Moving Object Segmentation for LiDAR-based Localization and Mapping

Xieyuanli Chen
University of Bonn
Email: xieyuanli.chen@igg.uni-bonn.de

I. Motivation And Related Work

The ability to identify which parts of the environment are static and which ones are moving is key to safe and reliable autonomous navigation. It supports the task of predicting the future state of the surroundings, collision avoidance, and planning. This knowledge can also improve and robustify pose estimation, sensor data registration, and simultaneous localization and mapping (SLAM). Thus, accurate and reliable moving object segmentation (MOS) in sensor data at frame rate is a crucial capability supporting most autonomous mobile systems. Depending on the application domain and chosen sensor setup, moving object segmentation can be a challenging task. While there has been a large interest in vision-based moving object segmentation [27, 23, 20] and approaches combining vision and LiDAR sensors [33, 22], we concentrate here on approaches using only LiDAR sensors. Below, we distinguish between map-based and map-free approaches.

Map-based approaches. Most of the existing LiDAR-based approaches target the cleaning of a point cloud map. These methods mostly run offline and rely on a prebuilt map. Some methods use time-consuming voxel ray casting and require accurately aligned poses to clean the dense terrestrial laser scans [12, 25]. To alleviate the computational burden, visibility-based methods have been proposed [21, 32]. These types of methods associate a query point cloud to a map point within a narrow field of view, e.g. cone-shaped used by Pomerleau et al. [21]. Recently, Pagad et al. [18] propose an occupancy map-based method to remove dynamic points in LiDAR scans. Kim et al. [14] propose a range-image-based method, which exploits the consistency check between the query scan and the pre-built map to remove dynamic points. Even though such map-based methods can separate moving objects from the background, they need a pre-built and cleaned map and therefore usually can not achieve online operation.

Map-free approaches. Recently, LiDAR-based semantic segmentation methods operating only on the sensor data have achieved great success [17, 8, 15]. Wang et al. [31] tackle the problem of segmenting things that could move from 3D laser scans of urban scenes, e.g. cars, pedestrians, and bicyclists. Ruchti and Burgard et al. [24] also propose a learning-based method to predict the probabilities of potentially movable objects. Dewan et al. [10] propose a LiDAR-based scene flow method that estimates motion vectors for rigid bodies. Based on that, they recently developed a semantic segmentation method [9], which exploits the temporally consistent information from the sequential LiDAR scans. Bogoslavskyi and Stachniss [2] propose a class-agnostic segmentation method for 3D LiDAR scans that exploits range images to enable online operation and leads to more coherent segments, but does not distinguish between moving and non-moving objects.

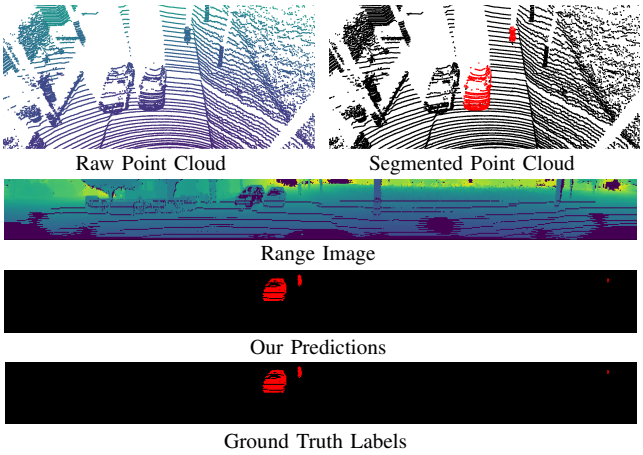
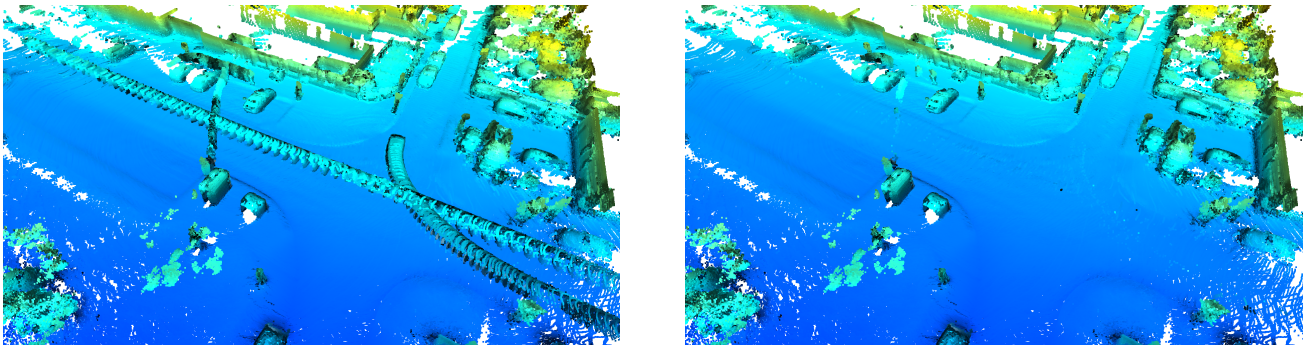


Fig. 1: Moving object segmentation using our current approach. Our method can detect and segment the currently moving objects given point cloud data exploiting its range projection. Instead of detecting all *potentially movable* objects such as vehicles or humans, our approach distinguishes between *actually moving* objects (labeled in red) and static or non-moving objects (black) in the upper row. At the bottom, we show the range image and our predictions in comparison to the ground truth labels.

II. Preliminary Results

Semantic information has been successfully used in multiple LiDAR-based applications, e.g., SLAM [6], loop closing [4] and localization [5]. Towards moving object segmentation, semantic segmentation can also be seen as a relevant step. Most existing semantic segmentation methods, however, only find *movable* objects, e.g. vehicles and humans, but do not distinguish between actually moving objects, like driving cars or walking pedestrians, and non-moving/static objects, like parked cars or building structures. There are also multiple 3D point cloud-based semantic segmentation approaches [29, 28, 26], which also perform well in semantic segmentation tasks. Among them, Shi et al. [26] exploit sequential point clouds and predict moving objects. However, based on networks operating directly on point clouds, these methods are usually heavy and difficult to train. Furthermore, most of them are both



(a) Raw point clouds

(b) Point clouds with moving segments removed

Fig. 2: Mapping results on KITTI Odometry dataset sequence 08, frame 3960-4070, where we show the accumulated point cloud (a) without removing segments and (b) when we remove the segments predicted as moving.

time-consuming and resource-intensive, which might not be applicable for autonomous driving.

Instead of using 3D point cloud-based approaches, we investigate the usage of range projection-based semantic segmentation methods to achieve real-time capability and operation beyond the frame rate of the LiDAR sensor. To distinguish the moving and non-moving object, we first combined the semantic segmentation with a SLAM method and proposed a semantic SLAM [6]. The proposed semantic SLAM method allows us to generate high-quality semantic maps, while at the same time improve the geometry of the map and the quality of the odometry by filtering out moving objects, like moving cars, while keeping static objects, like parked cars. To this end, we proposed a dynamic filter within the SLAM pipeline. We filter dynamics by checking semantic consistency between the new observation and the world model, when we update the map. If the labels are inconsistent, we assume those surfels belong to an object that moved between the scans. Therefore, we add a penalty term to the computation of the stability term in the recursive Bayes filter. After several observations, we can remove the unstable surfels. In this way, we achieve the detection of dynamics and finally the removal.

Though our semantic SLAM method distinguishes the moving and non-moving objects to improve odometry and mapping results, it needs to maintain a map. To achieve non-map-based MOS, we recently proposed a novel approach [7], which exploits sequential range images combined with a convolutional neural network. We investigate the usage of three recent range projection-based semantic segmentation methods proposed by Milioto et al. [17], Cortinhal et al. [8], and also ours [15] to tackle MOS with the prospect of real-time capability and operation beyond the frame rate of the LiDAR sensor. Our method does not rely on a pre-built map and operates online. We exploit residuals between the current frame and the previous frames as an additional input to the investigated semantic segmentation networks to enable class-agnostic moving object segmentation. Note that the proposed method does not depend on a specific range projection-based semantic segmentation architecture. For training, we reorganize the SemanticKITTI [1] dataset and merge the original labels into two classes, moving and static. By training the network with proposed new binary masks, our method distinguishes between moving cars and parked cars in an end-to-end fashion. As shown in Fig. 2, we compare the aggregated point cloud maps (a) directly with the raw LiDAR scans, (b) with the cleaned LiDAR scans by applying our MOS predictions as masks. As can be seen, there are moving objects present that pollute the map, which might have adversarial effects, when used for localization or path planning. By using our MOS predictions as masks, we can effectively remove these artifacts and get a clean map. Furthermore, our method operates online in real-time.

III. Future Work

Joint Segmentation. Since MOS and semantic segmentation are highly related, a joint framework will improve the performance of both tasks. In my next step, I will combine both tasks by proposing a joint network, which fulfills semantic segmentation together with moving object segmentation.

Spatio-temporal Architecture. Recently, several well-studied techniques in natural language processing, e.g., long short-term memory (LSTM) [13] and Transformer [30], have been successfully transferred to computer vision tasks and achieve very promising results, e.g., image classification [11], image completion [19], object detection [3] and semantic segmentation [16]. Instead of using sequential information with 2D CNN, I plan to also exploit such spatio-temporal architectures to encode the information of LiDAR frames to better distinguish between moving and non-moving objects.

Unsupervised Learning and Uncertainty Estimation. Traditional supervised learning methods have achieved reasonable results in moving object segmentation, however, it highly depends on human involvement in annotating and does not guarantee a good generalization in unseen environments. Furthermore, traditional learning-based segmentation methods have not been derived from a probabilistic framework that can offer uncertainty estimates. I plan to investigate an unsupervised learning method with uncertainty estimation to better describe the probability of an object to be moving or not, rather than a simple binary classification. The uncertainty representation is needed for long-term map maintenance and localization.

REFERENCES

[1] J. Behley, M. Garbade, A. Milioto, J. Quenzel, S. Behnke, C. Stachniss, and J. Gall. SemanticKITTI: A Dataset for Semantic Scene Understanding of LiDAR Sequences. In *Proc. of the IEEE/CVF Intl. Conf. on Computer Vision*, 2019.

[2] I. Bogoslavskyi and C. Stachniss. Fast range image-based segmentation of sparse 3d laser scans for online operation. In *Proc. of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems*, 2016.

[3] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko. End-to-end object detection with transformers. In *Proc. of the Europ. Conf. on Computer Vision*, 2020.

[4] X. Chen, T. Läbe, A. Milioto, T. Röhling, O. Vysotska, A. Haag, J. Behley, and C. Stachniss. OverlapNet: Loop Closing for LiDAR-based SLAM. In *Proc. of Robotics: Science and Systems*, 2020.

[5] X. Chen, T. Läbe, L. Nardi, J. Behley, and C. Stachniss. Learning an Overlap-based Observation Model for 3D LiDAR Localization. In *Proc. of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems*, 2020.

[6] X. Chen, A. Milioto, E. Palazzolo, P. Giguère, J. Behley, and C. Stachniss. SuMa++: Efficient LiDAR-based Semantic SLAM. In *Proc. of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems*, 2019.

[7] X. Chen, S. Li, B. Mersch, L. Wiesmann, J. Gall, J. Behley, and C. Stachniss. Moving object segmentation in 3d lidar data: A learning-based approach exploiting sequential data. *arXiv preprint*, 2021.

[8] T. Cortinhal, G. Tzelepis, and E. Aksoy. SalsaNext: Fast, Uncertainty-Aware Semantic Segmentation of LiDAR Point Clouds. In *Proc. of the IEEE Vehicles Symposium (IV)*, 2020.

[9] A. Dewan and W. Burgard. DeepTemporalSeg: Temporally Consistent Semantic Segmentation of 3D LiDAR Scans. In *Proc. of the IEEE Intl. Conf. on Robotics & Automation*, 2020.

[10] A. Dewan, T. Caselitz, G. Tipaldi, and W. Burgard. Rigid scene flow for 3d lidar scans. In *Proc. of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems*, 2016.

[11] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *Proc. of the Int. Conf. on Learning Representations (ICLR)*, 2021.

[12] J. Gehrung, M. Hebel, M. Arens, and U. Stilla. An approach to extract moving objects from mls data using a volumetric background representation. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 4, 2017.

[13] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[14] G. Kim and A. Kim. Remove, then revert: Static point cloud map construction using multiresolution range images. In *Proc. of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems*, 2020.

[15] S. Li, X. Chen, Y. Liu, D. Dai, C. Stachniss, and J. Gall. Multi-scale interaction for real-time lidar data segmentation on an embedded platform. *arXiv preprint arXiv:2008.09162*, 2020.

[16] J. Liang, N. Homayounfar, W. Ma, Y. Xiong, R. Hu, and R. Urtasun. Polytransform: Deep polygon transformer for instance segmentation. In *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, 2020.

[17] A. Milioto, I. Vizzo, J. Behley, and C. Stachniss. RangeNet++: Fast and Accurate LiDAR Semantic Segmentation. In *Proc. of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems*, 2019.

[18] S. Pagad, D. Agarwal, S. Narayanan, K. Rangan, H. Kim, and G. Yalla. Robust Method for Removing Dynamic Objects from Point Clouds. In *Proc. of the IEEE Intl. Conf. on Robotics & Automation*, 2020.

[19] N. Parmar, A. Vaswani, J. Uszkoreit, L. Kaiser, N. Shazeer, A. Ku, and D. Tran. Image transformer. In *Proc. of the Int. Conf. on Machine Learning (ICML)*, 2018.

[20] P. Patil, K. Biradar, A. Dudhane, and S. Murala. An end-to-end edge aggregation network for moving object segmentation. In *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, 2020.

[21] F. Pomerleau, P. Krüsiand, F. Colas, P. Furgale, and R. Siegwart. Long-term 3d map maintenance in dynamic environments. In *Proc. of the IEEE Intl. Conf. on Robotics & Automation*, 2014.

[22] G. Postica, A. Romanoni, and M. Matteucci. Robust moving objects detection in lidar data exploiting visual cues. In *Proc. of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems*, 2016.

[23] H. Rashed, M. Ramzy, V. Vaquero, A.E. Sallab, G. Sistu, and S. Yogamani. FuseMODNet: Real-Time Camera and LiDAR Based Moving Object Detection for Robust Low-Light Autonomous Driving. In *The IEEE/CVF Intl. Conf. on Computer Vision Workshops*, 2019.

[24] P. Ruchti and W. Burgard. Mapping with dynamic-object probabilities calculated from single 3d range scans. In *Proc. of the IEEE Intl. Conf. on Robotics & Automation*, 2018.

[25] J. Schauer and A. Nüchter. The peopleremover—removing dynamic objects from 3-d point cloud data by traversing a voxel occupancy grid. *IEEE Robotics and Automation Letters*, 3(3):1679–1686, 2018.

[26] H. Shi, G. Lin, H. Wang, T. Hung, and Z. Wang. Spsequencenet: Semantic segmentation network on 4d point clouds. In *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, 2020.

[27] M. Siam, H. Mahgoub, M. Zahran, S. Yogamani, M. Jagersand, and A. El-Sallab. MODNet: Moving Object Detection Network with Motion and Appearance for Autonomous Driving. *arXiv preprint arXiv:1709.04821*, 2017.

[28] H. Tang, Z. Liu, S. Zhao, Y. Lin, J. Lin, H. Wang, and S. Han. Searching Efficient 3D Architectures with Sparse Point-Voxel Convolution. In *Proc. of the Europ. Conf. on Computer Vision*, 2020.

[29] H. Thomas, C. Qi, J. Deschaud, B. Marcotegui, F. Goulette, and L. Guibas. KPConv: Flexible and Deformable Convolution for Point Clouds. In *Proc. of the IEEE/CVF Intl. Conf. on Computer Vision*, 2019.

[30] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In *Proc. of the Advances in Neural Information Processing Systems (NIPS)*, 2017.

[31] D. Wang, I. Posner, and P. Newman. What could move? finding cars, pedestrians and bicyclists in 3d laser data. In *Proc. of the IEEE Intl. Conf. on Robotics & Automation*, 2012.

[32] W. Xiao, B. Vallet, M. Brédif, and N. Paparoditis. Street Environment Change Detection from Mobile Laser Scanning Point Clouds. *ISPRS Journal of Photogrammetry and Remote Sensing*, 107:38–49, 2015.

[33] J. Yan, D. Chen, H. Myeong, T. Shiratori, and Y. Ma. Automatic Extraction of Moving Objects from Image and LIDAR Sequences. In *Proc. of the Intl. Conf. on 3D Vision*, 2014.