Fusion of Optical Flow and Inertial Measurements for Robust Egomotion Estimation

Michael Bloesch, Sammy Omari, Péter Fankhauser, Hannes Sommer, Christian Gehring,
Jemin Hwangbo, Mark A. Hoepflinger, Marco Hutter, Roland Siegwart
Autonomous Systems Lab, ETH Zürich, Switzerland, bloeschm@ethz.ch

***Abstract*— In this paper we present a method for fusing optical flow and inertial measurements. To this end, we derive a novel visual error term which is better suited than the standard continuous epipolar constraint for extracting the information contained in the optical flow measurements. By means of an unscented Kalman filter (UKF), this information is then *tightly* coupled with inertial measurements in order to estimate the egomotion of the sensor setup. The individual visual landmark positions are not part of the filter state anymore. Thus, the dimensionality of the state space is significantly reduced, allowing for a fast online implementation. A nonlinear observability analysis is provided and supports the proposed method from a theoretical side. The filter is evaluated on real data together with ground truth from a motion capture system.**

I. INTRODUCTION

The use of cameras as light-weight egomotion sensors has been studied very broadly in the past few decades. The main advantage of a camera is that rich information can be obtained at relatively low power consumption. However, this information richness also poses the main difficulty, as the vast amount of information needs to be handled properly before the egomotion can be inferred.

Within the computer vision community, Davison [3] presented one of the first algorithms that is able to accurately track the 3D pose of a monocular camera. His idea was to design an Extended Kalman Filter (EKF) which simultaneously tracks the pose of the camera as well as the 3D position of points of interest, whereby the reprojection errors of the perceived features serve as innovation term. In the following, different authors presented adaptations in order to tackle different weaknesses of this approach, such as feature initialization [15] and limited map size [2].

Compared to the above mentioned *non-delayed* approaches, *delayed* methods also take past robot poses and measurements into account. The delayed approaches have become popular with the work of Klein and Murray [11]: Based on a subset of camera frames (keyframes) a bundle adjustment algorithm [20] optimizes a map, while the actual pose of the camera is tracked by minimizing the reprojection error between map and camera. Strasdat et al. [19] argued that in terms of accuracy and computational costs it would be more beneficial to increase the number of tracked features rather than the number of frames they are tracked in. In the following, the limits of vision-only state estimation and mapping where pushed even further by various other elaborate delayed frameworks [13], [18], [9].

In parallel to the "vision-only" based approaches, other researchers started including inertial measurements into their estimation algorithms. Relying on a known visual pattern, Mirzaei and Roumeliotis [14] showed one of the first online methods for extrinsic IMU-camera calibration and IMU bias estimation. Later, Kelly and Sukhatme [10], Jones and Soatto [7], as well as Weiss et al. [21] presented different frameworks for visual-inertial navigation including the co-estimation of calibration parameters. All of these authors emphasize the importance of analyzing the observability characteristics of the underlying system and discuss the related issues. Recently, Leutenegger et al. [12] presented a delayed framework in which the authors included visual and inertial error terms into a nonlinear optimization in order to estimate the motion of a visual-inertial multi-camera system as well as the landmarks in the map.

Efforts have also been done in order to find other visual error terms for combining the image information with inertial measurements. For example, Diel et al. [4] directly used the epipolar constraint between two matching features in subsequent frames as innovation term for their Kalman filter and thereby fused the visual information with the accelerometer measurements (the gyroscopes and attitude are handled separately). By making the assumption that all features lie on a single plane, Omari et al. [17] derived a visual error term for optical flow measurements and combined it with inertial measurements by means of an UKF. Both approaches have in common that the 3D position of the features are not included into the state of the filter which significantly reduces the computational costs. Similarly, Mourikis and Roumeliotis [16] also excluded the position of the features from the states of their filter and introduced a measurement model in order to account for the information when a feature is measured in multiple camera frames.

The primary goal of the present work is to propose a simple and reliable framework for the estimation of quantities which are critical for the safe operation of autonomous robots. We want to emphasize that we do not focus on achieving high-precision position and attitude accuracy, rather, our goal is to achieve a robust estimation of the velocity and inclination angle of the robot. This is especially important for systems which are controlled through dynamic motion,

This research was supported in part by the Swiss National Science Foundation (SNF) through project 200021_149427/1 and the National Centre of Competence in Research Robotics.

such as legged robots or quadrocopters. For this reason, we introduce visual error term which can directly extract information from a single feature match and does not rely on repeated measurements of the same feature. The above mentioned work of Diel et al. [4] is the closest to the present approach. In contrast to it, we propose the use of a different visual error term and co-estimate the inverse scene depth. By means of an UKF, we carry out a *tight* fusion of the visual and inertial measurements, whereby gyroscope and accelerometer measurements are included during the prediction step and the visual error terms serve as innovation during the update step. Furthermore, avoiding the inclusion of the feature positions into the filter state allows for a very fast online implementation of the method. The presented approach is supported by a full nonlinear observability analysis and evaluated on data from real experiments.

The remainder of this paper is structured as follows: After introducing the most important notations and conventions in section II, we describe the structure of the filter including the prediction and update steps in section III. In section IV we show and discuss the result of the nonlinear observability analysis. The experimental setup is described in section V. Finally, we discuss the obtained results in section VI and conclude with section VII.

II. Prerequisites

For better readability we give a short overview on the employed notations and conventions. The coordinates, expressed in a frame A, of a vector from a point P to a point Q are denoted by ${}_A\boldsymbol{r}_{PQ}$. If B is a second coordinate frame, then $\boldsymbol{C}_{BA}$ maps the coordinates expressed in A to the corresponding coordinates in B. The rotation between both frames is generally parametrized by the unit quaternion $\boldsymbol{q}_{BA}$, with the corresponding mapping $\boldsymbol{C} : \boldsymbol{q}_{BA} \mapsto \boldsymbol{C}_{BA}$. Throughout the paper, we add a subscript k to a quantity v, if we want to talk about its value at a time t_k, i.e., $v_k = v(t_k)$. Two coordinate frames are of interest: the world fixed coordinate frame W and the sensor frame B. For the sake of simplicity the following derivation assumes that the camera and the IMU coordinate frames are aligned with B.

We handle rotations as elements of $SO(3)$, where, together with the exponential and logarithm map, difference and derivatives are defined on $\mathbb{R}^3$. This is of high importance for the setup of the filter as well as for the corresponding observability analysis. Please note, that for this reason, also derivatives containing quaternions will be three dimensional in the corresponding directions, e.g. $\dot{\boldsymbol{q}} = -\boldsymbol{\omega} \in \mathbb{R}^3$ [1].

III. Filter Setup

A. Optical Flow and Visual Error Term

Based on the assumption of a static scene the following identity can be directly derived using kinematics relations only:

$$0 = {}_B\boldsymbol{v}_B + ({}_B\boldsymbol{w}_B^{\times}\boldsymbol{m}_i + \boldsymbol{u}_i)\lambda_i + \boldsymbol{m}_i\dot{\lambda}_i, \tag{1}$$

where ${}_B\boldsymbol{v}_B$ and ${}_B\boldsymbol{w}_B$ are the robot-centric velocity and rotational rate. The quantities $\boldsymbol{m}_i$, $\boldsymbol{u}_i$ and λ_i are related to the optical flow of a static feature i and represent the unit length bearing vector, the optical flow vector, and the depth of the feature. The challenge here is to find a way to properly extract information out of the equation without having to co-estimate the depth (and it's derivative) for each single optical flow measurement. A very common approach is to employ the continuous epipolar constraint which results from the above equation if left-multiplied by $\boldsymbol{m}_i^T({}_B\boldsymbol{w}_B^{\times}\boldsymbol{m}_i + \boldsymbol{u}_i)^{\times}$:

$$0 = \boldsymbol{m}_i^T({}_B\boldsymbol{w}_B^{\times}\boldsymbol{m}_i + \boldsymbol{u}_i)^{\times}{}_B\boldsymbol{v}_B. \tag{2}$$

This corresponds to an analytical elimination of the depth and its derivative. The problem is that this reduction does not consider the stochastic nature of the system and draws the estimation process towards singularities, e.g. zero velocity, which don't correspond to the maximum likelihood estimate (which is in general a desirable goal for estimation). As a trade-off we propose to eliminate the derivative of the depth analytically by left-multiplying the equation by a 2×3 matrix $\boldsymbol{M}_i$ which fulfills:

$$\boldsymbol{M}_i\boldsymbol{m}_i = 0 \quad \wedge \quad \boldsymbol{M}_i\boldsymbol{M}_i^T = \boldsymbol{I}_2. \tag{3}$$

Additionally we make use of an inverse-depth parametrization, $\alpha_i = 1/\lambda_i$, and obtain

$$0 = \boldsymbol{M}_i\left({}_B\boldsymbol{v}_B\alpha_i + ({}_B\boldsymbol{w}_B^{\times}\boldsymbol{m}_i + \boldsymbol{u}_i)\right). \tag{4}$$

In comparison to the continuous epipolar constraint, this term retains more of the original constraint and is less susceptible to singularities. However, it also still contains one additional unknown, α_i, per visual feature. In order to cope with this, we will assume that the inverse depths α_i exhibit a Gaussian distribution around a mean α with standard deviation σ_α. The new parameter α corresponds to the inverse scene depth and will be co-estimated in the estimation process.

B. Filter States and Prediction Equations

The states of a filter have to be selected such that appropriate prediction and measurement equation can be derived. We define the following filter states:

$$\boldsymbol{x} := \left(\boldsymbol{r}, \boldsymbol{v}, \boldsymbol{q}, \boldsymbol{c}, \boldsymbol{d}, \alpha\right), \tag{5}$$

$$:= \left({}_W\boldsymbol{r}_{WB}, {}_B\boldsymbol{v}_B, \boldsymbol{q}_{WB}, {}_B\boldsymbol{b}_f, {}_B\boldsymbol{b}_\omega, \alpha\right), \tag{6}$$

where $\boldsymbol{r}$ is the world position of the sensor, $\boldsymbol{v}$ represents its robot-centric velocity, $\boldsymbol{q}$ parametrizes the rotation between the sensor and the world coordinate frame, and $\boldsymbol{c}$ and $\boldsymbol{d}$ are the biases of the accelerometer and gyroscope. The additional state α is the inverse scene depth which is used for incorporating the optical flow measurements. The advantage of the robot-centric choice of states is that we thereby partition the state into non-observable states (absolute position and yaw) and observable states and thus avoid numerical problems related to non-observable states. A small drawback is that the noise of the gyroscope propagates onto the velocity state as well. Since, as will be shown later, the robot-centric velocity is fully observable, the additional noise can be compensated by the filter.

Analogous to other fusion algorithms including inertial measurements, we embed the proper acceleration measurement $\tilde{\boldsymbol{f}}$ and the rotational rate measurement $\tilde{\boldsymbol{\omega}}$ of the IMU directly into the prediction step of the proposed filter. Assuming that both measurements are affected by white Gaussian noise, $\boldsymbol{w}_f$ and $\boldsymbol{w}_\omega$, and additive bias terms, $\boldsymbol{c}$ and $\boldsymbol{d}$, we can write down

$$\tilde{\boldsymbol{f}} = \boldsymbol{f} + \boldsymbol{c} + \boldsymbol{w}_f, \tag{7}$$

$$\tilde{\boldsymbol{\omega}} = \boldsymbol{\omega} + \boldsymbol{d} + \boldsymbol{w}_\omega. \tag{8}$$

Both quantities are related to the kinematics of the sensor by

$$\boldsymbol{f} = \boldsymbol{C}(\boldsymbol{q}_{BW})\left({}_W\dot{\boldsymbol{v}}_B - \boldsymbol{g}\right), \tag{9}$$

$$\boldsymbol{\omega} = -\dot{\boldsymbol{q}}_{BW}, \tag{10}$$

where $\boldsymbol{g}$ is the gravity vector in W. By evaluating the total derivative of the filter states and combining it with the inertial measurements we obtain the following continuous time differential equations:

$$\dot{\boldsymbol{r}} = \boldsymbol{C}(\boldsymbol{q})\boldsymbol{v} + \boldsymbol{w}_r, \tag{11}$$

$$\dot{\boldsymbol{v}} = -\left(\tilde{\boldsymbol{\omega}} - \boldsymbol{d} - \boldsymbol{w}_\omega\right)^\times \boldsymbol{v} + \tilde{\boldsymbol{f}} - \boldsymbol{c} - \boldsymbol{w}_f + \boldsymbol{C}^T(\boldsymbol{q})\boldsymbol{g}, \tag{12}$$

$$\dot{\boldsymbol{q}} = \boldsymbol{C}(\boldsymbol{q})(\tilde{\boldsymbol{\omega}} - \boldsymbol{d} - \boldsymbol{w}_\omega), \tag{13}$$

$$\dot{\boldsymbol{c}} = \boldsymbol{w}_c, \tag{14}$$

$$\dot{\boldsymbol{d}} = \boldsymbol{w}_d, \tag{15}$$

$$\dot{\alpha} = w_\alpha. \tag{16}$$

The additional continuous white Gaussian noise processes $\boldsymbol{w}_c$ and $\boldsymbol{w}_d$ model a certain drift affecting the bias terms. w_α is included in order to handle varying inverse scene depths and $\boldsymbol{w}_r$ is included for being able to excite the full filter state and for modeling errors caused by the subsequent discretization of the states. For all white Gaussian noise processes, the corresponding covariance parameters, $\boldsymbol{R}_r$, $\boldsymbol{R}_f$, $\boldsymbol{R}_\omega$, $\boldsymbol{R}_c$, $\boldsymbol{R}_d$, and $\boldsymbol{R}_\alpha$ describe the magnitude of the noise. Except for $\boldsymbol{R}_r$ and $\boldsymbol{R}_\alpha$ which are tuning parameters, all covariance parameters can be identified by considering the Allan plots of the IMU measurements [5].

The discretization is based on a simple Euler forward integration scheme. Please note that for the rotational states, the step forward can be taken on the corresponding sigma algebra and then be mapped back onto $SO(3)$. This corresponds to (with $\Delta t_k = t_k - t_{k-1}$):

$$\boldsymbol{q}(t_k) = \exp\left(\Delta t_k \dot{\boldsymbol{q}}(t_{k-1})\right) \otimes \boldsymbol{q}(t_{k-1}). \tag{17}$$

This leads to:

$$\boldsymbol{r}_k = \boldsymbol{r}_{k-1} + \Delta t_k \left(\boldsymbol{C}_{k-1}\boldsymbol{v}_{k-1} + \boldsymbol{w}_{r,k}\right), \tag{18}$$

$$\boldsymbol{v}_k = \left(I - \Delta t_k \left(\tilde{\boldsymbol{\omega}}_k - \boldsymbol{d}_{k-1} - \boldsymbol{w}_{\omega,k}\right)^\times\right)\boldsymbol{v}_{k-1} + \Delta t_k \left(\tilde{\boldsymbol{f}}_k - \boldsymbol{c}_{k-1} - \boldsymbol{w}_{f,k} + \boldsymbol{C}_{k-1}^T \boldsymbol{g}\right), \tag{19}$$

$$\boldsymbol{q}_k = \exp\left(\Delta t_k \boldsymbol{C}_{k-1}(\tilde{\boldsymbol{\omega}}_k - \boldsymbol{d}_{k-1} - \boldsymbol{w}_{\omega,k})\right) \otimes \boldsymbol{q}_{k-1}, \tag{20}$$

$$\boldsymbol{c}_k = \boldsymbol{c}_{k-1} + \Delta t_k \boldsymbol{w}_{c,k}, \tag{21}$$

$$\boldsymbol{d}_k = \boldsymbol{d}_{k-1} + \Delta t_k \boldsymbol{w}_{d,k}, \tag{22}$$

$$\alpha_k = \alpha_{k-1} + \Delta t_k w_{\alpha,k}. \tag{23}$$

C. Measurement Equations

The measurement equations are directly based on the findings of section III-A. For each available optical flow measurement i, we directly define the corresponding 2D innovation term for the filter:

$$\boldsymbol{y}_i = \boldsymbol{M}_i \left(\boldsymbol{v}\ \alpha_i + \left(\boldsymbol{\omega}^\times \boldsymbol{m}_i + \boldsymbol{u}_i\right)\right). \tag{24}$$

As discussed above, we introduced the inverse scene depth as a filter state and thus model deviations of the single inverse depths α_i as measurement noise:

$$\alpha_i = \alpha + n_{\alpha,i}, \quad n_{\alpha,i} \sim \mathcal{N}(0, \sigma_\alpha^2). \tag{25}$$

Furthermore, we also have to model noise on the bearing vectors $\boldsymbol{m}_i$ and optical flow vectors $\boldsymbol{u}_i$. For typical scenarios the major part of the uncertainties originate through $\boldsymbol{u}_i$, which lies in the orthogonal subspace of $\boldsymbol{m}_i$. Thus, we can introduce an additive lumped noise term on $\boldsymbol{u}_i$, whereby it is sufficient to excite directions orthogonal to $\boldsymbol{m}_i$ only. This can be achieved by means of the previously defined matrix $\boldsymbol{M}_i$ ($\boldsymbol{n}_u$ is two dimensional):

$$\tilde{\boldsymbol{u}}_i = \boldsymbol{u}_i - \boldsymbol{M}_i^T \boldsymbol{n}_u, \tag{26}$$

$$\boldsymbol{n}_u \sim (0, \boldsymbol{R}_u). \tag{27}$$

With this the innovation term becomes:

$$\boldsymbol{y}_i = \boldsymbol{M}_i \left(\boldsymbol{v}(\alpha + n_{\alpha,i}) + \left(\boldsymbol{\omega}^\times \boldsymbol{m}_i + \tilde{\boldsymbol{u}}_i\right)\right) + \boldsymbol{n}_u. \tag{28}$$

The parameter $\boldsymbol{R}_u$ describes the accuracy of the visual measurements and the parameter σ_α^2 depends on the variance of the inverse depths in the scene.

An interesting effect is that whenever the velocity is small or when the inverse scene depth tends towards zero (i.e. the scene is far away), the innovation term will be equivalent to a visual gyroscope:

$$\boldsymbol{y}_i^* = \boldsymbol{M}_i \left(\left(\boldsymbol{\omega}^\times \boldsymbol{m}_i + \tilde{\boldsymbol{u}}_i\right)\right) + \boldsymbol{n}_u. \tag{29}$$

D. Unscented Kalman Filter and Outliers Detection

An unscented Kalman filter (UKF) is employed as filtering framework. The main reason for this is that the UKF can handle correlated noise between prediction and update by using a single set of augmented sigma points for both steps. All equations required for its implementation are the prediction equation (18)-(23) and the update equation (29), whereby the single innovation terms of the multiple features are stuck together. The twofold use of the gyroscope measurement can be directly seen in these equations. Please note that the implementation has to take into account that, although the attitude is parametrized by a unit quaternion, the corresponding noise and perturbations are always on a 3D subspace. For a detailed discussion on the employed UKF itself please refer to [8].

In order to handle the high sensitivity of Kalman filters to outliers, we implement a simple outliers detection method on the innovation terms. Using an analogous approach as Mirzaei et al. [14], we reject a visual measurement whenever the Mahalanobis distance of the corresponding innovation

terms exceeds a certain threshold. The predicted covariance of the innovation is used as weighting for the Mahalanobis distance and the threshold is chosen in such a manner that, in theory, 1% of the inliers are rejected. Considering that the underlying probability distribution is a χ^2-distribution with two degrees of freedom the threshold is set to $p = 9.21$. In summary, the criteria for rejecting a measurement i is given by (where $\boldsymbol{S}_i$ is the predicted covariance matrix):

$$\boldsymbol{y}_i^T \boldsymbol{S}_i^{-1} \boldsymbol{y}_i > p. \tag{30}$$

IV. Observability Analysis

A nonlinear observability analysis is carried out for the proposed system. A detailed discussion of the theory behind it was provided by Hermann and Krener [6]. In the scope of this paper we only outline the rough procedure of the analysis. Based on the nonlinear representation of the system an observability matrix is derived in order to assess the observability characteristics of the system. The system can be written as follows, whereby the noise quantities can be ignored since they don't affect the observability analysis:

$$\dot{\boldsymbol{x}} = \begin{pmatrix} \boldsymbol{C}\boldsymbol{v} \\ \hat{\boldsymbol{\omega}}^\times \boldsymbol{v} - \hat{\boldsymbol{f}} + \boldsymbol{C}^T \boldsymbol{g} \\ -\boldsymbol{C}\hat{\boldsymbol{\omega}} \\ 0 \\ 0 \\ 0 \end{pmatrix}, \tag{31}$$

$$\boldsymbol{h}_i(\boldsymbol{x}) = \boldsymbol{M}_i \left(\boldsymbol{v} \ \alpha + (-\hat{\boldsymbol{w}}^\times \boldsymbol{m}_i + \tilde{\boldsymbol{u}}_i) \right), \tag{32}$$

with the shortcuts $\hat{\boldsymbol{f}} = -\tilde{\boldsymbol{f}} + \boldsymbol{c}$ and $\hat{\boldsymbol{\omega}} = -\tilde{\boldsymbol{\omega}} + \boldsymbol{d}$.

The observability matrix is composed of the gradient of the Lie derivatives of the above system, whereby $\tilde{\boldsymbol{f}}$ and $\tilde{\boldsymbol{\omega}}$ are, in the context of this analysis, the inputs to the system. We can show, that if there are three optical measurements with non-coplanar bearing vectors and if the inverse scene depth is not zero we can simplify the observability matrix to the following term (if $\alpha = 0$ only the gyroscope bias and the inverse scene depth itself (if $\boldsymbol{v} \neq 0$) are observable):

$$\boldsymbol{O} = \begin{bmatrix} 0 & \boldsymbol{I} & 0 & 0 & 0 & \frac{1}{\alpha}\boldsymbol{v} \\ 0 & 0 & 0 & 0 & \boldsymbol{I} & 0 \\ 0 & 0 & \boldsymbol{C}^T\boldsymbol{g}^\times & -\boldsymbol{I} & 0 & \boldsymbol{C}^T\boldsymbol{g} - \hat{\boldsymbol{f}} \\ 0 & 0 & \hat{\boldsymbol{\omega}}^\times \boldsymbol{C}^T \boldsymbol{g}^\times & 0 & 0 & \hat{\boldsymbol{\omega}}^\times \boldsymbol{C}^T \boldsymbol{g} \end{bmatrix}. \tag{33}$$

Throughout the analysis only rank-preserving row operations are carried out which keeps the relation between each column and a specific state of the filter. We also have to keep in mind, that $\tilde{\boldsymbol{f}}$ and $\tilde{\boldsymbol{\omega}}$ represent system inputs in this analysis, and thus a single line in the matrix can be duplicated by inserting different values for $\tilde{\boldsymbol{f}}$ and $\tilde{\boldsymbol{\omega}}$ (see [6]). By inserting two non-colinear values for $\tilde{\boldsymbol{\omega}}$ (through $\hat{\boldsymbol{\omega}}$) in the last row of the matrix we can further simplify the matrix to:

$$\boldsymbol{O} = \begin{bmatrix} 0 & \boldsymbol{I} & 0 & 0 & 0 & \frac{1}{\alpha}\boldsymbol{v} \\ 0 & 0 & 0 & 0 & \boldsymbol{I} & 0 \\ 0 & 0 & 0 & -\boldsymbol{I} & 0 & -\hat{\boldsymbol{f}} \\ 0 & 0 & \boldsymbol{C}^T\boldsymbol{g}^\times & 0 & 0 & \boldsymbol{C}^T\boldsymbol{g} \end{bmatrix}. \tag{34}$$

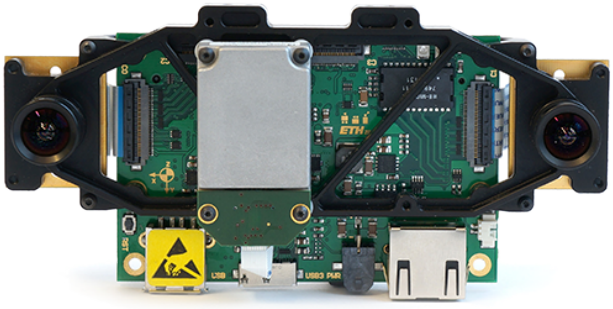

Fig. 1. ASL visual-inertial SLAM sensor employed for evaluating the presented optical flow and inertial measurement fusion approach.

The rank of this matrix is 12 (independent of the choice of $\boldsymbol{C}$, $\boldsymbol{v}$, or $\hat{\boldsymbol{f}}$) and the dimension of the right null-space is consequently 4, which is spanned by the following matrix:

$$\boldsymbol{N} = \begin{bmatrix} \boldsymbol{I} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \boldsymbol{g}^T & 0 & 0 & 0 \end{bmatrix}^T. \tag{35}$$

In an informal way, the perturbations along the directions spanned by $\boldsymbol{N}$ cannot be perceived at the filter output. While the first column corresponds to the absolute position of the system, the second column represents a rotation around the gravity axis, i.e., global position and yaw angle are not observable. Mathematically this can be written as:

$$\boldsymbol{r}^* = \boldsymbol{r} + \delta\boldsymbol{r}, \tag{36}$$

$$\boldsymbol{q}^* = \exp(\boldsymbol{g}\delta\psi) \otimes \boldsymbol{q}, \tag{37}$$

where $\delta\boldsymbol{r}$ and $\delta\psi$ are perturbations. $\boldsymbol{r}$ and $\boldsymbol{q}$ cannot be distinguished from $\boldsymbol{r}^*$ and $\boldsymbol{q}^*$, respectively.

All in all, the above nonlinear observability analysis allows us to state that for all points in the state-space (except if $\alpha = 0$) there *exists* some inputs $\tilde{\boldsymbol{f}}$ and $\tilde{\boldsymbol{\omega}}$ (corresponding to a certain motion of the sensor) such that all states are locally weakly observable, except for the global position and yaw angle.

V. Experimental Setup

To validate the proposed scheme, the Unscented Kalman filter was implemented in C++. The filter was tested on data that were recorded using the ASL visual-inertial SLAM sensor (see fig. 1), with synchronized global-shutter camera (Aptina MT9V034 at 20 Hz) and IMU (Analog Devices ADIS16488 at 200 Hz). The pose of the sensor was additionally tracked using a Vicon motion tracking system at 100 Hz.

The image features are tracked using a Lukas-Kanade-based tracker. Salient image features that are used for tracking are extracted by first applying a FAST corner detector, computing the Shi-Tomasi score for each extracted corner and then selecting those corners which have the highest score while ensuring a uniform distribution of the features in the image. A uniform feature distribution is ensured by masking parts of the images that are already populated with

	Attitude (rad)			Velocity (m/s)		
	Roll	Pitch	Yaw	X	Y	Z
Prop.	0.012	0.005	0.464	0.057	0.070	0.087
Epi.	0.020	0.008	0.437	0.162	0.121	0.200

TABLE I

RMS VALUES OF PROPOSED FILTER VS. FILTER WITH STANDARD CONTINUOUS EPIPOLAR CONSTRAINT.

strong features and by only adding new, weaker features in unpopulated image regions.

Feature extraction and LK-tracking for 150 features is taking less than 2.5 ms in total on a single core of an Intel i7-3740QM processor for one frame. Equivalently, a measurement update step using 50 optical flow features is performed in 10 ms. During the experiments an average feature count of 50 features was used. The rather bad scalability of the filter update can be easily overcome by changing to the information form of the filter, which will be part of future work.

VI. Results and Discussion

The presented approach was evaluated on different datasets from an indoor environment where the feature depths range between 0.5 m and 5 m. The motion of the sensor included rotational rate of up to 3 rad/s. Our main goal was to develop a filter for delivering high-rate and reliable state estimates rather than being mainly focused on estimation accuracy. Furthermore, the main states of interest are the velocities and the inclination angles since they are of major importance if it comes to control of dynamic robot motions. Using a 2 minute long dataset where the sensor was excited along its different degrees of freedom, the RMS values depicted in table I were computed. In order to evaluate the proposed visual error term, we implemented the same filter setup with the standard epipolar constraint (equation (2)) and observed an increase of the RMS values by a factor 2.

The estimated IMU biases converge relatively fast depending on the motion of the system. While we have no ground truth values for the bias terms, figure 2 shows the typical convergence of the biases when the system is being excited along its different directions. Figuring out which direction needs to be excited for improving the estimation of a certain state can be a very difficult problem and is not within the scope of this paper. The 3σ-bounds of the covariance matrix are plotted as dashed lines.

Figure 3, 4, and 5 present the results from a dataset where after some initial motion the sensor holds still for awhile before being moved again. This can be clearly seen between $33-43$ seconds. In contrast to the standard epipolar constraint, the employed visual error term still extracts information from the optical flow measurements analogous to a visual gyroscope. Still, during this phase additional uncertainty accumulates in the different states. However, as soon as the sensor is moved again, the observable states very quickly converge back to the reference. This can be nicely observed for the velocity estimates. Note as well, that although the position of the sensor is unobservable, it can be corrected and loose uncertainty to some extent

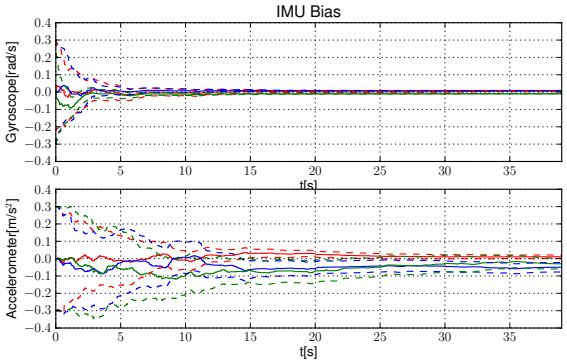

Fig. 2. Estimated IMU biases. Red: x-, blue: y-, green: z-coordinate. Dashed lines: 3σ-bounds. The initial converges is supported by motion of the sensor. The estimate of the accelerometer bias is more accurate along the x-axis because it is more often aligned with the gravity axis. The gyroscope biases converge faster since the optical flow measurement have a direct impact on the angular rates.

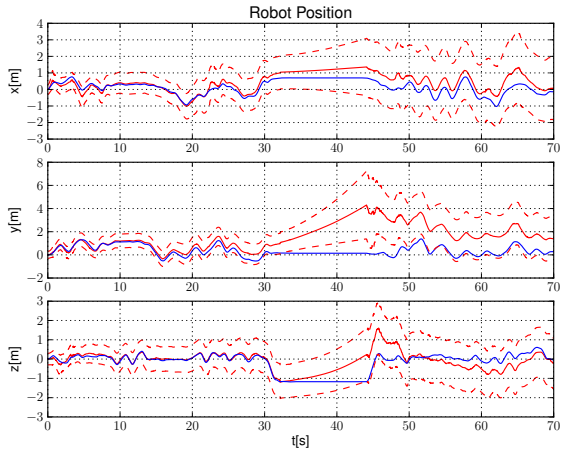

Fig. 3. Estimated sensor position. Red: estimated values. Red dashed line: 3σ-bound. Dashed blue line: motion capture ground truth. The position state is affected by increasing uncertainty since it is not observable and represents the integration of the velocity estimate.

through the cross-correlation it maintains with the other states. Furthermore, the initial inclination error (roll and pitch) of about 20° can be corrected within 1-2 seconds.

All in all, the filter exhibits a rather average performance in terms of accuracy when compared with the state of the art visual-inertial algorithms. However, when considering that only frame to frame (20 Hz) information is included into the filter, the obtained results are relatively surprising, especially since other quantities like the IMU biases have to be co-estimated simultaneously. A major advantage of this approach is that the filter is free of any complex initialization procedure and only relies on single feature matches between subsequent frames. With this, it does not require long term tracking of features and is thus much less affected by fast motions.

VII. Conclusion and Future Work

In this paper we presented a relatively simple approach for fusing optical flow and inertial measurements. By deriving a

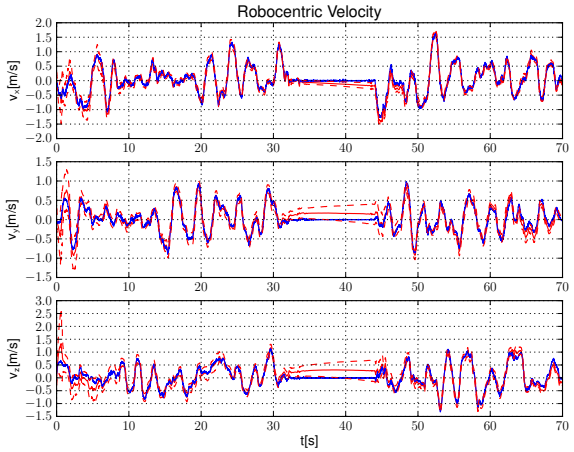

Fig. 4. Estimated sensor velocity expressed in the sensor coordinate frame itself. Red: estimated values. Red dashed line: 3σ-bound. Dashed blue line: motion capture ground truth. The robot-centric velocity is fully observable and consequently has a bounded uncertainty. Even after a phase of increased uncertainty it is able to recover if sufficient excitation is available.

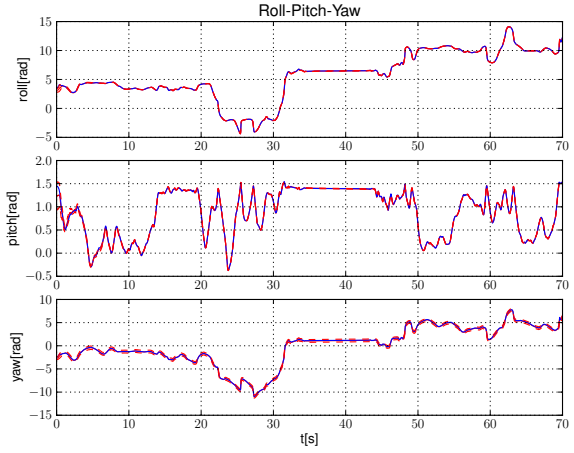

Fig. 5. Roll, pitch, and yaw angle of the sensor. Red: estimated values. Red dashed line: 3σ-bound. Dashed blue line: motion capture ground truth. Pitch and roll are observable and consequently exhibit a nice tracking behavior. Yaw is not observable and slowly drifts away.

special optical flow error term and embedding it into an UKF framework, we were able to derive a filter for estimating the egomotion of the sensor, the IMU biases as well as the inverse scene depth. By carrying out a nonlinear observability analysis we showed that all states except for the global position and yaw angle are locally weakly observable. The results obtained on a real dataset confirmed that the filter was able to estimate the different observable states.

One important aspect of future work will be the combination of the presented approach with other visual localization methods. While the strength of the presented approach lies in its robustness and speed, it could be combined together with some static feature tracking in order to improve its accuracy and long term stability. Other possible extensions include the implementation on multiple cameras or the combination with further sensor modalities.

REFERENCES

[1] M. Bloesch, C. Gehring, P. Fankhauser, M. Hutter, M. A. Hoepflinger, and R. Siegwart, "State Estimation for Legged Robots on Unstable and Slippery Terrain," in *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, 2013.

[2] L. Clemente, A. Davison, I. Reid, J. Neira, and J. Tardós, "Mapping Large Loops with a Single Hand-Held Camera," in *Proceedings of Robotics: Science and Systems*, Atlanta, GA, USA, June 2007.

[3] A. J. Davison, "Real-Time Simultaneous Localisation and Mapping with a Single Camera," in *IEEE Int. Conference on Computer Vision*, 2003, pp. 1403–1410 vol.2.

[4] D. D. Diel, P. DeBitetto, and S. Teller, "Epipolar Constraints for Vision-Aided Inertial Navigation," in *IEEE Workshops on Application of Computer Vision*, vol. 2, 2005, pp. 221–228.

[5] N. El-Sheimy, H. Hou, and X. Niu, "Analysis and Modeling of Inertial Sensors Using Allan Variance," *IEEE Trans. on Instrumentation and Measurement*, vol. 57, no. 1, pp. 140–149, 2008.

[6] R. Hermann and A. Krener, "Nonlinear controllability and observability," *IEEE Trans. on Automatic Control*, vol. 22, no. 5, pp. 728–740, Oct. 1977.

[7] E. S. Jones and S. Soatto, "Visual-inertial navigation, mapping and localization: A scalable real-time causal approach," *The International Journal of Robotics Research*, vol. 30, no. 4, pp. 407–430, 2011.

[8] S. J. Julier, "The scaled unscented transformation," in *Proc. of the American Control Conference*, vol. 6, May 2002, pp. 4555–4559.

[9] M. Kaess, H. Johannsson, R. Roberts, V. Ila, J. J. Leonard, and F. Dellaert, "iSAM2: Incremental smoothing and mapping using the Bayes tree," *The International Journal of Robotics Research*, vol. 31, no. 2, pp. 216–235, 2012.

[10] J. Kelly and G. S. Sukhatme, "Visual-Inertial Sensor Fusion: Localization, Mapping and Sensor-to-Sensor Self-calibration," *Int. Journal of Robotics Research*, vol. 30, no. 1, pp. 56–79, Nov. 2011.

[11] G. Klein and D. Murray, "Parallel Tracking and Mapping for Small AR Workspaces," in *IEEE and ACM Int. Symposium on Mixed and Augmented Reality*, 2007, pp. 225–234.

[12] S. Leutenegger, P. Furgale, V. Rabaud, M. Chli, K. Konolige, and R. Siegwart, "Keyframe-Based Visual-Inertial SLAM using Nonlinear Optimization," in *Proceedings of Robotics: Science and Systems*, Berlin, Germany, June 2013.

[13] C. Mei, G. Sibley, M. Cummins, P. Newman, and I. Reid, "RSLAM: A System for Large-Scale Mapping in Constant-Time Using Stereo," *Int. Journal of Computer Vision*, vol. 94, no. 2, pp. 198–214, 2011.

[14] F. M. Mirzaei and S. I. Roumeliotis, "A Kalman filter-based algorithm for IMU-camera calibration: Observability analysis and performance evaluation," *IEEE Trans. on Robotics*, vol. 24, no. 5, pp. 1143–1156, 2008.

[15] J. Montiel, J. Civera, and A. Davison, "Unified Inverse Depth Parametrization for Monocular SLAM," in *Proceedings of Robotics: Science and Systems*, Philadelphia, USA, Aug. 2006.

[16] A. I. Mourikis and S. I. Roumeliotis, "A Multi-State Constraint Kalman Filter for Vision-aided Inertial Navigation," in *IEEE Int. Conf. on Robotics and Automation*, 2007, pp. 3565–3572.

[17] S. Omari and G. Ducard, "Metric visual-inertial navigation system using single optical flow feature," in *European Control Conference*, July 2013, pp. 1310–1316.

[18] H. Strasdat, A. J. Davison, J. M. M. Montiel, and K. Konolige, "Double window optimisation for constant time visual SLAM," in *IEEE Int. Conference on Computer Vision*, 2011, pp. 2352–2359.

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

[20] B. Triggs, P. McLauchlan, R. Hartley, and A. Fitzgibbon, "Bundle Adjustment A Modern Synthesis," in *Vision Algorithms: Theory and Practice*, ser. Lecture Notes in Computer Science, B. Triggs, A. Zisserman, and R. Szeliski, Eds. Springer Berlin Heidelberg, 2000, vol. 1883, pp. 298–372.

[21] S. Weiss, M. Achtelik, S. Lynen, L. Kneip, M. Chli, and R. Siegwart, "Monocular Vision for Long-term Micro Aerial Vehicle State Estimation: A Compendium," *Journal of Field Robotics*, vol. 30, no. 5, pp. 803–831, 2013.