Highly Accurate Video-Based Train Localization - replacing Balises with Natural Reference Points

Darius Burschka*, **Christian Robl****

*Department of Computer Science, Technical University Munich
Munich, Germany
email: burschka@tum.de

**M2C ExpertControl GmbH
Offenberg, Germany
email: christian.robl@m2cec.com

Abstract: *Recently, visual odometry has been successfully applied as a video-based approach across domains. We adapted this approach to railways achieving excellent results without using any other conventional rail sensors. Herewith, we propose an extension to our visual rail odometry approach that allows to visually compensate for the inevitable odometry drifts based on sporadically visible local scene structures and that provides means for a highly accurate train localization based on existing geo-referenced infrastructure of the rail system. The specific conditions of the visual rail navigation require an adaptation of the conventional VSLAM (Video-based Simultaneous Localization and Mapping) systems to cope with the limited and self-similar property of the observed area. We show how this extension can be used to replace the currently used* train report system *with a significantly increased global accuracy and reduced drift in the estimation between the geo-referenced rail structures like balises. Furthermore, a migration scenario is proposed which overcomes the issue of the approval of new localization systems.*

Area: *Rail Navigation*

1. Motivation

Current train control systems locate trains in a block based manner using track occupancy units (e.g. axle counter). However, in order to increase the capacity on the tracks, moving blocks will have to be introduced, running successive trains within their absolute braking distance. With ETCS Level 2, trains locate themselves with track-side balises and their wheel odometry with a required accuracy of 5m + 5% of the distance to the last balise group. Realizing moving block (ETCS Level 3) requires a continuous, highly accurate and safe (up to SIL 4) localization of the head of each train. Therefore new localization systems for trains are required. The goal of the Swiss programme ”smartrail 4.0” [1] is that all track bound objects locate themselves track selective at anytime. Possible sensor combinations for the different use cases are shown in the related feasibility study [8] focusing on GNSS-IMU-wheel odometry [3]. In the Austrian project ”Greenlight” GNSS is used as the main means of localization in addition to fiber optic sensing and also IMU [9]. Video localization is a supporting sensor in these projects [2]. According to

978-3-944976-28-0

[8], [9] the most demanding use cases are ”Fast and precise clearing of track, switch, barrier (Use case 1.1)” and ”Precise train end position (Use case 1.2)”. Its requirements with regard to functional, technical and operational properties of the individual objects to be localized are shown in Fig. 1.

Use Case Nr	Item of Localization	SIL	Accuracy	MTTF	MTTR	Availability	Latency or Periode
1.1	Track occupancy	SIL3/ SIL4	x ≤66.4 m R Track selective y ≤+/-1.5 m R	1 Year	0,5h	0,9999375	1s
1.2 – 1.7 8.1 – 8.2	Train Position	SIL 3	10 m R Track selective y ≤+/-1.5 m R	1 Year	0,5h	0,9999375	1s

Figure 1: Requirements in most demanding use cases from [8]

Conventional visual odometry achieves the reported high accuracy and small drift through selection of well distributed features in the entire camera image, which are visible over long frame sequences. The larger the angular field covered by the light rays entering the camera, the more accurate is the resulting accuracy of the system [5]. However, the large field of view required for such systems is not available in many situations during train navigation because of the occlusions in the scene through trains running on neighboring tracks. In many cases. the only reliable motion information can be reconstructed from the areas in the ground area of the tracks in front of the train (Fig. 2).

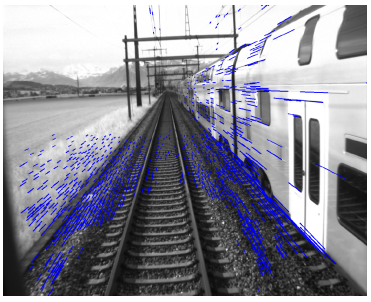

Figure 2: Ground area around the tracks - the only reliable source of information for navigation.

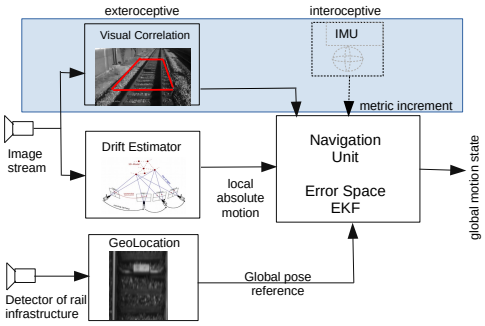

Figure 3: System architecture.

The frame-rate of a camera system (e.g. 1280x1024 @ 60Hz) is often too low to capture the high-dynamics of the sensed motion. Tilting trains adapting the train body to the curvature of the track or trains with a flexible suspension may require a fast motion capture unit. The resulting high frequency swings of the train may drastically change the orientation of the camera to the ground. This can be solved with an *interoceptive sensor*, like an inertial unit (IMU) (Fig. 3) or wheel odometry. Interoceptive sensors do not rely on external information from the environment. This simplifies their processing, makes them independent on the outside conditions, and allows a fast update rate ($>$ 1kHz). In our system, we extend the incremental motion sensing with *exteroceptive sensing* in form of a visual correlation unit that senses directly the relative motion to the static scene and avoids errors like wheel slip (especially for powered axles), wheel wear or temperature drifts that are difficult to model.

The incremental systems sense just incremental changes of motion and the occurring errors are accumulated to drifts of the localization unit. The unavoidable drifts are compensated by a visual drift compensation unit that observes external structures that may be temporarily occluded by passing trains. The results are used to estimate the current drift in the position estimated by the incremental stage. The final unit is a dedicated detector of geo-located rail infrastructure, e.g. balises, masts for the catenary, bridges, points or other surveyed references, that is used to provide a global geo-location of the train.

It is still an open issue how to get approval for such a new localization system based on mainly new technology for rail applications. The necessary ground truth as reference needs to be one dimension more accurate than the localization system. Since track selectivity needs to be achieved with SIL 4, such a reference is hard to obtain at anytime [11], especially for and with GNSS based systems [10] . However this will be necessary for getting approval by the authorities.

2. Approach

We propose to modify the conventional visual odometry approaches to cope with the uniqueness problems of the train navigation. Since typical structure from motion approaches are not feasible in a continuous manner, we base our system on integration of the incremental units (Figure 3) that can operate at update frequencies higher than a typical camera frame-rate. We use the structure from motion approaches to estimate the drifts in an error state Kalman Filter, which can be done with much lower update frequencies.

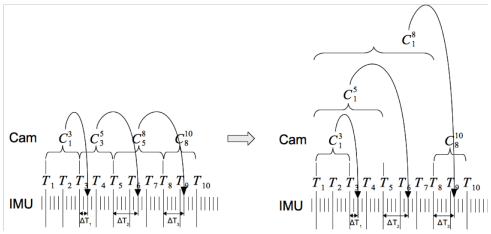

Figure 4: (Left) fusion of the incremental units, (right) Keyframe update of the drift when reference structures become visible.

The navigation unit (Fig. 3) can further optimize the calculation of the distance by freezing the reference frame $\mathcal{I}'_{\sqcup}$ (key-frame) for a number of following frames, if the estimated velocity is slow. Since the distance traveled is the integral of the responses from the optical correlation, small detection errors usually integrate to increasing drifts in the distance. Switching to the key-frame-processing results in the detection errors appearing as noise overlayed over the true distance instead of appearing as accumulated drift.

2.1. Robust Estimation of Metric Motion Parameters

Conventional Visual SLAM approaches use the information from a sparse point matching system in the camera images. The points are *tracked* between the image pairs from the sequence or *matched* based on the local information in the neighborhood of the points. The difference is that while *tracking* assumes a local search around the expected position, in which a local image patch is searched, *matching* allows larger changes in the image position, because each point is described by a more or less complex description (SIFT or AGAST).

While this processing works in most flying and automotive environments, we need to be able to match the information in the area of the tracks with a very strong self similarity that leads to many mismatches between the frames. We increase the uniqueness of the local environment by growing the local region to a large area shown in the Fig. 5. We try to match this template in the consecutive image using a Sum-of-Squared-Differences (SSD) method from OpenCV. We refer to this module because of the similarity to an optical computer mouse as "Train Mouse".

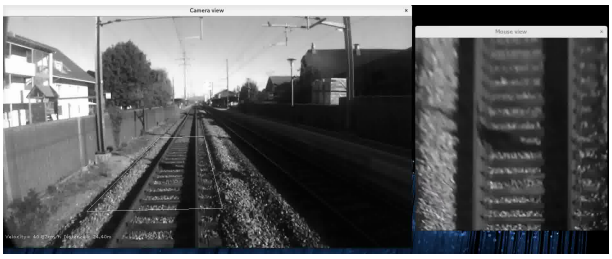

Figure 5: The rectangular region shown in the left image is rectified to the "top-view" image shown on the right. A template in this image is searched in the consecutive image rectified in the same way.

A homography matrix $\tilde{H}$ that was used to calculate the rectified image $\mathcal{I}'$ in Fig. 5 right $\tilde{H} = \left(\tilde{R} + \frac{\vec{T}\vec{n}^T}{d}\right)$. The rotation matrix $\tilde{R}$ describes the rotation between the current orientation of the physical camera and the top-view orientation of the rectified view. The vector $\vec{T}$ describes the translation between the images, which is zero in our case.

We search for a rectangular template with the size (x', y') from the ${\mathcal{I}'}_t$ region of the first image in the corresponding region ${\mathcal{I}'}_{t+1}$ using the SSD template matching method that searches for the maximum of the function (1):

$$f(x_p, y_p) = \sum_{x',y'} ({\mathcal{I}'}_t(x', y') - {\mathcal{I}'}_{t+1}(x_p + x', y_p + y'))^2 \tag{1}$$

The estimated displacement $(x_p, y_p)_t$ from the maximum response of $f(x_p, y_p)$ estimates the horizontal and vertical image motion of the template between the images. This measures a pixel accurate shift of the template between the images. The search for the correct displacement for the current $(x_p, y_p)_t$ can be accelerated by using a prediction of these values. In a generic case, the system needs to check the entire possible range of $\{x_p, y_p\}$ that covers the entire possible velocity profile. This is a computationally intensive operation. Due to the high inertia of the train, these value change only little between consecutive frames. We can reduce the search for the correct placement of the template only to a small band around the previous $(x_p, y_p)_{t-1}$ values.

We can calculate a more accurate displacement of the template between the images by applying a sub-pixel alignment of the templates. If the remaining change between both images is under 1 [pixel] then we can use the Taylor series expansion to explain the brightness change at a specific pixel $\mathcal{I}'(x, y)$ to:

$${\mathcal{I}'}_t(x + \delta x.y + \delta y) \approx {\mathcal{I}'}_t(x, y) + \frac{\partial {\mathcal{I}'}_t(x, y)}{\partial x}\delta x + \frac{\partial {\mathcal{I}'}_t(x, y)}{\partial y}\delta x \tag{2}$$

If we assume that the new image ${\mathcal{I}'}_{t+1}$ is a result of a sub-pixel motion $(\delta x, \delta y)$ then we can

estimate from the equation:

$$\mathcal{I'}_{t+1}(x.y) - \mathcal{I'}_t(x,y)\frac{\partial \mathcal{I'}_t(x,y)}{\partial x}\delta x + \frac{\partial \mathcal{I'}_t(x,y)}{\partial y}\delta x = \vec{\mathcal{G}}^T \cdot \delta\vec{p} = ||\vec{\mathcal{G}}|| \cdot ||\delta\vec{p}|| \tag{3}$$

$$\text{with} \quad \vec{\mathcal{G}} = \left(\frac{\partial \mathcal{I'}_t(x,y)}{\partial x}, \frac{\partial \mathcal{I'}_t(x,y)}{\partial y}\right)^T$$

We see that once we calculated the gradient vector $\mathcal{G}$ from the previous image, we can calculate the sub-pixel update of the motion in horizontal and vertical direction $(\delta x.\delta y)$ by decomposing the motion $||\vec{\delta p}||$ along the gradient according to the horizontal and vertical ratios of $\vec{\mathcal{G}}$.

We calculate the resulting shift as an average of responses within the template. It is obvious from (3) that only pixels with a difference in brightness between the images contribute to the motion estimation. We reduce the sensitivity to noise by using only pixels with the gradient above a threshold $||\vec{\mathcal{G}}|| > \epsilon_G$, which is tuned depending on the expected camera noise.

The resulting average image motion $(\Delta x, \Delta y)$ can be linearly scaled to the forward and side-wards metric velocity with knowledge about the mounting height L above the ground. The metric values of the forward velocity v_l and the side-wards motion v_s (due to curves in the route) can be computed from *similar triangles* relation between the camera projection on the image plane and the relation of the height L of a rectified camera providing the image $\mathcal{I'}$ to:

$$\Delta x_i = x_p + \delta x, \quad \Delta y_i = y_p + \delta y, \quad v_l = \frac{L \cdot p_y}{f \cdot t_f}\Delta y_i, \quad v_s = \frac{L \cdot p_x}{f \cdot t_f}\Delta x_i \tag{4}$$

Possible changes in the orientation of the camera image $\mathcal{I'}$ scale it with the focal length f, the metric pixel-size (p_x, p_y) and the time interval between two frames t_f as it is shown in (4).

2.2. Keyframe State Update from Rail Infrastructure

In the recent years, the development of Vision Aided Inertial Navigation Systems (VINS) [6] showed great progress. Mourikis et al. [7] demonstrated a hard-realtime capable mono vision/IMU fusion algorithm using an Extended Kalman Filter (EKF). In their approach a certain window over past poses is kept within the filter state vector to process feature measurements taken from different locations along the traveled trajectory. Using limited data windows makes real-time implementation possible but turns the system into an odometry system as trajectory loop closures cannot be integrated (Figure 4).

System state estimation for safety critical systems requires sensor data fusion in hard real-time. Probabilistic filters are often used for this purpose due their simplicity, low computational complexity and deterministic timing behavior. A numerically robust filter implementation as well as full system state observability are fundamental to guarantee long-term stable state estimation. While numerically stable algorithms are well established, a state estimation formulation with full observability can not always be guaranteed. This situation is critical in two aspects: firstly, unbounded filter covariances can cause numerical instability. Secondly, linearization of non-linear systems often assumes small state errors.

A transformation function between the unobservable states and a local reference is defined while the local reference is augmented to the filter state vector. The prediction step is used to switch the filter states and its covariance to the new local reference. It is marginalized out at the same time. We propose the concept of a vision-aided inertial navigation filter. We sporadically change the filter reference frame to a new frame with a lower relative uncertainty compared to the current system state. We separate local real-time state estimation from global navigation and relax timing constraints on the latter one. The implementation is realized as a square root UD filter [4] to improve numerical stability. Stochastical cloning [7] is used in (5) for the correct fusion of two updates with varying latency time ΔT (Fig. 4).

$$\tilde{x}_{k+1} = \begin{pmatrix} x_{aug} \\ x_{k+1} \end{pmatrix} = \begin{pmatrix} I & 0 \\ \tilde{0} & \tilde{A}_k \end{pmatrix} \begin{pmatrix} x_{aug} \\ x_k \end{pmatrix} \quad \tilde{P}_{k+m|k} = \begin{pmatrix} \tilde{P}_{kk} & \tilde{P}_{kk}\tilde{F}^T \\ \tilde{F}\tilde{P}_{kk} & \tilde{P}_{k+m|k} \end{pmatrix}, \quad \text{with } \tilde{F} = \prod_{i=1}^{m} \tilde{A}_{k+i} \tag{5}$$

In SIL 4 approved ETCS Level 2 operation, a train is sending every 6 seconds a train position report (TPR) to the Radio Block Center (RBC). The TPR consist of the last passed balise (group) and the traveled distance from there as well as the direction of travel. To overcome the issues of certification we propose that in a 1st stage we prove that the new localization system has the same performance regarding quality and safety than the certified one (GAMAB principle). Therefore with the new localization system (regardless of the technology) the same TPRs have to be generated with at least the same quality. Successfully comparing a statistical relevant number of TPRs from the current and new systems can be used to get approval according to e.g. CSM 2013/402/EC [11]. Within the presented video localization we will use a global drift compensation with natural reference points and rail infrastructure other than balises (e.g. points, bridges, tunnels, catenary masts). However we will use visual balise detection - without the need for a separate sensor - in order to trigger the TPR generation for comparison purposes. With a successful proof of the same or better quality of the new system we can introduce ”artificial balises” wherever required to meet the requirements for realizing moving block. At the same time, this will lead to a lean and promising migration strategy (no change to the ETCS interface to the RBC).

3. Results

Figure 6: Installation of the camera system.

The mobile system for the image acquisition in real trains is composed of a NIR camera with a focal length of 8mm and resolution of 1280x1024 pixels @ 60Hz, a GPS receiver for time synchronization with GPS time, a rapid prototyping computer (i7-6700 with 8 cores @ 2.40 GHz) and a battery pack for an independent power supply. The camera system was mounted on the windscreen of the locomotive with a pitch angle towards the track of $8°$ (SBB setup) or $21.5°$ (OeBB setup) resp.. The bigger tilt angle was chosen in order to have a better pixel resolution of the track width, since according to [2] the track detection in the

image is the dominant contribution to the uncertainty. Tilt and yaw angles will be identified and compensated for during image processing as proposed in [2].

The installation in a Swiss locomotive Re 420 and an Austrian locomotive/baggage car class 4061 can be seen in Fig. 6. On the Re 420 only the camera system was installed where the reference GPS/IMU system was installed on a measurement coach that was coupled to the Re 420. This leads to some synchronization issues in time (latencies) and place (locomotive starts/stops moving earlier than the coach).

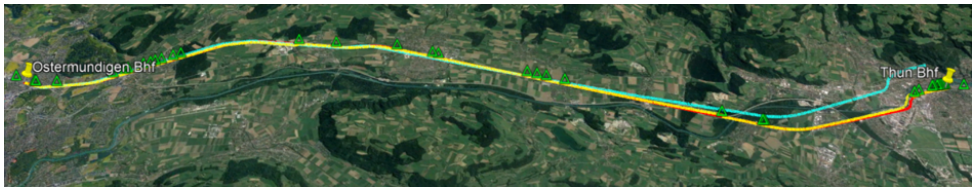

Figure 7: Comparison of the estimated paths between Ostermundingen and Thun:(cyan) camera w/o correction, (red) GNSS/IMU reference, and (yellow) camera with update with railway point frogs (green triangles).

Figure 7 shows the calculated 3D path without drift compensation (cyan) of one of the measurements taken from Ostermundingen to Thun in Switzerland using only data collected from the camera located in the locomotive. The calculated path is compared with the GNSS/IMU combination (red). Small deviations are observed in both the integrated distance (the image-to-real world distance scale shall be refined) and in behavior in the curve (optical flow to be refined). The measured path does not rely on global reference, meaning that a drift of the measured position is expected and should increase with the traveled distance. The corrected 3D path (yellow) by using railway point frogs (green triangles) as global references reduces the deviations to a precision below 0.25m and the accuracy below 1m (see [2] for details).

In ETCS a 1D coordinate system (traveled distance along the track) is used instead of a world global 3D coordinate system (e.g. longitude, latitude, altitude). Applying a 1D coordinate system to the visual odometry results in reduced deviations to the track topography (GTG) and GNSS/IMU combination. [2] shows that the drift is below 1% for all eight test runs between Ostermundigen and Thun (route length ∼26 km). Figure 8 shows the result when comparing to GPS/IMU for an exemplary test run on this route.

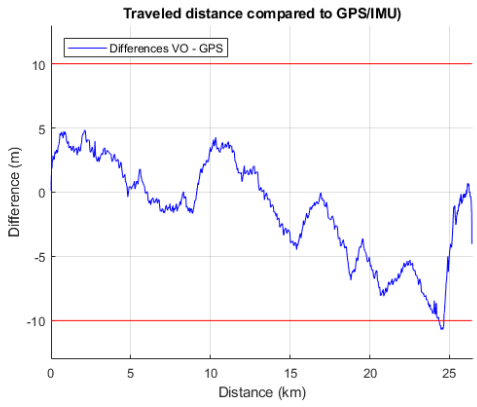

Figure 8: Comparison of the traveled distance between visual odometry and GPS.

In [2] the calculated distance is also compared to the distance between balises. The balises can be identified in the acquired images. The first balise of each balise group is taken for the calculation of the distances. Figure 9 shows the nom-

inal distance (black) from the track topography, the distance calculated by using the GNSS/IMU combined measurement (red) and the distance calculated by Visual Odometry (blue). The drift between the nominal values and the values measured by Visual Odometry are below 0.7%.

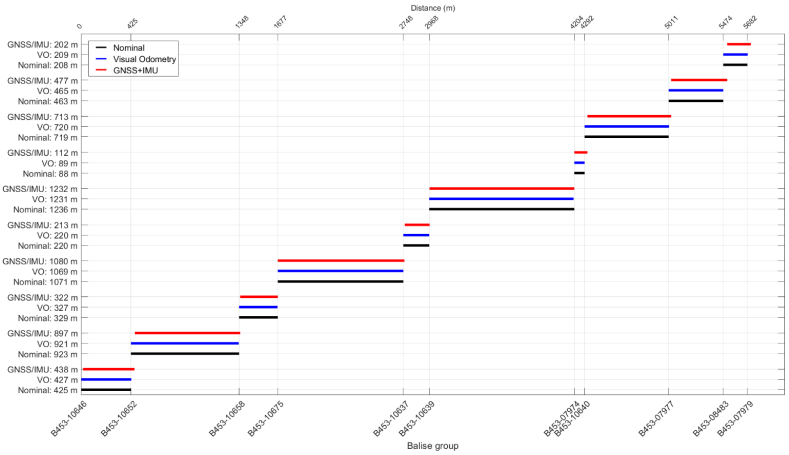

Figure 9: Comparison of the traveled distance for consecutive balises[2].

The longterm goal in railways is to remove costly infrastructure elements in the track (e.g. balises, axle counters). Therefore, balises can only be used as global references for a migration scenario and for proving that the new localization is at least as good as the current approved localization with balises and wheel odometry. The TPRs of the current system can be compared with ”virtual” TPRs from the new system generated in the background before using them. Railway points frogs can also be used as global references for visual odometry. The identification of the frogs in the acquired images is a side product of the track/rail identification needed for visual odometry. As can be seen in Figure 10 the frog is the intersection of the inner rails of adjacent tracks.

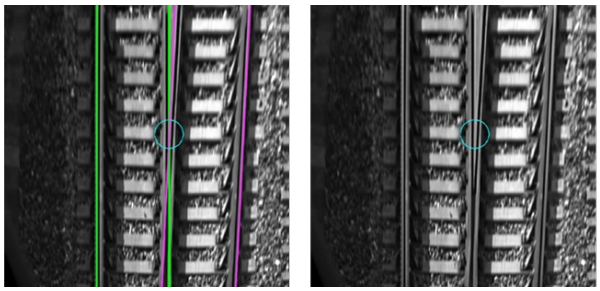

Figure 10: Identification of the frog

Figure 11 shows the difference between the traveled distance between consecutive points frogs from the track topography (GTG) and visual odometry. It can be seen that for most of the measurements, the drift is smaller than 1%. However there a still a number of measurements with higher drifts. The determination of the intersection between the rails is sensitive to uncertainties in the rail identification and therefore the identification has to be improved (e.g. higher pixel resolution of track width) in order to be more robust.

The mobile camera setup cannot use an external light source e.g. IR illuminator that has to be mounted outside the locomotive. Therefore a computation of the visual odometry within in tunnels (or at night) is not possible with this setup. However tunnel entries/exits could be used as global references. Figure 12 shows a measurement run in Austria between Vienna Heiligenstadt

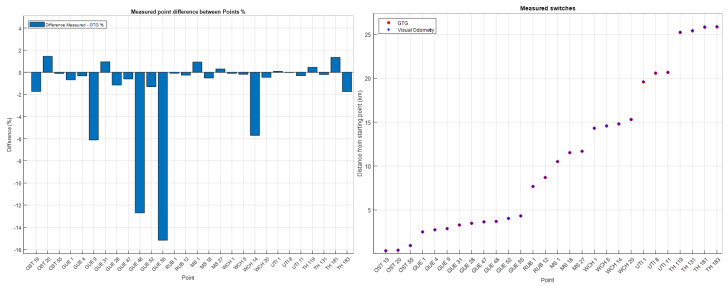

Figure 11: Traveled distance between consecutive points frogs

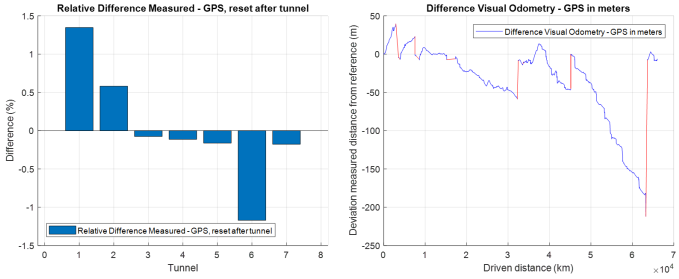

Figure 12: Traveled distance between tunnels - reset after tunnel (red)

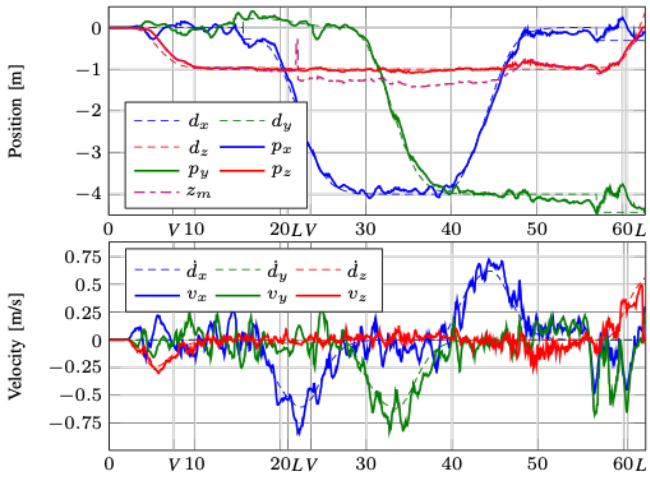

Figure 13: Accuracy of the pose estimation with drift compensation.

and with 6 tunnels. The tunnel exits were used a global references. The first results show that the drift of visual odometry compared to GPS/IMU for this ∼6.5 km route is always below 1.5%.

In Figure 13, we tested our framework on a shorter distance with a cheap IMU as the navigation unit. The camera was updating the drift estimate with every 4s. This also allows for a better compensation of the train navigation.

4. Conclusions

Visual odometry together with global references can be used for localization of trains. In order to reach the necessary SIL4 for train localization, a combination with other sensors (e.g. GNSS/IMU, FOS, map) is still required. Currently, the system is one of several sensors in the architecture of the Greenlight project of OeBB. In this project it will also be used to synchronize

the different sensors and rail infrastructure (e.g. balise, axle counters) with GPS time in order to create virtual TPRs that could be compared with the real TPRs. With the measured drift compared to track topography it could be determined the maximum traveled distance before a global reference is required. Thus, analyzing the railway network with regard to points and tunnels will show where additional global references will be required. Improvements in track width, rail and frog detection will lead to even better results in the future.

References

[1] SBB AG. smartrail 4.0. http://www.smartrail40.ch/index.asp.

[2] SBB AG and M2C ExpertControl GmbH. Technology Report PoC GLAT Supporting Technologies Video/FOS. https://www.smartrail40.ch/service/download.asp?mem=0&path=\download\downloads\TechnologyReport-PoC_GLAT_Video-FOS_v01-01.pdf.

[3] SBB AG, M2C ExpertControl GmbH, Schild & Partner, and TU Braunschweig IVA. Zwischenbericht Technologie PoC Lokalisierung. https://www.smartrail40.ch/service/download.asp?mem=0&path=\download\downloads\Integrierter_ZwischenberichtTechPocGLAT_v1.2_web.pdf.

[4] G. J. Bierman and C. L. Thornton. Numerical comparison of discrete Kalman filter algorithms - Orbit determination case study. In *15th Conference on Decision and Control and Symposium on Adaptive Processes*, pages 859–872, Jan 1976.

[5] Elmar Mair, Michael Suppa, and Darius Burschka. Error Propagation in Monocular Navigation for Zinf Compared to Eightpoint Algorithm. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'13)*, November 2013.

[6] Agostino Martinelli. Vision and IMU data fusion: Closed-form solutions for attitude, speed, absolute scale, and bias determination. *IEEE Trans. Robotics*, 28(1):44–60, 2012.

[7] Anastasios I. Mourikis and Stergios I. Roumeliotis. A multi-state constraint kalman filter for vision-aided inertial navigation. *Proceedings 2007 IEEE International Conference on Robotics and Automation*, pages 3565–3572, 2007.

[8] Christian Robl, Eckehard Schnieder, Raoul Schild, and Uwe Becker. Machbarkeitsstudie Lokalisierung smartrail 4.0. https://www.smartrail40.ch/service/download.asp?mem=0&path=\download\downloads\SR40_Machbarkeitsstudie_Lokalisierung.pdf.

[9] Christian Sagmeister, Manfred Staettner, and Christian Robl. Safe, Precise and Reliable Localization of Track Bound Railway Objects. In *ZEVrail: 45. Tagung Moderne Schienenfahrzeuge, Graz*, 04 2019.

[10] Eckehard Schnieder and Alex Brand. Sicherheit und Zulassung satellitengestützter Ortung im Schienenverkehr. In *safe.tech, München*, 04 2018.

[11] Eckehard Schnieder, Raoul Schild, Uwe Becker, Alex Brand, and Thomas Freissler. Safe and Precise localisation of Railway Objects GNSS Multisensor based Architecture for highly accurate and safe Object Localisation in Railways. In *Railways, Barcelona*, 09 2018.