Experimental Comparison of Visual-Aided Odometry Methods for Rail Vehicles

Florian Tschopp[1], Thomas Schneider[1], Andrew W. Palmer[2], Navid Nourani-Vatani[2], Cesar Cadena[1], Roland Siegwart[1], and Juan Nieto[1]

***Abstract*— Today, rail vehicle localization is based on infrastructure-side Balises (beacons) together with on-board odometry to determine whether a rail segment is occupied. Such a coarse locking leads to a sub-optimal usage of the rail networks. New railway standards propose the use of moving blocks centered around the rail vehicles to increase the capacity of the network. However, this approach requires accurate and robust position and velocity estimation of all vehicles. In this work, we investigate the applicability, challenges and limitations of current visual and visual-inertial motion estimation frameworks for rail applications. An evaluation against RTK-GPS ground truth is performed on multiple datasets recorded in industrial, sub-urban, and forest environments. Our results show that stereo visual-inertial odometry has a great potential to provide a precise motion estimation because of its complementing sensor modalities and shows superior performance in challenging situations compared to other frameworks.**

I. Introduction

In recent years, the need for public transportation has risen dramatically. Rail transportation alone has increased by over $60\,\%$ in the last 16 years in Switzerland [1]. However, current infrastructure is reaching its capacity limits. To keep up with this growth, there is a need to improve the system efficiency.

In train applications, a crucial part of the current infrastructure is the traffic control system. Most of current rail control systems divide the railroad tracks into so-called blocks [2]. The block size is determined by the worst case braking distance of every vehicle that is likely to operate on this track. Vehicle localization and interlocking of the blocks is performed using infrastructure-side beacons. Such a fixed block strategy results in very conservative interlocking and thus, decreases the overall efficiency of the system.

The new European Train Control System (ETCS) Level 3 aims to replace the fixed blocks with moving blocks centered around the vehicle. This concept has the potential of increasing the capacity of train networks by a factor of $190\,\%$ to $500\,\%$ [3]. Furthermore, fixed track-side sensing infrastructure (e.g. axle-counters, Balises) may be replaced with on-board sensors, leading to a more cost-effective solution in the long-run. Even with the vast amount of research in related applications (e.g. autonomous cars), the success of ETCS Level 3 is subject to the development of new algorithms that are able to precisely and reliably estimate both the position and velocity of all rail vehicles [4], [5].

Fig. 1. Datasets recorded with a custom sensor setup for visual-aided odometry in sub-urban and industrial environments are used for evaluation of popular visual-aided odometry frameworks for rail applications. We show that high accuracy motion estimation can be achieved using stereo vision. Furthermore, incorporating inertial measurements increases accuracy and robustness.

In rail applications only few restrictions exist in regard to weight and power consumption of the localization solution. Therefore, one is pretty open in choosing suitable sensor modalities and estimation algorithms. Current research in train localization mainly focuses on the fusion of global navigation satellite system (GNSS) with inertial measurements coupled with infrastructure-side beacons. In safety critical application such as train localization, a high level of reliability can only be achieved using redundant and complementary sensors.

Recently, the robotics and computer vision communities have reported visual motion estimation and localization systems with an impressive accuracy and robustness [6]–[10]. We believe that synchronized visual and inertial sensors have the right properties to be an ideal extension to the currently used sensor modalities. A continuous global localization is often not feasible using vision sensors due to ambiguous environments or drastic appearance changes. However, combining incremental odometry information with localization to reduce drift accumulated by the odomertry method can provide a continuous and high-accuracy pose estimation. For this reason, as a first step towards such a system, we want to investigate current state-of-the-art visual(-inertial) motion estimation frameworks for their applicability on train applications. The main challenges include high speeds, constrained motion leading to potential observability issues of IMU biases [11], challenging lighting conditions and highly

[1]Authors are members of the Autonomous Systems Lab, ETH Zurich, Switzerland; {firstname.lastname}@mavt.ethz.ch

[2]Authors are with Siemens Mobility, Berlin, Germany; {firstname.lastname}@siemens.com

repetitive patterns on the ground.

Our contribution consists of the first evaluation of popular generic visual-aided odometry frameworks for rail applications. We use a real time kinematics (RTK) Global Position System (GPS) device as ground truth, in order to evaluate the pose estimates on datasets recorded on two trajectories in an industrial, sub-urban and forest environment. Furthermore, we identify, investigate, and discuss specific challenges of current methods.

II. Related Work

Motion estimation is the backbone of many established small autonomous ground robots and has had an increasing presence due to the rise of autonomous driving. Current solutions in commercial products rely on the use of sensor fusion of GNSS (in outdoor scenarios), wheel odometry, and additional sensors such as Light Detection And Ranging devices (LiDARs) or cameras, e.g. [12]–[14].

In contrast to generic ground robots, trains have a distinct feature: their motion is constrained to the rail network. This paper studies advantages and disadvantages this constraint implies to the motion estimation performance compared to more generic approaches.

A. Rail vehicle odometry and localization

The current research goal is to increase the accuracy and robustness of motion estimation and localization. The approaches are split into improving the in-place infrastructure (e.g. track-side Balises, odometer and speed sensors) or investigating new sensor modalities.

Mourillas and Poncet [15] and Palmer and Nourani-Vatani [16] describe measures on how to increase the robustness and reliability of the currently used on-board odometry measurements using wheel encoders and ground speed radars. Recent works also include using machine learning approaches such as least squares support vector machines (LSSVMs) to improve the localization accuracy [17].

To decrease the dependency on track-side infrastructure, new sensor modality research is highly focused on the use of an inertial measurement unit (IMU) together with a tachometer [18] or GNSS. Otegui et al. [19] summarizes many works fusing IMU and GPS signals employing an extended Kalman filter (EKF) or a particle filter (PF).

To further improve accuracy, Hasberg et al. [20] and Heirich et al. [21] introduce Simultaneous Localization and Mapping (SLAM) for path-constrained motion. Fusing GNSS with IMU measurements and a path constraint, high accuracies in position retrieval ($< 1\,\mathrm{m}$) and map building with a precision of around $10\,\mathrm{cm}$ are presented [20].

All these existing methods rely heavily on GPS, IMU and wheel encoders / ground speed radars, each of which have their own failure cases. For instance GPS has denied areas, suffers from multi-path effects near structures and is easy to jam, IMU bias may become unobservable [11], [22], encoders suffer from wheel slippage or mis-calibration [16] and radars have problems with reflectance off the ground [15], [23]. To improve the robustness and achieve high safety levels through redundancies, additional sensing modalities such as cameras will be critical.

Wohlfeil [24] performs visual track recognition based on edge extraction with a known track gauge width. Failure cases were observed where switches were missed or confused, especially in challenging lighting conditions (e.g. bright sky). Furthermore, a continuous position estimate is not provided, only the direction of travel after a switch. Bah et al. [25] present a pure vision-based odometry method for trains by transforming a front facing camera image to a birds-eye view of the track and finding correspondences on two consecutive frames. This method might fail with low-texture or repetitive grounds which often occur in train environments.

The mentioned visual-aided odometry and localization algorithms are not directly suitable for continuous motion estimation as they only provide information near switches [24] or require manual data association [25]. Furthermore, by not considering specific constraints, the visual odometry can later be fused with this information to get even more reliable and accurate pose estimation, to detect when a method is failing or to detect changes in the expected environment [26]. The goal of this paper is to study the performance of generic visual-aided ego-motion estimation for the rail application.

B. Generic visual-aided ego-motion estimation

State-of-the-art approaches in odometry estimation and SLAM using vision sensors can be classified into filter approaches (mostly EKF), where typically all the past robot poses are marginalized and only the most recent pose is kept in the state [6], [27], and sliding-window approaches, where multiple poses are kept in an optimization [7], [8].

Sliding-window based approaches are studied in detail by Strasdat et al. [28], proving to outperform filter-based approaches when employing the same computational resources. Furthermore, sliding-window based approaches are very flexible for incorporating measurements from different sensing modalities with different propagation models. To keep the computational costs within hardware limits, these schemes typically limit the state to within a sliding window. Efforts to unify both approaches are presented by Bloesch et al. [29].

The most prominent visual odometry methods are probably *ORB-SLAM* [7] and *Direct Sparse Odometry (DSO)* [8]. *ORB-SLAM* extracts and tracks keypoint features in an image stream while *DSO* uses a direct approach based on image gradients. These methods cannot recover the metric scale of the map, which is critical in the given application.

One prominent method to recover the metric scale is adding an IMU. Extending the previous works in [30] and [31] respectively, the scale can be observed by incorporating inertial measurements, often referred to as visual-inertial odometry (VIO). However, depending on the performed motion, the IMU biases are not fully observable [11], [22], [30] resulting in errors in scale estimation.

Another method to recover the scale is stereo vision shown in [32] and [33]. Leutenegger et al. [34] proposes

the combination of stereo vision and IMU measurements, resulting in a reliable feature-based sparse map and accurate camera pose retrieval.

In automotive applications, many of the challenges such as high velocities and constraint motion are similar to the rail domain. There, odometry is often solved by using wheel encoders [35], [36] as they do not suffer from high slip as in rail applications. Furthermore, stereo vision [37] and monocular visual odometry (VO) with learned depth [38] have also been used successfully for ego motion estimation. A multitude of state-of-the-art stereo-visual odometry frameworks are tested for automotive applications in the visual odometry part of the KITTI Vision Benchmark Suite [39]. One popular and well performing open-source pipeline is *ORB-SLAM2* [32]. Unfortunately, the KITTI dataset does not include synchronized IMU measurements and therefore does not allow in-depth insights into the benefits inertial data could provide. Finally, the scale of the motion can also be retrieved by exploiting non-holonomic constraints [40] which, however, relies on frequent turns.

In contrast to the mentioned approaches for automotive applications, this paper aims to investigate the benefit inertial data can provide for motion estimation in the rail domain and compares it to already successfully deployed stereo visual odometry.

III. Visual-aided odometry pipelines

In order to evaluate the performance of visual-aided ego-motion estimation for rail applications, we made a selection of the most promising available pipelines summarized in Table I.

A. Visual-inertial odometry algorithms

The goal of VIO is to increase robustness and observe the scale of the motion using inertial measurements. Advantages of VIO are gravity aligned maps and complementing sensor modalities. One disadvantage is the dependency on specific motion patterns in order to make the biases observable.

In this paper, the following state-of-the-art algorithms are introduced and further evaluated.

1) ROVIO: In [6], a light-weight visual-inertial odometry algorithm based on an EKF is presented. It shows high robustness even in very fast rotational motions. *ROVIO* directly uses pixel intensity errors on patches and can therefore be considered a direct method.

2) VINS-Mono: Qin et al. [9] proposes a VIO algorithm based on indirect tightly coupled non-linear optimization. Compared to a stereo visual-inertial pipeline like *OKVIS* [34] (see Section III-C) which can also deal with stereo cameras, *VINS-Mono* is specifically designed for monocular VIO with main differences in the initialization procedure. Furthermore, *VINS-Mono* reports slightly better accuracy results in unmanned aerial vehicle (UAV) applications [43] compared to *OKVIS* when used in monocular mode.

3) Batch optimization: Using *ROVIO* [6] as an estimator, *ROVIOLI* is an online front-end to build maps in the *maplab* [41] format. The created map can be post-processed using *maplab* batch bundle-adjustment to reduce drift and correct for linearization errors.

B. Stereo visual odometry algorithms

In addition to using inertial measurements, the metric scale of the motion can also be immediately retrieved using depth measurements of a stereo camera pair. In contrast to VIO, stereo-visual odometry does not require specific motions. However, as the method is purely visual, it will fail whenever the visual system faces challenges in tracking landmarks.

1) ORB-SLAM2: Mur-Artal and Tardos [32] provide a complete visual SLAM system for monocular, stereo or RGB-D cameras called *ORB-SLAM2*. The odometry front-end of the system is based on matching ORB features. The optimization is performed on a pose graph only containing the most relevant keyframes. Stereo constraints are incorporated in the cost function by projecting the landmarks with successful stereo matches to an augmented keypoint measurement including the coordinates of both cameras.

C. Stereo visual-inertial algorithms

In order to compensate for failure cases of the previously mentioned approaches, stereo vision and inertial measurements can be combined into a unified framework.

1) OKVIS: Leutenegger et al. [34] introduce tight-coupling of inertial and indirect visual measurements in a keyframe based approach optimizing inertial and re-projection errors in a sliding-window. In addition to the previously mentioned algorithms, *OKVIS* is able to deal with both stereo cameras by fusing landmarks visible in both frames and inertial data by using pre-integrated factors [44].

2) Stereo-SWE: Fusing landmarks, such as in *OKVIS*, can result in problems if the stereo matches contain wrong matches or outliers. Even if a robust cost function could avoid taking them into account, all additional information this landmark could provide is lost after a wrong merge. Alternatively, stereo matches could also be used as additional independent measurements for each landmark observation instead of fusing the landmarks.

Due to the lack of an available implementation for this approach, we extended the visual-inertial *Sliding Window Estimator (SWE)* presented by Hinzmann et al. [42]. In addition to the mentioned re-projection error and inertial error, a weighted depth error is introduced for each landmark observation with stereo matches. The depth error is the difference of the measured depth obtained by triangulating the stereo matches and the depth of the corresponding landmark projected to the camera. Inspired by [32], depth error uncertainties are scaled by their distance to the cameras and only considered up to a certain distance relative to the baseline between the cameras.

TABLE I

OVERVIEW OF VISUAL-AIDED ODOMETRY APPROACHES.

	Estimator type			Sensor measurements			Comment
	EKF	Sliding-window	Batch	Monocular	Stereo	IMU	
ROVIO [6]							Light-weight EKF based VIO using patch tracking.
VINS-Mono [9]							Tightly coupled indirect monocular VI fusion.
Batch optimization [41]							Offline global batch VI bundle-adjustment.
ORB-SLAM2 [32]							Indirect stereo visual SLAM framework.
OKVIS [34]							Keyframe based tight coupling of stereo VI fusion.
Stereo-SWE [42]							Tightly coupled VI fusion using depth as independent measurement.

Fig. 2. Customized Siemens Combino test vehicle [45] for data collection in Potsdam, Germany.

IV. EXPERIMENTAL EVALUATION

In this section, the experimental evaluation of the mentioned algorithms is shown[1].

We start with describing the datasets, explain recoding and evaluation procedure, and show the results and some in-depth analysis.

A. Datasets

Due to the lack of suitable available datasets, the estimators are tested on data recorded in Potsdam, Germany on a Siemens Combino tram (see Figure 2), which is customized for autonomous driving tests. `Trajectory1` is a short 780 m low-velocity (up to 25:5 km=h) track around the depot in an industrial environment and close-by structures as shown in Figure 3. `Trajectory2` is along a public tram-line about 2900 m long with speeds up to 52:4 km=h representing a real-life scenario. This trajectory includes rural, sub-urban, urban, and woody environments. The datasets were captured on a sunny day in August 2018 as dealing with extreme conditions for visual sensing (rain, fog, nighttime) is beyond the scope of this paper.

B. Hardware setup

1) VIO setup: For the data collections, we deployed a custom-built stereo visual-inertial sensor which is synchronizing all measurements in hardware similar as in [46]. To feature higher accuracy, exposure time compensation is utilized [47].

The sensor consists of two global-shutter cameras arranged in a fronto-parallel stereo setup and a compact, precision six degrees of freedom IMU. The camera was selected to provide a high frame-rate to be able to get a reasonable number of frames per displacement, even at higher speeds, and also to feature a high dynamic range to deal with the challenging

[1]All evaluations were performed on an Intel Xeon E3v5, 48 GB RAM laptop but not in real-time (2 4 fps).

Fig. 3. `Trajectory1` around the depot in an industrial environment with speeds up to 25:5 km=h.

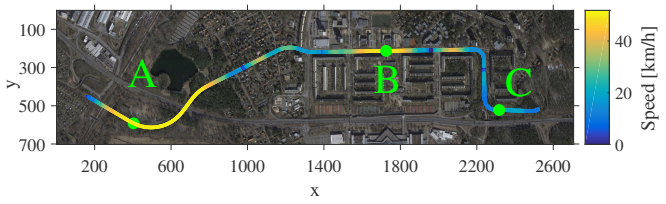

Fig. 4. `Trajectory2` following a public tram-line in Potsdam, Germany featuring a rural, sub-urban, urban, and woody environment and speeds up to 52:4 km=h. The green letters indicate challenging scenarios discussed in Section IV-D.

lighting conditions. The IMU was chosen to feature low noise values and also to support temperature calibration, as direct sunlight can highly change the temperature of the sensors. The sensor specifications are summarized in Table II. In order to investigate the influence of the baseline distance of the stereo setup, data was collected with three baselines of 31 cm, 71 cm and 120 cm. Those baselines were chosen to have a wide variety from baselines common in automotive applications up to those used in fixed wing UAVs. Figure 1 shows the deployed sensor in a 31 cm baseline configuration

[2]The hardware is able to capture up to 164 fps.

TABLE II

SENSOR SPECIFICATIONS DEPLOYED FOR DATA COLLECTION.

Device	Type	Specification
Camera	Basler acA1920-155uc	Frame-rate 20 fps[2], Resolution 1920 1200, Dynamic range 73 dB
Lens	Edmund Optics	Focal length 8 mm 70 deg opening angle; Aperture f=5:6
IMU	ADIS16445	Temperature calibrated, 300 Hz, 250 deg=s, 49 m=s^2

mounted behind the windshield inside the front cabin of the test vehicle.

2) Calibration: Sensor intrinsic and extrinsic calibration was performed using the *Kalibr* toolbox [47]. The transformation of the IMU with respect to the master camera (camera 1) is constant and calibrated in a lab environment. The transformation between the two cameras is then determined separately in-situ using a 7 5 checkerboard with tile sizes of 10:8 cm. Even though a larger calibration target might be beneficial to enable mutual observations, for the larger baselines we needed a board of 1:5 2 m and 3 4 m, respectively, which are more difficult to manufacture and handle and were not available during data collection.

3) Ground truth: Ground truth data is acquired using the high-precision RTK GNSS device OTXS RT3005G. Typical accuracies of 0:05 m and 0:1 deg are possible after post-processing.

C. Evaluation

The main metrics used in this paper to evaluate the performance of visual-aided odometry pipelines are incremental distance and heading errors.

We use the segment-based approach introduced by Nourani-Vatani and Borges [48] to deal with unbound errors in odometry [49]. Thereby, the trajectory is divided in segments. Two different segment lengths 10 m and 50 m are evaluated to test the evolution of errors. Each segment is aligned with the corresponding ground truth trajectory segments using the first 10 % of the segment. The distance error then corresponds to the distance between the endpoints of the segments. The heading error is the difference in heading estimation between the two segment ends.

D. Results and discussion

All ego-motion estimation pipelines investigated in Section III are tested on both trajectories using the different baselines. To enable a high level of comparability, the state-of-the-art pipelines are used out-of-the-box with only minor tuning. As pure odometry is under investigation here, all loop closing capabilities are disabled.

Figure 5 and Figure 6 show aligned paths of the different estimated trajectories to the ground truth.

1) Estimator performance: Using the 31 cm baseline, a good calibration can be ensured. Table III shows the evaluation results comparing the different estimation pipelines.

Both *ROVIO* and *VINS-Mono* fail to work properly for rail applications. Due to a very constrained motion and frequent constant velocity scenarios, the IMU biases cannot be estimated correctly locally. This results in significant scale drift, especially visible in `trajectory2` with longer sections of constant velocity. In addition, unobservable biases lead to inconsistent estimator state and the need to re-initialize multiple times during the trajectory. While *ROVIO* can partly recover from such a reset, *VINS-Mono* cannot, resulting in a somehow unfair comparison to the others. However, *ROVIO* also shows bad performance as it is highly dependent on a good knowledge of the IMU biases.

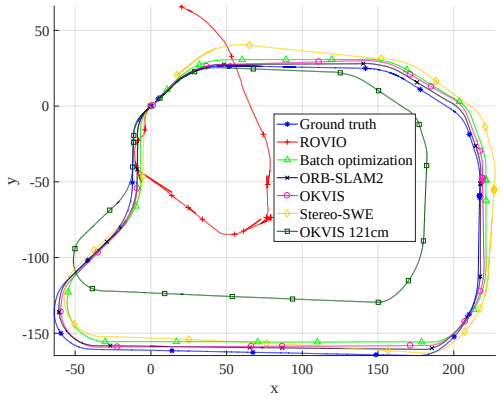

Fig. 5. Path alignment of trajectory estimations of the different motion estimation pipelines on `trajectory1`. If not stated otherwise, the 31 cm baseline is displayed. Due to the unrecoverable resets of the estimator, *VINS-Mono* is omitted here.

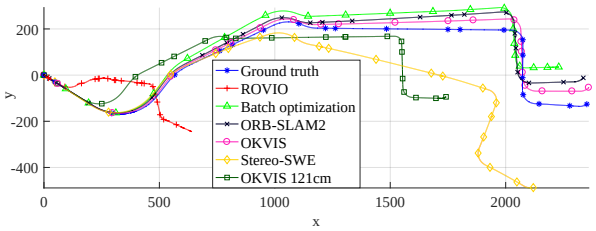

Fig. 6. Path alignment of trajectory estimations of the different motion estimation pipelines on `trajectory2`. If not stated otherwise, the 31 cm baseline is displayed. Due to the unrecoverable resets of the estimator, *VINS-Mono* is omitted here.

Using *ROVIOLI* to build a map and *maplab* [41] to globally batch bundle-adjust the maps, the scale and IMU biases can partially be recovered. This suggests that in both EKF and sliding-window approach, the bias estimation problem suffers significantly from linearization issues if there is not enough axis excitement in the window. A further hint for this is the improved performance of *VINS-Mono* compared to *ROVIO* which, among other possible causes, could be due to the difference in window size.

In comparison, when using stereo constraints, metric scale can be estimated correctly during the whole trajectory. On `trajectory1`, both *OKVIS* and *ORB-SLAM2* show very similar performance. `Trajectory2` is more challenging due to faster motion, more challening lighting conditions and more dynamic objects such as cars, pedestrians and other

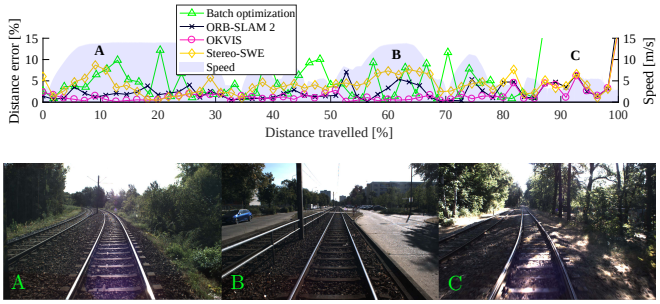

Fig. 7. Top: Errors of the best performing pipelines during `trajectory2` with a 31 cm baseline. The letters indicate selected challenging scenarios discussed in Section IV-D. Bottom: Camera images of the challenging scenarios.

TABLE III

RESULT OVERVIEW OF ESTIMATION ERRORS (DISTANCE IN % / HEADING IN $^{\mathrm{deg}}/_{\mathrm{m}}$) FOR THE 31 cm BASELINE CONFIGURATION. THE BEST PIPELINE IN THE RESPECTIVE SCENARIO AND ERROR METRIC (MEDIAN OR ROOT MEAN SQUARED ERROR (RMSE)) IS EMPHASIZED IN BOLD LETTERS.

[3] CONTAINS RESETS OF THE ESTIMATOR.

	Trajectory Segment length		Trajectory1 10 m	 50 m	Trajectory2 10 m	 50 m
Visual-inertial	ROVIO [6][3]	Median	66:570=0:0490	67:723=0:0578	75:292=0:0269	75:149=0:0210
		RMSE	74:620=0:1471	67:468=0:1119	77:297=0:0632	74:035=0:0511
	VINS-Mono [9][3]	Median	5:060 = 0:1033	10:589 = 0:4093	43:552 = 0:0408	45:339 = 0:0525
		RMSE	783:40 = 0:5966	250:9 = 0:5741	685:78 = 0:2412	274:54 = 0:1805
	Batch optimization [41]	Median	7:092 = 0:0322	2:899 = 0:0066	12:361 = 0:0153	4:239 = 0:0084
		RMSE	9:050 = 0:0685	4:396 = 0:0111	17:336 = 0:0302	10:90 = 0:0143
Stereo visual	ORB-SLAM2 [32]	Median	**2:138** = **0:0204**	3:054 = 0:0093	1:786 = 0:0078	1:829 = **0:0033**
		RMSE	3:751 = 0:0605	5:026 = 0:0436	4:526 = 0:0126	3:956 = 0:0073
Stereo visual-inertial	OKVIS [34]	Median	2:152 = 0:0219	**2:850** = **0:0070**	**1:428** = **0:0074**	**1:110** = 0:0038
		RMSE	**3:732** = 0:0336	**4:295** = **0:0103**	**3:361** = **0:0116**	**2:907** = **0:0055**
	Stereo-SWE [42]	Median	2:845 = 0:0249	4:029 = 0:0128	3:710 = 0:0099	3:840 = 0:0087
		RMSE	7:640 = **0:0332**	5:742 = 0:0113	5:552 = 0:0151	4:998 = 0:0116

trams in the scene. There, *OKVIS* is able to outperform *ORB-SLAM2* in most cases. This is especially visible in the RMSE in Table III which suggests that *OKVIS* also has a higher robustness compared to *ORB-SLAM2*. The complementing sensor modalities show benefits for motion estimation, most prominently in dynamic environments and at higher speeds. For *Stereo-SWE*, using the depth as an independent part in the optimization problem does not show an increase in accuracy. Also, it has a drawback of increasing in tuning parameter number, which is the weighting factor between depth errors and re-projection and inertial errors. This increases the tuning effort.

The distance errors along `trajectory2` for the best four performing estimators are shown in Figure 7. Three distinctive challenging scenarios *A*, *B* and *C* can be identified. They are also indicated in Figure 4. These challenging scenarios give evidence to the difference in estimator performances, and are summarized in Table V.

2) Challenging scenarios: Scenario *A* is visible approximately 10 % of the way through the trajectory. There, the tram is moving with high velocity, which increases the complexity of feature tracking. While the optimized feature tracking of *ORB-SLAM2* and *OKVIS* are able to deal with this, *Stereo-SWE* and *ROVIOLI* (used to build the map for batch optimization), which share the same feature-matching algorithm, have trouble finding enough feature matches.

Around scenario *B*, the tram is also moving at high speeds as shown in Figure 7. However, in contrast to scenario *A*, there is no curve and the optical flow of all nearby structure is in the same direction as the repetitive patterns on the ground such as railings or railroad ties. This leads to aliasing and wrong feature matches. More evidence can also be found at the beginning of `trajectory2` before entering the curve. Using the IMU as a complementing sensor modality, as in *OKVIS*, seems to be beneficial in this case. *Stereo-SWE* still has troubles with feature tracking as in scenario *A*. By masking out the area in-front of the vehicle where most visual aliasing is happening, different behaviours for the estimators can be observed as shown in Figure 8. While *OKVIS* behaves almost the same, the increase in error in

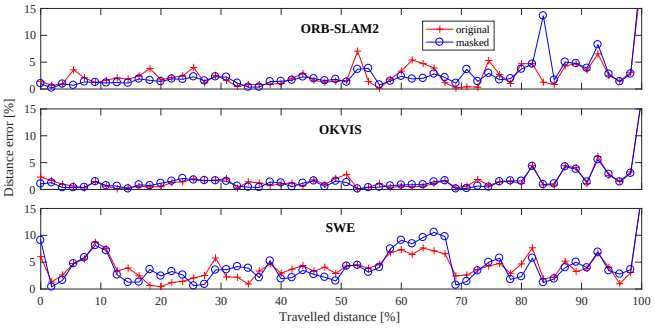

Fig. 8. Distance errors using the whole image (red) and masked out visual aliasing areas (blue).

scenario *B* for *ORB-SLAM2* can be reduced. However, by removing some of the visual information in slow sections and no other available measurement source, the level of robustness is decreased. This can be observed after about 83 % of the trajectory where short heavy reflections in the upper part of the image can cause huge errors. In this scenario, one could use Nyquist theory to detect and neglect affected regions instead of neglecting static parts of the image. The *SWE* shows higher errors in scenario *B* due to the reduced visual information in an already challenging feature-matching scenario.

Finally, scenario *C* occurs in the woods at the end of the trajectory. Fast switches from shadows to direct sunlight seem to be a challenge for all investigated pipelines. Using improved auto-exposure control [51] and shadow compensation [50] might be beneficial.

3) Baseline comparison: In general, a larger baseline should provide more reliable depth information as the quantization error in the disparity is less prominent. However, the quality of depth calculation is highly sensitive to a good calibration. Using conventional methods, calibration is much more challenging using a higher baseline as it is harder to guarantee mutual observations of the calibration target for both cameras. This results in a degrading calibration quality with higher baselines. Table IV summarizes evaluation results using the stereo visual(-inertial) pipelines with different baselines.

TABLE IV

RESULT OVERVIEW OF MEDIAN ERRORS (DISTANCE IN % / HEADING IN $\frac{\text{deg}}{\text{m}}$) CHANGING THE BASELINES. THE BEST PIPELINE IN THE RESPECTIVE SCENARIO IS EMPHASIZED IN BOLD LETTERS. [4] LOSES TRACK AT HIGHER SPEEDS AFTER ABOUT 7 8 % OF THE TRAJECTORY.

	Segment length	10 m			50 m		
	Baseline	31 cm	71 cm	120 cm	31 cm	71 cm	120 cm
trajl	ORB-SLAM2 [32]	**2:138 = 0:0204**	**2:701** = 0:0199	30:128 = 0:1140[4]	3:054 = 0:0093	**5:002** = 0:0110	33:862 = 2:8459[4]
	OKVIS [34]	2:152 = 0:0219	3:077 = 0:0198	20:733 = **0:0154**	**2:850 = 0:0070**	5:049 = **0:0055**	21:184 = **0:0075**
	Stereo-SWE [42]	2:845 = 0:0249	4:543 = **0:0177**	**18:340** = 0:0225	4:029 = 0:0128	6:649 = 0:0081	**19:247** = 0:0206
traj2	ORB-SLAM2 [32]	1:786 = 0:0078	**3:247** = 0:0094	32:121 = 0:0478[4]	1:829 = **0:0033**	3:783 = 0:0049	26:772 = 0:0305[4]
	OKVIS [34]	**1:428 = 0:0074**	3:465 = **0:0068**	30:947 = **0:0072**	**1:110** = 0:0038	**3:609 = 0:0036**	29:045 = **0:0035**
	Stereo-SWE [42]	3:710 = 0:0099	5:621 = 0:0096	**24:299** = 0:0097	3:840 = 0:0087	6:152 = 0:0060	**25:271** = 0:0054

TABLE V

CHALLENGING SCENARIOS OF THE MOTION ESTIMATION PIPELINES.

Scenario	Affected estimators	Cause	Solution
A	*Stereo-SWE*	High speed	Optimize feature tracking.
B	*ORB-SLAM2*, *Stereo-SWE*	High speed & Aliasing	Detect and neglect affected region; use IMU fusion.
C	all	Lighting conditions	Shadow compensation [50]; improve auto exposure [51].

For the stereo algorithms shown in the scenarios of `trajectory1` and `trajectory2`, it seems more important to have an accurate calibration using a small baseline than to be able to reliably incorporate further away landmarks. Using IMU measurements, motion estimation is still possible up to a fixed scale error, visible in Figure 5 and Figure 6, while *ORB-SLAM2* loses track at higher speeds. In order to benefit from the advantages of higher baselines, improved calibration procedures such as online calibration [52] could have a high benefit.

V. CONCLUSIONS

Being able to accurately localize a rail vehicle has a high potential to improve infrastructure efficiency. In a real-world application, high safety levels are typically achieved using redundant systems. This paper studies the contribution visual systems can provide to getting closer to robust high accuracy odometry. We did an in-depth experimental comparison using real-world rail datasets of several state-of-the-art visual-aided odometry pipelines.

It was observed that the monocular visual-inertial odometry methods *ROVIO* and *VINS-Mono* experience severe scale drift and are not able to keep a consistent estimator state due to locally unobservable IMU biases. Using stereo vision, accurate motion estimation is achievable, especially using the stereo visual-inertial pipeline *OKVIS*. Specific challenging scenarios for the individual pipelines can result from high speeds, aliasing with repetitive patterns on the ground, and challenging lighting conditions.

In conclusion, even without enforcing specific motions, visual-aided odometry can achieve high accuracies for rail vehicles, but is not reliable enough for use in isolation for safety critical applications. However, in combination with other odometry and localization methods such as GNSS, wheel odometry or ground radars, vision can complement for failure cases of other sensors. Motion constraints can be incorporated either as a separate part in the estimation pipeline or directly into the investigated algorithms using a motion model in the propagation state for EKF based algorithms or additional motion model error term in optimization based algorithms. High-speed scenarios will require higher framerates to ensure feature tracking and a larger baseline for reliable depth estimation of unblurred landmarks implying calibration challenges. Furthermore, all tested datasets are recorded during nice weather. Like most visual pipelines, the investigated approaches will suffer from limited visibility. However, using cameras with extended spectral band sensitivity such as long-wave infrared (LWIR) shows potential to enable also good performance in bad weather conditions [53].

ACKNOWLEDGEMENT

This work was generously supported by Siemens Mobility, Germany. The authors would like to thank Andreas Pfrunder for his help in initial data collections and evaluations.

REFERENCES

[1] Bundesamt für Statistik, "Verkehrsleistungen im Personenverkehr - 1960-2016," 2017. [Online]. Available: https://www.bfs.admin.ch/bfs/de/home/statistiken/mobilitaet-verkehr/personenverkehr/leistungen.assetdetail.4083119.html

[2] O. Stalder, *ETCS for Engineers*, 1st ed. Hamburg, Germany: DVV Media Group GmbH, 2011.

[3] C. Williams, "The next ETCS Level?" in *2016 IEEE International Conference on Intelligent Rail Transportation, ICIRT 2016*, 2016.

[4] J. Beugin and J. Marais, "Simulation-based evaluation of dependability and safety properties of satellite technologies for railway localization," *Transportation Research Part C: Emerging Technologies*, 2012.

[5] J. Marais, J. Beugin, and M. Berbineau, "A Survey of GNSS-Based Research and Developments for the European Railway Signaling," *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, no. 10, pp. 2602–2618, 2017.

[6] M. Bloesch, S. Omari, M. Hutter, and R. Siegwart, "Robust visual inertial odometry using a direct EKF-based approach," in *IEEE International Conference on Intelligent Robots and Systems*, Seattle, WA, USA, 2015, pp. 298–304.

[7] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardós, "ORB-SLAM: A Versatile and Accurate Monocular SLAM System," *IEEE TRANSACTIONS ON ROBOTICS*, vol. 31, no. 5, 2015.

[8] J. Engel, V. Koltun, and D. Cremers, "Direct Sparse Odometry," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.

[9] T. Qin, P. Li, and S. Shen, "VINS-Mono: A Robust and Versatile Monocular Visual-Inertial State Estimator," *IEEE Transactions on Robotics*, vol. 34, no. 4, pp. 1004–1020, 2018.

[10] M. Burki, I. Gilitschenski, E. Stumm, R. Siegwart, and J. Nieto, "Appearance-based landmark selection for efficient long-term visual localization," in *EEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. Daejeon, Korea: IEEE, 10 2016, pp. 4137–4143.

[11] A. Martinelli, "Closed-form solution of visual-inertial structure from motion," *International Journal of Computer Vision*, vol. 106, no. 2, pp. 138–152, 2014.

[12] G. Hitz, F. Bloechliger, M. Buerki, K. Egger, and A. Pfrunder, "Sevensense - Camera-Based Navigation Solutions for the Next Generation of Autonomous Robots," 2018. [Online]. Available: https://sevensense.ch/

[13] M. Rendall, B. Webb, and R. Gariepy, "Clearpath Robotics: Autonomous Mobile Robots," 2018. [Online]. Available: https://www.clearpathrobotics.com/

[14] B. SA, "BlueBotics - Mobile Robots at Your Service," 2018. [Online]. Available: http://www.bluebotics.com/

[15] D. H. Murillas and L. Poncet, "Safe odometry for high-speed trains," in *IEEE International Conference on Intelligent Rail Transportation, ICIRT*, Birmingham, UK, 2016, pp. 244–248.

[16] A. W. Palmer and N. Nourani-Vatani, "Robust Odometry using Sensor Consensus Analysis," in *IEEE International Conference on Intelligent Robots and Systems (accepted)*, Madrid, Spain, 2018.

[17] R. Cheng, Y. Song, D. Chen, and L. Chen, "Intelligent Localization of a High-Speed Train Using LSSVM and the Online Sparse Optimization Approach," *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, no. 8, pp. 2071–2084, 8 2017.

[18] B. Allotta, P. D'Adamio, M. Malvezzi, L. Pugi, A. Ridolfi, and G. Vettori, "A localization algorithm for railway vehicles," in *Conference Record - IEEE Instrumentation and Measurement Technology Conference*, vol. 2015-July, 2015.

[19] J. Otegui, A. Bahillo, I. Lopetegi, and L. E. Díez, "A Survey of Train Positioning Solutions," *IEEE SENSORS JOURNAL*, vol. 17, no. 15, 2017.

[20] C. Hasberg, S. Hensel, and C. Stiller, "Simultaneous localization and mapping for path-constrained motion," *IEEE Transactions on Intelligent Transportation Systems*, vol. 13, no. 2, pp. 541–552, 6 2012.

[21] O. Heirich, P. Robertson, and T. Strang, "RailSLAM - Localization of rail vehicles and mapping of geometric railway tracks," in *Proceedings - IEEE International Conference on Robotics and Automation*. IEEE, 5 2013, pp. 5212–5219.

[22] S. Du, W. Sun, and Y. Gao, "Improving Observability of an Inertial System by Rotary Motions of an IMU," *Sensors (Basel, Switzerland)*, vol. 17, no. 4, 2017.

[23] M. Spindler, D. Stein, and M. Lauer, "Low Power and Low Cost Sensor for Train Velocity Estimation," in *2016 IEEE International Conference on Intelligent Rail Transportation (ICIRT)*, Birmingham, UK, 2016, pp. 0–5.

[24] J. Wohlfeil, "Vision based rail track and switch recognition for self-localization of trains in a rail network," in *IEEE Intelligent Vehicles Symposium, Proceedings*. IEEE, 6 2011, pp. 1025–1030.

[25] B. Bah, E. Jungabel, M. Kowalska, C. Leithäuser, A. Pandey, and C. Vogel, "ECMI-report Odometry for train location," 2009.

[26] R. Altendorfer, S. Wirkert, and S. Heinrichs-Bartscher, "Sensor Fusion as an Enabling Technology for Safety-critical Driver Assistance Systems," *SAE International Journal of Passenger Cars - Electronic and Electrical Systems*, vol. 3, no. 2, pp. 183–192, 2010.

[27] A. J. Davison, I. D. Reid, N. D. Molton, and O. Stasse, "MonoSLAM: Real-Time Single Camera SLAM," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 6, pp. 1052–1067, 6 2007.

[28] H. Strasdat, J. M. Montiel, and A. J. Davison, "Real-time monocular SLAM: Why filter?" in *Proceedings - IEEE International Conference on Robotics and Automation*, 2010, pp. 2657–2664.

[29] M. Bloesch, M. Burri, H. Sommer, R. Siegwart, and M. Hutter, "The Two-State Implicit Filter - Recursive Estimation for Mobile Robots," *IEEE Robotics and Automation Letters*, vol. 3, no. 1, pp. 1–1, 2017.

[30] L. von Stumberg, V. Usenko, and D. Cremers, "Direct Sparse Visual-Inertial Odometry using Dynamic Marginalization," in *IEEE International Conference on Robotics and Automation (ICRA)*, Madrid, 2018.

[31] R. Mur-Artal and J. D. Tardos, "Visual-Inertial Monocular SLAM with Map Reuse," *IEEE Robotics and Automation Letters*, vol. 2, no. 2, pp. 796–803, 2017.

[32] ——, "ORB-SLAM2: An Open-Source SLAM System for Monocular, Stereo, and RGB-D Cameras," *IEEE Transactions on Robotics*, pp. 1–8, 2017.

[33] R. Wang, M. Schwörer, and D. Cremers, "Stereo DSO: Large-scale direct sparse visual odometry with stereo cameras," in *International Conference on Computer Vision*, Venice, Italy, 2017.

[34] S. Leutenegger, S. Lynen, M. Bosse, R. Siegwart, and P. Furgale, "Keyframe-based visualinertial odometry using nonlinear optimization," *The International Journal of Robotics Research*, vol. 34, no. 3, pp. 314–334, 2015.

[35] U. Schwesinger, M. Burki, J. Timpner, S. Rottmann, L. Wolf, L. M. Paz, H. Grimmett, I. Posner, P. Newman, C. Hane, L. Heng, G. H. Lee, T. Sattler, M. Pollefeys, M. Allodi, F. Valenti, K. Mimura, B. Goebelsmann, W. Derendarz, P. Muhlfellner, S. Wonneberger, R. Waldmann, S. Grysczyk, C. Last, S. Bruning, S. Horstmann, M. Bartholomaus, C. Brummer, M. Stellmacher, F. Pucks, M. Nicklas, and R. Siegwart, "Automated valet parking and charging for e-mobility," *IEEE Intelligent Vehicles Symposium, Proceedings*, vol. 2016-Augus, pp. 157–164, 2016.

[36] P. Merriaux, Y. Dupuis, P. Vasseur, and X. Savatier, "Wheel odometry-based car localization and tracking on vectorial map," in *17th International IEEE Conference on Intelligent Transportation Systems (ITSC)*, Qingdao, China, 2014, pp. 1890–1891.

[37] W. Churchill and P. Newman, "Experience Based Navigation: Theory, Practice and Implementation," Ph.D. dissertation, University of Oxford, 2012.

[38] D. Barnes, W. Maddern, G. Pascoe, and I. Posner, "Driven to Distraction: Self-Supervised Distractor Learning for Robust Monocular Visual Odometry in Urban Environments," 2017.

[39] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, Providence, RI, USA, 2012.

[40] D. Scaramuzza, F. Fraundorfer, M. Pollefeys, and R. Siegwart, "Absolute scale in structure from motion from a single vehicle mounted camera by exploiting nonholonomic constraints," *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1413–1419, 2009.

[41] T. Schneider, M. Dymczyk, M. Fehr, K. Egger, S. Lynen, I. Gilitschenski, and R. Siegwart, "maplab: An Open Framework for Research in Visual-inertial Mapping and Localization," *IEEE Robotics and Automation Letters*, vol. 3, no. 3, pp. 1418–1425, 11 2018.

[42] T. Hinzmann, T. Schneider, M. Dymczyk, A. Schaffner, S. Lynen, R. Siegwart, and I. Gilitschenski, "Monocular Visual-Inertial SLAM for Fixed-Wing UAVs Using Sliding Window Based Nonlinear Optimization," in *International Symposium on Visual Computing (ISVC)*, Las Vegas, USA, 2016.

[43] J. Delmerico and D. Scaramuzza, "A Benchmark Comparison of Monocular Visual-Inertial Odometry Algorithms for Flying Robots," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2018.

[44] T. Lupton and S. Sukkarieh, "Visual-inertial-aided navigation for high-dynamic motion in built environments without initial conditions," *IEEE Transactions on Robotics*, vol. 28, no. 1, pp. 61–76, 2012.

[45] Siemens Mobility GmbH, "Siemens Mobility presents worlds first autonomous tram," 2018. [Online]. Available: https://www.siemens.com/press/PR2018090290MOEN

[46] J. Nikolic, J. Rehder, M. Burri, P. Gohl, S. Leutenegger, P. T. Furgale, and R. Siegwart, "A synchronized visual-inertial sensor system with FPGA pre-processing for accurate real-time SLAM," in *IEEE International Conference on Robotics and Automation*, Hong Kong, China, 2014, pp. 431–437.

[47] P. Furgale, J. Rehder, and R. Siegwart, "Unified temporal and spatial calibration for multi-sensor systems," in *IEEE International Conference on Intelligent Robots and Systems*, Tokyo, Japan, 2013, pp. 1280–1286.

[48] N. Nourani-Vatani and P. V. K. Borges, "Correlation-based visual odometry for ground vehicles," *Journal of Field Robotics*, vol. 28, no. 5, pp. 742–768, 2011.

[49] A. Kelly, "Linearized error propagation in odometry," *International Journal of Robotics Research*, vol. 23, no. 2, pp. 179–218, 2004.

[50] C. McManus, W. Churchill, W. Maddern, A. D. Stewart, and P. Newman, "Shady dealings: Robust, long-term visual localisation using illumination invariance," *Proceedings - IEEE International Conference on Robotics and Automation*, pp. 901–906, 2014.

[51] N. Nourani-Vatani and J. M. Roberts, "Automatic Camera Exposure Control," *Proceedings of the Australasian Conference on Robotics and Automation*, pp. 1–6, 2007.

[52] T. Schneider, M. Li, M. Burri, J. Nieto, R. Siegwart, and I. Gilitschenski, "Visual-inertial self-calibration on informative motion segments,"

in *IEEE International Conference on Robotics and Automation*, Singapore, 2017, pp. 6487–6494.

[53] N. Pinchon, M. Ibn-Khedher, O. Cassignol, A. Nicolas, F. Bernardin, P. Leduc, J. P. Tarel, R. Bremond, E. Bercier, G. Julien, N. Pinchon, O. Cassignol, F. Bernardin, P. Leduc, J.-P. Tarel, R. Brémond, and E. Bercier, "All-weather vision for automotive safety: which spectral band?" in *International Conference Night Drive Tests and Exhibition*, Paris, France, 2016.