2018 IEEE Intelligent Vehicles Symposium (IV)
Changshu, Suzhou, China, June 26-30, 2018

Dense 3D Semantic SLAM of traffic environment based on stereo vision

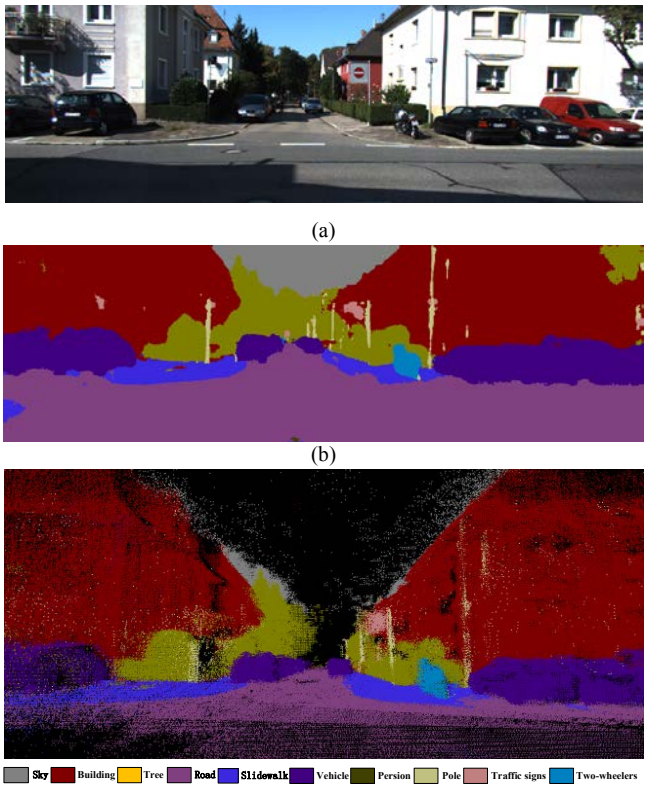
Linhui Li, Zhijie Liu, Ümit Özgüner, Jing Lian, Yafu Zhou, Yibing Zhao

***Abstract*—To solve the intelligent vehicles' problems of 'where am I?' and 'what is around me?', a dense 3D sematic Simultaneous Localization and Mapping (SLAM) system is proposed to evaluate the pose of the intelligent vehicles and build the dense 3D semantic map. We address these challenges by combining a state of art Stereo-ORB-SLAM system and Convolutional Neural Networks. Firstly, we build a dense 3D point cloud map by using a four thread Stereo-ORB-SLAM system. Subsequently, a fully convolutional neural network architecture which uses RGB-D image as input is used to obtain pixel-wise segmentation. Finally, we fuse the geometric information and semantic information to get the semantic map. We test our method on the KITTI dataset and our dataset made with the Fpgalena stereo camera. Results indicate the system was effective in the real-time building of a semantic map, the speed of the entire system is about 10Hz, and the loop closing function can eliminate most of the drifting errors.**

***Index Terms*—Semantic SLAM, convolutional neural network, stereo vision.**

I. INTRODUCTION

Self-driving vehicles are a very hot topic of research in recent years. We believe self-driving vehicles can serve the public's need for safer transportation and better mobility. In a safety report on intelligent vehicles produced by Google company Waymo [1], it's stated that fully self-driving vehicles need to be able to solve four key problems: "where am I ?", "what is around me?", "what will happen next?", and " what should I do next?". This viewpoint puts forward the more detailed and specific requirement s for intelligent vehicles. The first problem is essentiall one about the localization of vehicles, and the second is about perceiving the environment environments. Solving these two problem allows the vehicles to obtain the geometric and semantic information from the environment which lets the car understand the environments like a human does. These are vital for cars' path planning [2] and autonomous navigation. Furthermore, they are also the basics of vehicle-vehicle and vehicle-human interaction [3]. In view of the first two problems, we propose a Dense 3D Semantic SLAM system which combines the geometric information from a state-of-art Stereo-ORB-SLAM system[4] with recent advances in semantic segmentation using Convolutional Neural Networks [5] (CNNs).



(a)

(b)

(c)

Fig. 1. Semantic maps. Fig. (a) is one of the color image in the KITTI dataset. Fig. (b) is corresponding segmentation results and Fig. (c) is the dense 3D point cloud semantic map generated by our system

Our approach is to use Stereo-ORB-SLAM to build a globally consistent dense 3D octomap. This is the basic process of fusing the pixel-wise semantic information from CNNs prediction into the semantic map, as shown in Figure 1. Stereo-ORB-SLAM [4] is a real-time accurate and effective SLAM system for outdoor environments. It has localization and loop closing functions to help intelligent vehicles be precisely located while outdoors. However, Stereo-ORB-SLAM builds a sparse point cloud map which does not meet the demands of a dense 3D semantic map. A fourth semantic mapping thread is added into the original ORB-SLAM system to solve this problem. In the fourth thread we obtain smooth disparity by using the semi-global stereo matching algorithm [6], and then we build a dense 3D map based on the disparity and the pose calculated in the SLAM system. The detailed information from the semantic map allows an intelligent

This project is supported by the National Natural Science Foundation of China (Grant Nos. 51775082, 61473057, 61203171) and the China Fundamental Research Funds for the Central Universities (Grant Nos. DUT17LAB11, DUT15LK13).

Linhui Li, Zhijie Liu, Jing Lian, Yafu Zhou and Yibing Zhao are with the School of Automotive Engineering, Faculty of Vehicle Engineering and Mechanics, Dalian University of Technology, Dalian 116024, China. And Jing Lian is the corresponding author. (e-mail:lilinhui@dlut.edu.cn; liuzhijie@mail,dlut.edu.cn;lianjing@dlut.edu.cn;dlzyf@dlut.edu.cn;zhaoyibing005@163.com)

Ümit Özgüner is with the Department of Electrical and Computer Engineering, The Ohio State University, Columbus, OH, 43210USA (e-mail: ozguner.1@osu.edu).

vehicle to know its location and what is around it.

The geometric information obtained from the dense map is helpful for improving the accuracy of the segmentation, and the geometric outline of the objects is beneficial to the segmentation of things such as vehicles, humans, trees, and so on. Obtaining depth information has become especially easily with the development of depth sensors such as surface scanning lasers and stereo vision. Therefore, we use the four channel RGB-D fused by disparity and the RGB as the input of the CNNs. The CNNs proposed in our previous work [5] are a new deep fully convolutional neural network architecture made by modifying the AlexNet network architecture. In this work we divide the traffic environment into 10 classes: sky, building, sidewalk, road, tree, car, pedestrian, pole, traffic sign, and two-wheelers. In our experiment, the results show that by using RGB-D as the input of CNNs, the accuracy of the segmentation can be improved the accuracy of the segmentation compared with just taking RGB as the input.

We evaluate our Semantic SLAM System with the KITTI dataset [7], and the entire system speed is about 10Hz. When compared to the real environment, the dense 3D semantic map that is built indicates that our system can precisely locate vehicles and is efficient mapping. In addition, to test the universality and scalability of the Stereo-ORB-SLAM, we run the stereo-ORB-SLAM system on our dataset captured by the stereo camera Fpgalena and rectify the epipolar lines using the Bouguet algorithm. The experimental results are shown in the Part of V.

II. Related Work

In the past, a significant amount of work has been done in semantic SLAM. Javier Civera proposed a monocular SLAM system [8] while Renato F. Salas-Moreno proposed a SLAM++ [9] system which maps indoor scenes. These two ways obtain the pose of the image using an Extended Kalman Filter (EKF) and find the semantic information on the level of the objects by building pre-defined object models. However, the pose calculated by EKF is not accurate and effective enough when compared to ORB-SLAM [4], and the application of the system is limited because of the pre-defied object models. Abhijit Kundu et al. joined the CRF [10] model with SFM to produce a 3D volumetric semantic + occupancy map. However, the segmentation of the CRF has lower robustness and lower generalization. Additionally, the SFM system has no loop-closing nor relocation functions to tackle the position and posture of the intelligent vehicles.

With the development of GPU and deep learning, CNN has showed better robustness, better generalization, and high accuracy in segmentation such as in AlexNet [11], VGGNet [12], GoogleNet [13] etc. It has also become the trend for vehicles to perceive their environments using CNNs. John McCormac et al. [14] fused the CNN with ElasticFusion SLAM to build an annotated 3D map using a RGB-D camera. However, the RGB-D camera doesn't meet the demand of the outdoor environment because of the perception distance limitation. Furthermore, depth is easily affected by the level of illumination. Our approach is like [14], we obtain the pixel-wise semantic information by using new fully convolutional networks which can directly use the depth information generated from the fourth thread of Stereo-ORB-SLAM. Furthermore, fully CNNs are suitable for extensive environments. In addition, this is different from the work [14] because we use the Stereo-ORB-SLAM system to meet the requirements of the urban environment. Compared to the other similar SLAM system presented above, the ORB-SLAM system shows better robustness, higher accuracy , and a lower computational cost. In addition, our semantic SLAM system also has the relocalization and loop-closing functions to accurately solve the problem of the localization of vehicles. It is suitable for the intelligent vehicles' environmental perception, and it is also valuable to other researchers of intelligent vehicles.

III. Method Overview

Our semantic SLAM pipeline is composed of two units: a real-time Stereo-ORB-SLAM system which obtains the geometric information from the environment and a fully Convolutional Neural Network system to obtain the semantics of the scenes. The whole system pipeline is illustrated in Fig. 2.

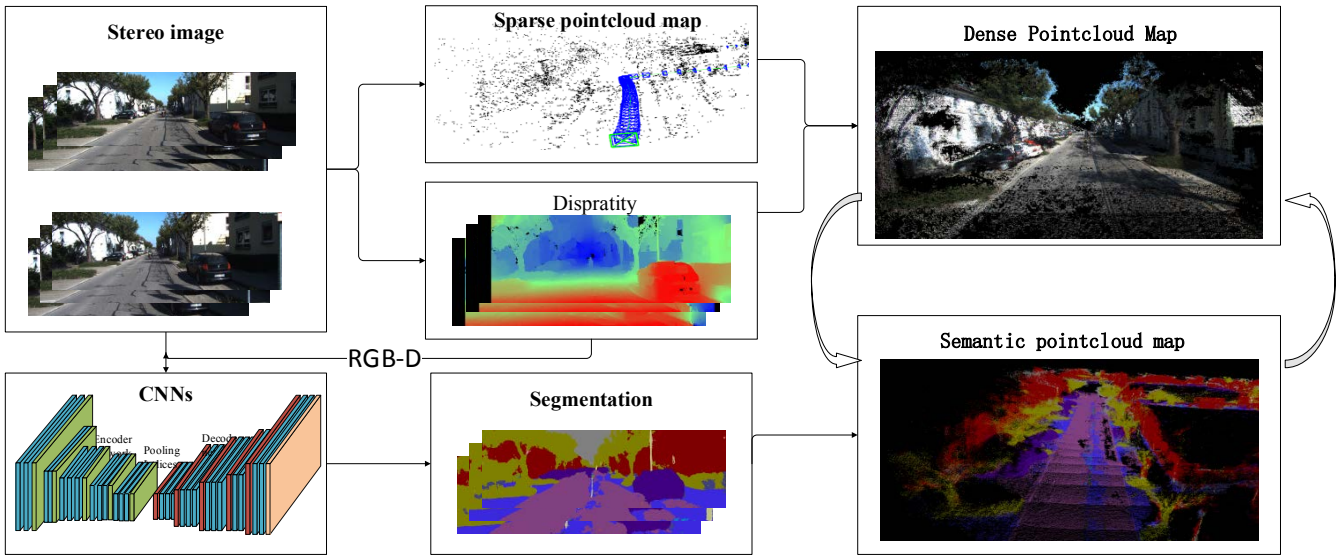


Fig. 2. System pipline

At first, Stereo–ORB- SLAM calculates the pose of each frame and builds the sparse map by tracking, local mapping, and the loop closing third thread. Then, in the fourth thread, the disparity of the keyframes is generated using the semi-global stereo matching algorithm. We make use of the pose and disparity of the keyframes to build the dense 3D point cloud map. Subsequently, the CNNs system takes the RGB-D as input and returns a pixel-wise segmentation image. Finally, we fuse the 3D point cloud map with the semantics of each point to obtain the 3D point cloud map. We transform the 3D point cloud map to a 3D octomap [15] which is more flexible and compact.

IV. Semantic Stereo-ORB-SLAM System

The ORB-SLAM [4] is carefully extended for our method since it is robust to camera motion and motion blur. The extended Stereo-ORB-SLAM system has four main threads based on ORB [16] features: tracking, local mapping, loop closing, and semantic mapping. The first three threads are same as ORB-SLAM [4], and we build the new semantic mapping thread to build the semantic map. The first three threads are mainly used to gain the accurate pose Twc (W denotes the World frame and C denotes camera frame) of every frame and optimize it. The initial pose is estimated in the tracking thread by a velocity model prediction or matching the features of the frame with local map. Subsequently, several steps are taken to optimize the pose. These steps consist of motion-only bundle adjustment [17] (BA), local BA in the local mapping thread, and full BA after loop closing. The accurate pose of camera is the key to the basics of mapping and relocalization. A survival strategy selects the points and keyframes that lead to unnecessary redundancy and low computational cost. We also choose the keyframe as the representation of the environment for dense 3D mapping. The details of the first three threads are described in [4] [18].

In the fourth thread we apply a semi-global stereo matching algorithm to obtain a smooth disparity map of the keyframe. Next, we gain the dense 3D point cloud map from back projecting the disparity with the pose of the keyframes. Finally, considering the huge storage cost of reusing the point cloud map, we transform the point cloud map to an octomap which uses a probabilistic occupancy estimation based on octrees. When we obtain the dense 3D map, we fuse the semantic information from the segmentation into a map to obtain the semantic map. The following section outlines each of the components in more detail.

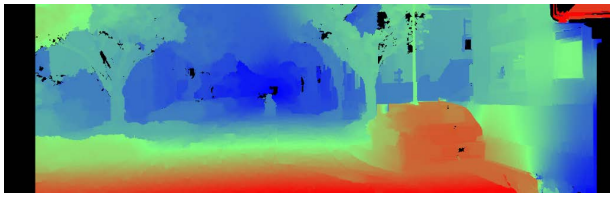
A. Obtaining the Disparity Map

We choose the keyframes selected in the first three threads as the representation of the environment to build the main structure of the world. Doing it this way is efficient and helps avoid unnecessary redundancy leading to low computational costs. This is especially the case when vehicles operate at the same place since the completed map will not have needless expansion. Therefore, in the fourth thread we only operate the keyframes to get its disparity and segmentation results.

Disparity is obtained though the semi-global stereo matching step. In addition, we make use of fast global image smoothing [19] to optimize the coarse disparity and make thedepth more continuous. These are the basic steps of building the globally consistent map. Fig. 3 shows the final matching results for a pair of stereo vision images in the KITTI dataset. In the disparity map, a larger gray value means there is a larger disparity value. We also obtain the colorful disparity for better visualization. Excellent disparity is the basis of the 3D reconstruction, and in Section 5 the depth image is also obtained through this semi-global matching algorithm.

(a) RGB

(b) Disparity



(c) Colorful disparity

Fig. 3 Disparity map

B. Octomap

The point cloud map has three problematic issues: 1) the format of the map is not compact enough, 2) errors of the pose lead to overlap of the map, and 3) the point cloud map is not effective for navigation. Octomap can represent full 3D models including free and unknown areas in a volumetric way , and furthermore, it is updatable, flexible, and compact. Octomap uses a tree-based representation to offer maximum flexibility regarding the mapped area and resolution. It performs a probabilistic occupancy estimation to ensure updatability and for coping with camera noise. An octree [20] is a hierarchical data structure composed of a voxel. Each voxel is recursively subdivided into eight subvoxels until a given minimum voxel size is reached. The minimum voxel size determines the resolution of the map. The probability $P(n|z_{1:t})$ of the leaf node being occupied given the measurements $z_{1:t}$ is estimated according to Function (1):

$$P(n|z_{1:t}) = \left[1 + \frac{1-P(n|z_t)}{P(n|z_t)} \frac{1-P(n|z_{1:t-1})}{P(n|z_{1:t-1})} \frac{P(n)}{1-P(n)}\right]^{-1} \quad (1)$$

C. Segmentation

In previous work [5], we proposed fully Convolutional Neural Networks (CNN) architecture for traffic scene segmentation. In work [5] we also compared our CNNs with

Segnet [21] and found it achieves better real-time performance at 22ms per image and has a competitive segmentation accuracy of 73.1%. In this work, we also take RGB-D as the input of CNNs [5] to gain the semantic information of each pixel. The network architecture consists of an encoder network and a corresponding decoder network. The details of the network architecture and training methods are introduced more clearly in our previous work [5]. The network is trained and tested using the deep learning framework Caffe [22]. The depth image is obtained by the stereo matching mentioned above. The segmentation results are fused into the dense 3D map to obtain the semantic map.

V. Experiments And Analysis

A. Network training

CNNs are computed through the GPU and others through the CPU. The software and hardware configurations are shown in Table I.

TABLE I
COMPUTER CONFIGURATION

Project	Content
CPU	Intel Xeon E5-2620
RAM	32GB
GPU	GeForce GTX TITAN X
Operating System	Ubuntu 14.04 LTS
Cuda	Cuda7.5 with Cudnn v5
Data Processing	Python 2.7, C++, etc.
Deep Learning Framework	Caffe

Cityspace data set is comprised of a large, diverse set of stereo video sequences recorded in streets from 50 different cities. 5,000 of these images have high quality pixel-level annotations. This proves that we can obtain the disparity using the stereo matching mentioned above and test the performance of CNNs using RGB-D image as input. We divide a traffic scene into the 10 dominant classes mentioned above. The original dataset is resized from 2048×1024 resolution to 400×200 resolution. This helps to avoid high memory costs at training time. The generated dataset consists of 2,975 RGB-D images in the training set, 500RGB-D images in the validation set, and 1,525 RGB-D images in the test set.

We use the cross-entropy loss function [23] as the objective function for training the network. The loss is summed up over all the pixels in a mini-batch (10 images). The segmentation accuracy of the model is tested on the validation set after each round of 300 iterations until training loss converges. To verify the influence of the introduction of the disparity map on segmentation results, we use the RGB images corresponding to the RGB-D images as input to train and test the network. The training loss and validation accuracy curve up to 10,000 iterations are shown in Fig. 4. We obtain competitive global accuracy of 73.1% and the accuracy of every class is showed in work [5]. Fig. 4 clearly show when using the RGB-D images the validation accuracy is higher than that obtained when using the RGB images. In addition, the training loss using RGB-D images is smaller than that when using RGB images. These results clearly show that the disparity maps help improve the segmentation results. In addition, both roads and vehicles have a big influence on vehicle navigation. Our CNNs show better performance in vehicle and road classification, with accuracies at 88.7% and 91.6% respectively. We use the trained CNNs to obtain the pixel-wise semantic information of the KITTI dataset for the semantic SLAM. The experimental results indicate our network performs generalization well.

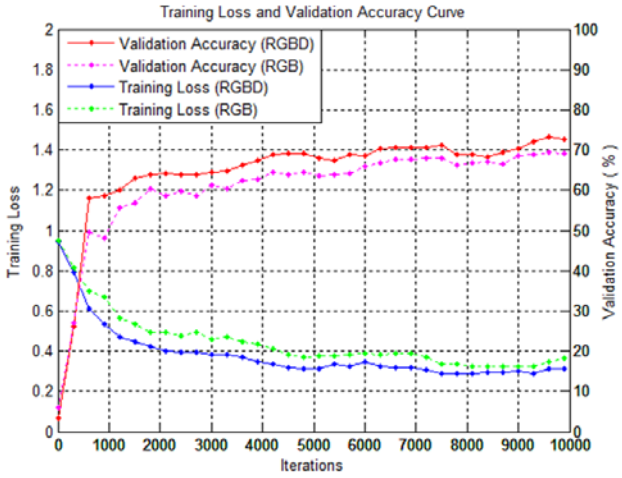


Fig. 4. Training loss and validation accuracy curve

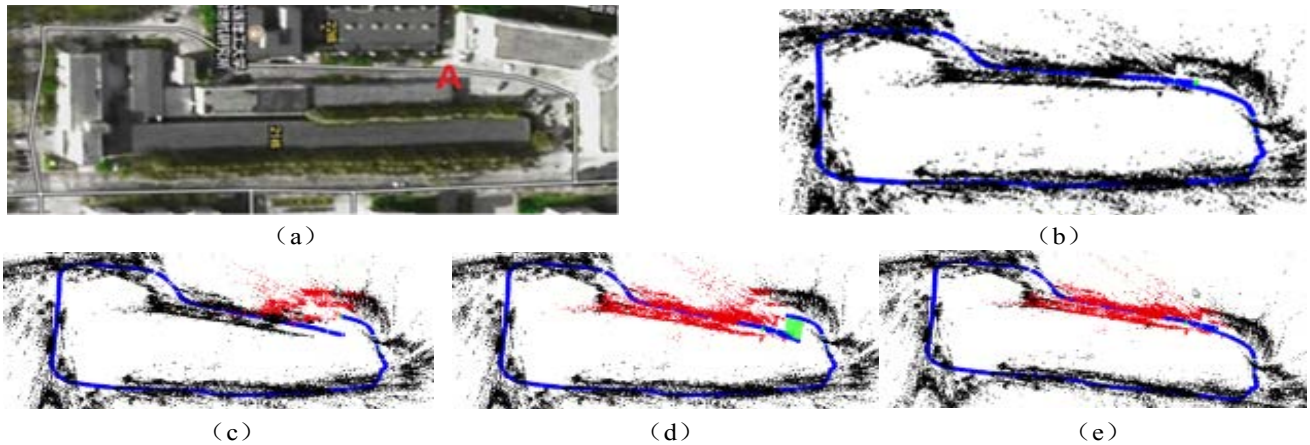


Fig. 5. Loop Closing Results: Fig. (a) is the satellite map of our experiment area. The vehicle with the stereo camera starts at place A and runs a circle counterclockwise to come back to place A. Fig. (b) is the sparse point cloud map and the blue line is the trajectory of the SLAM system，dark points are the feature points saved in map, red points are the feature points of the current frame. Drifting errors cause the trajectory to not be closed in Fig. (c). Loop closing eliminates the drifting errors in Fig. (d) and Fig. (e).

B. *Loop Closing*

Loop Closing [24] plays an important role in eliminating the drifting errors of the SLAM system because of camera observation errors and calculation errors. Loop closing is essentially one part of the semantics. It makes the vehicles understand the world consistently rather than as an endless corridor. Furthermore, it also makes vehicles remember the places where they have gone previously so that they can eliminate the drifting errors when the car comes back to the same place again. We take the stereo camera Fpgalena to capture the stereo image at the school in 30Hz and obtain the epipolar rectified image using Bouguet algorithm. We find the parameters of the stereo camera using Zhang Zhengyou calibration [25]. The original image is 1280 X 640 resolution, and we cut the image into 1240 x 360 resolution to obtain the necessary effective information and decrease the computational cost. We test our SLAM system on this data to obtain the trajectory of the car, illustrated in Fig. 5. In our experiment, the polar line of stereo image is not completely aligned because of the camera calibration errors which lead to the depth of the matching points not being accurate enough. These are the main things that cause the drifting errors in Fig. 5 (c). However, in Fig. 5 (d), loop closing is detected, and loop closing eliminates most of drifting errors making the trajectory closing, such as in Fig. 5 (e). This is a significant step for vehicles to use the semantic information.

C. *Semantic map*

The stereo sequences in the KITTI dataset are captured in urban and highway environments. The stereo sensor has a 54cm baseline and works at 10Hz. We run our system on the sequence 00 and obtain a semantic map such as the one in Fig. 6. The 00 sequence contains 4,541 stereo images, and it has many loops. It is representative and challenging to the large-scale scenes for the SLAM system. In our system, the mean tracking time of every frame is 0.1024s, and the entire speed of the system is about 10Hz. The semantic map is saved in the octomap format for better application. In addition, we also show our results using the point cloud map for better visualization, such as in Fig 7. The segmentation results indicate our trained CNNs perform generalization well. The semantic maps which are closed to the real street scenes indicate our system can build a globally consistent dense 3D semantic map in large-scale outdoor environments. The strong definition and straight contours of the cars and trees prove the high accuracy localization of our system.

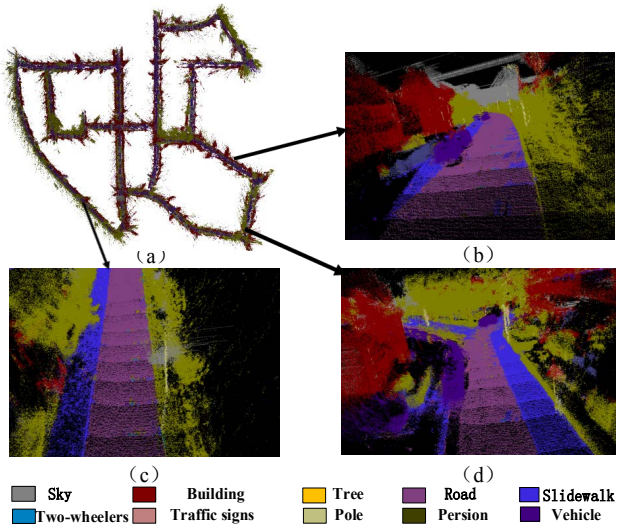


Fig.6. Semantic maps. Fig. (a) is the globally consist semantic map of the KITTI sequence 00 saved as the octomap. Figs. (b) (c) and (d) are the details of the map directed by the arrows, they are showed using point cloud format for better visualization.

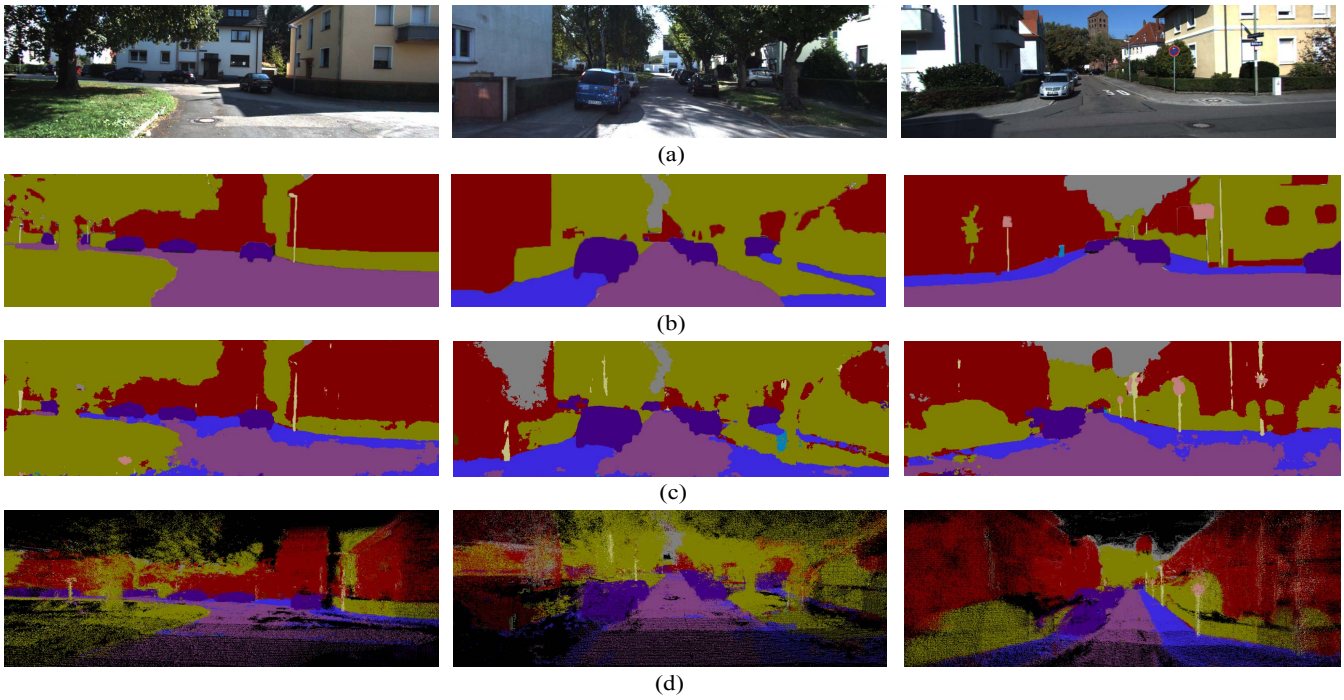


Fig. 7. Semantic point cloud maps: Subfigure (a) is the original set of color images of the test samples. The ground truth corresponding to the color images are shown in subfigure (b). Subfigure (c) shows the segmentation results using our trained CNNs. Subfigure (d) shows the dense 3D point cloud map generated using our Semantic SLAM.

VI. CONCLUSION

In this paper, we propose a semantic SLAM based on stereo vision and CNNs for intelligent vehicle localization and environment perception. A four thread Stereo-ORB- SLAM system for dense semantic mapping is presented to calculate the pose of the vehicles and build a 3D octomap. We take the RGB-D as the input of our CNNs to get semantic information of each pixel. Finally, we fuse the geometric and semantic information to obtain the semantic map. We test the performance of the proposed semantic SLAM system using the KITTI dataset. Results indicate the system was effective and able to build the semantic map in real time. The speed of the entire system is about 10Hz. In addition, using our dataset captured with the stereo camera Fpgalena, we proved the universality and validity of the system.

We believe this is just the start of how geometrical information from SLAM and segmentation can be brought together to enable truly powerful semantic and object-aware mapping. In essence, combining semantics and geometry is a win-win relationship. Semantics can help SLAM tackle the data association problem and decrease dependency on image features. Pose and a globally consistent map also help with segmentation. In future work, we will make deeper level discussions about semantic SLAM. Furthermore, by using semantic information, we will decrease the wrong matching points of the tracking thread and make up for the unavailable depth through stereo matching.

REFERENCES

[1] Waymo company, "On the road to fully self-driving," *https://news.ycombinator.com/from?site=googleapis.com*, 2017.

[2] R.H. Zhang, Z.C. He, H.W. Wang, F. You and K.N. Li, "Study on self-tuning tyre friction control for developing main-servo loop integrated chassis control system," IEEE Access, vol. 5, pp. 6649-6660, 2017.

[3] Yang, D., Kurt, A., Redmill, K., and Özgüner, Ümit. . "Agent-based microscopic pedestrian interaction with intelligent vehicles in shared space." *The, International Workshop* 2017:69-74.

[4] Mur-Artal, Raúl, and J. D. Tardós. "ORB-SLAM2: An Open-Source SLAM System for Monocular, Stereo, and RGB-D Cameras." *IEEE Transactions on Robotics* 33.5(2016):1255-1262.

[5] Li, Linhui, Qian, B., Lian, J., Zheng, W., and Zhou, Yl. "Traffic Scene Segmentation Based on RGB-D Image and Deep Learning." *IEEE Transactions on Intelligent Transportation Systems* PP.99(2017):1-6.

[6] Achmad, M. S. H., Findari, W. S., Ann, N. Q., Pebrianti, D., and Daud, M. R. "Stereo camera — Based 3D object reconstruction utilizing Semi-Global Matching Algorithm." *International Conference on Science and Technology-Computer* IEEE, 2017.

[7] Geiger, A., Lenz, P., Stiller, C., and Urtasun, R . "Vision meets robotics: The KITTI dataset." *International Journal of Robotics Research* 32.11(2013):1231-1237.

[8] Civera, J., Gálvez-López, D., Riazuelo, L., Tardós, J. D., and Montiel, J. M. M.. "Towards semantic SLAM using a monocular camera." *Ieee/rsj International Conference on Intelligent Robots and Systems* IEEE, 2011:1277-1284.

[9] Salas-Moreno, R. F., Newcombe, R. A., Strasdat, H., Kelly, P. H. J., and Davison, A. J.. "SLAM++: Simultaneous Localisation and Mapping at the Level of Objects." *IEEE Conference on Computer Vision and Pattern Recognition* IEEE Computer Society, 2013:1352-1359.

[10] A Kundu, Y Li , F Dellaert , F Li , and JM Rehg l. "Joint Semantic Segmentation and 3D Reconstruction from Monocular Video." *European Conference on Computer Vision* Springer, Cham, 2014:703-718.

[11] Krizhevsky, Alex, I. Sutskever, and G. E. Hinton. "ImageNet classification with deep convolutional neural networks." *Communications of the Acm* 60.2(2012):2012.

[12] Simonyan, Karen, and A. Zisserman. "Very Deep Convolutional Networks for Large-Scale Image Recognition." *Computer Science* (2014).

[13] Szegedy, Christian, et al. "Going deeper with convolutions." Computer Vision and Pattern Recognition IEEE, 2015:1-9.

[14] Mccormac, J., Handa, A., Davison, A., and Leutenegger, S. "SemanticFusion: Dense 3D Semantic Mapping with Convolutional Neural Networks." (2016).

[15] Wurm, K. M., Hornung, A., Bennewitz, M., Stachniss, C., and Burgard, W. "OctoMap: A probabilistic, flexible, and compact 3D map representation for robotic systems." Proc. of the ICRA Workshop on Best Practice in 3D Perception and Modeling for Mobile Manipulation 2010.

[16] E Rublee , V Rabaud , K Konolige, and G Bradski "ORB: An efficient alternative to SIFT or SURF." *IEEE International Conference on Computer Vision* IEEE, 2012:2564-2571.

[17] Bill Triggs, Philip F. McLauchlan, Richard I. Hartley, and Andrew W. Fitzgibbon,"Bundle adjustment: a modern synthesis," in *Vison algorithms*: *theory and practice*, pp. 298-372, Springer, 2000.

[18] Mur-Artal, Raúl, J. M. M. Montiel, and J. D. Tardós. "ORB-SLAM: A Versatile and Accurate Monocular SLAM System." *IEEE Transactions on Robotics* 31.5(2015):1147-1163.

[19] D. Min, S. Choi, J. Lu, and B. Ham, "Fast Global Image Smoothing Based on Weighted Least Squares," *IEEE Transactions on Image Processing A Publication of the IEEE Signal Processing Society,* vol. 23, no. 12, pp. 5638-5653, 2014.

[20] Meagher, Donald. "Geometric modeling using octree encoding." *Computer Graphics & Image Processing* 19.2(1982):129-147.

[21] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation,"*Computer Science*, 2015.

[22] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional Architecture for Fast Feature Embedding," in *Proc. 22nd ACM Int. Conf. Multimedia,* Nov. 2014, pp. 675-678.

[23] E. Shelhamer, J. Long, and T. Darrell, "Fully Convolutional Networks for Semantic Segmentation," vol. 79, no. 10, pp. 1337-1342, 2014.

[24] Galvez-López D, Tardos J D. Bags of Binary Words for Fast Place Recognition in Image Sequences[J]. IEEE Transactions on Robotics, 2012, 28(5):1188-1197..

[25] Zhngyou Zhang. A flexible new technique for camera calibration [J]. Technical Report MSR-TR-98-71, Microsoft Research, December 1998.