

Recognition of Human Characteristics Using Multiple Mobile Robots With 3D LiDARs

Koki Makita¹, Dražen Brščić¹ and Takayuki Kanda¹

Abstract—Mobile robots are gradually entering our human spaces. Apart from being able to accurately localize themselves, in populated environments robots also need to detect humans and recognize their characteristics. In this work we focused on using 3D LiDARs for sensing, and in particular on the question whether they can be used to classify characteristics of people around the robot. Moreover, since using sensors from multiple robots is expected to give more accurate recognition, we compared several ways how to combine multiple 3D LiDARs. We evaluated these methods using simulator data as well as actual real-world data. For the real-world data we created a novel dataset taken with multiple robots equipped with 3D LiDARs in a shopping center, which also includes manually labeled characteristics of the detected pedestrians. The results show that combining several 3D LiDARs makes the recognition more accurate. However, we were not able to achieve satisfactory accuracy with real-world data.

I. INTRODUCTION

In the near future, a number of different mobile robots are expected to share with us our daily environments and public spaces. Robots will carry out service jobs such as the delivery of objects, route guiding, patrolling, or providing personal mobility. As more robots work in the same space, the cooperation between multiple robots will allow even better services.

In order to move and work around humans, robots require advanced capabilities. They need to understand their surroundings and determine their location, but also to detect humans and recognize their state and behavior. Various onboard sensors can be used for that, and among them the 3D LiDAR is a very promising one. 3D LiDARs are attracting much attention as they are frequently used as the main sensor for autonomous driving. There is currently driving a large push for the development of better and less expensive 3D LiDARs, which is expected to lead to a further spread of its use beyond just autonomous cars, and in particular for mobile robots. While laser range finders (2D LiDARs) are still frequently used on robots, especially for robot localization, 3D LiDARs give accurate 3D measurements and thus also allow gathering more detailed information about surrounding people. Moreover, in comparison to other sensors used on mobile robots to observe people, such as RGB cameras, 3D LiDARs have a number of advantages, such as being robust against lighting condition. And importantly for public spaces where privacy issues can be an important concern,

3D LiDARs are better option than cameras as they do not allow easy identification of surrounding persons. Yet, a disadvantage of the currently most often used 3D LiDAR sensors is that their measurements are sparse. This provides a challenge for the sensing.

For robots that cohabit the space with humans, recognizing the characteristics of surrounding people can make the interaction between robots and humans smoother. For instance, if the robot knows that a person carries a bag, it can offer to help carrying or it might change the way how to hand over something to the person; on the other hand, when suggesting places to visit in a mall, information about the person's gender and age can be useful. Since 3D LiDARs can measure the 3D shape of the people in the vicinity, this can be used to recognize characteristics that can influence the interaction, such as their gender, age, height or carried objects. However, even though 3D LiDARs have sometimes been used for detecting and tracking humans, recognizing humans' characteristics has rarely been studied so far.

In addition, since in the near future more than one robot will be present in the space, we expect that combining the information from 3D LiDARs on multiple robots could result in even better recognition accuracy. However, how to combine the measurements, and how much improvement can actually be achieved is still an open question.

This paper provides two main contributions towards the study of the above issues. First, we perform an evaluation of several approaches to estimate people's characteristics based on data from multiple 3D LiDARs. These methods are evaluated using both simulated data and actual pedestrian data. Second, we created a unique dataset of pedestrian measurements in public space using three mobile robots equipped with 3D LiDARs. To our knowledge, no such dataset existed until this study.

II. RELATED WORK

A. Application of 3D LiDARs for mobile robots

In mobile robotics 3D LiDARs have most often been studied for the tasks of self localization and building of the environment map [1]–[3]. In these works, 3D LiDARs are used to build a map of the environment using some type of SLAM (simultaneous localization and mapping) algorithm. Once the 3D map is built, it can then be used to localize the robot, which is done by comparing the map with the 3D LiDAR measurements. Even though localization is not the main topic of study in this work, we use the localization with 3D LiDARs to simplify the detection of humans and extracting their characteristics.

This work was supported by JST CREST Grant Number JPMJCR17A2, Japan

¹All authors are with HRI Laboratory, Kyoto University, Kyoto, Japan. makita@robot.soc.i.kyoto-u.ac.jp, {drazen, kanda}@i.kyoto-u.ac.jp

B. Sensing of humans based on depth data

Depth information can also be used to recognize humans or objects around robots. For example, the 3D LiDAR can be used for the detection and tracking of humans and other objects around robots. Especially in the field of autonomous driving, many studies have been done on the detection of cars or pedestrians. For autonomous cars this is a crucial capability as it is needed to avoid collisions. For example, in [4] and [5], 3D LiDARs are used in order to detect pedestrians around the autonomous vehicles. Tracking people around the robot is also often studied [6]–[8]. Nevertheless, there are very few studies about obtaining humans' characteristics with 3D LiDARs. An exception is our recent work in [9], where we studied the use of a (single) onboard 3D LiDAR for the estimation of people's gender and body orientation. In the current work we expand this research to multiple 3D LiDARs.

A related sensor that also gives depth information is the depth camera. It typically has a better resolution than a 3D LiDAR, but a smaller range. The depth camera is often combined with an RGB camera, and this combined sensor is called an RGB-D camera (an example is the Microsoft Kinect sensor). There are several studies that use either depth or RGB-D cameras for the person detection [10] or tracking [11]–[13], but in a number of works they are also used for the recognition of their characteristics. For example, Linder et al. studied using RGB-D data for recognition of gender in [14], and other human characteristics in [15]. In part of the results in [14] only depth information (i.e. without RGB) was used, which is similar to this study.

C. Recognition with multiple sensors

Sensing using multiple sensors is widely studied. In general, how to fuse measurements from several sensors is an essential question when using multiple sensors for classification. Roughly speaking, there are 3 main approaches to fusion (see e.g. [16]):

- *input fusion* (measurement fusion): simply combine the raw sensor measurements and do classification on the combined data (note that for 3D LiDAR (i.e. depth) data this can be simply done by joining the point clouds, but for other sensing modalities, like RGB, it may not be clear how to do it)
- *deep fusion* (early fusion, feature-level fusion): partially process data from each sensor, fuse the obtained features, and do the final classification using the fused features
- *score fusion* (late fusion, output fusion, decision-level fusion): do classification on each sensor data separately and fuse the recognition scores

There are a few works where multiple 3D LiDARs are used. One example is [17], which presents a dataset that has two Velodyne HDL-32E LiDARs, and also several other sensors. However, in that work all the sensors were mounted on the same car. This means they were always in the same fixed spatial relation to each other, and the straightforward

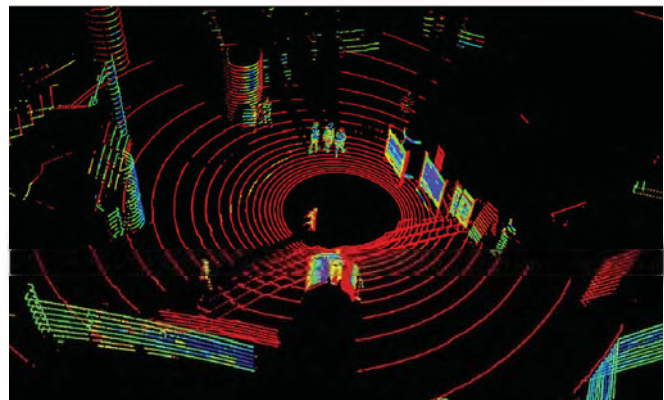


Fig. 1. Point cloud observed from Velodyne HDL-32E mounted on a robot.

way to use them would be to combine their measurements, i.e. to use input fusion. In contrast, in this study we were interested in the case where each 3D LiDAR is on a different robot, and such direct combination is not easily applicable (see Section IV).

The ways how to do deep fusion are frequently studied. Nowadays the most common approach to do classification is to use neural networks, where deep fusion corresponds to fusing the data (features) in a middle layer of the neural network. One example of such work is presented in [18], which introduces a way to fuse image-like outputs of convolutional layers in a convolutional neural network. The fusion method is called *view pooling*, and in [18] it is used to fuse data from multiple RGB cameras. This is the method that we use in this study.

Deep fusion and score fusion are compared in a number of works [16], [19]. Their results mostly suggest that fusion in the middle of neural network achieves higher recognition accuracy than score fusion.

Apart from using same sensor modalities, combining sensors with different modalities is also possible, and similar approaches to fusion are used. For example, several works studied the use of data from a 3D LiDAR and an RGB camera and combining them using a variant of deep fusion [20], [21].

III. 3D LiDAR DATA AS INPUT

In this study, we employed 3D LiDARs to measure the surrounding of the robot. The data from 3D LiDARs contains a large number of 3D points which represents the surface of the target objects ("3D point cloud"). Figure 1 shows an example 3D LiDAR output.

The 3D LiDAR sensor we used is Velodyne HDL-32E¹. It is a so-called rotational multi-layer scanner, which uses 32 laser units to measure the distance to objects. The laser units are arranged vertically and rotate around the central axis at 10 times a second. As a result, we obtain a scan of the environment like the one shown in Fig. 1. As there are only 32 sensing units the vertical angular resolution of the sensor is quite low, about 1.33°. On the other hand the

¹<https://velodynelidar.com/products/hdl-32e/>

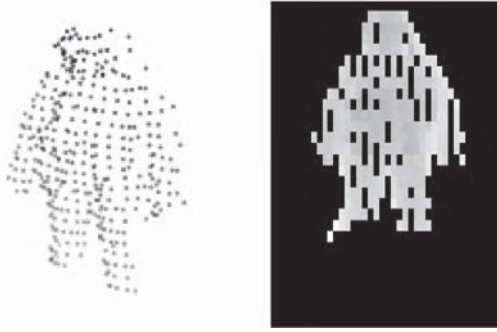


Fig. 2. Example conversion of 3D point cloud (left) into 2D depth image (right)

horizontal angular resolution is much higher ($\approx 0.16^\circ$). The low vertical resolution is the reason why measurement data is relatively sparse.

We are interested in the 3D point cloud of each person. However, as the 3D point cloud data is sparse, following the approach from [9] we first converted it into two-dimensional depth images. The depth images were then used as input to the neural network for recognition.

Figure 2 shows an example point cloud and the corresponding converted depth image. The converted images are 40 pixels in width and 30 pixels in height. This size was experimentally determined to be appropriate, as it results in a neural network with relatively small number of parameters in the input layer, but at the same time it does not lead to too much data loss during conversion.

IV. CLASSIFICATION METHODS

As mentioned in Section II-C, the combination of data from multiple sensors can be done in 3 basic ways: input fusion, deep fusion and score fusion.

It is important to note that both the robots and people are moving in the environment. Therefore the number of LiDARs that detect one specific person can change, i.e. at a specific instant a person can be seen by one, two or many LiDARs. As a consequence, all the methods for fusing the data need to be able to work with a changing number of sensors. This is a distinctive feature of our problem, and it is often not considered in other works, which assume that the number of sensors that need to be combined is fixed – e.g. [19] presupposes having exactly 2 camera views.

Combining raw data (i.e. input fusion) from 3D LiDARs on different robots is difficult in practice, because it requires that the self localization of each robot is very accurate. Even small errors in the estimated robot pose can lead to badly aligned point clouds. Combining such data would result in a distorted point cloud, which would strongly influence the recognition performance.

We therefore tested the other 2 approaches – fusing data from multiple 3D LiDARs in the middle layer of the neural network (*deep fusion*) or at its output (*score fusion*). In addition, we also tested the recognition based on data from only one sensor as a baseline. The details of the methods are presented below.

A. Using a single 3D LiDAR

In order to have a baseline with which to compare the results of methods that fuse multiple sensor data, we used a simple method that uses only data from one 3D LiDAR.

The classification was performed with a convolutional neural network. Figure 3a shows the network structure. The input of the network is a 30x40 pixel depth images, as described above. The network has three times convolution layers and max pooling layer, followed by a fully connected layer, a batch normalization layer and finally a softmax output layer. In all of the convolutional layers of this networks, zero-padding was used so that the input size and the output size are the same. As shown in Fig. 3, the filter sizes are respectively 5x5 pixels, 3x3 and 3x3 from the nearest one to the input, the numbers of filters are 8, 16 and 16 respectively.

B. Score fusion

A straightforward way to fuse data from multiple 3D LiDARs is to do recognition for each sensor separately and fuse the resulting scores. The neural network structure for such fusion is shown in Fig. 3b. Data from each LiDAR first goes through a neural network to calculate the recognition scores for each sensor. The network for all sensors have the same weights and structure (they have the same structure and weights as the network for single sensor recognition).

During prediction, scores for each 3D LiDAR are first calculated, which produces 2-dimensional vectors. The number of such vectors for one person will be equal to the numbers of 3D LiDARs that could observe that person. The final fused scores were calculated by taking the maximum (element wise) of the score vectors. (We also tested with using the average, but the results were very similar so we do not report them.)

C. Deep fusion

In order to fuse data in the middle of the neural network, we used a layer called "view pooling", introduced in [18]. The structure in Fig. 3c shows the structure of the network with the view pooling layer. The view pooling layer fuses outputs from pooling layers and calculates the element wise maximum of each input as its output. The parameters in convolutional layers before this layer are shared among each input branch. (Note that instead of maximum, for view pooling value averaging can also be used. But the results in [18] show that maximum achieves slightly higher recognition accuracy than average, so we used the maximum operation in the view pooling layer.)

This method has several good characteristics. First, since the weights are shared, the number of parameters does not increase with the number of inputs. In fact, since we used exactly the same layers as in the single sensor and score fusion networks, the number of parameters was the same (view pooling does not have parameters). Second, the maximum operation in the pooling layer insures that removing one input will not result in a sudden change of the output. As a result, the accuracy gracefully degrades with the reduction of the number of sensors, and the recognition

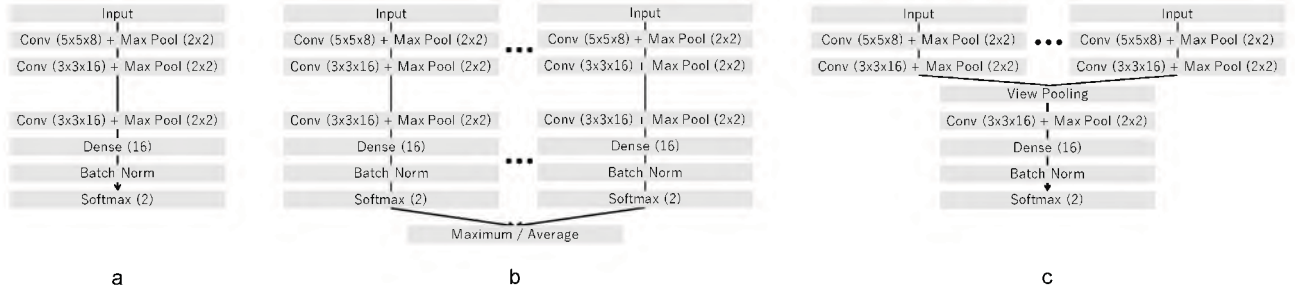


Fig. 3. The networks structures compared in this study: a) using single sensor (baseline method); b) fusion of the per sensor recognition scores ("score fusion"); c) fusion in a network middle layer ("deep fusion").

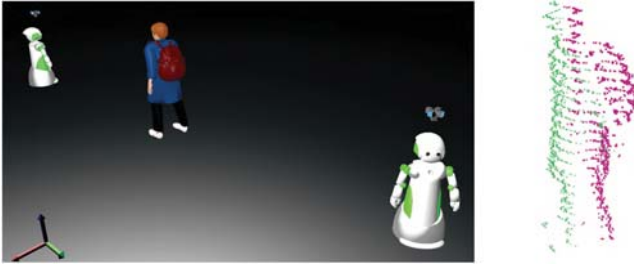


Fig. 4. Example of two simulated robots and a person (left), and the resulting point clouds from the two 3D LiDARs (right).

performance is stable even if the number of the sensors is constantly changing.

D. Network training and evaluation

For training the deep fusion network the input was the data from all LiDARs that were detecting a specific person at an instant. In contrast, the neural network for single sensor recognition, which is the same as the network in the input branches in score fusion, was trained on data from a single LiDAR. To make the results comparable, we did the training and evaluation for each sensor's data separately, and then averaged the resulting accuracy values.

The training of neural networks in all three approaches converged quickly, so the maximum number of epochs was set to 20. In addition we used early stopping to quit learning when the value of error function stops improving.

In all results the recognition accuracy was evaluated using 10-fold cross-validation. In each step of the cross-validation, the frames were chosen in a way that train and validation data do not contain the same models.

V. EVALUATION ON SIMULATED DATA

A. Data collection

We first compared the fusion methods based on simulated 3D LiDAR measurements of pedestrians. The simulations were done with the MORSE simulator [22]. We used a custom extended version of the MORSE simulator which allows the simulation of multiple animated people, and also provides an improved simulation of 3D LiDAR measurements [23].

In the simulation one human pedestrian and six robots with Velodyne LiDARs were placed in the simulated world. Each

TABLE I
THE RECOGNITION ACCURACY ON DATA FROM SIMULATION

Network type	Bag task	Gender task
Single sensor	73.5 %	75.1 %
Score fusion	97.1 %	97.8 %
Deep fusion	89.7 %	95.7 %

simulation ran for 4 to 5 seconds, after which the simulation was restarted with a different human model. In total 98 different 3D human models were used in the simulation (52 male, 46 female). In each simulation run, the locations and orientations of the six robots were chosen at random in the vicinity of the pedestrian. An example scene of the simulation (for simplicity using only two robots) is shown in Fig. 4.

The simulation was set up so that only the pedestrian could be observed from the sensors. Moreover, the absolute positions of all robots were known. This simplified the preprocessing of the point clouds and combining multiple sensors. The created simulated dataset contains in total 71,300 frames, where one frame contains all measurements at a single instant, i.e. the point clouds from all 6 sensors.

B. Evaluation results

We performed learning and prediction with the data from simulation and measured the recognition accuracy. The accuracy was measured on two tasks: classifying whether a person carries a bag and classifying their gender. The only type of bag that simulated human models had were backpacks, like the one in Fig. 4.

Table I shows the obtained recognition accuracy on the two tasks. As can be seen, that the recognition accuracy using multiple 3D LiDARs is higher than when using a single 3D LiDAR, as expected. The results also show that the recognition accuracy of score fusion was higher than that of deep fusion, in particular for the bag recognition task.

C. Discussion

The result that score fusion performs better than deep fusion seems to contradict the earlier reported findings, e.g. in [19]. This could be due to the difference between depth data from 3D LiDARs and RGB data used in [19]. Also, since the view pooling layer outputs the maximum value per



Fig. 5. Scene during data collection.

pixel and thus selects only one of the multiple inputs, during learning the error for each pixel propagates to only one of the multiple inputs. We suspect this might make the learning of the deep fusion network less efficient than score fusion learning, and thus requires more data to train well. However, further study would be necessary to confirm this point.

VI. EVALUATION ON ACTUAL LiDAR DATA

In the simulated data presented above the number of human models was limited and their movements simple, there was always only one human and the view from all sensors was not occluded, and the simulated measurements were idealized. This is obviously insufficient to draw conclusions about the performance of the fusion methods with actual sensors.

While there are datasets which provide 3D LiDAR measurements taken in the real world, in particular autonomous driving datasets such as [17] or [24], there was no suitable dataset for our purpose, i.e. one that contains LiDAR measurements from multiple robots (or cars) moving in the same area. Therefore we made this data collection ourselves.

A. Data collection

We collected data using four robots: 3 Robovie R3s² mounted on omnidirectional platforms, and one Robovie II. However, due to a technical problem with one of them during the data collection, we used only data from three robots. Each robot had one 3D LiDAR mounted in the back, approximately 1.5 meters above the floor, see Fig. 5. The data was collected in ATC, a shopping mall in Osaka, Japan. During the data collection the robots were driven manually by human operators using joysticks. This was done in a way to resemble the movement of guards or information staff, i.e. the robots would be stopped or turn in place, and from time to time they would move to a nearby location. In this study we used the data collected in three runs totaling about 25 minutes.

B. Extracting of person point clouds

In contrast to the data from simulation, extracting the point cloud of each person from the actual measurement is not as straightforward, since now all objects in the environment are included in the LiDAR measurement. We used the same method as in [9], which we briefly summarize here.

We localized the robots using a prebuilt 3D map of the environment. All robots were localized using the same map. The parts of the LiDAR measurement which did not match the 3D map were assumed to correspond to moving objects, i.e. the surrounding humans, so we clustered them into separate parts belonging to different humans. Clusters with too few points were discarded. Finally, the extracted clusters were used to track each person's position. This was done by keeping track of all tracked humans and applying a Kalman filter to updated their positions using the extracted clusters from all sensors.

C. Labeling

During the collection we also recorded videos of the pedestrian around the robots from cameras fixed in the field. Combining the videos with tracking results, allowed us to manually label the characteristics of the people that were tracked.

The characteristics that we labeled were the gender (male or female), age (small child, adult or elderly) and whether the person had a bag (no bag, or bag in front, back, on right side, left side or both on left and right). These characteristics could easily be determined from the camera views. After an initial check with multiple labelers we decided that for these characteristics there was little possibility of confusion, so the labeling was done by one person.

The obtained dataset contains 53,510 frames of labeled data in total. About 64.8% of the data corresponded to men and 36.2% to women. The percentage of people carrying some sort of bag was 42.2% (percentages for front, back, left, right, and both left and right were 7.5%, 16.3%, 12.1%, 17.7%, and 4.0% respectively). Finally, the age distribution was children 4.7%, adults 89.0%, elderly 6.3%.

The dataset with both person point clouds and corresponding labels are made freely available³.

D. Evaluation results

For some of the labeled characteristics the classes were not well balanced – e.g. there was only a small number of children or elderly compared to the number of adults. Dealing with highly unbalanced classes is challenging for learning and it was not our main concern in this study, so in the evaluation we limited ourselves to the recognition of characteristics for which the classes were relatively well balanced. Thus the learning and prediction were performed on two tasks: classifying whether a person has a bag (having no bag vs. having any type of bag) and classifying their gender.

Table II shows the obtained result. Like with the simulated data, the results show that when fusing data from multiple LiDARs it is possible to achieve higher accuracy than when using a single 3D LiDAR. However, in contrast to the simulation data, comparing the two ways of data fusion did not show significant difference in recognition accuracy. Moreover, the obtained recognition accuracy is much lower

²https://www.vstone.co.jp/products/robovie_r3/index.html

³<http://www.robot.soc.i.kyoto-u.ac.jp/en/research/dataset-multiple-lidars/>

TABLE II
RECOGNITION ACCURACY ON DATA FROM ACTUAL PEDESTRIANS

Network type	Bag task	Gender task
Single sensor	57.1 %	58.9 %
Score fusion	60.4 %	61.2 %
Deep fusion	61.7 %	61.1 %

than those that were obtained with simulated pedestrians, and are probably not good enough for any actual application.

E. Discussion

There could be several reasons why there was a large difference between the obtained recognition accuracy for simulated and real-world data. One probable reason is that human models in the simulator do not reflect the actual pedestrians. The number of simulated models was limited and their shape (clothes, hairstyle) was fixed. Also the characteristics of the models in the simulator were probably easier to recognize; for example all bags in the simulator were backpacks, whereas recognizing whether in the real-world dataset a person has any type of bag is a more difficult task.

Yet another reason could be the number of samples in the actual LiDAR data. Since there is much more variation in the shapes of real people, the size of the dataset might have simply not been enough.

Finally, an important difference is that in the simulation the robots were surrounding the person at a fairly close distance. This was difficult to achieve in the real world without bothering the movement of the people. At most times the measurements were taken with all LiDARs on the same side of the person and with only one or two of them being close to the person, like the situation shown in Fig. 5.

The recognition accuracy could be improved by making a larger dataset. In addition, in this work we only tried to estimate the person's characteristics based on single measurements, but it is also possible to fuse the recognition results from multiple instants. As the results from [9] show, this can give a significantly better recognition accuracy.

VII. CONCLUSION

In this study, we focused on the estimation of people characteristics using multiple 3D LiDARs. In order to do this, we created a unique dataset of actual pedestrians using three robots with onboard 3D LiDARs, and we evaluated several methods of data fusion using both this dataset as well as simulated data. Results show that using multiple 3D LiDARs does indeed increase recognition accuracy. In addition, score fusion gave better accuracy than deep fusion on simulation data, and with six 3D LiDARs the achieved accuracy was quite high. However, the accuracy obtained on real-world data was low. This leaves open the question why exactly, so in future we will focus on ways to improve the accuracy.

REFERENCES

- [1] J. Zhang and S. Singh, "Loam: Lidar odometry and mapping in real-time." in *Robotics: Science and Systems*, vol. 2, no. 9, 2014.
- [2] F. Moosmann and C. Stiller, "Velodyne slam," in *2011 IEEE Intelligent Vehicles Symposium*, 2011, pp. 393–398.
- [3] A. Nüchter, K. Lingemann, J. Hertzberg, and H. Surmann, "6d slam—3d mapping outdoor environments," *Journal of Field Robotics*, vol. 24, no. 8-9, pp. 699–722, 2007.
- [4] K. Kidono, T. Miyasaka, A. Watanabe, T. Naito, and J. Miura, "Pedestrian recognition using high-definition lidar," in *2011 IEEE Intelligent Vehicles Symposium*, 2011, pp. 405–410.
- [5] K. Li, X. Wang, Y. Xu, and J. Wang, "Density enhancement-based long-range pedestrian detection using 3-d range data," *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 5, pp. 1368–1380, 2015.
- [6] L. Spinello, M. Luber, and K. O. Arras, "Tracking people in 3d using a bottom-up top-down detector," in *2011 IEEE International Conference on Robotics and Automation*, 2011, pp. 1304–1310.
- [7] M. Häselich, B. Jöbgen, N. Wojke, J. Hedrich, and D. Paulus, "Confidence-based pedestrian tracking in unstructured environments using 3d laser distance measurements," in *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2014, pp. 4118–4123.
- [8] Z. Yan, T. Duckett, and N. Bellotto, "Online learning for human classification in 3d lidar-based tracking," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2017, pp. 864–871.
- [9] D. Bršćić, R. W. Evans, M. Rehm, and T. Kanda, "Using a rotating 3d lidar on a mobile robot for estimation of person's body angle and gender," *Sensors*, vol. 20, no. 14, p. 3964, 2020.
- [10] L. Spinello and K. O. Arras, "People detection in rgb-d data," in *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2011, pp. 3838–3843.
- [11] M. Luber, L. Spinello, and K. O. Arras, "People tracking in rgb-d data with on-line boosted target models," in *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2011, pp. 3844–3849.
- [12] D. Bršćić, T. Kanda, T. Ikeda, and T. Miyashita, "Person tracking in large public spaces using 3-d range sensors," *IEEE Transactions on Human-Machine Systems*, vol. 43, no. 6, pp. 522–534, 2013.
- [13] M. Munaro and E. Menegatti, "Fast rgb-d people tracking for service robots," *Autonomous Robots*, vol. 37, no. 3, pp. 227–242, 2014.
- [14] T. Linder, S. Wehner, and K. O. Arras, "Real-time full-body human gender recognition in (rgb)-d data," in *2015 IEEE International Conference on Robotics and Automation*, 2015, pp. 3039–3045.
- [15] T. Linder and K. O. Arras, "Real-time full-body human attribute classification in rgb-d using a tessellation boosting approach," in *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2015, pp. 1335–1341.
- [16] J. Sanchez-Riera, K.-L. Hua, Y.-S. Hsiao, T. Lim, S. C. Hidayati, and W.-H. Cheng, "A comparative study of data fusion for rgb-d based visual recognition," *Pattern Recognition Letters*, vol. 73, pp. 1–6, 2016.
- [17] Z. Yan, L. Sun, T. Krajník, and Y. Ruichek, "EU long-term dataset with multiple sensors for autonomous driving," *CoRR*, vol. abs/1909.03330, 2019. [Online]. Available: <http://arxiv.org/abs/1909.03330>
- [18] H. Su, S. Maji, E. Kalogerakis, and E. Learned-Miller, "Multi-view convolutional neural networks for 3d shape recognition," in *IEEE international conference on computer vision*, 2015, pp. 945–953.
- [19] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," in *IEEE conference on computer vision and pattern recognition*, 2016, pp. 1933–1941.
- [20] X. Chen, H. Ma, J. Wan, B. Li, and T. Xia, "Multi-view 3d object detection network for autonomous driving," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1907–1915.
- [21] Z. Wang, W. Zhan, and M. Tomizuka, "Fusing bird's eye view lidar point cloud and front view camera image for 3d object detection," in *2018 IEEE Intelligent Vehicles Symposium*, 2018, pp. 1–6.
- [22] G. Echeverria, N. Lassabe, A. Degroote, and S. Lemaignan, "Modular open robots simulation engine: Morse," in *2011 IEEE International Conference on Robotics and Automation*, 2011, pp. 46–51.
- [23] S. Satake, T. Kaczmarek, D. Bršćić, and T. Kanda, "Facilitating software development for mobile social robots by simulating interactions between a robot and pedestrians," in *14th ACM/IEEE International Conference on Human-Robot Interaction*, 2019, pp. 626–627.
- [24] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1231–1237, 2013.