SEMANTIC ROAD SCENE KNOWLEDGE FOR ROBUST SELF-CALIBRATION OF ENVIRONMENT-OBSERVING VEHICLE CAMERAS

Alexander Hanel* , Uwe Stilla

Photogrammetry and Remote Sensing, Technical University of Munich (TUM), Germany
- (alexander.hanel, stilla)@tum.de

KEY WORDS: Camera calibration, self-calibration, structure from motion, semantic segmentation, road scene understanding

ABSTRACT:

Environment-observing vehicle camera self-calibration using a structure from motion (SfM) algorithm allows calibration over vehicle lifetime without the need of special calibration objects being present in the calibration images. Scene-specific problems with feature-based correspondence search and reconstruction during the SfM pipeline might be caused by critical objects like moving objects, poor-texture objects or reflecting objects and might have negative influence on camera calibration. In this contribution, a method to use semantic road scene knowledge by means of semantic masks for a semantic-guided SfM algorithm is proposed to make the calibration more robust. Semantic masks are used to exclude image parts showing critical objects from feature extraction, whereby semantic knowledge is obtained by semantic segmentation of the road scene images. The proposed method is tested with an image sequence recorded in a suburban road scene. It has been shown that semantic guidance leads to smaller deviations of the estimated interior orientation and distortion parameters from reference values obtained by test field calibration compared to a standard SfM algorithm.

1. CALIBRATION OF ON-BOARD VEHICLE CAMERAS

Environment perception is one of the key enablers of advanced driver assistance systems in vehicles and especially on the way to autonomous driving. Different types of sensors for environment perception can be found on board of vehicles (Winner et al., 2015) (Ziebinski et al., 2016). Ultrasonic, radar, LiDAR sensors and cameras are together covering applications ranging from centimeters till few hundred meters, like parking assistance (Zhang et al., 2014), object detection (Hanel et al., 2018) or adaptive cruise control. Compared to other environment-perceiving sensors, cameras are small, cheap (Janai et al., 2017), provide high-resolution data and are therefore of special interest for mass-produced vehicles. While thermal-infrared cameras are typically used for applications in nighttime scenarios, are RGB cameras operating in the visible spectrum complementary for daytime scenarios (Zhang et al., 2014) (Janai et al., 2017). For many applications, especially on-board forward-looking cameras are highly relevant as they are observing the upcoming driveway of the vehicle. Though nowadays forward-looking cameras have been used in a stereo camera system (Dang et al., 2009) (Keller et al., 2011), a mono camera (Enzweiler and Gavrila, 2009) is easier to integrate into the vehicle design, even cheaper (Dubey, 2016) and synchronization problems between multiple sensors are not relevant (Azzopardi et al., 2010).

Having a calibrated camera is essential for various automotive applications of cameras like measuring distances or speed from images (Ribeiro et al., 2006) (Dubey, 2016), when 3D reasoning is required (Janai et al., 2017) or for multi sensor fusion (Geiger et al., 2012) (Heng et al., 2014). Today, on-board cameras are often calibrated by test field calibration, often using special test fields for automotive end-of-production-line calibration (Ernst et al., 1999) (Hella Gutmann Solutions GmbH, 2016) or by self-calibration during calibration drives on public roads (Thatcham Research and ADAS Repair Group, 2016) with appropriate road objects, for example lane markings (Ribeiro et al., 2006) (Paula et al., 2014) or traffic signs (Hanel and Stilla, 2018). One limiting aspect of these approaches is that they require certain objects to be available in the scene. Another limiting factor is that these approaches typically provide only a small number of reference points obtained from a test field or from single road objects, additionally often covering only parts of the complete image, which could lead to negative influence on the quality of camera calibration (Luhmann et al., 2006).

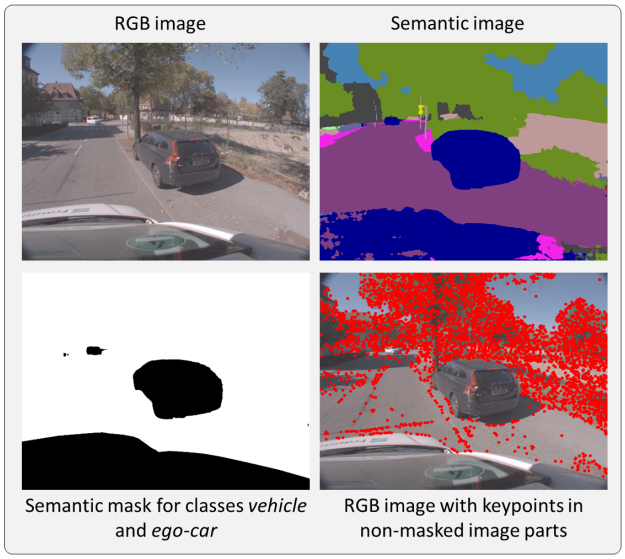

Figure 1. Top left: RGB road scene image as acquired by a vehicle camera. Top right: Semantic image obtained by semantic segmentation of the RGB image. Bottom left: Semantic mask obtained from semantic segments of the class *vehicle* and from the fix-pixel segment of the *ego-car*. Bottom right: SIFT feature keypoints detected in the RGB image considering only the non-masked image parts.

*Corresponding author

3D reconstruction and localization algorithms like visual SLAM (e.g. (Mur-Artal and Tardós, 2017)) or structure from motion (e.g. (Schönberger and Frahm, 2016)) allow for self-calibration of cameras given an image sequence. In a typical workflow of these approaches, correspondences between different images of the sequence are established and both camera pose and a 3D representation of the scene captured in the images obtained consecutively. For both approaches, even sparse 3D reconstructions have - depending on the number of images - typically several thousand 3D points that can be used as reference points for calibration, which has been reported to be an important factor for camera calibration (Stamatopoulos and Fraser, 2014). Furthermore, for mono cameras (Bertozzi et al., 2010) and especially for multi camera systems (Heng et al., 2014), (Zabatani et al., 2017), thermal of mechanical effects in vehicles could effect on-board cameras and deteriorate the quality of camera calibration over vehicle lifetime. These effects could make parameters estimated by previous camera calibration invalid and life-long repetitive calibration desirable, wherefore a self-calibration approach working in all kinds of road scenes is highly favorable.

Applied to images of road scenes, 3D reconstruction and localization algorithms solely based on appearance features might suffer from problems caused by road scene-specific groups of critical objects. As first group, objects with reflecting surfaces like building windows, vehicle window panes or metallic vehicle paint could lead to incorrect feature correspondences. As second group, poor-texture objects with large homogeneous areas like sky or with small-scale repetitive textures like tarmac could lead to incorrect correspondences as well. As third group, moving objects like pedestrians or other vehicles or trees in the wind could cause a bad reconstruction as 2D image points of the same feature in different images might not be related to the same 3D road scene point even for correct matches. These scene-specific problems with correspondence search and reconstruction could subsequently have negative influence on vehicle camera self-calibration. As these problems can be related to certain groups of objects, a priori knowledge about object groups shown in road scene images could be used to avoid negative influences on camera calibration. As correspondence search covers the complete image, semantic segmentation methods providing pixel-wise semantic knowledge about the objects shown in an image can be used.

The main contributions of this paper are:

- A new method for on-board vehicle camera self-calibration by a semantic-guided structure from motion algorithm, whereby semantic knowledge about objects being present in images is applied to create semantic masks disabling feature extraction in image parts showing critical objects like moving objects in order to make calibration more robust.
- An analysis of the effect of different semantic classes used for creating semantic masks on the structure from motion algorithm and on the consecutive global bundle adjustment.
- An analysis and statistical evaluation of the interior and distortion parameters estimated by the proposed method in comparison to a reference calibration with a 3D test field.

2. RELATED WORK

2.1 3D reconstruction and localization

3D reconstruction of the environment and localization can be achieved by approaches like structure from motion (SfM), visual SLAM and visual odometry based on data from different kinds of sensors, like monocular (Engel et al., 2014), (Mur-Artal et al., 2015), stereo (Wang et al., 2017), or RGB-D (Henry et al., 2012), (Nießner et al., 2013) cameras, or LiDAR (Jiang et al., 2016), (Graeter et al., 2018), for example. SfM can be seen as the more general approach compared to visual odometry and visual SLAM aiming at a globally consistent system of a 3D reconstruction of the environment including all camera poses. Unlike the other two kinds of approaches, SfM methods typically don't aim at incremental processing or real-time performance, whereby computationally expensive offline optimization steps (i.e. bundle adjustment) allow to obtain global consistency even without requiring looped trajectories (Scaramuzza and Fraundorfer, 2011). As there are no real-time requirements and extensive optimization is favorable for camera calibration, the proposed method relies on a SfM method for 3D reconstruction. Further, the proposed method uses an indirect method for 3D reconstruction and localization, i.e. a method relying on feature descriptors like SIFT (Lowe, 1999) or SURF (Bay et al., 2006), in contrast to a direct method comparing intensities between different image patches, as for the later ones problems have been reported for auto exposure cameras and in the case of vignetting (Bergmann et al., 2018), which both could apply for on-board cameras, besides often suffering in the case of large motions between consecutive images (Younes et al., 2019), which could apply for images recorded at high vehicle velocities.

2.2 Automotive camera self-calibration

On-board cameras in vehicles might be installed in different kinds of sensor settings, therefore there have been calibration approaches proposed for mono cameras (Miksch et al., 2010), for multi camera systems or for multi sensor systems (Dang et al., 2009) (Rehder et al., 2017) like camera and LiDAR (Geiger et al., 2012), (Levinson and Thrun, 2013). Though this contribution proposes a method to be used with a mono camera, it can be easily extended for a multi camera or multi sensor setting. Approaches for automotive camera self-calibration are either aiming at intrinsic calibration (Houben, 2014), (Keivan and Sibley, 2015) (Hanel and Stilla, 2018), i.e. estimating the interior orientation and distortion parameters of the cameras, at extrinsic calibration (Ruland et al., 2010), (Heng et al., 2014), i.e. estimating the parameters of the exterior orientation, or at both (Heng et al., 2013). The proposed method allows for both intrinsic and extrinsic calibration simultaneously.

Some approaches for automotive camera self-calibration rely on object-specific features on road objects like road markings (Ribeiro et al., 2006) or traffic signs (Hanel and Stilla, 2018) and are often constrained by scene conditions like a flat ground plane (e.g. (Catala-Prat et al., 2006)), the presence of the objects, or by special driving behavior, like the need to driving nearly parallel to the markings (Ribeiro et al., 2006). In contrast, approaches relying on image features detected by descriptors like SIFT (Dang et al., 2009) or SURF (Heng et al., 2014) are typically not subject to the aforementioned constraints. Though, robustness might be hard to achieve by a SfM method for camera self-calibration due to problems with correspondence search in road scene images, as reported by (Ruland et al., 2010) for the cases of poor illumination or poor road textures. Additionally it has been reported that moving cars or pedestrians could have negative influence on camera calibration, wherefore some methods include special outlier removal steps to make calibration more robust, like (Dang et al., 2009). To overcome both problems, the proposed method uses additional semantic knowledge to exclude critical objects

from correspondence search and subsequently from the reference points for calibration.

2.3 Semantic segmentation

Semantic segmentation (Ronneberger et al., 2015), (Chen et al., 2016), (Shelhamer et al., 2017), (Badrinarayanan et al., 2017), (Chen et al., 2018a) aims at predicting a semantic image (figure 1 top right) giving pixel-wise information about the semantic class of objects shown in the corresponding RGB image (figure 1 top left). For semantic segmentation of road scene images, often classes like vegetation, vehicle, building or road are distinguished, often based on the class definition by (Cordts et al., 2016). While a segment obtained by semantic segmentation can contain multiple individual objects of this class, do approaches for instance segmentation (Hariharan et al., 2014), (He et al., 2017) distinguish individual objects by pixel-wise segments. In contrast, methods for object detection (Ren et al., 2017), (Liu et al., 2016), (Redmon and Farhadi, 2017)), (Girshick et al., 2014) typically describe objects only by an enclosing rectangle. Both methods for instance segmentation and object detection do typically address only object classes of interest, like pedestrians or cars, wherefore no knowledge about the semantics of image parts showing other objects is obtained. As image features might be detected in all image parts, pixel-wise knowledge for the complete image as obtained by semantic segmentation is favorable and therefore used in the proposed method.

2.4 Semantic 3D reconstruction and localization

Semantic knowledge has been already used in previous work together with 3D reconstruction and localization approaches. As one example (Li and Belaroussi, 2016), (Mahe et al., 2018) and (Runz et al., 2018) enrich the 3D reconstruction by assigning semantic information to the 3D points. As another example, (Hirzer et al., 2017), (Schönberger et al., 2018) use semantic knowledge for localization within a 3D reconstruction. More similar to the proposed method, (Murali et al., 2017) integrate semantic knowledge into the 3D reconstruction and localization pipeline at different stages: during feature tracking, 3D reconstruction and the navigation process. In contrast to our work, they are working with gray value images only, are aiming at real-time navigation and are assuming sensors that have been already calibrated. Also similar to the proposed method, (Wang et al., 2018) and (Yu et al., 2018) are using knowledge about the presence of dynamic objects in images to remove feature outliers to make the 3D reconstruction more robust. Both methods are treating movable objects by special processing steps, underlining their potential for causing problems when using 3D reconstruction and localization approaches for self-calibration in dynamic environments like public road scenes. Most similar to the proposed method, (Kaneko et al., 2018) propose to exclude image parts from feature extraction based on semantic knowledge obtained by image-based semantic segmentation. In empirical studies, they have found that *sky* and *car* are the most relevant classes for exclusion. In contrast to their work, the method proposed in this paper relies on SfM instead of visual SLAM, focuses on camera calibration and is applied to real images in contrast to synthetic images only.

3. SEMANTIC ROAD SCENE KNOWLEDGE AS A PRIORI INFORMATION FOR SELF-CALIBRATION

In this section, the proposed approach for using semantic road scene knowledge as a priori information for a 3D reconstruction algorithm in order to make vehicle camera self-calibration more robust is proposed. The workflow (figure 2) covers the following steps, which will be explained in detail in the following paragraphs. First, semantic segmentation is applied to all recorded images to obtain a semantic mask for each image. Second, an exclusion mask is derived from each semantic mask consisting of image segments belonging to semantic classes of critical objects which should be excluded from correspondence search to avoid the aforementioned problems. Third, a 3D point cloud of the road scene is obtained by a structure form motion algorithm using the recorded road scene image sequence together with the exclusion masks for correspondence search and incremental reconstruction. Forth, after having completed the reconstruction, the projective 3D point cloud of the road scene is transformed by spatial similarity transform using vehicle position data available for the image recording time points from GPS to obtain a Euclidean 3D point cloud with metric scale. Fifth, a global bundle adjustment is performed to obtain optimized estimates for the image and scene points as well as the interior and exterior camera parameters. The method is intended to be performed with an image sequence recorded from a road scene with a vehicle camera that should be calibrated.

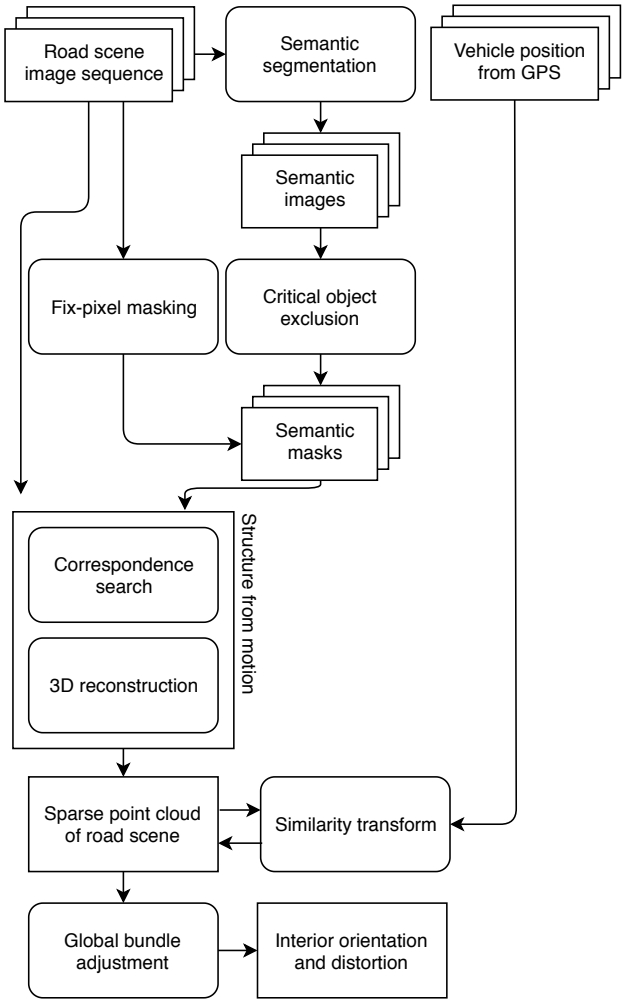

Figure 2. Workflow of the proposed method for robust self-calibration of a vehicle camera by using semantic road scene knowledge in a semantic-guided structure from motion algorithm.

3.1 Semantic segmentation

Semantic knowledge is obtained by image-based semantic segmentation. Inference with a trained model is done for each RGB image of the calibration image sequence, resulting in pixel-wise semantic images (figure 1 top right). The model has been trained on road scene images using ground truth annotations distinguishing typical semantic classes for road scenes, like *road*, *vehicle* or *building*. As a model is used that has been trained by others (e.g. (Chen et al., 2018a)), the need for own time-intensive hyperparameter tuning to optimize the model performance is avoided, and a model can be selected that has been trained using large-scale datasets requiring special high performance computing facilities to be available (e.g. 370 GPUs as reported by (Chen et al., 2018b). Furthermore, by using a trained model, the often costly need to acquire ground truth annotation for own image sequences can be avoided.

3.2 Semantic masking

Semantic masking is intended to exclude image parts showing potentially critical objects from feature extraction to avoid unreliable reference points for camera calibration. Potentially critical objects are identified by their semantic class.

Based on the previously obtained semantic images, a binary semantic mask is created for each RGB image that should be used for feature extraction, whereby based on semantic segments one value (white color in figure 1 bottom left) is assigned to pixels that should be considered for feature extraction, while the other value (black color in figure 1 bottom left) is assigned to pixels that should not be considered for feature extraction. A list of critical semantic classes that allow to extract the semantic segments of the respective classes from the semantic images that should not be used for feature extraction has to be provided manually. Different combinations of classes have been investigated for their relevance with regard to camera self-calibration, see section 5.

3.3 Fix-pixel masking

Fix-pixel masking is intended to be applied for image parts showing the same objects in each image. In road scene images acquired by a forward-looking on-board camera, this might be true especially for parts of the ego-car like the bonnet or the windscreen. There are multiple reasons why fix-pixel masking seems reasonable especially for the case of the ego-car in road scene images. First, though feature matches between different images showing the ego-car could probably be found easily, are they undesired as the corresponding 3D scene points would not be in a consistent coordinate system with the 3D points triangulated from static parts of the scenes, besides that the baseline necessary for triangulation would be completely missing. Second, besides the aforementioned geometric problems, especially in the case of the ego-car having reflecting metal surfaces or reflecting windshields, there might be undesired feature matches being established between "real" objects and their reflections in RGB images. Third, as the semantic knowledge obtained by semantic segmentation relies on processing of these RGB images, reflections might cause wrong semantic classes to be assigned, which would consequently lead to errors in the subsequent steps like semantic masking or semantic matching.

Fix-pixel segments are created by manually defining the contour polygon of the segment covering the desired image part (e.g. showing the ego-car). Afterwards, the fix-pixel segments are fused with the semantic segments (see subsection 3.2) into a common semantic mask.

3.4 Remaining structure from motion pipeline

The remaining structure from motion pipeline used for this paper follows the method proposed by (Schönberger and Frahm, 2016), and covers the following steps: After feature extraction considering the semantic masks, exhaustive matching is performed, i.e. matches are searched in all image pairs, as no real-time requirements have to be met. Having correspondences for all images established after the matching step, a sparse 3D point cloud of the road scene is obtained by sparse reconstruction. The reconstruction is initialized with a random image pair, then the images are registered, 3D coordinates of scene points calculated by triangulation and optimized by bundle adjustment. These scene points define the reference points for calibration. For sparse reconstruction, the same camera model is used for all images, assuming that there have been no changes in the interior orientation and distortion parameters during the time the image sequence has been acquired. After sparse reconstruction, the point cloud is transformed to integrate metric scaling into the point cloud based on a similarity transformation. As the same camera model is used for all images, sparse reconstruction provides already an Euclidean point cloud instead of a projective one (Hartley, 1993), wherefore a similarity transformation is sufficient for scale integration. The transformation parameters are calculated using camera positions for which external position information from GPS is available. As the time points and frequency of GPS positions and images don't match, filtered camera positions are obtained by linear interpolation between raw camera positions from GPS. As last step, a global bundle adjustment is performed on the final transformed 3D point cloud reconstructed from all images of the sequence to obtain optimized estimates for the scene points and the interior and exterior camera parameters.

4. EXPERIMENTS

The proposed method is evaluated with a sequence of 300 road scene images recorded by the multi-sensor vehicle MODISSA (Borgmann et al., 2018) during a test drive through a suburban environment. Semantic segmentation is done using the Deeplabv3+ network (Chen et al., 2018a), which is in one of the leading places in the Cityscapes benchmark for semantic segmentation of road scene images (Cordts et al., 2019). The used model has been trained by the Deeplabv3+ developers on the Cityscapes dataset (Cordts et al., 2016). 3D reconstruction is done with COLMAP (Schönberger and Frahm, 2016), which has shown best performance in a comparison with other structure from motion algorithms (Bianco et al., 2018). A camera model with two focal length parameters, two principal point parameters and two radial and two tangential distortion parameters according to (Brown, 1971) is used.

4.1 Semantic masking experiments

The list of semantic classes used to derive the semantic masks follows the Cityscapes class definition (Cordts et al., 2016), whereby several classes are grouped by the following categories: First *vehicle: bicycle, bus, car, caravan, license plate, motorcycle, trailer, train, truck*, second *nature: terrain, vegetation*, third *human: person, rider*, forth *construction: bridge, building, fence, guard rail, tunnel, wall*, fifth *flat: parking, rail track, road, sidewalk*, sixth *object: pole, pole group, traffic light, traffic sign*; last, the class *sky* forms its own category. Additionally, the *ego car* mask is given.

The following experiments have been performed with different sets of semantic classes that have been excluded from feature extraction by applying the appropriate mask:

- Experiment 1: No semantic masking, defines the baseline
- Experiment 2: Additionally to 1, *ego car*, category *vehicle*
- Experiment 3: Additionally to 1, *ego car*, category *nature*
- Experiment 4: Additionally to 1, *ego car*, category *sky*
- Experiment 5: Additionally to 1, *ego car*, category *human*
- Experiment 6: Additionally to 1, *ego car*, category *construction*
- Experiment 7: Additionally to 1, *ego car*, category *flat*
- Experiment 8: Additionally to 1, *ego car*, category *object*
- Experiment 9: Additionally to 1, *ego car*, all road users, i.e. categories *human, vehicle*
- Experiment 10: Additionally to 1, *ego car*, all movable objects, i.e. categories *human, sky, vegetation, vehicle*

4.2 Reference calibration

For obtaining reference values for a comparison with the outcome of the proposed method, test field camera calibration is done. According to (Luhmann et al., 2016), calibration with 2D test fields can be disadvantageous in terms of accuracy and correlations between parameters, wherefore the reference calibration relies on a 3D test field consisting of three orthogonal planes forming an "open cube" with checkerboard patterns on each plane; the checkerboard corners define the reference points. Reference calibration is done by the following steps. First, The pixel coordinates of the corners are extracted by the method described by (Geiger et al., 2012). The world coordinates are obtained based on the known size of the checkerboard squares. Second, initial values for the interior orientation and distortion parameters are obtained by single-plane calibration using openCV (OpenCV, 2017). Third, updated values for the interior orientation and the distortion parameters as well as the relative orientation between the test field and the calibration images are obtained by multi-plane calibration using openCV. For both the second and third step, the complete set of calibration images is divided based on random separation into multiple subsets to filter out outlier images and to make a single calibration step computationally less expensive. The final interior parameter and distortion coefficients values are obtained by averaging over successful multi-plane calibrations with the aforementioned subsets, whereby a low reprojection error is used as measure for success. In average, a reprojection error of 0.666 px is obtained.

5. EVALUATION WITH AN SUBURBAN ROAD SCENE IMAGE SEQUENCE

The following description of the results is three-folded, taking data statistics, feature distributions and a comparison between the proposed method and the reference calibration into account. Each experiment is run multiple times with different initial image pairs as an important non-deterministic factor in a SfM pipeline, leading to different 3D reconstructions. Thereby, precision values are obtained that are used for statistical hypothesis tests.

Exp.	#F	#M	#3D	R [1]	RPE [px]
1	1580586	17172	69083	785084	0.477
2	1466027	13239	66562	760181	0.477
3	429435	8755	29032	264624	0.527
4	1492879	13200	68219	772647	0.476
5	1509078	13447	68809	788617	0.481
6	1381974	13089	62895	688276	0.477
7	1284702	12451	47771	640743	0.490
8	1497729	13398	68156	779162	0.480
9	1464801	13211	66591	759214	0.477
10	384137	6630	28079	245081	0.447

Table 1. Dataset statistics with number of features (#F), number of matches (#M), number of 3D points (#3D), redundancy (R) and mean reprojection error of all feature points (RPE). #3D, R and RPE are averaged across all test runs of each experiment.

5.1 Data statistics

Dataset statistics (table 1) show that the baseline method (experiment 1) has a larger number of features than any of the experiments with semantic masks (experiments 2 - 10), what is intuitive through the nature of excluding image parts from feature extraction. While for the most experiments the feature number drops by less than 20 percent, is the drop larger than 70 percent for experiments 3 and 10. Interestingly, experiment 3 shows the worst reprojection error, while experiment 10 shows the best one across all experiments. Excluding experiments 3 and 10, the drop in the number of matches relative to the baseline method is larger than the drop in the number of extracted features. More interestingly, except for experiment 7, the drop in the number of 3D points relative to the baseline is smaller than the drop in the number of features, which might be interpreted that the features of semantic-guided SfM are better suitable for triangulating 3D points. For every experiment, the redundancy is larger than the redundancy of the reference test field calibration, bringing one major advantage compared to test field calibration into account. For most experiments, the reprojection error deviates around 1/100 px from the baseline method; only for experiment 10, the reprojection error decreases by approximately 3/100 px. All experiments show a smaller reprojection error than reference calibration (see subsection 4.2).

5.2 Feature distribution

The feature distribution across the complete image has special importance for camera calibration with regard to the validity of estimated parameters only for certain image parts or the complete image (Luhmann et al., 2006). By visual inspection of figure 3, an uneven distribution of the features can be seen for all experiments, ranging from areas with no or only a few features (dark blue color) until areas with around 100 features; though, more than approximately 60 features occur only in very small image parts (e.g. bonnet in experiment 1). For example, in the lower left image part showing the road (cf. figure 1), the number of features is lower compared to other image parts typically showing surrounding objects. For some experiments, individual observations can be made. For experiment 7, the center part of the lower image half contains less features than for most other experiments. For experiments 3 and 10, remarkably larger image parts contain no or only few features (dark blue color) compared to all other experiments; this observation matches with the lower numbers in table 1. For experiment 1, features on the car bonnet are detected very often; though, they are not reliable for SfM as there are no

motions between images for these features. At least from visual inspection, negative impact of semantic masking on the feature distribution cannot be seen for the most experiments.

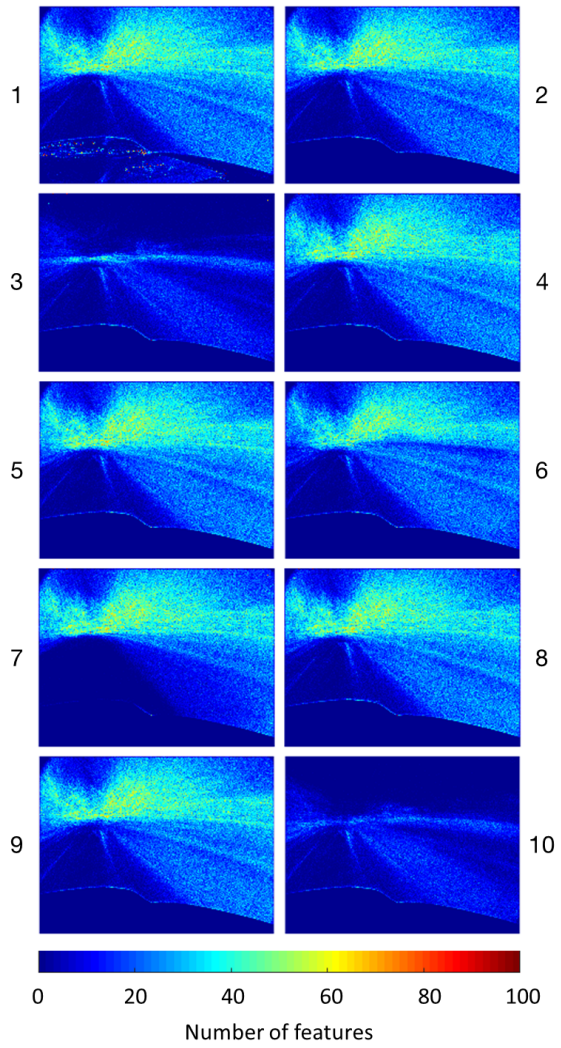

Figure 3. Distribution of features across the complete image for the different experiments. For only very few pixels, features occur more than 60 times. Values are referred to the complete sequence of 300 images.

5.3 Deviation from reference calibration

From the deviations between interior orientation and distortion parameter values estimated by the proposed method and the reference method, several observations can be made by visual (figure 4) and numerical analysis. First, experiment 6 shows the smallest deviations for three parameters (x component of focal length f_x and tangential distortion parameters p_1, p_2), while experiments 3 (y component of focal length f_y, y component of the principal point c_y) and 10 for two parameters (radial distortion parameters k_1, k_2) and experiment 7 for one (x component of the principal point c_x). The baseline method never shows the smallest deviations from reference. Second, remarkably, experiment 7 shows visually larger deviations than other experiments for almost all parameters. Third, the radial distortion parameters k_1, k_2 have a different sign for experiment 10. Forth, all deviations for the y component of both focal length f_y and principal point c_y are remarkably larger than the deviations of the corresponding x components f_x and c_x. Fifth, experiment 6 shows the best precision (least values) for three parameters (f_y, c_y and p_1), experiments 2 (k_1, k_2) and 4 (c_x, p_2) for two parameters and experiment 8 for one parameter (f_x). Again, the baseline method never shows the least deviation. Sixth, statistical hypothesis testing shows that no deviation between reference and the experiments is significant for the parameters f_x, c_x, and all distortion parameters. For f_y, c_y, only the deviations for experiment 3 are significant; though, it has to be considered that the precision for experiment 3 is worse up to a factor of 100 compared to the other experiments. Seventh, further tests show that deviations for some of the estimated interior orientation and distortion parameters from the baseline method are significant for each experiment except number three; for no experiment, all deviations are significant.

5.4 Discussion

The proposed method relies on good performing semantic segmentation; though, the proposed method currently does not assess the quality of the segmentation nor includes special processing steps to reduce or avoid the influence of errors like averaging semantic masks from consecutive time points.

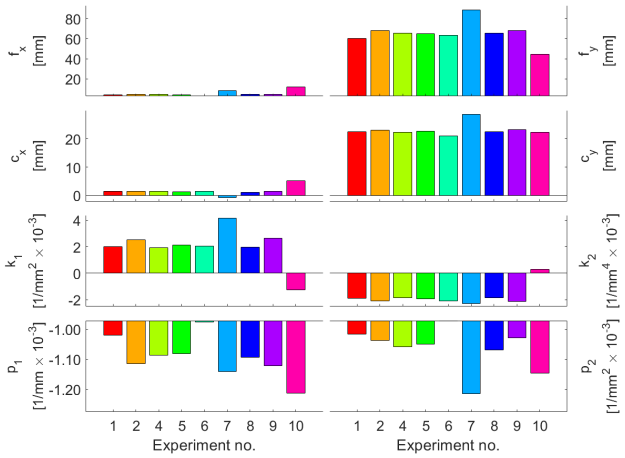

Figure 4. Deviations of interior orientation and distortion parameter values estimated by the proposed method from values estimated by reference calibration. Same scale for each left-right pair. Experiment 3 is not shown as some of its deviation values are by magnitudes larger compared to the other experiments.

Regarding camera calibration, a more through evaluation of the estimated interior orientation and distortion parameters could give more insights into the benefits and limitations of the proposed method; rather for self-calibration than for test field calibration, strong correlations between estimated camera parameters might occur. This applies for vehicle camera calibration in particular, as some degrees of freedom between camera and reference points are fixed, like the rotation around the optical axis, making more thorough evaluation beneficial. Additionally, the reference calibration was obtained by averaging over several image subsets using openCV-based calibration, no bundle adjustment using the complete image set has been performed. Nevertheless, the proposed method has shown to be able to achieve smaller deviations from reference calibration and better precision values than the baseline method. The selection of the appropriate semantic classes for generating the semantic mask appears to be important with regard to the deviations and the precision. Best performance (see discussion of deviations and precision in subsection 5.3) is

observed for semantic masks considering the ego-car and the semantic category *construction*.

6. CONCLUSION

In this contribution, a method for on-board vehicle camera self-calibration relying on a semantic-guided structure from motion algorithm has been proposed. Image-based semantic segmentation is applied to create semantic masks excluding image parts showing critical objects like moving objects with potential negative influence on camera calibration from feature extraction in the structure from motion pipeline. The method has been tested on an image sequence recorded in an suburban road scene. Results show that for semantic masks obtained for selected semantic classes, smaller deviations from a reference camera calibration can be obtained compared to not using the semantic guidance in the structure from motion algorithm. Obviously, future work should aim at integrating the semantic guidance into other steps of the structure from motion pipeline, like keypoint matching or sparse reconstruction. Furthermore, more comprehensive statistical information about the estimated parameters, especially covariance matrices, should be derived from bundle adjustment and be used for a more thorough evaluation of the proposed method. Last, to prove the applicability for use in automotive applications, the robustness of the proposed method in various road scenes with different kinds of buildings and vegetation or in bad weather conditions, for example, should be investigated.

ACKNOWLEDGEMENT

The authors would like to thank Dr. Marcus Hebel, Björn Borgmann and Joachim Gehrung from the Fraunhofer Institute of Optronics, System Technologies, and Image Exploitation (Fraunhofer IOSB) for acquiring and providing both test data for the proposed method as well as data for the reference calibration.

REFERENCES

Azzopardi, M.A., Grech, I., Leconte, J., 2010. A high speed trivision system for automotive applications. *European Transport Research Review*, 2(1), 31–51.

Badrinarayanan, V., Kendall, A., Cipolla, R., 2017. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*,39(12), 2481–2495.

Bay, H., Tuytelaars, T., Van Gool, L., 2006. SURF: Speeded Up Robust Features. In: A. Leonardis, H. Bischof and A. Pinz (eds), *Computer Vision – ECCV 2006*, Springer Berlin Heidelberg, Berlin, Heidelberg, 404–417.

Bergmann, P., Wang, R., Cremers, D., 2018. Online Photometric Calibration of Auto Exposure Video for Realtime Visual Odometry and SLAM. *IEEE Robotics and Automation Letters*, 3(2), 627–634.

Bertozzi, M., Bombini, L., Broggi, A., Grisleri, P., Porta, P. P., 2010. *Smart Cameras*. Springer US, Boston, MA, chapter Camera-Based Automotive Systems, 319–338.

Bianco, S., Ciocca, G., Marelli, D., 2018. Evaluating the Performance of Structure from Motion Pipelines. *Journal of Imaging*.

Borgmann, B., Schatz, V., Kieritz, H., Scherer-Klöckling, C., Hebel, M. , Arens, M., 2018. Data processing and recording using a versatile multi-sensor vehicle. *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, IV-1, 21–28.

Brown, D. C., 1971. Close-range camera calibration. *Photogrammetric Engineering*, 37(8), 855–866.

Catala-Prat, A., Rataj, J., Reulke, R., 2006. Self-Calibration Systemfor the Orientation of a Vehicle Camera. In: *ISPRS Comision V Symposium: Image Engineering and Vision Metrology*.

Chen, L., Barron, J. T., Papandreou, G., Murphy, K., Yuille, A. L., 2016. Semantic Image Segmentation with Task-Specific Edge Detection Using CNNs and a Discriminatively Trained Domain Transform. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, 4545–4554.

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H., 2018a. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. In: V. Ferrari, M. Hebert, C. Sminchisescu and Y. Weiss (eds), *Computer Vision – ECCV 2018*, Springer International Publishing, Cham, 833–851.

Chen, L., Collins, M. D., Zhu, Y., Papandreou, G., Zoph, B., Schroff, F., Adam, H., Shlens, J., 2018b. Searching for Efficient Multi-Scale Architectures for Dense Image Prediction. In: *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada.*, 8713–8724.

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S. and Schiele, B., 2016. The Cityscapes Dataset for Semantic Urban Scene Understanding. In: *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S. and Schiele, B., 2019. Cityscapes Dataset - Benchmark Suite - Pixel-Level Semantic Labeling Task. Website. https://www.cityscapes-dataset.com/benchmarks/pixel-level-results.

Dang, T., Hoffmann, C., Stiller, C., 2009. Continuous Stereo Self-Calibration by Camera Parameter Tracking. *IEEE Transactions on image processing*, 18(7), 1536–1550.

Dubey, A., 2016. Stereo vision - Facing the challenges and seeing the opportunities for ADAS applications. Brochure.

Engel, J., Schöps, T., Cremers, D., 2014. LSD-SLAM: Large-Scale Direct Monocular SLAM. In: *European Conference on Computer Vision (ECCV)*.

Enzweiler, M., Gavrila, D. M., 2009. Monocular Pedestrian Detection: Survey and Experiments. *IEEE Transactions on Pattern Analysisand Machine Intelligence*, 31(12), 2179–2195.

Ernst, S., Stiller, C., Goldbeck, J., Roessig, C., 1999. Camera calibration for lane and obstacle detection. In: *Proceedings 199 IEEE/IEEJ/JSAI International Conference on Intelligent Transportation Systems (Cat. No.99TH8383)*, 356–361.

Geiger, A., Moosmann, F., Car, Ö., Schuster, B., 2012. Automatic camera and range sensor calibration using a single shot. In: *2012 IEEE International Conference on Robotics and Automation*, 3936–3943.

Girshick, R., Donahue, J., Darrell, T., Malik, J., 2014. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In: *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 580–587.

Graeter, J., Wilczynski, A., Lauer, M., 2018. LIMO: Lidar-Monocular Visual Odometry. In: *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IEEE, 7872–7879.

Hanel, A. and Stilla, U., 2018. Iterative Calibration of a Vehicle Camera using Traffic Signs detected by a Convolutional Neural Network. In: *Proceedings of the 4th International Conference on Vehicle Technology and Intelligent Transport Systems*.

Hanel, A., Kreuzpaintner, D., Stilla, U., 2018. Eevaluation of a traffic sign detector by synthetic image data for advanced driver assistance systems. *ISPRS - Internation-al Archives of the Photogrammetry, Remote Sensing and Spatial Informa-tion Sciences*, XLII-2, 425–432.

Hariharan, B., Arbeláez, P., Girshick, R., Malik, J., 2014. Simultaneous Detection and Segmentation. In: D. Fleet, T. Pajdla, B. Schiele and T. Tuytelaars (eds), *Computer Vision – ECCV 2014*, Springer Internation-al Publishing, Cham, 297–312.

Hartley, R. I., 1993. Euclidean reconstruction from uncalibrated views. In: *Joint European-US workshop on applications of invariance in computer vision*, Springer, 235–256.

He, K., Gkioxari, G., Dollár, P., Girshick, R., 2017. Mask R-CNN.In: *2017 IEEE International Conference on Computer Vision (ICCV)*, 2980–2988.

Hella Gutmann Solutions GmbH, 2016. CSC-Tool. Operating Instructions.

Heng, L., Bürki, M., Lee, G. H., Furgale, P., Siegwart, R., Pollefeys, M., 2014. Infrastructure-based calibration of a multi-camera rig. In: *2014 IEEE International Conference on Robotics and Automation (ICRA)*, 4912–4919.

Heng, L., Li, B., Pollefeys, M., 2013. CamOdoCal: Automatic intrin-sic and extrinsic calibration of a rig with multiple generic cameras and odometry. In: *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IEEE, 1793–1800.

Henry, P., Krainin, M., Herbst, E., Ren, X., Fox, D., 2012. RGB-D mapping: Using Kinect-style depth cameras for dense 3D modelingof indoor environments. *The International Journal of Robotics Research*, 31(5), 647–663.

Hirzer, M., Roth, P. M., Lepetit, V., 2017. Efficient 3D Tracking in Urban Environments with Semantic Segmentation. In: *BMVC*, BMVA Press.

Houben, S., 2014. Towards the intrinsic self-calibration of a vehicle-mounted omni-directional radially symmetric camera. In: *2014 IEEE Intelligent Vehicles Symposium Proceedings*, 878–883.

Janai, J., Güney, F., Behl, A., Geiger, A., 2017. Computer Vision for Autonomous Vehicles: Problems, Datasets and State-of-the-Art. *CoRR*.

Jiang, C., Paudel, D. P., Fougerolle, Y., Fofi, D., Demonceaux, C., 2016. Static-Map and Dynamic Object Reconstruction in Outdoor Scenes Using 3-D Motion Segmentation. *IEEE Robotics and Automation Letters* , 1(1), 324–331.

Kaneko, M., Iwami, K., Ogawa, T., Yamasaki, T., Aizawa, K., 2018. Mask-SLAM: Robust Feature-Based Monocular SLAM by Masking Using Semantic Segmentation. In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 371–379.

Keivan, N., Sibley, G., 2015. Online SLAM with any-time self-calibration and automatic change detection. *2015 IEEE International Conference on Robotics and Automation (ICRA)*, 5775–5782.

Keller, C. G., Enzweiler, M., Rohrbach, M., Llorca, D. F., Schnorr, C., Gavrila, D. M., 2011. The Benefits of Dense Stereo for Pedestrian Detection. *IEEE Transactions on Intelligent Transportation Systems*, 12 (4),1096–1106.

Levinson, J., Thrun, S., 2013. Automatic Online Calibration of Cameras and Lasers.

Li, X., Belaroussi, R., 2016. Semi-Dense 3D Semantic Mapping from Monocular SLAM. *CoRR*.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S. E., Fu, C., Berg, A. C., 2016. SSD: Single Shot MultiBox Detector. In: *European Conference on Computer Vision (ECCV)*, Lecture Notes in Computer Science, 9905, Springer, 21–37.

Lowe, D. G., 1999. Object recognition from local scale-invariant features. In: *Proceedings of the Seventh IEEE International Conference on Computer Vision*, 2, 1150–1157.

Luhmann, T., Fraser, C., Maas, H.-G., 2016. Sensor modelling and camera calibration for close-range photogrammetry. *ISPRS Journal of Photogrammetry and Remote Sensing*, 115, 37–46. Theme issue 'State-of-the-art in photogrammetry, remote sensing and spatial information science'.

Luhmann, T., Robson, S., Kyle, S., Harley, I., 2006. *Close Range Photogrammetry. Principles, Methods and Applications*. Whittles Publishing.

Mahe, H., Marraud, D. and Comport, A. I., 2018. Semantic-only Visual Odometry based on dense class-level segmentation. *2018 24th International Conference on Pattern Recognition (ICPR)* pp. 1989–1995.

Miksch, M., Yang, B., Zimmermann, K., 2010. Homography-based extrinsic self-calibration for cameras in automotive applications. In: *Workshop on Intelligent Transportation*, 17–22.

Mur-Artal, R. Tardos, J. D., 2017. ORB-SLAM2: an Open-Source SLAM System for Monocular, Stereo and RGB-D Cameras. *IEEE Transactions on Robotics*, 33(5), 1255–1262.

Mur-Artal, R., Montiel, J. M. M., Tardós, J. D., 2015. ORB-SLAM:a Versatile and Accurate Monocular SLAM System. *IEEE Transactions on Robotics*, 31(5), 1147–1163.

Murali, V., Chiu, H., Samarasekera, S., Kumar, R. T., 2017. Utilizing semantic visual landmarks for precise vehicle navigation. In: *ITSC*, IEEE, 1–8.

Nießner, M., Zollhöfer, M., Izadi, S., Stamminger, M., 2013. Real-time 3D Reconstruction at Scale Using Voxel Hashing. *ACM Trans. Graph.*, 32(6), 169:1–169:11.

OpenCV, 2017. Camera Calibration and 3D Reconstruction. Website, http://opencv.org. 2017-01-30.

Paula, M. B. D., Jung, C. R., Silveira, L. G. D., 2014. Automatic on-the-fly extrinsic camera calibration of onboard vehicular cameras. *Expert Systems with Applications: An International Journal*, 41(4), 1997–2007.

Redmon, J., Farhadi, A., 2017. YOLO9000: Better, Faster, Stronger.In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 6517–6525.

Rehder, E., Kinzig, C., Bender, P., Lauer, M., 2017. Online stereo camera calibration from scratch. *2017 IEEE Intelligent Vehicles Symposium (IV)*, 1694–1699.

Ren, S., He, K., Girshick, R., Sun, J., 2017. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(6), 1137–1149.

Ribeiro, A. A. G. A., Dihl, L. L., Jung, C. R., 2006. Automatic cam-era calibration for driver assistance systems. In: *Proceeedings of 13th International Conference on Systems, Signals and Image Processing*, 173–176.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. In: N. Navab, J. Hornegger, W. M. Wells and A. F. Frangi (eds), *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, Springer International Publishing, Cham, 234–241.

Ruland, T., Loose, H., Pajdla, T., Krüger, L., 2010. Hand-eye autocalibration of camera positions on vehicles. In: *13th International IEEE Conference on Intelligent Transportation Systems*, 367–372.

Runz, M., Buffier, M., Agapito, L., 2018. MaskFusion: Real-Time Recognition, Tracking and Reconstruction of Multiple Moving Objects. In: *2018 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, 10–20.

Scaramuzza, D., Fraundorfer, F., 2011. Visual Odometry [Tutorial]. *IEEE Robotics & Automation Magazine*, 18(4), 80–92.

Schönberger, J. L., Frahm, J., 2016. Structure-from-Motion Revisited.In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4104–4113.

Schönberger, J., Pollefeys, M., Geiger, A., Sattler, T., 2018. Seman-tic Visual Localization. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE Computer Society.

Shelhamer, E., Long, J., Darrell, T., 2017. Fully Convolutional Networks for Semantic Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4), 640–651.

Stamatopoulos, C., Fraser, C. S., 2014. Automated Target-Free Network Orienation and Camera Calibration. *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, II-5, 339–346.

Thatcham Research and ADAS Repair Group, 2016. Code of Practice For the Replacement & Refitting of Automotive Glazing for vehicles fitted with screen mounted Advanced Driver Assistance Systems (ADAS).

Wang, K., Lin, Y., Wang, L., Han, L., Hua, M., Wang, X., Lian, S., Huang, B., 2018. A Unified Framework for Mutual Improvement of SLAM and Semantic Segmentation. *CoRR*.

Wang, R., Schwörer, M., Cremers, D., 2017. Stereo DSO: Large-Scale Direct Sparse Visual Odometry with Stereo Cameras. In: *International Conference on Computer Vision (ICCV)*, Venice, Italy.

Winner, H., Hakuli, S., Lotz, F., Singer, C., 2015. *Handbook of Driv-er Assistance Systems: Basic Information, Components and Systems for Active Safety and Comfort*. 1st edn, Springer Publishing Company, Incorporated.

Younes, G., Asmar, D., Zelek, J., 2019. FDMO: Feature Assisted Direct Monocular Odometry. In: *14th International Conference on Computer Vision Theory and Applications*.

Yu, C., Liu, Z., Liu, X.-J., Xie, F., Yang, Y., Wei, Q., Fei, Q., 2018. DS-SLAM: A Semantic Visual SLAM towards Dynamic Environments. In: *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 1168–1174.

Zabatani, A., Bareket, S., Menashe, O., Sperling, E., Bronstein, A., Bronstein, M., Kimmel, R., Surazhsky, V., 2017. Online compensation of thermal distortions in a stereo depth camera. Patent.

Zhang, B., Appia, V., Pekkucuksen, I., Liu, Y., Batur, A. U., Shastry, P., Liu, S., Sivasankaran, S., Chitnis, K., 2014. A Surround View Camera Solution for Embedded Systems. In: *2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 676–681.

Ziebinski, A., Cupek, R., Erdogan, H., Waechter, S., 2016. A Survey of ADAS Technologies for the Future Perspective of Sensor Fusion. In: N. T. Nguyen, L. Iliadis, Y. Manolopoulos and B. Trawiński (eds), *Computational Collective Intelligence*, Springer International Publishing, Cham, 135–146.