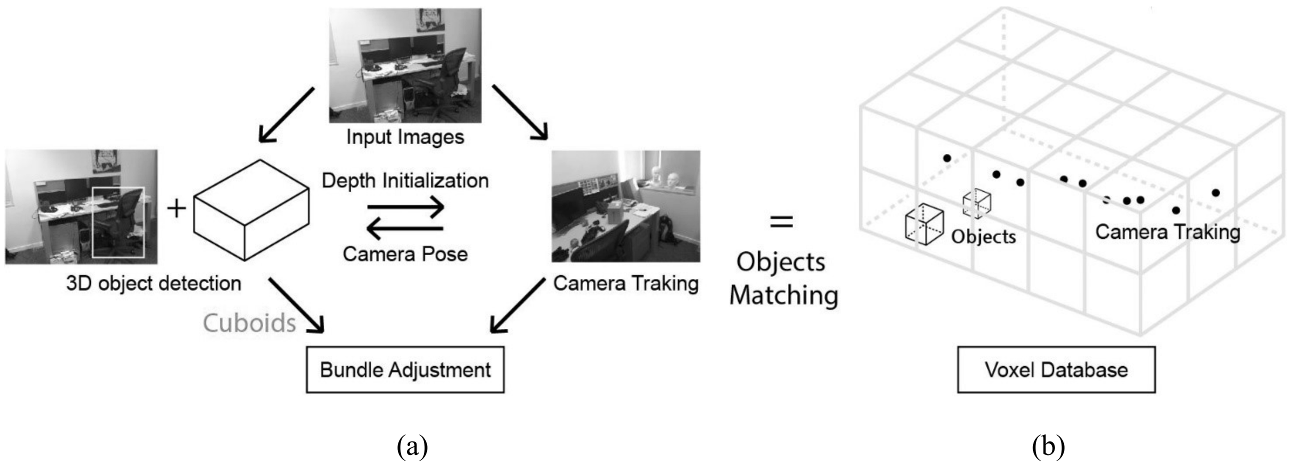


Figure 1. VPS of based on voxel labeling using object SLAM. (a) Cuboids generated after 2D deep learning network detection, and (b) voxel labeling-based database

3.2 Single Image 3D Object Detection

Efficiently create 3D cuboids using 2D bounding boxes in 3D space. A typical 3D cuboid is created with 9 degree of freedom parameters. As shown in Figure 2, the three vanishing points of the cuboid generated by our method are generated after being projected by the calibration matrix of the object and the rotated camera frame in a 2D bounding box. 8 corners of 2D bounding boxes are obtained based on the generated vanishing points. After creating a cuboid edge in 2D image space, the 3D pose of the cuboid is estimated. Calculate the 3D position and dimensions of the cuboid using the Perspective-n-Point (PnP) solver.

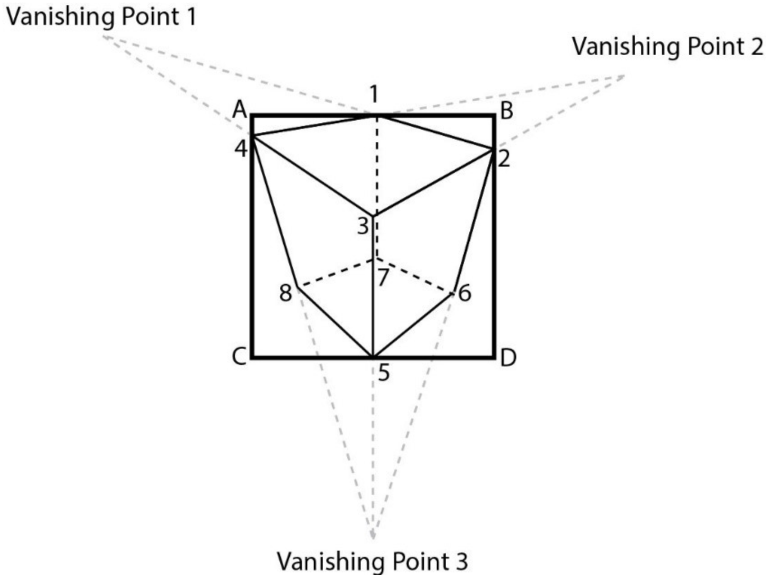
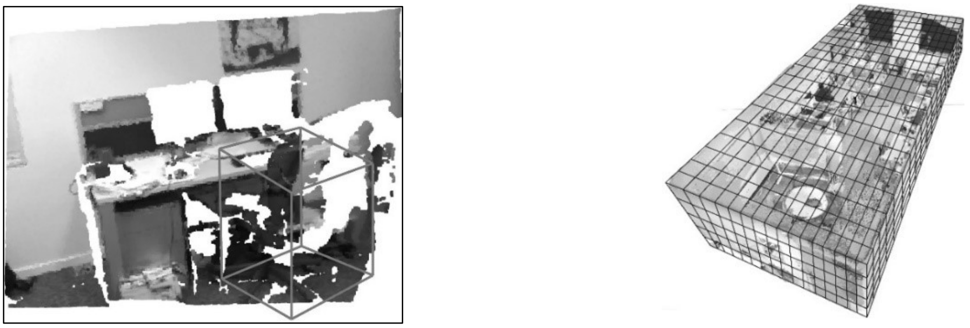


Figure 2. Cuboid proposals generation from 2D object box

3.3 VPS based on voxel labeling

The indoor space is defined and constructed as a database of voxel addresses. As shown in Figure 3(a), the 3D pose of the object and the position of the camera can be tracked. The user's movement path and location are determined by matching the estimated object pose information and the camera location information with the database object pose information of the voxel address as shown in Figure 3(b).



(a) (b)

Figure 3. The architecture of the proposed approach. It has two main components: (a) a cuboid proposed in a point cloud, and (b) a voxel labeling database composed of box coordinates of objects

4. CONCLUSION

Our method is a VPS that matches the position and pose of an object in a defined space after estimating the cuboid of an object by learning the image acquired from the camera through deep learning. Generating and optimizing 3D cuboids with bounding boxes and vanishing points of objects in 2D image object detection using deep learning networks. We build a database on a voxel-by-voxel basis to determine the user's location and show that structured voxels can efficiently obtain high positioning accuracy and orientation estimation. We propose is a deep learning-based VPS using fixed objects indoors when GPS is not enough to provide customized augmented reality content to users. In the future, it will be applied to the augmented reality service system by estimating the position and pose of an object using a graph neural network with sufficient scalability and research value of deep learning. It is necessary to expand to navigation and scene analysis in indoor spaces without GPS using mobile devices.

REFERENCES

[1] Florian Chabot, Mohamed Chaouch, Jaonary Rabarisoa, Celine Teuliere and Thierry Chateau, "Deep MANTA: A coarse-to-fine many-task network for joint 2d and 3D vehicle analysis from monocular image," In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2040–2049, 2017.

[2] Wadim Kehl, Fabian Manhardt, Federico Tombari, Slobodan Ilic and Nassir Navab, "SSD-6D: Making rgb-based 3D detection and 6d pose estimation great again," In *IEEE International Conference on Computer Vision*, 2017.

[3] Jianxiong Xiao, Bryan Russell and Antonio Torralba, "Localizing 3D cuboids in single-view images," In *Advances in neural information processing systems (NIPS)*, pp. 746–754, 2012.

[4] Xiaozhi Chen, Kaustav Kundu, Ziyu Zhang, Huimin Ma, Sanja Fidler and Raquel Urtasun, "Monocular 3D object detection for autonomous driving," In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2147–2156, 2016.

[5] S. Yang and S. Scherer, "Cubeslam: Monocular 3-d object slam," *IEEE Transactions on Robotics,* vol. 35, no. 4, pp. 925–938, 2019.

[6] Raul Mur-Artal, J.M.M. Montiel and Juan D. Tardos, "ORB-SLAM: a versatile and accurate monocular SLAM system," *IEEE Transactions on Robotics*, 31(5):1147–1163, 2015.

[7] Jakob Engel, Vladlen Koltun and Daniel Cremers, “Direct sparse odometry,” *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 2017.

[8] Renato F. Salas-Moreno, Richard A. Newcombe, Hauke Strasdat, Paul H.J. Kelly and Andrew J. Davison, “SLAM++: Simultaneous localisation and mapping at the level of objects,” In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1352–1359, 2013.

[9] Dorian Galvez-Lopez, Marta Salas, Juan D. Tardos and J.M.M. Montiel, “Real-time monocular object SLAM,” Robotics and Autonomous Systems, 75:435–449, 2016.

[10] Nan Yang, Rui Wang, Jorg Stuckler and Daniel Cremers, “Leveraging deep depth prediction for monocular direct sparse odometry,” In *European Conference on Computer Vision*, pp. 835–852. Springer, 2018.