CONCEPT ON LANDMARK DETECTION IN ROAD SCENE IMAGES TAKEN FROM A TOP-VIEW CAMERA SYSTEM

C. Albrecht[1,*], S. Kraus[2], U. Stilla[1]

[1] Photogrammetry and Remote Sensing, Technical University of Munich, Munich, Germany - (christian.albrecht, stilla)@tum.de
[2] MAN Truck & Bus SE, Munich, Germany - sven.kraus@man.eu

KEY WORDS: Semantic Segmentation, Computer Vision, Automated Driving, Top-View System

ABSTRACT:

In this paper, we demonstrate the inclusion of a top-view camera system mounted on a city bus in an existing sensor setup. A novel sensor setup with five down-facing cameras is mounted on the roof of a MAN Lion's City 12 city bus to extract landmarks in road scene images. Its positioning is validated by an exemplary detection of lane markings. The concept for further landmark detection with the help of the presented camera system is explained in this paper and sensor data fusion methods are proposed. Based on our previous findings (Albrecht et al., 2019), strengths of the novel sensor system are introduced to improve the current environment perception system. For now, only a qualitative observation of the capability to detect lane markings and other landmarks can be presented. Future work will use the current findings for landmark detection for a vehicle self-localization system.

1. INTRODUCTION

Semantic segmentation is one of the main tasks for automated vehicles on public roads. Not only the detection of objects, but also their semantic information help perceiving the environment. As semantics can only be determined by optical sensors like cameras, a later fusion of semantics and range data is preferable. Usually cameras in automated vehicles are mounted horizontally to have a big field of view in the vertical direction. (He et al., 2016) and (Chen et al., 2018) use these road scene images for semantic segmentation. Also, some datasets (Geiger et al., 2013) (Caesar et al., 2019) (Cordts et al., 2016) provide horizontal camera images in combination with radars and laser scanners.

In contrast, we are using a top-view system with cameras facing downwards. From that sensor setup, we expect higher detection rates of lane markings and curbs next to the vehicle. These are areas where current sensor systems have their blind spots. The area on the ground that can be observed increases with the height of the sensor system. The accuracy on the other hand decreases as a bigger area is covered with the same amount of image pixels. As the top-view system is mounted on a bus, the proposed method shall evaluate if the high mounting position also increases overall detection rates.

Automated vehicles are not able to cover all possible driving situations only by using camera data, so a fusion with ranging sensors like laser scanners and radars are needed. Our previous concept (Albrecht et al., 2019) focused on the sensor data fusion of laser scanner point clouds with segmented road images from horizontal cameras in an low-level fusion approach. In this extension, we include objects detected by the top-view camera system in a late fusion approach. For acquisition of sensor data, we will make use of a neural network proprietary pre-trained for semantic segmentation and a proprietary dataset consisting of images from a passenger car. In future, we will extend the data to also cover our current sensor setup.

* Corresponding author

Figure 1. Mounting of top-view cameras on a city bus - exemplary field of view is given by the green lines perpendicular to the busses outer hull

In this paper, we estimate the benefits of including a set of cameras facing down into a conventional sensor setup consisting of laser scanners, radars and horizontal cameras. By projecting the camera images to an assumed ground plane, there is no need to estimate the distance in the image but to calculate that by extrinsic calibration. We state this leads to improvements in motion estimation by visual odometry. Furthermore, the position of given landmarks like lane markings and curbs can very well be estimated in this plane-projection. The found static landmarks will be included into an existing sensor data fusion method and combined in a late fusion approach. Preliminary results of semantic segmentation will be shown and the transferability to the sensor system of the bus will be evaluated.

2. RELATED WORK

Having a variety of different sensor modalities is essential for a robust and reliable self-driving system. Introducing a novel camera system into a sensor setup of mainly radars and laser scanners has multiple advantages:

- Dense sensor output
- High horizontal and vertical resolution
- Field of view selectable by choice of lens
- Semantic information extractable

Especially the first two points enable research in the field of (deep) neural networks. There are different approaches to evaluate camera images dependent on the requested output. In the following paragraphs, an overview of object detection, semantic segmentation, general use of top-view systems and their impact on localization systems will be given.

For object detection one of the most well-known algorithms is YOLO (Redmon et al., 2016) and its proceeding improvements. With that, a simultaneous detection and classification take place as in (Liu et al., 2015). (Meyer et al., 2019) are using convolutional neural networks (CNN) to detect objects in a laser data using similar techniques as known from computer vision. By treating the laser scanner's raw data as a range view images, they also have a dense sensor output with high resolution depending on the used sensor type. Using similar input formats for different sensors will make it easier to fuse their data in future.

In contrast, semantic segmentation is used to assign a class label to each pixel in the input image. One example of a semantic segmentation framework is given by (Ronneberger et al., 2015). The use a fully convolutional approach to classify cells in a microbiology application. With an evaluation time of one second, this algorithm is too slow for robotic applications or automated driving. (Zhu et al., 2019) show a video-based approach to further improve the segmentation process by propagating labels between two frames jointly. In contrast to that, (Chen et al., 2018) show a network capable of close to real-time semantic segmentation. Another big benefit of their method is to be able to classify object at different scales, what is very important for our use case. A variety of methods are furthermore explained in (Long et al., 2015).

Systems with down-facing cameras have recently been used especially for advanced driver assistance systems. (Lin, Wang, 2010) present a model for top-view transformation to visualize the back of a car. In contrast to automated driving functions, the exact position in world coordinates and with the the intrinsic and extrinsic parameters are not relevant. A similar approach is given in (Li, Hai, 2011), again only to help the manual driver while in a parking maneuver. (Geppert et al., 2019) use a multi-camera system for visual odometry as well as localization in a given map. In contrast to our system, the cameras are mounted horizontally, and lack information of the near proximity of the car.

In this paper, we present a concept on landmark detection with a top-view camera system. Found landmarks are to be included into a graph-based localization system similar to (Wilbers et al., 2019). The constructed factor graph will be optimized by g2o (Kümmerle et al., 2011) and the landmarks have to be added to the graph accordingly. In contrast to (Gao et al., 2018), landmarks are explicitly calculated. A direct approach will not be used for this paper, as landmarks will be associated with a given map. In addition to that, a localization approach simply relying on camera images is not required.

In contrast to using a full simultaneous localization and mapping (SLAM) method, our method is only focused on a localization problem. Current SLAM algorithms are evaluated in (Bresson et al., 2017) and will be further investigated if needed. Currently we assume to have a globally referenced map. Detection of landmarks in the vehicle's sensor image and co-registration will be used to enable a city bus to self-localize in a given high-definition map.

3. DATA ACQUISITION

The used test vehicle is a MAN Lion's City 12 city bus equipped with various sensor systems. Besides radars and laser scanners, there are cameras designated mainly for object detection in the front of the vehicle. Furthermore, there are five cameras built up as a top-view system as can be seen in Figure 1. The positions and the covered field of view (FOV) projected to the ground plane are depicted in Figure 2. The combined FOV is used to perceive the near field up to 10 meters all around the bus. To cover the entire 12 meters length of the bus, the opening angle of the used cameras has to be sufficiently large and the cameras have to be mounted high above the ground. Our system is placed at approximately 2.85 meters above ground level. With increasing height the area is enlarged, whereas accuracy is reduced because of fewer pixels on the same surface. Nevertheless, calibration and multiple sensor data fusion are eased with less cameras.

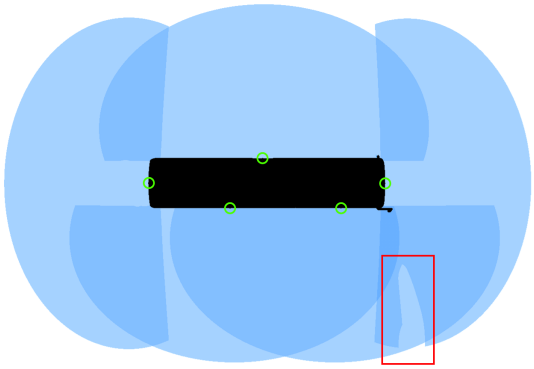

Figure 2. Positioning (green) and field of view (blue) of the five top-view cameras on the bus (facing right) - the right front mirror causes the cut-out in the red box

The opening angle of the used cameras is 190° in horizontal direction. With that, the camera's output is a distorted raw image with a resolution of 1928 x 1208 pixels on the sensor. An exemplary image is shown in Figure 3. Not only distortions, but in addition, self-reflections of the outer hull reduce the usable region in the image. For further evaluation of landmark detectors, different methods like pinhole calibration or methods from (Scaramuzza et al., 2006) will be used for calibration and compared with each other. Currently, only a deep learning based

Figure 3. Exemplary raw-data from one of the bus-mounted top-view cameras

method was used for extracting semantic labels from an input image, so camera calibration was not necessary yet. Nonetheless, a region of interest (ROI) in the center of the image was extracted to minimize the input of said self-reflections. The data from five cameras is collected simultaneously and can be fused in future work. Currently only one camera was used for detection at a time.

With the given pixel size on the sensor (μ_0) and effective focal length, the pixel's size on the ground can be calculated as:

$$\mu_{ground} = \frac{d * \mu_0}{f_{lens}} \tag{1}$$

Given $\mu_0 = 3\mu m$, $f_{lens} = 2.09mm$ and $d = 2.85m$, the size of one pixel on the road surface is $4.1mm$. Future work will evaluate if this accuracy is sufficient and can be used for future research.

Due to temporary governmental regulations, data acquisition with the measurement vehicle could not be performed before the paper deadline. To get an estimation of the usability of described algorithms, data from the same camera type on a similar test truck could be used for first tests. Figure 5a shows a picture from that truck. The lower position of the camera makes it comparable to a top-view system for passenger cars, but the self reflections are more common to a bus. Because of lacking calibration parameters, only one method for semantic segmentation with deep neural networks was further investigated.

4. PROPOSED METHOD

As described in section 3, we show a novel positioning strategy of our sensor setup higher above the ground than usual in automated driving. Section 2 shows some examples how cameras are usually used in road data sets and respective algorithms. In contrast to their methods, we are using the top-view camera system as an additional sensor for vehicle self-localization. In (Albrecht et al., 2019), we proposed a concept on low-level sensor data fusion as a combination of semantic labels from camera images and a laser point cloud. The concept is shown in Figure 4 and extended by the inputs of our novel top-view camera system.

Instead of combining the raw image data from these cameras, first of all landmarks that are only to be recognized by camera systems have to be extracted. In our case, these include lane

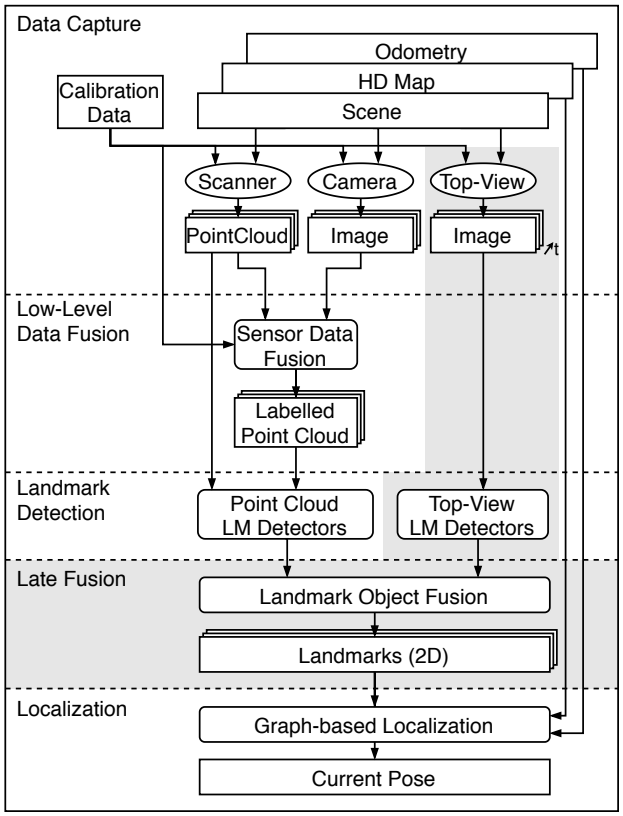

Figure 4. Concept on vehicle self-localization with low-level data fusion of camera and laser scanner data extended by the presented top-view system. The base of this method is given in (Albrecht et al., 2019), our novel impact is highlighted in gray.

markings and curbs in close proximity to the city bus. As already mentioned in Section 3, the camera system is closing the gap in the field of view of the other sensors that are mainly placed horizontally. The images don't have to be intrinsically calibrated, but their extrinsic calibration parameters have to be known to be able to define the position of the detected landmark relative to the reference coordinate frame. In our case, this is the center of the busses rear axle.

To give an example of landmark detection from top-view images, we will show the procedure for lane markings in this section. To localize candidates for lane markings, the given raw image from a top-view camera was semantically segmented by a pre-trained deep neural network used for a similar camera system. This outputs the regions of interest (ROI) in the input image, where lane markings can be found. In this case, there might also be false positive class labels.

In the next step, the ROIs are projected to an assumed ground plane in world space. By applying the transformation from projection space to a birds eye view (BEV), the lines can be detected by line-fitting algorithms. Since only lane markings and curbs in close proximity to the test vehicle are regarded, the search model can be a straight line for simplicity. The intrinsic and extrinsic camera parameters have to be determined before to calculate the transformation matrix. We hypothesize that having the cameras facing down, tracking of feature points and later landmarks will be more robust because of the lacking need of a depth estimation. The distance from a camera to the ground should not change over time. This will only be valid for lane markings close to the vehicle.

(a) Original image

(b) Semantic Segmentation of the image

Figure 5. Semantic Segmentation representation of an exemplary image taken from a similar sensor setup. The image is taken from a camera on the front right corner of the truck and facing right. It was segmented pixel-wise to detect lane markings (yellow) on the road.

For each ROI, the best fitting line in the BEV is found and validated against given parameters. Outliers that are too wide or their spatial area is under a specific threshold can be discarded. Furthermore, lines not in the direction of travel or perpendicular to it will also not be regarded.

Found landmarks are registered for all five cameras. These are then combined with the landmarks detected by the previously used detectors focusing on point clouds. Since the data fusion takes place with landmark objects, this procedure is called a late fusion. Landmarks can be detected by different sensors and even by multiple sensors of the same sensor system. The landmark object fusion algorithm has to combine objects that are spatially close to each other in world coordinates and find the best way to represent the landmark. Future work will evaluate the best metric for keeping, rejecting or combining landmark candidates.

5. EXPERIMENTS

The camera system as presented in Section 3 is currently mounted on a test bus. An exemplary image is shown in Figure 3. The provided FOV of each camera has been verified and is in accordance to Figure 2, but data collection still has to be performed with this specific sensor system as mentioned in Section 3. The Proof of Concept for the method proposed in Section 4 can also be shown with a similar sensor system, although accuracies for future use cases have to have to be evaluated for the real system again.

To evaluate the given concept with real data, a surround view system mounted on an MAN TGX was used to generate sample data. These cameras usually are used to help developers match their data sets to specific situations in traffic. They are neither intrinsically not extrinsically calibrated and are positioned similar to surround view systems in passenger cars. Therefore, the concept can only be tested up to the point that ROIs of lane marking candidates are defined. A stitching of all top-view cameras is not needed as already a single camera will give a good estimation of the minimum overall system performance.

As there is no publicly available data for this type of sensor setup, we can only start experiments with the collected truck data. Without having neither intrinsic nor extrinsic camera parameters for the given data, only a qualitative estimation of the camera setup is possible. For that, we performed semantic segmentation with a pre-trained deep neural network on images taken on an intersection in Wolfsburg. Results will be presented in Section 6 and further discussed there. A qualitative evaluation is given by the accuracy of semantic ROIs around lane markings in data collected by the test truck. These will be transferred into world space and lane markings will be detected on the ground plane representation.

The found lane markings will be used for self-localization in future work, so a relative accuracy of about 10 cm will be needed according to (Albrecht et al., 2019). As stated above, the accuracy of objects will be evaluated with the originally presented top-view system mounted on the test bus. If the landmarks can be detected accurately and robustly enough, they will be used for our graph-based localization system based on (Wilbers et al., 2019).

6. RESULTS

Figure 5 shows the potential of semantic segmentation for a top-view camera system with fish-eye lenses. The image is taken from a camera on the front right corner of the truck and facing right. The direction of travel would be to the left in this case. Although there was no given intrinsic camera calibration, the results look quite reasonable. The ROIs found in the original image are highlighted yellow in the right image. It can be seen that most of the lane markings are detected correctly. On the left side, there are spots where no ROIs are generated. This is caused by the fact that the neural network is trained with passenger car input data and there would be the car's outer hull.

All possible lane markings were detected and highlighted, including arrows on the tarmac. Furthermore, a manhole was falsely detected as lane marking on the right in Figure 5b. All other candidates are either in the direction of travel or in this case perpendicular to it. By filtering the ROIs transformed to the bespoke 2D ground plane, only these valid line segments would remain.

In this case, all found lane marking candidates would be inserted to a list of landmarks. Further research has to be performed to differentiate arrows from lane markings. Furthermore, there is no distinction between solid and dashed lines at the moment. Environmental knowledge as well as line models will be inferred for robust determination of line type. Not only lane markings but also curbs have to be recognized in the future. These are currently not detected and have to be added in future work.

For environment perception and localization, the detection of landmarks has to be performed in real-time. According to previous research (Albrecht et al., 2019) algorithms should run with at least 20 Hz. This image was generated in under 30 ms making it usable for real-time application in an autonomous vehicle, even though further processing has to be performed to transform these ROIs into line landmarks.

7. CONCLUSION AND OUTLOOK

In this paper, we presented a concept to include a top-view camera system into a sensor setup for environment perception consisting mainly of radar and laser scanners. The general capability of lane marking detection by deep neural networks is shown in this paper. By now, only a qualitative assessment of the detection accuracy could be given. Future work will further investigate the system's performance in real-time applications. The images used in this paper originate from a similar vehicle, but do not reflect the full potential of the presented sensor setup. Due to governmental regulations, further tests and data acquisition with the measurement vehicle could not be performed before the paper deadline.

This concept will be used as a basis for further research in vehicle self-localization. Different methods of sensor data fusion will extend this concept's scope. Not only the detection of landmarks in images generated by the presented top-view camera system but also the tracking will be further investigated. An example would be the realization of a visual odometry system to estimate the vehicles motion.

Furthermore, not only lane markings should be extracted from the camera images but also previously described landmarks should be further classified to increase robustness of matching while tracking landmarks. Sensor data fusion approaches will be used for that as well.

REFERENCES

Albrecht, C., Kraus, S., Zimmermann, A., Stilla, U., 2019. A Concept for an Automated Approach of Public Transport Vehicles to a Bus Stop. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, XLII-2/W16, 13–20.

Bresson, G., Alsayed, Z., Yu, L., Glaser, S., 2017. Simultaneous Localization and Mapping: A Survey of Current Trends in Autonomous Driving. *IEEE Trans. Intell. Vehicles*.

Caesar, H., Bankiti, V., Lang, A. H., Vora, S., Liong, V. E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O., 2019. nuScenes: A multimodal dataset for autonomous driving. *arXiv preprint arXiv:1903.11027*.

Chen, L., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A. L., 2018. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 40(4), 834-848.

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B., 2016. The cityscapes dataset for semantic urban scene understanding. *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*.

Gao, X., Wang, R., Demmel, N., Cremers, D., 2018. LDSO: Direct sparse odometry with loop closure. *Proc. IEEE/RSJ Int. Conf. Intelligent Robots and Syst. (IROS)*.

Geiger, A., Lenz, P., Stiller, C., Urtasun, R., 2013. Vision meets Robotics: The KITTI Dataset. *Int. J. Robot. Research*.

Geppert, M., Liu, P., Cui, Z., Pollefeys, M., Sattler, T., 2019. Efficient 2d-3d matching for multi-camera visual localization. *Proc. IEEE Int. Conf. Robotics and Automation (ICRA)*, 5972–5978.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 770–778.

Kümmerle, R., Grisetti, G., Strasdat, H., Konolige, K., Burgard, W., 2011. g2o: A general framework for graph optimization. *Proc. IEEE Int. Conf. Robotics and Automation (ICRA)*, 3607–3613.

Li, S., Hai, Y., 2011. Easy Calibration of a Blind-Spot-Free Fisheye Camera System Using a Scene of a Parking Space. *IEEE Trans. Intell. Transp. Syst.*, 12(1), 232-242.

Lin, C., Wang, M., 2010. Topview transform model for the vehicle parking assistance system. *Int. Comp. Symp.*, 306–311.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S. E., Fu, C., Berg, A. C., 2015. SSD: Single Shot MultiBox Detector. *arXiv preprint arXiv:1512.02325*.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 3431–3440.

Meyer, G. P., Laddha, A., Kee, E., Vallespi-Gonzalez, C., Wellington, C. K., 2019. Lasernet: An efficient probabilistic 3d object detector for autonomous driving. *cvpr*, 12677–12686.

Redmon, J., Divvala, S., Girshick, R., Farhadi, A., 2016. You only look once: Unified, real-time object detection. *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 779–788.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. *arXiv preprint arXiv:1505.04597*.

Scaramuzza, D., Martinelli, A., Siegwart, R., 2006. A toolbox for easily calibrating omnidirectional cameras. *Proc. IEEE/RSJ Int. Conf. Intelligent Robots and Syst. (IROS)*, 5695–5701.

Wilbers, D., Merfels, C., Stachniss, C., 2019. Localization with Sliding Window Factor Graphs on Third-Party Maps for Automated Driving. *Proc. IEEE Int. Conf. Robotics and Automation (ICRA)*.

Zhu, Y., Sapra, K., Reda, F. A., Shih, K. J., Newsam, S., Tao, A., Catanzaro, B., 2019. Improving semantic segmentation via video propagation and label relaxation. *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*.