USING OPTICAL FLOW FOR FILLING THE GAPS IN VISUAL-INERTIAL TRACKING

Gabriele Bleser and Gustaf Hendeby

Department Augmented Vision, German Research Center for Artificial Intelligence
Trippstader Str. 122, D-67663 Kaiserslautern, Germany
{gabriele.bleser, gustaf.hendeby}@dfki.de

ABSTRACT

During the last decades egomotion tracking has been an often addressed problem. Hybrid approaches evidentially have potential to provide accurate, efficient and robust results. *Simultaneous localisation and mapping* (SLAM) — in contrast to model-based approaches — is used to enable tracking in unknown environments. However, it also suffers from high computational complexity. Moreover, in many applications, the map itself is not needed and the target environment is partially known, *e.g.* in a few 3D anchor points. In this paper, rather than using SLAM, optical flow measurements are introduced into a model-based system. With these measurements, a modified visual-inertial tracking method is derived, which in Monte Carlo simulations reduces the need for 3D points and thus allows tracking during extended gaps of 3D point registrations.

1. INTRODUCTION

The past few decades extensive research has been conducted in hybrid tracking. In particular visual-inertial tracking, *i.e.* fusing visual information with kinematic data from miniature MEMS[1] inertial sensors, has gained in importance. The usage and advantages are extensively treated in literature. Many different methods with promising results have been proposed [28, 17, 2, 5]. Corke et al [8] give an introduction to the field.

Another successful line of work is visual *simultaneous localisation and mapping* (SLAM). New algorithms are constantly developed in order to tackle computational complexity and drift. The approaches utilise, *e.g.*, local bundle adjustment [25], parallelisation [20] and graph-based techniques [11]. Since the formalisation and first solutions to the visual SLAM problem [31, 9], a variety of extensions to the approach has emerged, *e.g.*, FastSLAM [24, 10], MPF-SLAM [29], Mini-SLAM [1], Divide and Conquer SLAM [26, 7] and Treemap [12]. Visual-inertial SLAM, the combination of both methodologies above is treated in, *e.g.*, [27, 29, 4].

SLAM has had great success in many applications. However, it is also often the case — *e.g.* in augmented reality — that (1) 3D structure is partly known or can easily be marked in the scene and (2) the expensive estimation of a denser map has no value in itself besides enabling stable pose estimation. This paper investigates an alternative strategy: instead of mapping additional structure, optical flow measurements are used to improve model-based tracking. This approach is suitable for applications where very few 3D anchor points are available or can be surveyed precisely in the scene. To do this, the method makes use of the mathematics epipolar constraints [22, 15]. Epipolar constraints are well-known in computer vision. However, in most cases they are used to initialise a visual SLAM process [10, 3, 19, 27] or to approximate relative camera motions [30, 15]. The idea of combining very few 3D anchor points with 2D optical flow measurements — without attempting to recover depth [19] — has hardly been considered in this context. This motivates the investigation in this paper.

[1] microelectromechanical systems

2. APPROACH

The contribution of this paper is to extend the model-based visual-inertial tracking system presented in [5] with the capability to exploit the information in 2D optical flow measurements.

In [5] it is shown that the demand for visible features is significantly reduced when inertial sensors are used. However, lacking vision measurements for more than half a second results in a quick deterioration of the pose estimate due to bias and noise inherent to *inertial measurement units* (IMU). As a consequence the active vision loop fails to recover reappearing features and the track is lost. This paper utilises that optical flow measurements can be obtained from the camera images at any time, without knowledge about the scene structure. It is well-known that the resulting constraints do not provide full observability. However, as will be shown, they improve the estimates in combination with the collinearity constraints provided by the features with known depth. Furthermore, they reduce the need for 3D features and allow tracking for extended periods of time without any 3D point registrations.

This paper puts focus on the sensor fusion used to combine measurements form an IMU and images. It does not address how to obtain the image measurements from the image data. The image processing literature contains many suitable algorithms for this step, see [6, 5, 20, 19] for some alternatives.

The paper is structured as follows: Section 3 presents the architecture and functionality of the extended sensor fusion system. The optical flow measurement model is derived in Section 3.3. Section 4 presents results from Monte Carlo simulations and Section 5 draws conclusions and outlines future work.

3. SENSOR FUSION

Recursive filters can estimate the pose and kinematics of a moving camera-IMU system — such as the one depicted in Figure 1 — from camera and inertial measurements. The *extended Kalman*

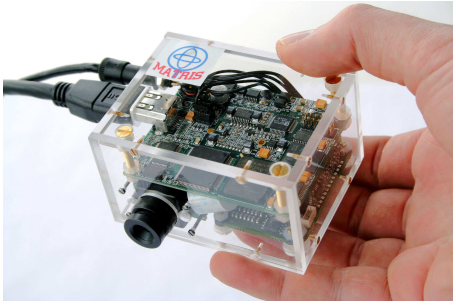


Figure 1: The integrated camera and inertial sensor package used for the experiments in [5].

filter (EKF) [18] is used to combine the information from the measurements — 2D/3D point correspondences and optical flow measurements from image processing, and 3D angular velocities and

 ISSN 2076-1465

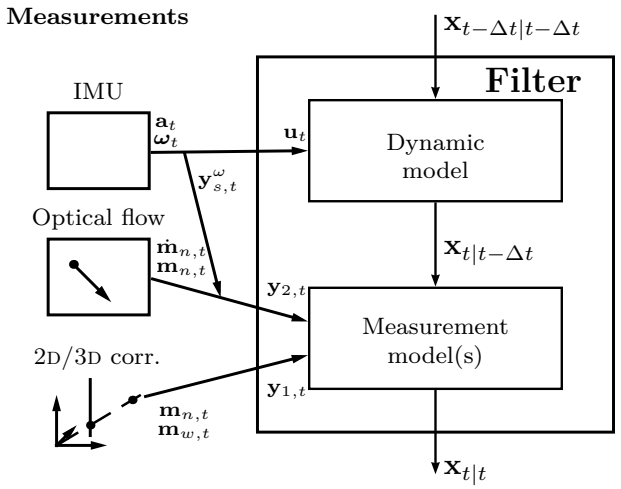


Figure 2: Filter architecture and data flow.

linear accelerations from the IMU. The architecture of the fusion system is outlined in Figure 2.

The core of the system is the state-space model, *i.e.*, the mathematical description of the problem under investigation. It has two components: The dynamic model, $\mathbf{x}_t = f(\mathbf{x}_{t-\Delta t}, \mathbf{u}_t, \mathbf{v}_t)$, describes the evolution of the state, $\mathbf{x}_t$, with time, t, subject to known control input, $\mathbf{u}_t$. The measurement model, $0 = h(\mathbf{x}_t, \mathbf{y}_t, \mathbf{e}_t)$, relates the noisy measurements, $\mathbf{y}_t$, to the state. Here, $\mathbf{v}_t$ and $\mathbf{e}_t$ denote time dependent stochastic process and measurement noise with $\mathbf{v}_t \sim \mathcal{N}(\mathbf{0}, Q_t)$ and $\mathbf{e}_t \sim \mathcal{N}(\mathbf{0}, R_t)$, respectively.

For the sake of completeness, the general EKF equations are given here. Let $\hat{\mathbf{x}}_t$ be the estimate of $\mathbf{x}_t$ at time t with $\mathbf{x}_t \sim \mathcal{N}(\hat{\mathbf{x}}_t, P_t)$. The equations for the time update are:

$$\hat{\mathbf{x}}_{t|t-\Delta t} = f(\hat{\mathbf{x}}_{t-\Delta t|t-\Delta t}, \mathbf{u}_t, \mathbf{0}) \tag{1a}$$

$$P_{t|t-\Delta t} = F_t P_{t-\Delta t|t-\Delta t} F_t^T + V_t Q_t V_t^T, \tag{1b}$$

with

$$F_t = \frac{\partial f}{\partial \mathbf{x}}(\hat{\mathbf{x}}_{t-\Delta t|t-\Delta t}, \mathbf{u}_t, \mathbf{0}) \tag{1c}$$

$$V_t = \frac{\partial f}{\partial \mathbf{v}}(\hat{\mathbf{x}}_{t-\Delta t|t-\Delta t}, \mathbf{u}_t, \mathbf{0}). \tag{1d}$$

The equations for the measurement update are:

$$S_t = H_t P_{t|t-\Delta t} H_t^T + E_t R_t E_t^T \tag{2a}$$

$$K_t = P_{t|t-\Delta t} H_t^T S_t^{-1} \tag{2b}$$

$$\hat{\mathbf{x}}_{t|t} = \hat{\mathbf{x}}_{t|t-T} - K_t h(\hat{\mathbf{x}}_{t|t-\Delta t}, \mathbf{y}_t, \mathbf{0}) \tag{2c}$$

$$P_{t|t} = P_{t|t-\Delta t} - K_t H_t P_{t|t-\Delta t}, \tag{2d}$$

with

$$H_t = \frac{\partial \mathbf{h}}{\partial \mathbf{x}}(\hat{\mathbf{x}}_{t|t-\Delta t}, \mathbf{0}) \tag{2e}$$

$$E_t = \frac{\partial \mathbf{h}}{\partial \mathbf{e}}(\hat{\mathbf{x}}_{t|t-\Delta t}, \mathbf{0}), \tag{2f}$$

where S_t is the innovation covariance and K_t is the Kalman filter gain.

As indicated in Figure 2, the system model takes the inertial measurements as control input. Moreover, the camera measurements, 2D/3D correspondences and optical flow, are modelled in two separate measurement equations. Since the camera measurements are assumed mutually independent the measurement updates

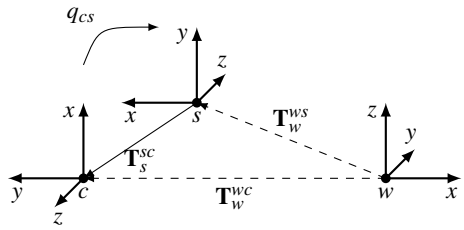


Figure 3: Illustration of the different 3D coordinate systems and how they are related. Rigid transformations are indicated by solid lines, flexible by dashed lines.

of the Kalman filter can be simplified by using sequential updates for each separate measurement. Compared to an all-in-one update with the entire set of observed features, this significantly reduces the computation time. With a minor modification to the standard EKF measurement update rule (2) [13], the sequential update gives an equivalent result to the all-in-one update, independently of the feature processing order.

Section 3.1 introduces the coordinate systems involved and the notation used subsequently. Section 3.2 presents the visual-inertial state-space model used to start with, and Section 3.3 extends this model with optical flow measurements.

3.1 Notation

Throughout the paper, $\mathbf{m}$ denotes points and $\tilde{\mathbf{m}}_n^T := [\mathbf{m}_n^T, 1]$ their homogenisation, $\mathbf{T}$ translations, and R rotations. First order time derivatives are indicated using a dot, *e.g.* $\dot{\mathbf{T}}$ denotes linear velocity. Angular velocity is denoted $\boldsymbol{\omega}$. Rotations are parametrised by unit quaternions, $\mathbf{q}$, with $R = \mathrm{rot}(\mathbf{q})$. The quaternion product is denoted $\odot$. See [31] for more information on quaternions and conversion formulas. The cross and dot product are denoted $\times$ and $\langle \cdot, \cdot \rangle$, respectively. Moreover, skew-symmetric matrices are denoted $\mathrm{S}(\cdot)$ with:

$$\mathrm{S}(\mathbf{u}) = \begin{bmatrix} 0 & -u_z & u_y \\ u_z & 0 & -u_x \\ -u_y & u_x & 0 \end{bmatrix}. \tag{3}$$

The following coordinate systems are used: the world frame, $\mathbf{w}$, (fixed to the target scene model), the camera frame, $\mathbf{c}$, (fixed to the moving camera), the sensor frame, $\mathbf{s}$, (fixed to the moving IMU), the pixel frame, $\mathbf{p}$, (fixed to the image plane) and the normalised image frame, $\mathbf{n}$, which is obtained from $\mathbf{p}$ using the relation $\tilde{\mathbf{m}}_n = K^{-1}\tilde{\mathbf{m}}_p$. The matrix K contains the intrinsic camera parameters [14]. Figure 3 illustrates the 3D coordinate systems and the transformations.

Subscripts (according to above) are used to indicate, in which frame a quantity is resolved. For transformations, subscripts with two letters denote the mapping and superscripts indicate direction.

3.2 Model-based Visual-Inertial Sensor Fusion

The inertial measurements, 3D angular velocities, $\mathbf{y}_s^{\omega}$, and 3D linear accelerations, $\mathbf{y}_s^{a}$, are considered to be known control input to the system model. Assuming a constant acceleration and constant angular velocity model, this gives a compact state vector,

$$\mathbf{x}^T = \begin{bmatrix} \mathbf{T}_w^{wsT} & \dot{\mathbf{T}}_w^{wsT} & \mathbf{q}_{sw}^T & \mathbf{b}_s^{\omega T} \end{bmatrix} \tag{4}$$

where $\mathbf{T}_w^{ws}$ denotes position, $\dot{\mathbf{T}}_w^{ws}$ linear velocity, and $\mathbf{q}_{sw}$ orientation of the IMU. Moreover, $\mathbf{b}_s^{\omega}$ denotes slowly time varying gyroscope biases. These parameters must be estimated in order to obtain reasonable tracking precision. The camera pose is obtained from the state vector using

$$\mathbf{q}_{cw} = \mathbf{q}_{cs} \odot \mathbf{q}_{sw} \tag{5a}$$

$$\mathbf{T}_w^{wc} = \mathbf{T}_w^{ws} + R_{ws}\mathbf{T}_s^{sc}, \tag{5b}$$

where $\mathbf{q}_{cs}$ and $\mathbf{T}_s^{sc}$ denote the hand-eye rotation and translation between the rigidly coupled camera and IMU. These quantities are calibrated once, *e.g.* using the method described in [16].

The inertial measurements are treated as control input. The system model is then

$$\begin{bmatrix} \mathbf{T}_{w,t+\Delta t}^{ws} \\ \dot{\mathbf{T}}_{w,t+\Delta t}^{ws} \\ \mathbf{q}_{sw,t+\Delta t} \\ \mathbf{b}_{s,t+\Delta t}^{\omega} \end{bmatrix} = \begin{bmatrix} \mathbf{T}_{w,t}^{ws} + \Delta t \dot{\mathbf{T}}_{w,t}^{ws} + \frac{\Delta t^2}{2} R_{ws,t}(\mathbf{y}_{s,t+\Delta t}^{a} - \mathbf{v}_{s,t}^{a}) + \frac{\Delta t^2}{2}\mathbf{g}_w \\ \dot{\mathbf{T}}_{w,t}^{ws} + \Delta t R_{ws,t}(\mathbf{y}_{s,t+\Delta t}^{a} - \mathbf{v}_{s,t}^{a}) + \Delta t \mathbf{g}_w \\ \exp\left(-\frac{\Delta t}{2}(\mathbf{y}_{s,t+\Delta t}^{\omega} - \mathbf{b}_{s,t}^{\omega} - \mathbf{v}_{s,t}^{\omega})\right) \odot \mathbf{q}_{sw,t} \\ \mathbf{b}_{s,t}^{\omega} + \mathbf{v}_{s,t}^{b^{\omega}} \end{bmatrix}, \tag{6a}$$

where $\mathbf{v}_{s,t}^{a}, \mathbf{v}_{s,t}^{\omega}$, and $\mathbf{v}_{s,t}^{b^{\omega}}$ are mutually independent process noises, $\mathbf{g}_w$ is the gravity vector, and $\exp(\mathbf{q})$ denotes the quaternion exponential with:

$$\exp(\mathbf{q})^T = \left[\cos\|\mathbf{q}\| \quad \frac{\mathbf{q}^T}{\|\mathbf{q}\|}\sin\|\mathbf{q}\|\right]. \tag{6b}$$

In a model-based camera tracking system, 2D/3D point correspondences are obtained by registering features with known 3D positions in the camera images. An appropriate measurement model for incorporating such measurements is given in the following. Let $\mathbf{y}_{1,t}^T := [\mathbf{m}_{n,t}^T, \mathbf{m}_{w,t}^T]$ be a 2D/3D point correspondence with mutually independent measurement noise $\mathbf{e}_{1,t}^T := [\mathbf{e}_{n,t}^T, \mathbf{e}_{w,t}^T]$. An implicit measurement equation based on the well-known perspective projection function:

$$\mathbf{m}_n = \begin{bmatrix} \frac{\mathbf{m}_{c,x}}{\mathbf{m}_{c,z}} & \frac{\mathbf{m}_{c,y}}{\mathbf{m}_{c,z}} \end{bmatrix}^T$$

with

$$\mathbf{m}_c = R_{cw}\mathbf{m}_w + T_c^{cw}$$

is then:

$$\mathbf{0}_2 = h_1(\mathbf{x}_t, \mathbf{y}_{1,t}, \mathbf{e}_{1,t}) = [\mathrm{I}_2 \quad -(\mathbf{m}_{n,t} + \mathbf{e}_{n,t})] \\ R_{cs}\left(R_{sw,t}(\mathbf{m}_{w,t} + \mathbf{e}_{w,t} - \mathbf{T}_{w,t}^{ws}) - \mathbf{T}_s^{sc}\right). \tag{7}$$

Here, I_2 denotes the 2×2 identity matrix and R_{cs}, $\mathbf{T}_s^{sc}$ are again the hand-eye rotation and translation, respectively. Note that in (7) the perspective division with the depth of the 3D point, $\mathbf{m}_{c,z}$, is avoided. This is preferred from a probabilistic point of view, since the division of two normal variables results in a Cauchy distribution, which has infinite second and higher order moments. As mentioned above, multiple feature observations are processed sequentially, each in a separate EKF measurement update step.

More details and variations of the above system can be found in [5].

3.3 Incorporation of Optical Flow Measurements

Optical flow is here defined as the velocity, $\dot{\mathbf{m}}_n$, of image location $\mathbf{m}_n$. It can be measured by computing the movement of a distinctive patch in subsequent camera images, *e.g.* by using a *Kanade-Lucas tracker* (KLT) [21] or other block matching methods, see for instance [9, 23, 20].

The pose and kinematics of a camera are directly related to the optical flow. Hence, optical flow measurements can be used to extract information about the camera movements. This section derives a measurement equation for optical flow measurements, which can be added directly to the state-space model introduced in Section 3.2.

Start by differentiating the point transformation $\mathbf{m}_c = R_{cw}\mathbf{m}_w + \mathbf{T}_c^{cw}$ — R_{cw} and $\mathbf{T}_c^{cw}$ form the camera pose — with respect to time:

$$\dot{\mathbf{m}}_c = \dot{R}_{cw}\mathbf{m}_w + R_{cw}\dot{\mathbf{m}}_w + \dot{\mathbf{T}}_c^{cw} \\ = \underbrace{\Omega_c^{cw}}_{\dot{R}_{cw}R_{wc}} R_{cw}\mathbf{m}_w + \dot{\mathbf{T}}_c^{cw} = \Omega_c^{cw}\underbrace{(\mathbf{m}_c - \mathbf{T}_c^{cw})}_{R_{cw}\mathbf{m}_w} + \dot{\mathbf{T}}_c^{cw}, \tag{8}$$

where $\Omega_c^{cw} := \mathrm{S}(\boldsymbol{\omega}_c^{cw})$ is the skew-symmetric matrix obtained from the angular velocity $\boldsymbol{\omega}_c^{cw}$.

Since optical flow is measured in the image plane and the depth of the 3D point, $\lambda := \mathbf{m}_{c,z}$, is unknown, it must be eliminated from (8). Let $\mathbf{m}_c = \lambda\tilde{\mathbf{m}}_n$ using the homogenisation of $\mathbf{m}_n$, then (8) can be reformulated as:

$$\dot{\tilde{\mathbf{m}}}_n = \boldsymbol{\omega}_c^{cw} \times \tilde{\mathbf{m}}_n + \frac{1}{\lambda}\mathbf{v}_c^{cw} - \frac{\dot{\lambda}}{\lambda}\tilde{\mathbf{m}}_n, \tag{9}$$

where $\mathbf{v}_c^{cw} := -\Omega_c^{cw}\mathbf{T}_c^{cw} + \dot{\mathbf{T}}_c^{cw}$ and per definition $\dot{\tilde{\mathbf{m}}}_n^T = [\dot{\mathbf{m}}_n^T, 0]$. Now, λ can be eliminated by applying $\langle\cdot, \mathbf{v}_c^{cw} \times \tilde{\mathbf{m}}_n\rangle$ to both sides of (9), resulting in the continuous epipolar constraint,

$$0 = \dot{\tilde{\mathbf{m}}}_n^T(\mathbf{v}_c^{cw} \times \tilde{\mathbf{m}}_n) + \tilde{\mathbf{m}}_n^T\left(\boldsymbol{\omega}_c^{cw} \times (\mathbf{v}_c^{cw} \times \tilde{\mathbf{m}}_n)\right). \tag{10}$$

The last step required to obtain a measurement equation is to rewrite (10) in terms of the known and estimated quantities, *i.e.*, the IMU pose and kinematics in (4), the IMU measurements and the hand-eye parameters in (5). Analogously to (8), using the fact that the hand-eye rotation and translation are rigid, *i.e.* $\dot{\mathbf{q}}_{cs} = \mathbf{0}_4$ and $\dot{\mathbf{T}}_s^{sc} = \mathbf{0}_3$, the expressions follow from (5) and its time derivative. The final relations between the quantities used in (10) and the known quantities are:

$$\boldsymbol{\omega}_c^{cw} = -\mathbf{q}_{cs} \odot (\mathbf{y}_s^{\omega} - \mathbf{b}_s^{\omega}) \odot \mathbf{q}_{sc} \tag{11a}$$

$$\mathbf{T}_c^{cw} = -R_{cs}\mathbf{T}_s^{sc} - R_{cs}R_{sw}\mathbf{T}_w^{ws} \tag{11b}$$

$$\dot{\mathbf{T}}_c^{cw} = -\mathrm{S}(\boldsymbol{\omega}_c^{cw})R_{cs}R_{sw}\mathbf{T}_w^{ws} - R_{cs}R_{sw}\dot{\mathbf{T}}_w^{ws}. \tag{11c}$$

In (11a), $\mathbf{y}_s^{\omega}$ denotes the angular velocity measured by the gyroscopes. This information is used as known input to the system model (6a), but is considered a noisy measurement for the optical flow model.

Let $\mathbf{y}_{2,t}^T := [\mathbf{y}_{s,t}^{\omega T}, \dot{\mathbf{m}}_{n,t}^T, \mathbf{m}_{n,t}^T]$ be a measurement made at time t comprising the angular velocity, $\mathbf{y}_{s,t}^{\omega}$, provided by the gyroscopes, and the optical flow, $\dot{\mathbf{m}}_{n,t}$, at image location, $\mathbf{m}_{n,t}$, provided by image processing. Moreover, let $\mathbf{e}_{2,t}^T := [\mathbf{e}_{s,t}^{\omega T}, \mathbf{e}_{\dot{n},t}^T, \mathbf{e}_{n,t}^T]$ denote mutually independent measurement noise. The final measurement equation is then obtained by inserting (11) into (10) and adding measurement noises:

$$0 = h_2(\mathbf{x}_t, \mathbf{y}_{2,t}, \mathbf{e}_{2,t}) = (\dot{\tilde{\mathbf{m}}}_{n,t} + \mathbf{e}_{\dot{n},t})^T\left(\mathbf{v}_{c,t}^{cw} \times (\tilde{\mathbf{m}}_{n,t} + \mathbf{e}_{n,t})\right) \\ + (\tilde{\mathbf{m}}_{n,t} + \mathbf{e}_{n,t})^T\left(\boldsymbol{\omega}_{c,t}^{cw} \times \left(\mathbf{v}_{c,t}^{cw} \times (\tilde{\mathbf{m}}_{n,t} + \mathbf{e}_{n,t})\right)\right). \tag{12}$$

By adding this equation to the state-space model of Section 3.2, optical flow measurements can be processed at any time as a complement to the registered image positions of known 3D feature points. Moreover, the computational complexity of processing such measurements is low, since the one-dimensional equation (12) reduces the matrix inversion in the EKF measurement update (2b) to a division with a scalar.

4. EXPERIMENTAL SETUP AND RESULTS

In order to evaluate the proposed method, *i.e.*, the value of extending the 2D/3D point correspondences with 2D optical flow measurements, the filter described in Section 3 is applied to a known data set.

The test simulates camera translations and rotations in all dimensions at various speeds. The trajectory is constructed in such a way that the camera focuses on one point of interest and then moves around it in the shape of an eight while rotating to keep the point in view direction. This is typical for close range camera localisation and visual servo applications. From the ground truth poses, inertial measurements are simulated and sampled at 100 Hz, by calculating angular velocities and linear accelerations — assuming $10\,\mathrm{m/s^2}$ acceleration due to gravity — and adding gyroscope biases. The trajectory is shown in Figure 5(a) and the inertial signals can be found

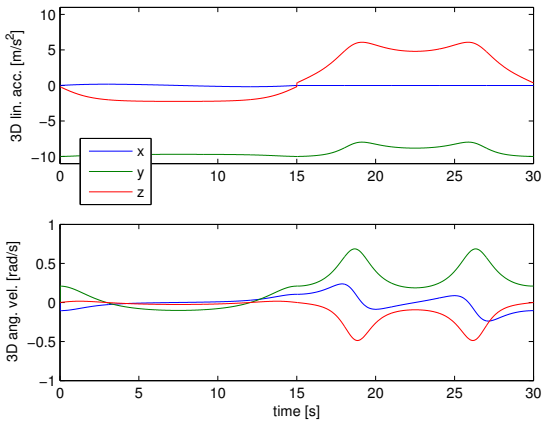


Figure 4: Inertial signals, sampled at 100 Hz.

Table 1: Simulation and estimation parameters: The 3D points are assumed known. The intrinsic camera parameters model a wide-angle lens with 95° horizontal field of view and VGA resolution. During estimation, the noise affecting the optical flow measurements has been approximated as additive measurement noise, $\mathbf{e}_t$, with standard deviation, $\sigma_{\mathbf{e}_t} = 0.3$. The noises in the table are given as standard deviations assuming equal noise in all dimensions.

	Simulation	**Estimation**	
$\sigma_{\mathbf{e}_{p,t}}$	0.5	1.5	[pixel]
$\sigma_{\mathbf{v}^a_{s,t}}$	–	0.1	[m/s^2]
$\sigma_{\mathbf{v}^\omega_{s,t}}$	–	0.01	[rad/s]

in Figure 4. Camera measurements are simulated by projecting 3D points using realistic intrinsic parameters, quantising the pixel coordinates to simulate the effects of digitalisation and feature extraction, adding noise, and transforming back into normalised image space. Optical flow measurements are simulated using (9).

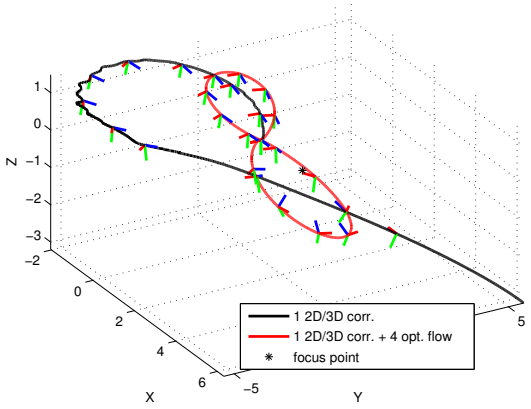
For evaluation, the inertial and camera measurements are utilised to recover the trajectory using the sensor fusion system outlined in Figure 2. The simulation and estimation parameters are provided in Table 1.

Figures 5 and 6 demonstrate promising improvements obtained by incorporating optical flow as proposed in Section 3.3. Figure 5 shows: when observing only one single 3D feature — the focus point — at 25 Hz, the filter fails to estimate the gyroscope biases, which results in a huge drift in the camera trajectory.[2] By adding four optical flow measurements, distributed over the corners of the camera images, the gyroscope biases are properly estimated and high accuracy is obtained.

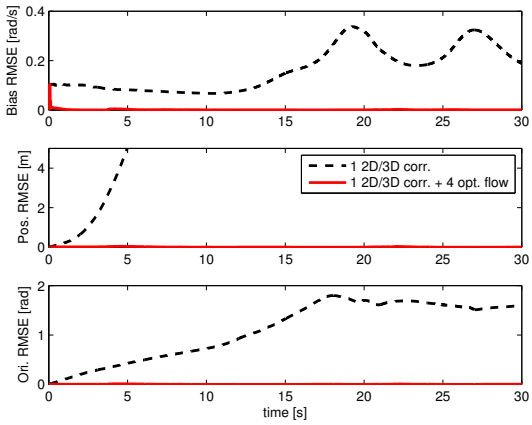
Figure 6 demonstrates another benefit. Here, two 3D features are observed, with different frequencies ranging from 10 to 1 Hz. The plot shows how the optical flow measurements significantly improve the results, as the velocity estimate otherwise degenerates rapidly with an observation rate below 2 Hz.

The results described above allow for the following conclusions. Though not individually providing full observability, optical flow measurements reduce the required quantity and frequency of observing features with known depth. Hence, they allow for tracking with minimal 3D knowledge about the target environment — especially when a wide-angle lens is used — and increase the robustness of the system against temporarily missing 3D feature observations, for instance due to occlusions.

[2] In [5] it is concluded that at least two to three visible 3D anchor points are required to obtain an accurate results.



(a) One example of estimated trajectory. The camera orientations are indicated by the coordinate systems.



(b) Root mean square error from 100 Monte Carlo simulations.

Figure 5: Tracking results from 100 Monte Carlo simulations. Note how in all cases the optical flow measurements reduce the error to almost zero, whereas the results quickly drift off when observing only one single 3D point as complement to IMU measurements.

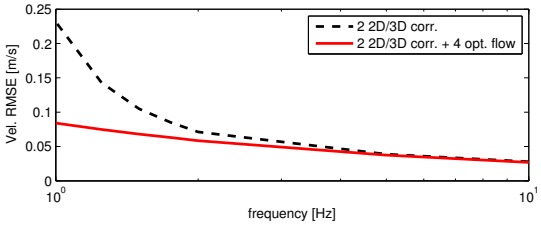


Figure 6: Root mean square velocity error vs. frequency of 2D/3D measurements from 100 Monte Carlo simulations. Two 3D anchor points are used. Note that these errors propagate time linearly into position.

5. CONCLUSION AND FUTURE WORK

This paper extends the model-based visual-inertial tracking system developed in [5] with optical flow measurements. Adding such measurements achieves two goals: First, camera pose and kinematics can be estimated correctly also during extended gaps of 3D point registrations. Second, under normal operation the need for 3D fea-

tures is reduced. This allows for tracking using minimal knowledge of the geometry of the scene. This way, robust and efficient tracking is obtained with very few 3D anchor points that could be installed or surveyed manually with reasonable effort. As such, the method provides an interesting alternative to a complete and computationally intense SLAM process. The benefits of using optical flow measurements have here been demonstrated in Monte Carlo simulations. Currently, the method is evaluated with experimental data. Observability of the camera pose and kinematics obtained from different configurations and numbers of such measurements is studied and outliers are handled.

6. ACKNOWLEDGEMENTS

The authors wish to thank the Automatic Control group at Linköping University, where this work was primarily performed.

REFERENCES

[1] H. Andreasson, T. Duckett, and A. Lilienthal. Mini-SLAM: Minimalistic visual SLAM in large-scale environments based on a new interpretation of image similarity. In *IEEE International Conference on Robotics and Automation*, Rome, Italy, Apr. 2007.

[2] M. Aron, G. Simon, and M.-O. Berger. Use of inertial sensors to support video tracking. *Computer Animation and Virtual Worlds*, 18:57 – 68, 2007.

[3] G. Bleser, M. Becker, and D. Stricker. Real-time vision-based tracking and reconstruction. *Journal of Real-Time Image Processing*, 2(2–3):161–175, Nov. 2007.

[4] G. Bleser and D. Stricker. Using the marginalised particle filter for real-time visual-inertial sensor fusion. In *International Symposium on Mixed and Augmented Reality*, Cambridge, UK, September 2008.

[5] G. Bleser and D. Stricker. Advanced tracking through efficient image processing and visual-inertial sensor fusion. *Computer & Graphics*, 33:59–72, Feb. 2009.

[6] G. Bleser, H. Wuest, and D. Stricker. Online camera pose estimation in partially known and dynamic scenes. In *International Symposium on Mixed and Augmented Reality*, pages 56–65, Santa Barabara, CA, Oct. 2006.

[7] L. A. Clemente, A. J. Davison, I. Reid, J. Neira, and J. D. Tardos. Mapping large loops with a single hand-held camera. In *Robotics: Science and Systems*, 2007.

[8] P. Corke, J. Lobo, and J. Dias. An introduction to inertial and visual sensing. *International Journal of Robotics Research*, 26:519–535, 2007.

[9] A. J. Davison. Real-time simultaneous localisation and mapping with a single camera. In *International Conference on Computer Vision*, volume 2, Nice, France, Oct. 2003.

[10] E. Eade and T. Drummond. Scalable monocular SLAM. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages 469–476, New York, NY, June 2006.

[11] E. Eade and T. Drummond. Monocular SLAM as a graph of coalesced observations. In *IEEE International Conference on Computer Vision*, Rio de Janeiro, Brazil, Oct. 2007.

[12] U. Frese. Treemap: An $O(\log n)$ algorithm for indoor simultaneous localization and mapping. *Autonomus Robots*, 21:103 – 122, 2006.

[13] F. Gustafsson. *Adaptive Filtering and Change Detection*. John Wiley & Sons, Sept. 2000.

[14] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2 edition, 2004.

[15] A. Heyden, F. Nyberg, and O. Dahl. *Recursive Structure and Motion Estimation Based on Hybrid Matching Constraints*, volume 4522/2007, pages 142–151. Springer Berlin / Heidelberg, July 2007.

[16] J. Hol, T. Schön, and F. Gustafsson. A new algorithm for calibrating a combined camera and IMU sensor unit. In *Proceedings of the 10th International Conference on Control, Automation, Robotics and Vision*, Hanoi, Vietnam, Dec. 2008.

[17] J. Hol, T. Schön, H. Luinge, P. Slycke, and F. Gustafsson. Robust real-time tracking by fusing measurements from inertial and vision sensors. *Journal of Real-Time Image Processing*, 2(2-3):149–160, Nov. 2007.

[18] A. H. Jazwinski. *Stochastic Processes and Filtering Theory*, volume 64 of *Mathematics in Science and Engineering*. Academic Press, Inc, 1970.

[19] F. Kendoul, I. Fantoni, and G. Dherbomez. Three nested Kalman filters-based algorithm for real-time estimation of optical flow, UAV motion and obstacles detection. In *IEEE International Conference on Robotics and Automation*, 2007.

[20] G. Klein and D. Murray. Parallel tracking and mapping for small AR workspaces. In *International Symposium on Mixed and Augmented Reality*, Nara, Japan, Nov. 2007.

[21] B. D. Lucas and T. Kanade. An iterative image registration technique with an application to Stereo Vision. In *International Joint Conference on Artificial Intelligence*, pages 674–679, Apr. 1981.

[22] Y. Ma, S. Soatto, J. Kosecka, and S. Sastry. *An Invitation to 3-D Vision*, volume 26. Springer Verlag, 2003.

[23] N. Molton, A. Davison, and I. Reid. Locally planar patch features for real-time structure from motion. In *British Machine Vision Conference*. BMVC, Sept. 2004.

[24] M. Montemerlo and S. Thrun. FastSLAM: A factored solution to the simultaneous localization and mapping problem. In *AAAI National Conference on Artificial Intelligence*, Edmonton, Canada, 2002.

[25] E. Mouragnon, F. Dekeyser, P. Sayd, M. Lhuillier, and M. Dhome. Real time localization and 3D reconstruction. In *Conference on Computer Vision and Pattern Recognition*, pages 363–370, 2006.

[26] L. M. Paz, P. Jensfelt, J. Tardos, and J. Neira. EKF SLAM updates in $o(n)$ with divide and conquer SLAM. In *IEEE International Conference on Robotics and Automation*, Rome, Italy, Apr. 2007.

[27] P. Pinies, T. Lupton, S. Sukkarieh, and J. D. Tardos. Inertial aiding of inverse depth SLAM using a monocular camera. In *IEEE International Conference on Robotics and Automation*, Rome, Italy, Apr. 2007.

[28] G. Reitmayr and T. Drummond. Going out: Robust model-based tracking for outdoor augmented reality. In *International Symposium on Mixed and Augmented Reality*, pages 109–118, Santa Barabara, CA, Oct. 2006.

[29] T. Schön, R. Karlsson, D. Törnqvist, and F. Gustafsson. A framework for simultaneous localization and mapping utilizing model structure. In *International Conference on Information Fusion*, Quebec, Canada, July 2007.

[30] S. P. N. Singh and J. W. Kenneth. Motion estimation by optical flow and inertial measurements for dynamic legged locomotion. In *In InerVis workshop at the IEEE International Conference on Robotics and Automation*, 2005.

[31] Z. Zhang and O. Faugeras. *3D Dynamic Scene Analysis*. Springer, 1992.