CNN-BASED PLACE RECOGNITION TECHNIQUE FOR LIDAR SLAM

Y. Yang *, S. Song, C. Toth

Dept. of Civil, Environmental and Geodetic Engineering, The Ohio State University, 470 Hitchcock Hall, 2070 Neil Avenue Columbus, OH 43210, USA - (yang.2695, song.1634, toth.2)@osu.edu

KEY WORDS: Lidar, Global descriptor, Deep learning, Place recognition

ABSTRACT:

Place recognition or loop closure is a technique to recognize landmarks and/or scenes visited by a mobile sensing platform previously in an area. The technique is a key function for robustly practicing Simultaneous Localization and Mapping (SLAM) in any environment, including the global positioning system (GPS) denied environment by enabling to perform the global optimization to compensate the drift of dead-reckoning navigation systems. Place recognition in 3D point clouds is a challenging task which is traditionally handled with the aid of other sensors, such as camera and GPS. Unfortunately, visual place recognition techniques may be impacted by changes in illumination and texture, and GPS may perform poorly in urban areas. To mitigate this problem, state-of-art Convolutional Neural Networks (CNNs)-based 3D descriptors may be directly applied to 3D point clouds. In this work, we investigated the performance of different classification strategies utilizing a cutting-edge CNN-based 3D global descriptor (PointNetVLAD) for place recognition task on the Oxford RobotCar dataset[1].

1. INTRODUCTION

One important aspect of SLAM algorithms is that the localization errors keep accumulating as the number of measurements keeps increasing, due to the errors in measurements caused by the noise of sensors (Dhiman et al., 2015). To handle this problem, SLAM algorithms rely on place recognition (PR), or loop closure detection (LCD) techniques, wherein the algorithms are able to recognize previously visited places and then use them as additional constraints for increasing the precision of localization estimation and solving the global localization problem. Therefore, a robust PR scheme could enhance the robustness and performance of SLAM algorithms. For the Lidar-SLAM, PR is still a challenging task and very few of the state-of the art algorithms has solved the loop closure problem (Singandhupe et al., 2019). Many methods have been proposed for this task, and a traditional solution is sensor integration with other sensors, such as camera (Olson, Edwin, 2009a) (Wu et al., 2016) or GPS (Emter, Thomas, 2012) (Emter et al., 2018). However, these techniques face challenges, such as vision based methods suffering from illumination changes, season-to-season based appearance changes and viewpoints differences, and poor GPS performance in urban areas.

Since Lidar data is invariant to lighting and appearance changes, the geometric methods for PR with 3D Lidar data, such as line feature-based scan matching, key point matching and 3D local feature-based strategies are widely investigated (Olson, Edwin, 2009b) (Bosse et al., 2013) (Dubé et al., 2017). Unfortunately, extracting and matching these features could be difficult in certain environments. To that end, CNN-based solutions have recently been proposed as effective learning tools to generate features from general environments. Due to the different ways to learn and extract descriptors, these solutions can be classified in two categories: semantic (local) level feature-based (Dubé et al., 2018) and frame (global) level feature-based (Angelina, Hee Lee, 2018) (Yin H et al., 2018) (Yin H et al., 2019) (Yin P et al., 2018a) (Yin P et al., 2018b). The major limitation for extracting semantic features is the assumption that there are enough static objects which have been adequately learned by the pretrained CNN model. However, this assumption may not always be satisfied in real-world practice. On the other hand, with the global descriptor, the PR task is handled as a similarity modeling problem in which Nearest Neighbor (NN) method is commonly used for classification. Additionally, one interesting task in the real-world PR practice is classification under the restriction that we may only observe a single example of each possible scenario before making a prediction about a test instance. This problem is known as one-shot learning (Koch et al., 2015), and the Siamese neural networks have been demonstrated as an effective solution for one-shot learning in imagery application (Yin W et al., 2015) and low dimensional 3D semantic segment descriptors classification (Cramariuc et al., 2018).

To efficiently generate reliable PR candidates by improving the performance of classification network, in this study we investigated a one-shot learning classification method, the CNN-based Siamese network with high dimensional global descriptors on 3D Lidar data (Figure 1). In the experiment, we compared the effectiveness of classification between our CNN-based classifier, a commonly used nearest neighbor (NN) method and random forests (RF) which is a typical nonlinear classic machine learning classifier. The details of proposed method are discussed in the remainder of this paper, structured as follows. Section 2 reviews the proposed method, including network for global feature descriptor extraction and CNN classifier model. The experiments, including training, testing and performance comparison are presented in Section 3. Finally, the conclusions are summarized in Section 4.

* Corresponding author

[1] https://robotcar-dataset.robots.ox.ac.uk/

2. METHOD

2.1 Global Descriptor

Compared to its image counterpart, applying a CNN model to 3D points is more challenging due to the fact points in a point cloud are generally unordered. Some works handled this challenge by projecting 3D point clouds into 2D image plane (Su et al., 2015) (Yin P et al., 2018b) or transforming point clouds into 3D volumetric representations (Qi et al., 2016) (Yin P et al., 2018a). The downside of these networks is that they cannot handle well the large-scale outdoor PR problems. Additionally, for these networks, Lidar data need to be preprocessed to provide proper input which is computationally expensive. To directly operate on an unordered points subset in a point cloud, (Angelina, Hee Lee, 2018) proposed the PointNetVLAD[2] network which integrates PointNet network and VLAD layer (Figure 2).

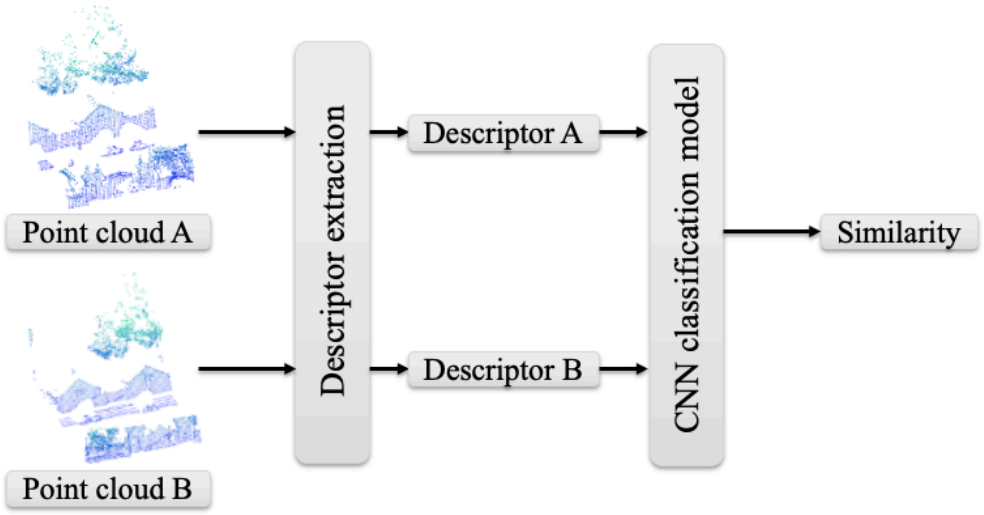


Figure 1. Proposed network for PR task

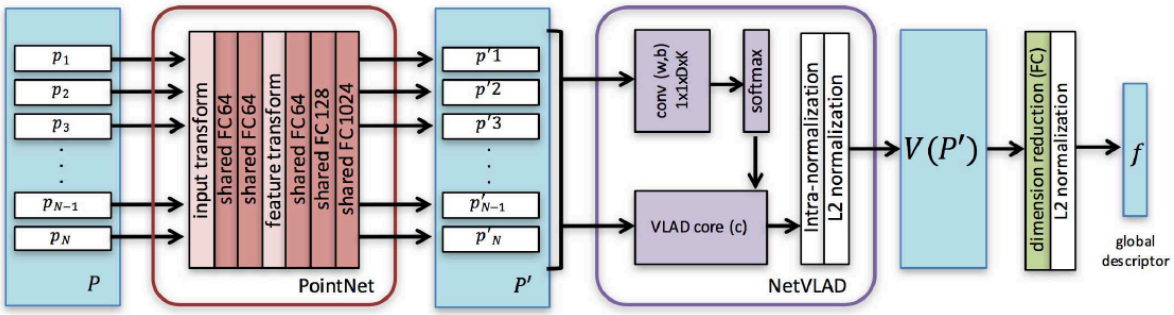


Figure 2. Network architecture of PointNetVLAD (Angelina, Hee Lee, 2018)

The PointNet extracts local feature descriptors for each input point by encoding points into vectors in a higher dimensional space. In the next phase, the NetVLAD layer aggregates local features into the VLAD bag-of-words (BoWs) global feature descriptor vectors. Additionally, since NetVLAD is a symmetric function and PointNet model transforms each point in the point cloud independently, the output global descriptor is invariant to the order of the points. In the training process, PointNetVLAD was trained with the lazy quadruplet loss in which the Euclidean distances between descriptors are used for calculating similarity. During inference (testing), NN method was used for classification. The lazy quadruplet loss is defined as:

$$L = \max_j \left(\left[\alpha + \delta_{pos} - \delta_{neg_j} \right]_+ \right) + \max_k \left(\left[\beta + \delta_{pos} - \delta_{neg_k^*} \right]_+ \right) \quad (1)$$

where α, β = margin
δ = Euclidean distances between global descriptor vectors

In this work, pretrained PointNetVLAD baseline network was used as global feature extractor.

2.2 CNN-based Classifier

As depicted in Fig 1, two input point clouds are firstly given to two branches of the Siamese network which are distinct PointNetVLAD networks and create global descriptors. In the next stage, these two descriptors are combined and processed by a CNN classification model in which the similarity score is calculated as the final output. The structure of CNN classification model is detailed in Figure 3.

[2] https://github.com/mikacuy/pointnetvlad.git

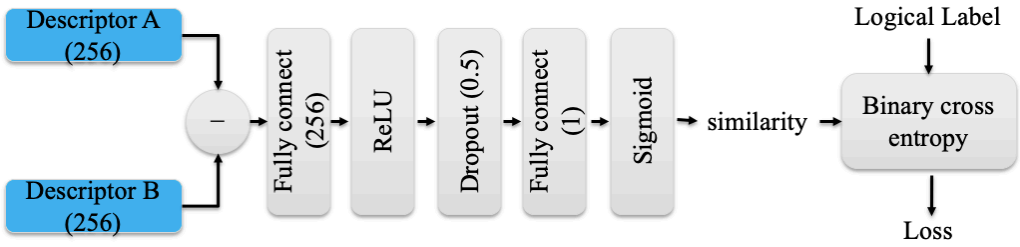


Figure 3. Network architecture of CNN classification model

One advantage of Siamese network is that the feature extraction model and classification model can be trained and used simultaneously or independently. In this work, we trained and used CNN classification model independently for a fair comparison of classification performance with respect to other classifiers. The binary cross entropy loss with stochastic gradient descent (SDG) is applied during training process. The loss function is formed as:

$$L = -t \log y - (1-t) \log(1-y) \quad (1)$$

where t = logic label
y = output of CNN classification model

3. EXPERIMENT

3.1 Training the Model

Since we only investigate the performance of classifiers in this work, the same training and test dataset, as used in original PointNetVLAD research, was applied to guarantee a consistent performance of feature extraction. The dataset was built from the Oxford RobotCar dataset (Maddern et al., 2017) in which 44 sets of full and partial runs were used. Training and testing reference maps are geospatially separated from each run with a proportion of 70% and 30%, respectively. Then submaps were segmented from reference maps following the rules: (1) each submap contains all Lidar points within a 20m trajectory of the vehicle, and (2) the intervals between submaps are 10m and 20m for training and testing datasets, respectively. Finally, 21,711 training submaps and 3,030 testing submaps were segmented out from original dataset. The submaps within 10m intervals in centroid coordinates are seen as structurally similar and labelled as "positive" and those with 50m are dissimilar and "negative".

The training results for the CNN classification model are shown in Figure 4. The loss quickly converged during the first few epochs and remains almost constant in subsequent epochs. Thus, we stopped training at 1000 epochs. The CNN classification model was trained on a Nvidia GeForce Titan Xp GPU.

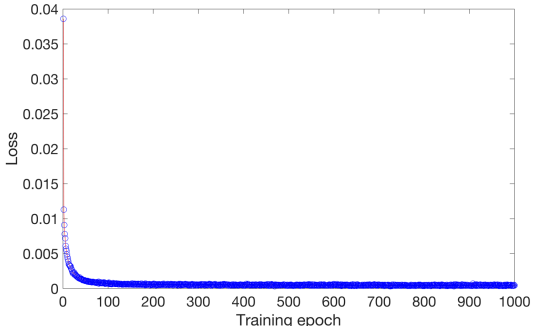


Figure 4. Training results of CNN classification model

3.2 Comparing Different Classification methods

In this section we compare the performance of different classification methods, e.g. CNN-based classifier, NN and RF by using the same set of global feature descriptors extracted from testing datasets. In terms of training the RF, the input is two concatenated descriptors and output is their matching probability. The closest neighbour in NN method is decided based on the Euclidian distance in descriptor vector space. The receiver operating characteristic (ROC) curves of the different classifiers are shown in Figure 5. The best accuracy is achieved by the CNN-based classifier. The numerical results are presented in Table 1. CNN-based classifier outperforms NN and RF in both general accuracy and true positive (recall) rate which are 95.3% and 70.1% respectively. Note that the commonly used NN method performs the worst, achieving a general accuracy of 61.1%, and true positive rate of 40.7%. The classic nonlinear classifier RF, used as a reference, outperforms NN with 88.5% in general accuracy and 57.6% in true positive.

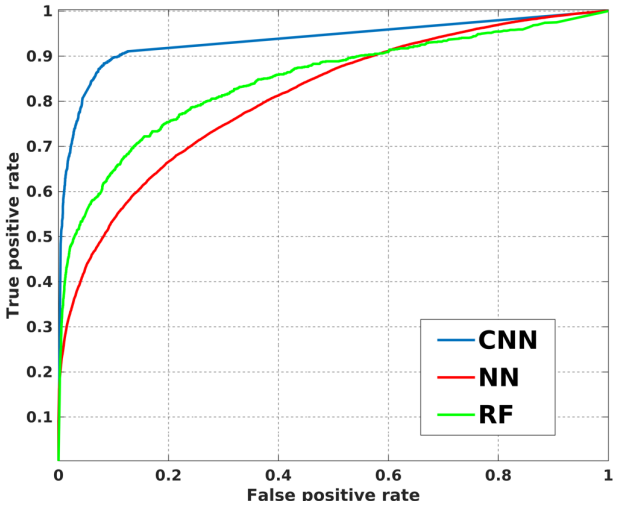


Figure 5. ROC curves for different classifiers with same set of global feature descriptors

Method	General Accuracy	True Positive Rate
NN	61.1%	40.7%
RF	88.5%	57.6%
CNN-based classifier	95.3%	70.1%

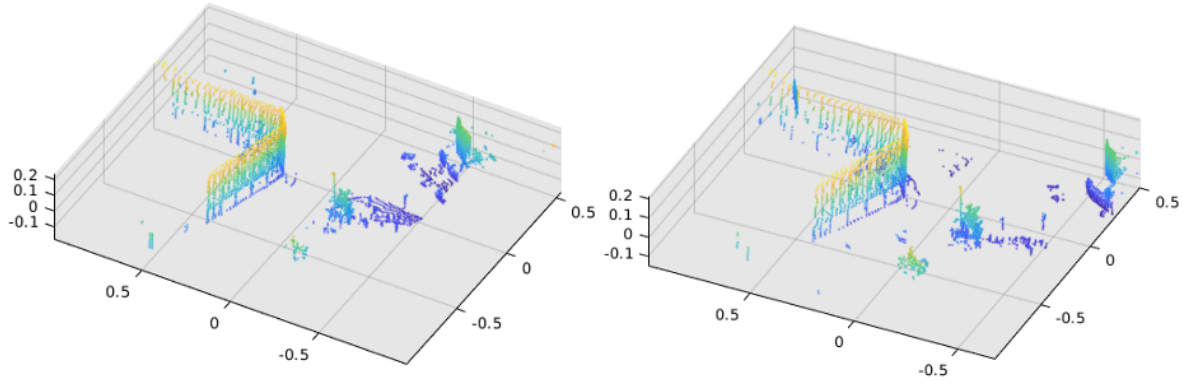
Table 1. Matching Accuracy statistics of different classifiers

Examples visualizing the matching results are presented in Figure 6 in which (a)~(c) are true positive matching, (d)~(f) are false positive matching. It can be seen that the proposed method shows robustness to noise, such as objects changing (Fig 6(a)),

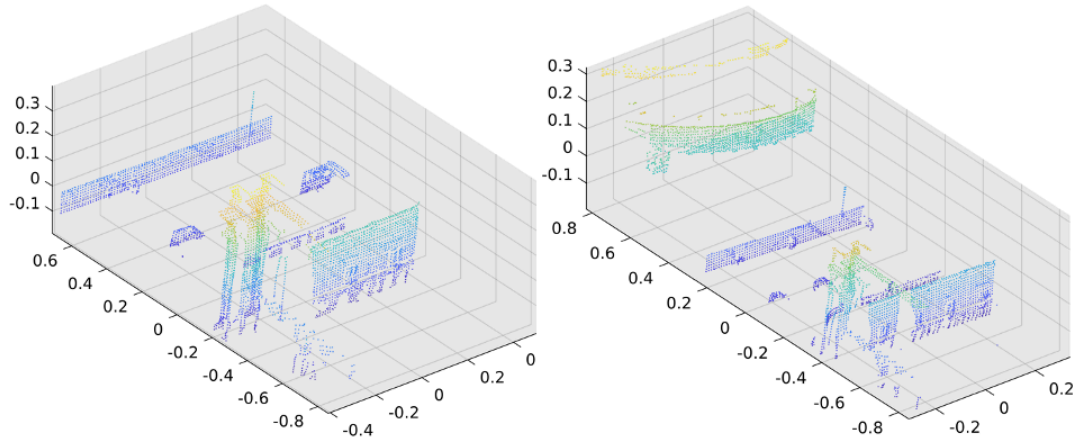
viewpoint changing (Fig 6(b)) and both objects and viewpoint changing (Fig 6(c)). On the other hand, Fig 6(d)~(f) reveal the fact that the CNN based classifier fails to distinguish dissimilar submap pairs when two scenarios contain very similar features, such as semblable building structures and trees.

4. CONCLUSION

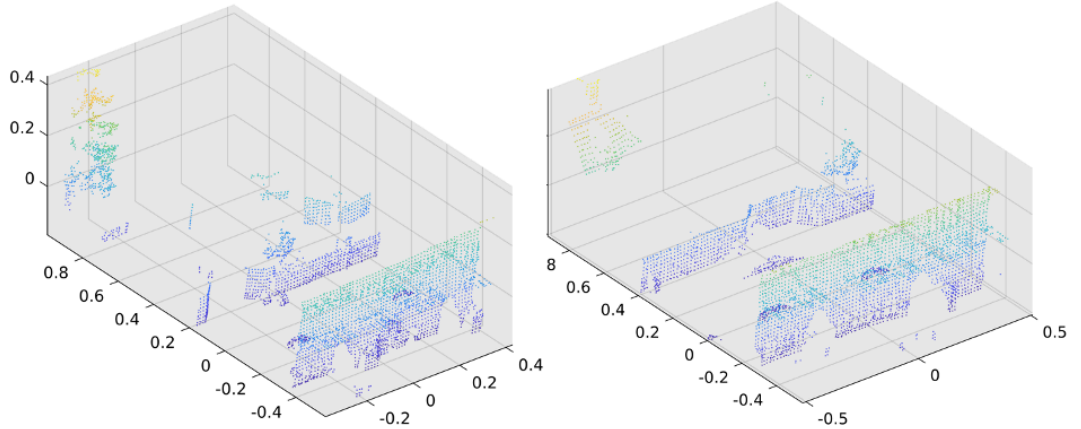
In this work, we investigated the performance of CNN based classifier with Lidar data for PR task. The testing results show that the proposed model outperforms both NN and RF methods and achieves true positive rate at 70.05%. However, many false matchings occur when scenarios contain very similar features. In the future work, we will try to (1) increase the performance in recall by using geometric or other constrains to reject false matches, and (2) integrate the proposed place recognition method into Lidar SLAM.



(a)



(b)



(c)

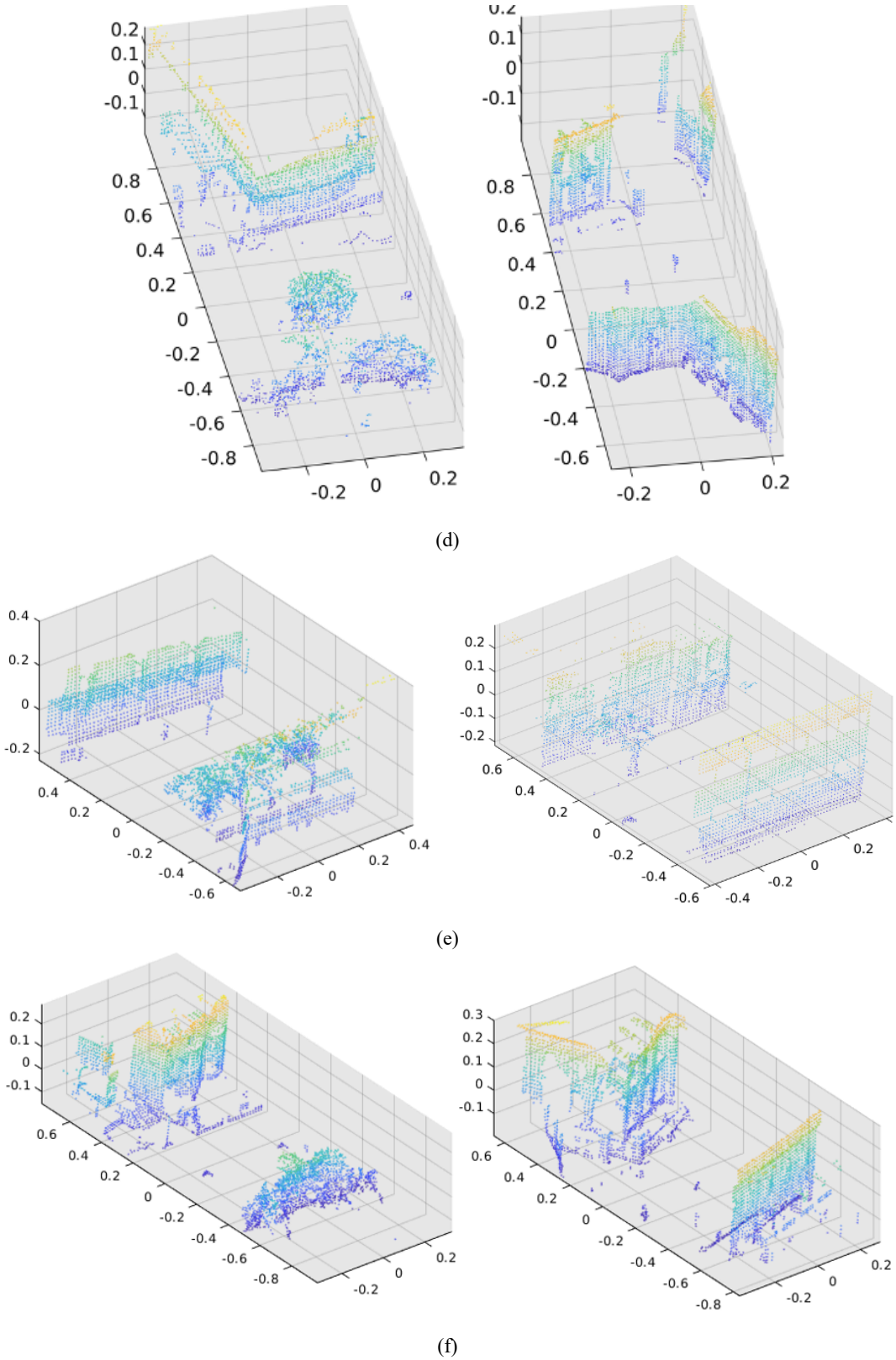


Figure 6. Examples of matching results, (a)~(c) are true positive matching, (d)~(f) are false positive matching.

REFERENCES

Angelina Uy, M. & Hee Lee, G. 2018, "Pointnetvlad: Deep point cloud based retrieval for large-scale place recognition", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4470.

Bosse, M. & Zlot, R. 2013, "Place recognition using keypoint voting in large 3D lidar datasets", *2013 IEEE International Conference on Robotics and Automation* IEEE, pp. 2677.

Cramariuc, A., Dubé, R., Sommer, H., Siegwart, R. & Gilitschenski, I. 2018, "Learning 3d segment descriptors for place recognition", *arXiv preprint arXiv:1804.09270*.

Dhiman, N.K., Deodhare, D. & Khemani, D. 2015, "Where am I? Creating spatial awareness in unmanned ground robots using SLAM: A survey", *Sadhana,* vol. 40, no. 5, pp. 1385-1433.

Dubé, R., Cramariuc, A., Dugas, D., Nieto, J., Siegwart, R. & Cadena, C. 2018, "SegMap: 3d segment mapping using data-driven descriptors", *arXiv preprint arXiv:1804.09557*.

Dubé, R., Dugas, D., Stumm, E., Nieto, J., Siegwart, R. & Cadena, C. 2017, "Segmatch: Segment based place recognition in 3d point clouds", *2017 IEEE International Conference on Robotics and Automation (ICRA)* IEEE, pp. 5266.

Emter, T. 2012, "Integrated Multi-Sensor Fusion and SLAM for Mobile Robots", *Proceedings of the 2011 Joint Workshop of Fraunhofer IOSB and Institute for Anthropomatics, Vision and Fusion Laboratory*KIT Scientific Publishing, pp. 91.

Emter, T. & Petereit, J. 2018, "Stochastic cloning and smoothing for fusion of multiple relative and absolute measurements for localization and mapping", *2018 15th International Conference on Control, Automation, Robotics and Vision (ICARCV)*IEEE, pp. 1508.

Koch, G., Zemel, R. & Salakhutdinov, R. 2015, "Siamese neural networks for one-shot image recognition", *ICML deep learning workshop*, Anonymous Lille.

Maddern, W., Pascoe, G., Linegar, C., Newman, P., 2017. 1 year, 1000 km: The Oxford RobotCar dataset. *The International Journal of Robotics Research* 36 (1), 3-15.

Olson, E. 2009a, "Recognizing places using spectrally clustered local matches", *Robotics and Autonomous Systems,* vol. 57, no. 12, pp. 1157-1172.

Olson, E.B. 2009b, "Real-time correlative scan matching", *2009 IEEE International Conference on Robotics and Automation*IEEE, pp. 4387.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R. & Dubourg, V. 2011, "Scikit-learn: Machine learning in Python", *the Journal of machine Learning research,* vol. 12, pp. 2825-2830.

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

Singandhupe, A. & La, H. 2019, "A review of slam techniques and security in autonomous driving", 2019 Third IEEE International Conference on Robotic Computing (IRC)IEEE, pp. 602.

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

Wu, Q., Sun, K., Zhang, W., Huang, C. & Wu, X. 2016, "Visual and LiDAR-based for the mobile 3D mapping", *2016 IEEE International Conference on Robotics and Biomimetics (ROBIO)*IEEE, pp. 1522.

Yin, H., Tang, L., Ding, X., Wang, Y. & Xiong, R. 2018, "LocNet: Global localization in 3D point clouds for mobile vehicles", *2018 IEEE Intelligent Vehicles Symposium (IV)*IEEE, pp. 728.

Yin, H., Wang, Y., Ding, X., Tang, L., Huang, S. & Xiong, R. 2019, "3d lidar-based global localization using siamese neural network", *IEEE Transactions on Intelligent Transportation Systems,* vol. 21, no. 4, pp. 1380-1392.

Yin, P., He, Y., Xu, L., Peng, Y., Han, J. & Xu, W. 2018a, "Synchronous adversarial feature learning for lidar based loop closure detection", *2018 Annual American Control Conference (ACC)*IEEE, pp. 234.

Yin, P., Xu, L., Liu, Z., Li, L., Salman, H., He, Y., Xu, W., Wang, H. & Choset, H. 2018b, "Stabilize an unsupervised feature learning for LiDAR-based place recognition", *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*IEEE, , pp. 1162.

Yin, W. & Schütze, H. 2015, "Convolutional neural network for paraphrase identification", *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp.901.