ROBUST VISUAL-INERTIAL ODOMETRY IN DYNAMIC ENVIRONMENTS USING SEMANTIC SEGMENTATION FOR FEATURE SELECTION

Patrick Irmisch*, Dirk Baumbach, Ines Ernst

Institute of Optical Sensor Systems, German Aerospace Center, Rutherfordstr. 2, 12489 Berlin
(Patrick.Irmisch, Dirk.Baumbach, Ines.Ernst)@dlr.de

KEY WORDS: Visual-Inertial Odometry, Dynamic Environment, Semantic Segmentation, Localization, Hand-Held, Vehicle

ABSTRACT:

Camera based navigation in dynamic environments with high content of moving objects is challenging. Keypoint-based localization methods need to reliably reject features that do not belong to the static background. Here, traditional statistical methods for outlier rejection quickly reach their limits. A common approach is the combination with an inertial measurement unit for visual-inertial odometry. Also, deep learning based semantic segmentation was recently successfully applied in camera based localization to identify features on common objects. In this work, we study the application of mask-based feature selection based on semantic segmentation for robust localization in high dynamic environments. We focus on visual-inertial odometry, but similarly investigate a state-of-the-art pure vision-based method as baseline. For a versatile evaluation, we use challenging self-recorded datasets based on different sensor systems. This includes a combined dataset of a real world system and its synthetic clone with a large number of humans for in-depth analysis. We further deploy large-scale datasets from pedestrian navigation in a mall with escalator scenes and vehicle navigation during the day and at night. Our results show that visual-inertial odometry performs generally well in dynamic environments itself, but also shows significant failures in challenging scenes, which are prevented by using the segmentation aid.

1. INTRODUCTION

The tasks for optical navigation become more and more challenging. Applied in different scenarios such as drone-, robotic-, driver assistance- or inspection systems, a reliable estimation of the current position and orientation of the system is mandatory. This can be realized by detecting and tracking features at salient keypoints in a sequence of camera images to estimate the ego-motion. Common implementation forms are Visual Odometry (VO) or Simultaneous Localization and Mapping (SLAM). The demands are versatile, ranging from low light conditions, fast camera movements to many moving objects. Traditional pure optical based navigation systems show difficulties in such scenarios, often fail completely in dynamic environments.

The compensation of these weaknesses has been a research topic for many years, while we concentrate on two main directions. First, optical systems are often fused with additional sensors. A common realization is the fusion with an Inertial Measurement Unit (IMU), known by the term Visual-Inertial Odometry (VIO). One representative is the Integrated Positioning System (IPS) (Börner et al., 2017), that is used for navigation, inspection, and 3D-modelling. Second, prior knowledge based on classification of point features to be located on static or moving objects is introduced. Due to the recent success of Deep Learning (DL), semantic segmentation has been frequently applied to mask specific areas of the image where features are unfavorable for pose estimation. This feature selection extension has shown to improve the performance of optical localization (Kaneko et al., 2018). Also, due to rapid progress in new hardware development for DL and similarly improving networks, DL based segmentation modules will most likely be available as additional hardware components in most future systems.

In this paper, we explore the limits of VIO in dynamic scenes and the potential of using semantic information for feature selection. This study is motivated by the frequent presence of moving objects in IPS related datasets and their unknown influence on the localization result. Due to its applications, long focal length cameras with a relatively small field of view are favored, which can lead to large occluded image areas by moving objects. Here, we use a straight-forward mask approach to ensure the exclusion of features also on slowly moving objects. The focus of this work is the evaluation based on challenging high dynamic datasets. Therefore, we deploy a hand-held system and its synthetic clone to create a dataset combining similar synthetic and real data with high dynamic human based content. Simulation is applied to evaluate the methods using ideal segmentation and substantial ground truth, and real data with DL

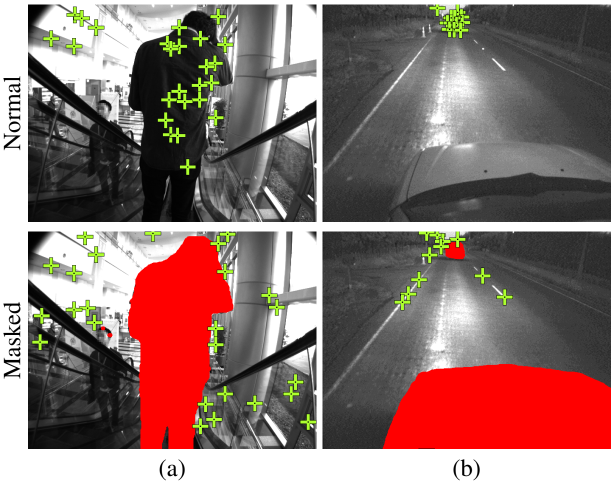


Figure 1. Illustration of selected features (green crosses) used for pose estimation in VIO without (Normal) and with (Masked) segmentation aid (red area) for a hand-held navigation on an escalator (a) and vehicle navigation at night (b, the brightness was increased for better visualization).

*Corresponding author

based segmentation is used to validate the observations. Further large-scale datasets from pedestrian- and vehicle navigation are utilized to generalize the observations. Exemplified in Figure 1, they include challenging scenes for localization on excalators and at night. Thus, our main contribution is a comprehensive evaluation of the application of a segmentation aid for feature selection in the existing IPS, a method for stereo camera based VIO. This contribution composes of the following aspects.

- We evaluate the influence of the segmentation aid on the performance of IPS using ground truth data.
- We extend and investigate ORBSLAM (Mur-Artal, Tardós, 2016) similarly, as baseline for pure visual localization.
- We consider different frame rates, which show to have a similar strong effect on IPS in dynamic scenes.
- We deploy different datasets from pedestrian- and vehicle navigation in dynamic indoor- and outdoor scenarios, recorded with different IPS sensor systems.

2. RELATED WORK

The susceptibility of methods for optical localization and odometry to moving objects is a well-known issue and different approaches are frequently proposed to tackle this problem. In the following, we summarize important approaches for feature-based optical localization and their improvements. A comprehensive overview of different approaches to increase the robustness of VO and SLAM is provided by (Saputra et al., 2018).

Feature-based methods for VO (Nister et al., 2004) and SLAM (Klein, Murray, 2007) typically rely on the same basic principles for robust localization. Geometric models in form of fundamental matrices and homographies are estimated based on matched image features, while random samples consensus (RANSAC) (Fischler, Bolles, 1981) is used to statistically exclude outliers and non-static feature points. Thereby, different geometric constraints are applied, such as epipolar or motion constraints. For instance, (Wu et al., 2017) used motion constraints in a stereo camera vehicle setup to predict feature positions in sequential images and use them as initialization for a feature tracker. The resulting camera pose is usually refined in an iterative cost minimization manner based on the reprojection error, while the choice of the cost function is important for robustness against outliers (MacTavish, Barfoot, 2015). Differently to VO that estimates an ego-motion only, SLAM provides precise long-term localization in known environments, but depends on well-defined 3D maps. (Mur-Artal, Tardós, 2016) proposed ORBSLAM that combines mapping of 3D points, keyframes, bundle adjustment, and loop closure, but has shown to be prone to dynamic environments (Kaneko et al., 2018).

A common approach to increase the robustness of navigation is sensor fusion, e.g. in combination with an IMU. A comprehensive overview for visual-inertial navigation is provided in (Chen et al., 2018a). The fusion can be realized using a Kalman filter (Grießbach et al., 2014) or in a combined minimization of the photogrammetric and IMUs measurement errors (Stumberg et al., 2018). Similar to motion constraints, the pose estimation of the IMU based on the strapdown mechanism can be used to predict feature positions to improve feature tracking. (Zhang et al., 2018a) found superior performance of visual-inertial navigation over pure visual methods in a dynamic office environment, by experimenting on datasets consisting of pedestrians and strong camera motion. We also investigate VIO in an office environment, but we concentrate on the application of semantic information and also use dynamic large scale datasets.

Due to the great success of DL based classification methods in recent years, different researchers introduced semantic information as prior knowledge into feature selection in the presence of moving objects. An overview of state of the art methods for semantic segmentation is provided by (Song et al., 2019). (Barnes et al., 2018) trained a DNN to learn the segmentation of static image areas in monocular VO. They used an additional 3D sensor setup and an offline mapping approach to automatically generate training data. (An et al., 2017) used semantic segmentation to assign higher weights to specific object classes during feature selection in VO. (Kaneko et al., 2018) developed Mask-SLAM and used semantic segmentation based on DeepLab v2 (Chen et al., 2018b) to create a mask for cars and the sky to exclude feature points in monocular ORBSLAM. (Bescos et al., 2018) combined multi-view geometry models and semantic segmentation to exclude features on moving objects for pose estimation and mapping. Also, (Yu et al., 2018) used semantic information to reject all keypoints belonging to an object, if a certain number of them where found to be moving by a consistency check, and during dense mapping. (Ganti, Waslander, 2019) investigated the application of uncertainties from a segmentation network in ORBSLAM. (Wang et al., 2019) simultaneously improved SLAM and semantic segmentation by distinguishing between features on moving, potentially moving and on the static background for SLAM and using the 3D pose information to refine the segmentation. (Schorghuber et al., 2019) distinguished between similar object states in a dynamic fashion, using a continuously updated confidence factor. In contrast, we decided to use the basic masking approach, since many slowly moving objects are only observed for a short time in our hand-held datasets. Representative for VIO, (Murali et al., 2017) used semantic information to classify visual landmarks as static in a tightly-coupled visual-inertial navigation system and evaluate this method on a self-recorded dataset for vehicle navigation. Similar to them, we use self-recorded datasets to evaluate our visual-inertial method, since current VIO benchmarks are less focused on high dynamic content. In contrast, our focus is on the evaluation of loosely-coupled VIO in highly dynamic environments, using vehicle datasets, but also hand-held datasets.

3. METHOD

The focus of this work is on IPS, a stereo-vision-aided inertial navigation system (Grießbach et al., 2014). It loosely-couples VO and inertial navigation. We extend IPS by introducing semantic segmentation based feature selection in the VO component. In the following, we review the method‘s main components, and our implemented segmentation aid extension in Subsection 3.3. All components are illustrated in Figure 2.

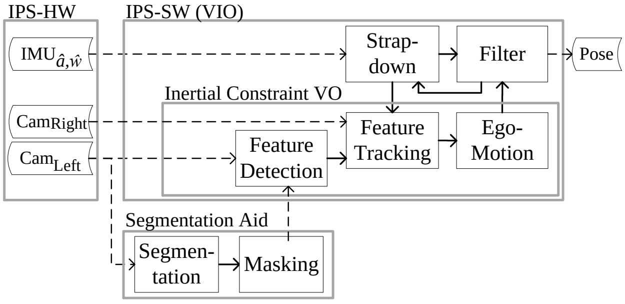


Figure 2. Illustration of the components of IPS and their relations, including sensor system *IPS-HW* and the navigation module *IPS-SW*, and the proposed segmentation aid.

3.1 Inertial constraint stereo visual odometry

To estimate VO, point features have to be detected, matched between both cameras (intra-matching) and tracked over consecutive frames (inter-matching). The matching is implemented as a patch-wise matching with a sliding window, which requires the detection of features in the left camera only. In this work, AGAST (Mair et al., 2010) is chosen as feature detector and Normalized Cross-Correlation as matching metric. To estimate the relative transformation $\Delta\mathbf{T}$ from time k to $k{+}1$, the intra-matched features at k are triangulated. The resulting 3D points $\hat{M}_k$ are projected into the stereo camera at $k{+}1$. $\Delta\mathbf{T}$ is then estimated through least square optimization of the reprojection error, which is the distance between the projection of $\hat{M}_k$ and the inter-matched feature points $\tilde{m}_{k+1}$. Equation (1) describes the objective for the left camera only, where $\mathbf{K}$ describes the camera matrix. The right camera is formulated accordingly, while respecting the calibrated stereo transformation.

$$\min_{\Delta\mathbf{T}} \|\mathbf{K}\Delta\mathbf{T}\hat{M}_k - \tilde{m}_{k+1}\|^2 \tag{1}$$

Single outliers in (1) can introduce significant errors or even cause the minimization to fail. Therefore, following approaches are applied in IPS to prohibit and reject outliers. First, epipolar constraints restrict the search space for the sliding window based intra-matching around the epipolar line with a distance threshold. Second, the inertial strapdown navigation solution is used to predict the image positions of $\hat{M}_k$ at time $k{+}1$. The search space around this position is restricted by the covariance, estimated through error propagation throughout the navigation pipeline. Third, RANSAC is applied to filter out mismatches.

3.2 Vision-aided inertial navigation

Next to the VO, IPS deploys an IMU consisting of each three mutually perpendicular accelerometers and gyroscopes. Using the strapdown mechanism (Wendel, 2011), an ego-motion based on the measured accelerations $\hat{a}$ and angular rates $\hat{\omega}$ is estimated. Unknown varying bias terms b_a, b_ω on the acceleration and angular rate measurements can lead to a strong drift if left uncompensated. Therefore, they are dynamically estimated and corrected during an initialization phase and navigation. The sensor fusion is implemented as an error state Kalman filter. Error propagation throughout the VO and the filtering process provides essential covariances for a robust solution. The estimated relative transformation of the VO is used as a measurement aid in the Kalman filter to update and correct the dynamic parameters of the strapdown equations. Next to the estimated pose and velocities, these parameters include the bias terms b_a, b_ω.

3.3 Masked based feature selection

Semantic segmentation is used to support the rejection of detected features on moving objects in the VO component. Based on pixel wise classification of defined object classes, a mask is generated that is used to accept or reject a point feature candidate during feature detection. We consider the classes human and car, which we assume to be constantly moving to be able to exclude all small movements. The implementation is inspired by (Kaneko et al., 2018). For pixel wise classification, we apply a DeepLab v2 ResNet implementation in Tensorflow with pretrained weights (Chen et al., 2018b, Nekrasov, 2016). It utilizes a deep convolutional neural network to obtain a pixel wise object assignment probability map for each considered class, while the final classification is given by the class with the highest score. Based on the segmentation of the target object classes, a mask is generated that defines the belonging to forbidden object classes. The mask is additionally edited based on a predefined default mask that covers static elements in the image, such as the own bonnet in case of vehicle navigation. To compensate inaccurate segmentation borders and difficult object-assignable or object-close image features, the mask is dilated by 4 pixels, oriented on the feature radius of AGAST. The application of the mask is implemented in the feature detection phase. After a point candidate is proposed by the specific corner detector, the image position is verified with the mask and accepted or rejected accordingly. Further selections of the features to use, e.g. with non-max suppression, follow and remain unchanged.

4. DATA ACQUISITION

IPS provides different hardware systems and a synthetic clone for data acquisition. The systems used in this work are illustrated in Figure 2. In this section, first the hardware systems are described and the simulation framework, which we extended to provide ideal semantic segmentation. Then, the datasets for the evaluation are introduced.

ID	R-V	R-HH	S-HH
Cameras	Prosilica GC1380H		synthetic
Sensor type	CCD-monochrome		monochrome
Resolution	1360×1024 $pixel$		
Pixel size	6.45 μm		
Focal length	8.2 mm	4.8 mm	
Baseline	0.45 m	0.20 m	

Table 1. IPS camera sensor parameters

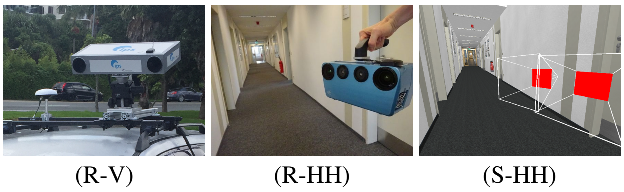

(R-V) (R-HH) (S-HH)

Figure 3. Used IPS sensor-heads: (R-V) Real world system on a vehicle, (R-HH) real world hand-held system, (S-HH) simulated hand-held system. (Ernst et al., 2018, Irmisch et al., 2019)

4.1 Real world

Two different IPS sensor systems are considered. Both are composed of the same IMU (ADIS-16488) and camera sensors, listed in Table 1, but differ in their focal length and baseline due to different target applications. Hardware system *R-HH* is a hand-held device that targets pedestrian navigation and 3D reconstruction in indoor, outdoor and underground environments. Hardware system *R-V* targets vehicle navigation and also provides a GNSS receiver (OEM625S). A FPGA is used to synchronously trigger the cameras and for accurate timestamp assignments of all sensor data.

4.2 Simulation

The simulation framework of (Irmisch et al., 2019) was deployed to generate synthetic datasets with substantial ground truth information. By transferring the estimated trajectory of

Dataset	**Distance**	**Velocity**		**Ang. Rate**	
	$[m]$	$[m/s]$		$[^\circ/s]$	
	3D	*mean*	*max*	*mean*	*max*
Corr-Sim	17.7	0.73	1.62	19.1	251
Corr-Real	18.9	0.63	1.64	19.5	148
Ipin-1	874	1.06	1.92	22.0	156
Ipin-2	902	0.9	1.63	18.3	129
Road-Day	6080	11.1	17.1	7.0	44.7
Road-Night	6260	10.1	16.1	6.7	37.2
Town-1	3060	4.6	11.3	4.8	48.1
Town-2	5190	3.3	12.4	3.4	41.7

Table 2. Camera motion characteristics for the different datasets. The distance describes the traveled 3D path. Each entry describes a separate run, only *Corr-Sim* and *Corr-Real* composes of 5 and 7 runs respectively, where the median distance is noted.

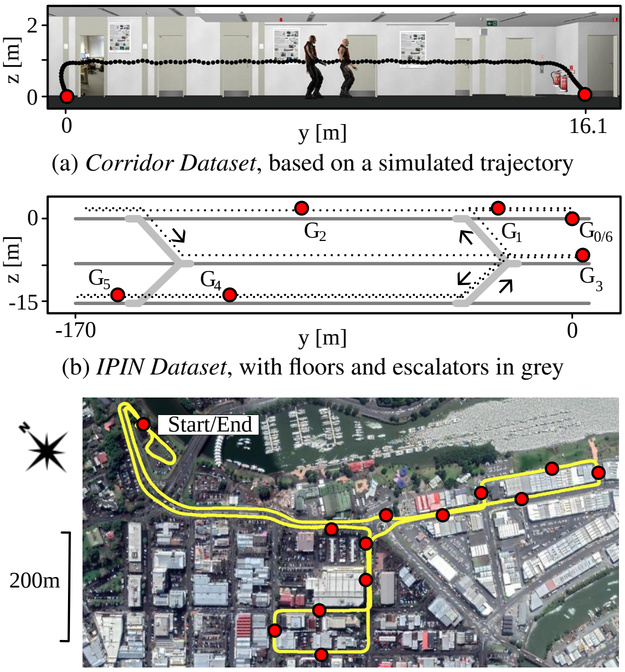
a real-world IPS into simulation, it allows to generate synthetic image- and IMU data with a realistic motion profile of the target platform. It can be used in any virtual environment with different camera configurations, including intrinsic-, extrinsic- and radiometric calibration, motion blur, exposure and frequency. In this work, the parameters of the real world device *R-HH* for camera simulation are used, including radiometric and geometric camera parameters. The synthetic clone is noted as *S-HH*. To ensure image quality, we used a super-sampling grid of 3 and accumulated 21 images to simulate motion blur with an exposure time of 5ms. The IMU is simulated with a frequency of 400Hz and noise parameters, oriented on the real-world IMU.

This simulation framework was extended to generate ideal semantic segmentation. Similar to (Gaidon et al., 2016), we render the scene a second time to generate per-pixel category- and instance-level ground-truth. During the second rendering, we disable all lighting-, shading and material effects and assign a unique label to each object, decoded in a RGB color value and set as ambient material property. An animated human model (Microsoft XNA, 2010) was used in the experiments, which is limited to straight walking with adjustable speed.

4.3 Datasets

The *Corridor Dataset* consists of 7 real and 5 synthetic recordings (*Corr-Real* and *Corr-Sim*) in a similar corridor environment with systems *R-HH* and *S-HH*. Table 2 provides information about camera dynamics and path length. The trajectories of both sources consist each of walking a short distance, illustrated in Figure 4 (a), but the individual recordings differ in the level of dynamic and presence of humans. For instance, the second session consists of two humans walking consistently in front of the camera. Or, the fourth sessions consists of two humans walk towards the camera, while another two are observed starting to walk slowly. The camera images were acquired with a frame rate of 30Hz and sorted out for 10Hz and 5Hz. The simulation provides complete ground truth, while for the real world dataset two Ground Control Points (GCPs) at the beginning and end of each session are used as reference.

The *IPIN Dataset* provides recordings in a mall like environment for the system *R-HH* with 10Hz frame rate. It was recorded in 2014 for the indoor navigation competition at the international conference on Indoor Positioning and Indoor Navigation (IPIN). This dataset is challenging for optical navigation due to the presence of densely crowded areas, strong light reflections, numerous escalators scenes between three floors, as illustrated in Figure 4 (b), and dynamic camera motion, as shown in Table 2. GCPs were provided by the IPIN team, from which we choose 6 as reference, where the system was hold for 5s.



(a) *Corridor Dataset*, based on a simulated trajectory

(b) *IPIN Dataset*, with floors and escalators in grey

(c) *NZ-Road Dataset*, exemplary for run *Town-1*

Figure 4. Illustration of exemplary trajectories (black dotted, yellow solid lines) and real world GCPs (red points).

The *NZ-Road Dataset* provides recordings for vehicle navigation on a high road and in an urban area. It was recorded in the context of *Digital Roads New Zealand* (Ernst et al., 2018, Zhang et al., 2018b) with focus on 3D-reconstruction using the system *R-V* with 10Hz frame rate. We reconsider four datasets for our systematic analysis of the segmentation aid. Two runs on a high road are selected, recorded on the same road at day and night. Both consist of a ride back and forth and another car driving consistently in front of the vehicle in safety distance (Figure 1 (b)). Two urban sets are used that contain pedestrians and numerous vehicles, both driving and parking (Figure 7 (e)). The recorded GPS data was utilized to select a number of GCPs per hand as reference (Figure 4 (c)).

5. EVALUATION

Based on the proposed datasets with dynamic content, we investigate the effect of the segmentation aid for VIO, represented by IPS. We apply the same segmentation aid in the stereo camera approach of ORBSLAM as pure vision-based baseline. For IPS we apply two different configurations. *IPS Fast* provides real-time localization at 10Hz by processing the images in half resolution (680×512 pixel), while the feature matching properties are optimized for maximum speed. *IPS Accurate* runs in near real-time and is usually used for accurate offline processing. *IPS Accurate* and *ORBSLAM* are applied using full image resolution (1360×1024 pixel). Depending on the method, the semantic segmentation is processed in different resolutions. For *IPS Accurate Masked* and *ORBSLAM Masked* full resolution was used. The approaches with segmentation at full resolution with the chosen neural network as well as localization at 30Hz are currently not real-time capable with our hardware

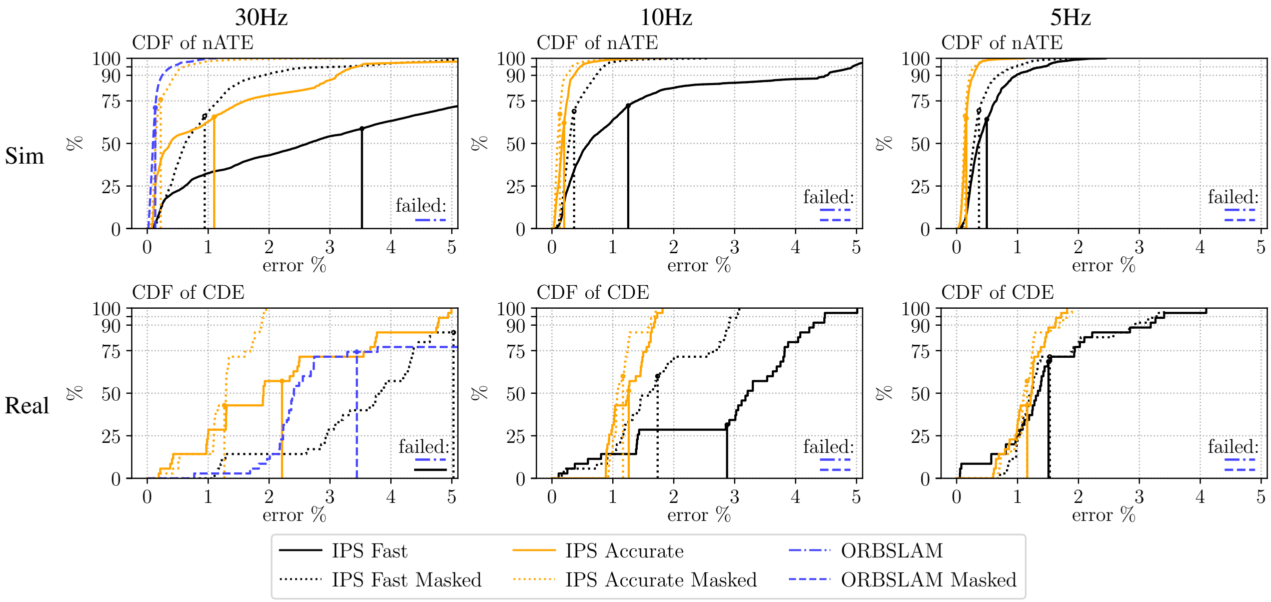


Figure 5. Results for the *Corridor Dataset* in CDF-representation, for different frame rates and for synthetic and real data respectively. The vertical lines mark the *mean* nATE or CDE respectively for each CDF. Navigation results that are considered as failed, are hidden.

setup, but still included in our experiments to exploit the full potential of the segmentation aid. In *IPS Fast Masked*, the segmentation was applied on a downscaled image of (320×256 pixel) to investigate our method with less-effort segmentation. In the following, we first introduce the applied metrics for evaluation. Then, we evaluate the methods on the *Corridor Dataset* in detail, where we also consider different camera frame rates. Finally, we apply the methods on the real-world datasets from pedestrian- and vehicle navigation.

5.1 Metrics

The availability of ground truth data varies for the different datasets. A complete ground truth is provided for the synthetic datasets, whereas only a few ground control points (GCPs) are available for the real world datasets.

In simulation, we use a *(normalized) Absolute Trajectory Error* (nATE) in (3) at timestamp i for the online localization result of IPS and ORBSLAM. It is based on the ATE from (Sturm et al., 2012) in (2), while we focus on the translational components using their norm $||.||_T$. We align the estimated trajectory $\mathbf{P}_0, ..., \mathbf{P}_n \in$ SE(3) with the ground truth trajectory $\mathbf{Q}_0, ..., \mathbf{Q}_n \in$ SE(3) by the transformation $\mathbf{S}$ that aligns the start pose $\mathbf{P}_0$ with the ground truth start pose $\mathbf{Q}_0$. This corresponds to the "alignment using a single state" (Zhang, Scaramuzza, 2018). Since IPS estimates an initial roll and pitch based on the IMU, only the heading is aligned, whereas for ORBSLAM all angles need to be aligned. To compensate increasing drift of the odometry method, we additionally normalize the error by the traveled distance d_i and use [%] as error unit.

$$\mathrm{ATE}_i := ||\mathbf{Q}_i^{-1}\mathbf{S}\mathbf{P}_i||_T, \tag{2}$$

$$\mathrm{nATE}_i := \frac{\mathrm{ATE}_i}{d_i}, \quad d_i = \sum_{j=1}^{i} ||\mathbf{Q}_{j-1}^{-1}\mathbf{Q}_j||_T \tag{3}$$

In real world, the ground truth is limited to GCPs $G_0, ..., G_m \in \mathbb{R}^3$, where we apply two metrics. First, we consider relative distances between GCPs and compare them to the estimated relative distances of the applied methods. Therefore, we define a *Cross Distance Error* (CDE) in (4), which considers the estimated distances between all GCPs, with $0 \le k < l \le m$. $\mathbf{P}_{[k]}$ describes the estimated pose when placed on G_k. The CDE is normalized by the traveled distance $d_{k,l}$, defined by the sum of distances between visited GCPs between G_k and G_l.

$$\mathrm{CDE}_{k,l} := \frac{\left| ||\mathbf{P}_{[k]}^{-1}\mathbf{P}_{[l]}||_T - ||G_k - G_l|| \right|}{d_{k,l}}, \tag{4}$$

$$d_{k,l} = \sum_{j=k+1}^{l} ||G_{j-1} - G_j|| \tag{5}$$

Second, we use the ATE by aligning the positions of the GCPS G_k with the corresponding trajectory positions of $\mathbf{P}_{[k]}$. For this, we use the implementation of the method in (Umeyama, 1991) provided by (Zhang, Scaramuzza, 2018).

For each metric we compute the *mean* based on the results of a dataset-specific number of runs and each 5 repetitions to account for the non-deterministic nature of the RANSAC component in VO. We further classify a method as succeeded or failed based on the following conditions. The method succeeded, if it was able to estimate a solution for at least 75% of the data, the mean of nATE or CDE is less than 5% (rounded) and the mean of ATE is less than 50m, if applied.

5.2 Evaluation of the Corridor Dataset

The *Corridor Dataset* was deployed to determine the limits of VIO and to exploit the potential of the segmentation aid. Concerning the former, the dataset contains many humans that are mostly walking or standing with small movements. For the latter, the ideal segmentation was used in the synthetic dataset. For comparability, we concentrate on the nATE and CDE, which share the same error unit [%]. The data is visualized using the Cumulative Distribution Function (CDF) in Figure 5, distinguished between simulation and real world with camera frame rates 30Hz, 10Hz, 5Hz. Each line shows the result of one method as a whole for 5 runs in simulation and 7 in real world (see section 4.3). Vertical lines visualize the CDF *mean*.

Run	Metric	IPS Fast		IPS Acc.		ORBSL.	
	mean	N	M	N	M	N	M
Ipin-1	ATE [m]	2.94	3.12	-	-	-	-
	CDE [%]	1.01	**0.91**	-	-	-	-
Ipin-2	ATE [m]	5.54	**2.52**	-	-	-	-
	CDE [%]	2.22	**0.79**	-	-	-	-
Road-Day	ATE [m]	20.45	20.91	20.17	20.94	f	f
	CDE [%]	0.60	0.61	0.44	0.45	f	f
Road-Night	ATE [m]	20.46	20.35	f	**18.6**	f	f
	CDE [%]	0.75	0.74	f	**0.61**	f	f
Town-1	ATE [m]	6.02	6.41	5.58	5.56	39.64	**9.4**
	CDE [%]	**0.49**	0.54	0.37	0.37	3.67	**0.70**
Town-2	ATE [m]	**9.18**	10.18	10.87	10.67	f	**20.64**
	CDE [%]	0.58	0.61	0.55	0.55	f	**0.78**

Table 3. Results for *IPIN*- and *NZ-Road Datasets* for the individual method without (N) and with (M) using a mask. Bolt numbers mark noticeable relative differences between (N) and (M) of at least 10 %. Particularities are marked with red background. If a result is not given, the method was either considered as failed (f) or was not applied (-).

First evaluating the pure vision-based method *ORBSLAM*, it only succeeded for high frame rates (30Hz) when using the mask, both in simulation and real world. Due to high dynamics of the hand-held system, noted in Table 2, ORBSLAM frequently failed for low frame rates. However, in simulation at 30Hz, *ORBSLAM Masked* performed the most accurate localization of all methods for this dataset.

For VIO, the results show that the segmentation aid increased navigation results in general for this specific dataset. This boost is distinctive at 30Hz and negligible at 5Hz. Due to higher frame rates, object motions are less pronounced in the image and features on slowly moving objects are more likely to be used in the VO estimation. Related, Figure 7 (b) shows single used points on slowly moving humans. Similar, when using a lower resolution in *IPS Fast*, this boost is still present at 10Hz, where it is comparatively small for *IPS Accurate*. It also appears that more features are used from moving- than from static objects, as in Figure 7 (a) for *IPS Fast*.

Comparing the results from simulation and real world, they show strong correlations despite different evaluation metrics. It is particularly striking that *IPS Accurate Masked* performs similar at all frame rates, in simulation and real world respectively.

5.3 Evaluation of the IPIN Dataset

For the *IPIN dataset*, we only consider *IPS Fast* since the images are only available in half resolution at 10Hz. The results for the two runs of the same trail are listed in Table 3. In *Ipin-1*, the segmentation aid only slightly improves the base method regarding the CDE. Figure 7 (c) depicts an example of a crowd area from this dataset. In contrast, *IPS Fast* shows large errors in *Ipin-2*, which do not occur when using the segmentation aid. The cause of this error is shown in Figure 1 (a), where a person stands in front of the camera during an escalator scene.

The behavior of the VIO during the escalator scene is analyzed in Figure 6, exemplary for the determined height in local coordinates. First, the body-frame up-axis z_b of the IMU is shown to delimit the escalator scene. Before and after the escalator, the IMU measures the walking motion profile and only measures the gravity during the ride. Second, the estimated normalized relative translation of the VO is shown. While the relative translation should be constantly high, *IPS Fast* VO frequently estimates zero-movement due to a high number of detected features on the person, as shown in Figure 1 (a). Using these VO estimations as aid in the navigation filter leads to wrongly estimated bias terms, exemplified for b_a on z_b in comparison with *IPS Fast Masked*. As a result, the navigation solution fails to estimate the change in height for this scene, compared with a barometer measurement as reference. Contrary, *IPS Fast Masked* is able to estimate the height almost similar to the reference.

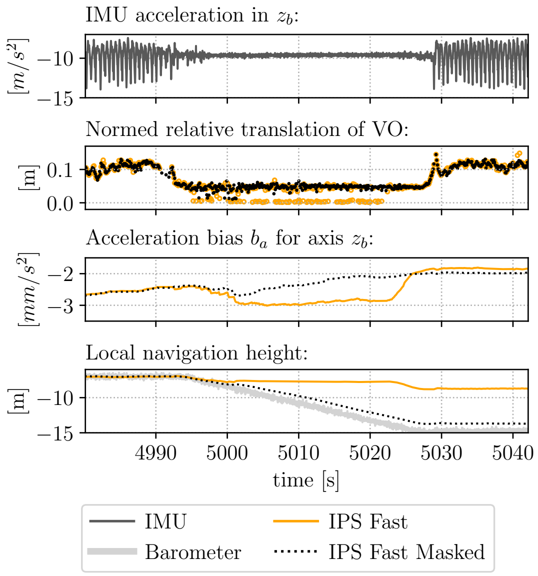


Figure 6. System- and localization parameters in local coordinates of *IPS Fast* (*Masked*) on an escalator of run *Ipin-2*.

5.4 Evaluation of the NZ-Road Dataset

The results show that ORBSLAM is struggling with high velocities at a low frame rate of 10Hz. Thus, both road runs are considered as failed. Figure 7 (d) shows that the segmentation aid helps in the beginning of *Road-Night* to differentiate between static and dynamic features, but the navigation still fails in the later course. In the urban area (*Town*), *ORBSLAM Masked* succeeded, while *ORBSLAM* failed in *Town-2*.

Regarding VIO, *IPS Accurate* struggled at the night dataset. Due to low light conditions and a car driving constantly in front of the cameras, comparatively many features are detected on this car at high image resolution and are frequently used for VO, exemplary shown in Figure 1 (b). This results in an unreliable navigation result, similar to the analyzed scene of the *IPIN Dataset*. For the other runs, the difference between the IPS methods with and without the mask is largely negligible. In the town datasets, the accuracy slightly decreased for *IPS Fast Masked*, which uses low image resolution and the segmentation aid that equally masks out parking cars (Figure 7 (e)).

6. DISCUSSION

Our results show that VIO naturally provides a good performance in dynamic scenes, even without a segmentation aid. Due to inertial constraints in the VO component, the method can be

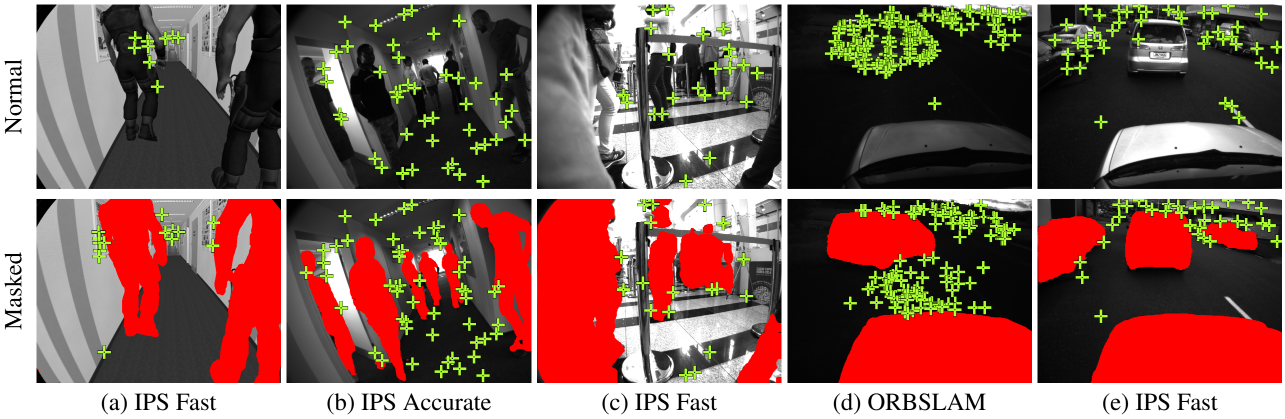


Figure 7. Illustration of selected features (green crosses) of the methods without (Normal) and with (Masked) applied segmentation aid (red area) for scenes from (a,b) the *Floor*-, (c) *IPIN*- and (d,e) *NZ-Road Dataset*. (a,b,c,e) show features used for pose estimation in IPS for the current frame. (d) shows the tracked keys that found *ORBSLAM* valid for the current frame.

effectively applied to low frame rates, such as 10Hz, where object motions are more pronounced and associated features are easier to sort out. However, inaccuracies and even complete failures of the navigation still arise. In direct comparison to ORBSLAM as pure visual baseline, we found, based on the presented datasets, that the used VIO method is mostly superior in terms of robustness and accuracy, despite ORBSLAMs keyframe and loop closure modules. However, this comparison is limited to the applied 10Hz in the vehicle datasets, the dynamic-only environments and limited amount of closed loops.

The combination with segmentation based feature selection has shown to prevent complete failures of VIO in selected scenarios, while the accuracy is generally consistent. Only in scenes with a high rate of parking cars and a low image resolution, the accuracy slightly decreased due to the chosen mask approach that we used in this study to ensure the exclusion of features on all slowly moving objects. The improvements were even more drastic for ORBSLAM with a similar segmentation aid. Thus, we can confirm the conclusion of (Bescos et al., 2018, Kaneko et al., 2018, Murali et al., 2017, Yu et al., 2018) that a segmentation aid can greatly improve optical localization and substantiate its applicability in the VIO-based localization system IPS.

For the evaluation based on different applications, we deployed a hand-held system, both in simulation and real world, and a sensor system for vehicle navigation. While the simulation provided complete and perfect ground truth, the real world experiments were limited by the number of reference points and sensor quality. Extensive public benchmarks (Blanco-Claraco et al., 2014, Geiger et al., 2013, Schubert et al., 2018, Sturm et al., 2012) exist, but met our requirements for this evaluation less, due to the necessary connection of a stereo camera with a synchronized IMU, the need of a short initialization phase and our focus on high dynamic content. Even though the effect of the segmentation aid was clearly visible, future experiments could benefit from extensive real world ground truth generation.

7. CONCLUSION

In this work, we evaluated the performance of Visual-Inertial Odometry (VIO), represented by the Integrated Positioning System (IPS), with an additional segmentation based feature selection in dynamic environments. We deployed ORBSLAM as pure-visual navigation baseline for comparison. For evaluation, we deployed different challenging large-scale datasets recorded with sensor systems for pedestrian- and vehicle navigation. For the former, we additionally created a combined real world and synthetic dataset with high dynamic content, to evaluate the methods at different frame rates. Using the segmentation aid for ORBSLAM, we could confirm an outstanding performance gain. While VIO at low frame rates has shown a relatively good performance in dynamic environments itself, we conclude that the segmentation aid mainly contributes in terms of robustness as it is able to prevent rare but significant failures. In selected scenarios however, the chosen basic mask approach lead to a slight decrease in localization accuracy, e.g. with many parking cars. In future, we will target this issue and explore this method for robust localization in environments with other dynamic elements, such as vegetation, steam, or water. We also plan to use such semantic information for 3D reconstruction of the static scene in these dynamic environments.

REFERENCES

An, L., Zhang, X., Gao, H., Liu, Y., 2017. Semantic segmentation–aided visual odometry for urban autonomous driving. *Int. Journal of Advanced Robotic Systems*, 14(5), 1–11.

Barnes, D., Maddern, W., Pascoe, G., Posner, I., 2018. Driven to distraction: Self-supervised distractor learning for robust monocular visual odometry in urban environments. *IEEE Int. Conf. on Robotics and Automation (ICRA)*.

Bescos, B., Fácil, J., Civera, J., Neira, J., 2018. DynaSLAM: Tracking, Mapping and Inpainting in Dynamic Scenes. *IEEE Robotics and Automation Letters*, 3, 4076–4083.

Blanco-Claraco, J.-L., Moreno-Dueñas, F.-Á., González-Jiménez, J., 2014. The Málaga urban dataset: High-rate stereo and LiDAR in a realistic urban scenario. *The Int. Journal of Robotics Research*, 33(2), 207–214.

Börner, A., Baumbach, D., Buder, M., Choinowski, A., Ernst, I., Funk, E., Grießbach, D., Schischmanow, A., Wohlfeil, J., Zuev, S., 2017. IPS – a vision aided navigation system. *Advanced Optical Technologies*, 6(2), 121–130.

Chen, C., Zhu, H., Li, M., You, S., 2018a. A Review of Visual-Inertial Simultaneous Localization and Mapping from Filtering-Based and Optimization-Based Perspectives. *Robotics*, 7(3), 45.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A. L., 2018b. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4), 834–848.

Ernst, I., Zhang, H., Zuev, S., Knoche, M., Dhiman, A., Chien, H.-J., Klette, R., 2018. Large-scale 3d roadside modelling with road geometry analysis: Digital roads new zealand. *15th International Symposium on Pervasive Systems, Algorithms and Networks (I-SPAN)*, 15–22.

Fischler, M. A., Bolles, R. C., 1981. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6), 381–395.

Gaidon, A., Wang, Q., Cabon, Y., Vig, E., 2016. Virtual worlds as proxy for multi-object tracking analysis. *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 4340–4349.

Ganti, P., Waslander, S. L., 2019. Network uncertainty informed semantic feature selection for visual slam. *16th Conf. on Computer and Robot Vision (CRV)*, IEEE, 121–128.

Geiger, A., Lenz, P., Stiller, C., Urtasun, R., 2013. Vision meets robotics: The KITTI dataset. *The Int. Journal of Robotics Research*, 32(11), 1231–1237.

Grießbach, D., Baumbach, D., Zuev, S., 2014. Stereo-vision-aided inertial navigation for unknown indoor and outdoor environments. *Int. Conf. on Indoor Positioning and Indoor Navigation (IPIN)*, IEEE, 709–716.

Irmisch, P., Baumbach, D., Ernst, I., Börner, A., 2019. Simulation framework for a visual-inertial navigation system. *IEEE Int. Conf. on Image Processing (ICIP)*, 1995–1999.

Kaneko, M., Iwami, K., Ogawa, T., Yamasaki, T., Aizawa, K., 2018. Mask-slam: Robust feature-based monocular slam by masking using semantic segmentation. *IEEE/CVF Conf. on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 3710–3718.

Klein, G., Murray, D., 2007. Parallel tracking and mapping for small ar workspaces. *6th IEEE and ACM Int. Symposium on Mixed and Augmented Reality (ISMAR)*, 25–234.

MacTavish, K., Barfoot, T. D., 2015. At all costs: A comparison of robust cost functions for camera correspondence outliers. *12th Conf. on Computer and Robot Vision (CRV)*, IEEE, 62–69.

Mair, E., Hager, G. D., Burschka, D., Suppa, M., Hirzinger, G., 2010. Adaptive and generic corner detection based on the accelerated segment test. *Computer Vision – ECCV 2010*, Lecture Notes in Computer Science, 6312, Springer, 183–196.

Microsoft XNA, 2010. Human skinned model.

Mur-Artal, R., Tardós, J. D., 2016. ORB-SLAM2: an Open-Source SLAM System for Monocular, Stereo and RGB-D Cameras. *IEEE Transactions on Robotics*, 33(5), 1255–1262.

Murali, V., Chiu, H.-P., Samarasekera, S., Kumar, R. T., 2017. Utilizing semantic visual landmarks for precise vehicle navigation. *IEEE 20th Int. Conf. on Intelligent Transportation Systems (ITSC)*, 1–8.

Nekrasov, V., 2016. Deeplab-resnet-tensorflow.

Nister, D., Naroditsky, O., Bergen, J., 2004. Visual odometry. *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 652–659.

Saputra, M. R. U., Markham, A., Trigoni, N., 2018. Visual SLAM and Structure from Motion in Dynamic Environments. *ACM Computing Surveys*, 51(2), 1–36.

Schorghuber, M., Steininger, D., Cabon, Y., Humenberger, M., Gelautz, M., 2019. Slamantic - leveraging semantics to improve vslam in dynamic environments. *IEEE/CVF Int. Conf. on Computer Vision Workshop (ICCVW)*, 3759–3768.

Schubert, D., Goll, T., Demmel, N., Usenko, V., Stückler, J., Cremers, D., 2018. The tum vi benchmark for evaluating visual-inertial odometry. *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, 1680–1687.

Song, W., Zheng, N., Zheng, R., Zhao, X.-B., Wang, A., 2019. Digital image semantic segmentation algorithms: A survey. *Journal of Information Hiding and Multimedia Signal Processing*, 10, 196–211.

Stumberg, L. v., Usenko, V., Cremers, D., 2018. Direct sparse visual-inertial odometry using dynamic marginalization. *IEEE Int. Conf. on Robotics and Automation (ICRA)*, 2510–2517.

Sturm, J., Engelhard, N., Endres, F., Burgard, W., Cremers, D., 2012. A benchmark for the evaluation of rgb-d slam systems. *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, 573–580.

Umeyama, S., 1991. Least-squares estimation of transformation parameters between two point patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(4), 376–380.

Wang, K., Lin, Y., Wang, L., Han, L., Hua, M., Wang, X., Lian, S., Huang, B., 2019. A unified framework for mutual improvement of slam and semantic segmentation. *International Conference on Robotics and Automation (ICRA)*, IEEE, 5224–5230.

Wendel, J., 2011. *Integrierte Navigationssysteme: Sensordatenfusion, GPS und inertiale Navigation*. 2 edn, Oldenbourg Wissenschaftsverlag GmbH, Germany, Munic.

Wu, M., Lam, S.-K., Srikanthan, T., 2017. A Framework for Fast and Robust Visual Odometry. *IEEE Transactions on Intelligent Transportation Systems*, 18(12), 3433–3448.

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. *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, 1168–1174.

Zhang, C., Liu, Y., Wang, F., Xia, Y., Zhang, W., 2018a. VINS-MKF: A Tightly-Coupled Multi-Keyframe Visual-Inertial Odometry for Accurate and Robust State Estimation. *Sensors*, 18(11), 4036.

Zhang, H., Ernst, I., Zuev, S., Börner, A., Knoche, M., Klette, R., 2018b. Visual odometry and 3d point clouds under low-light conditions. *Int. Conf. on Image and Vision Computing New Zealand (IVCNZ)*, IEEE, 1–6.

Zhang, Z., Scaramuzza, D., 2018. A tutorial on quantitative trajectory evaluation for visual(-inertial) odometry. *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, 7244–7251.