Direct Visual-Inertial Odometry with Stereo Cameras

Vladyslav Usenko, Jakob Engel, Jörg Stückler, and Daniel Cremers

***Abstract*— We propose a novel direct visual-inertial odometry method for stereo cameras. Camera pose, velocity and IMU biases are simultaneously estimated by minimizing a combined photometric and inertial energy functional. This allows us to exploit the complementary nature of vision and inertial data. At the same time, and in contrast to all existing visual-inertial methods, our approach is fully direct: geometry is estimated in the form of semi-dense depth maps instead of manually designed sparse keypoints. Depth information is obtained both from static stereo – relating the fixed-baseline images of the stereo camera – and temporal stereo – relating images from the same camera, taken at different points in time. We show that our method outperforms not only vision-only or loosely coupled approaches, but also can achieve more accurate results than state-of-the-art keypoint-based methods on different datasets, including rapid motion and significant illumination changes. In addition, our method provides high-fidelity semi-dense, metric reconstructions of the environment, and runs in real-time on a CPU.**

I. Introduction

Camera motion estimation and 3D reconstruction are amongst the most prominent topics in computer vision and robotics. They have major practical applications, well-known examples are robot navigation [30] [23] [28], autonomous or semi-autonomous driving [12], large-scale indoor reconstruction, virtual or augmented reality [24], and many more. In all of these scenarios, in the end one requires both the camera motion as well as information about the 3D structure of the environment – for example to recognize and navigate around obstacles, or to display environment-related information to a user.

In this paper, we propose a tightly coupled, direct visual-inertial stereo odometry. Combining a stereo camera with an inertial measurement unit (IMU), the method estimates accurate camera motion as well as semi-dense 3D reconstructions in real-time. Our approach combines two recent trends: *Direct image alignment* based on *probabilistic, semi-dense depth estimation* provides rich information about the environment, and allows for exploiting all information present in the images. This is in contrast to traditional feature-based approaches, which rely on hand-crafted keypoint detectors and descriptors, only utilizing information contained at, e.g., image corners – neglecting large parts of the image. Simultaneously, *tight integration* of inertial data into tracking provides accurate short-term motion constraints. This is of particular benefit for direct approaches: Direct image alignment is well-known to be heavily non-convex, and convergence can only be expected if a sufficiently accurate initial estimate is available. While in practice techniques like coarse-to-fine tracking increase the convergence radius, tight inertial integration solves this issue even more effectively, as the additional error term and resulting prior ensure convergence even for rapid motion. We show that it even allows for tracking through short intervals without visual information, e.g. caused by pointing the camera at a white wall. In addition, inertial measurements render global roll- and pitch observable, reducing global drift to translational 3D motion and yaw rotation.

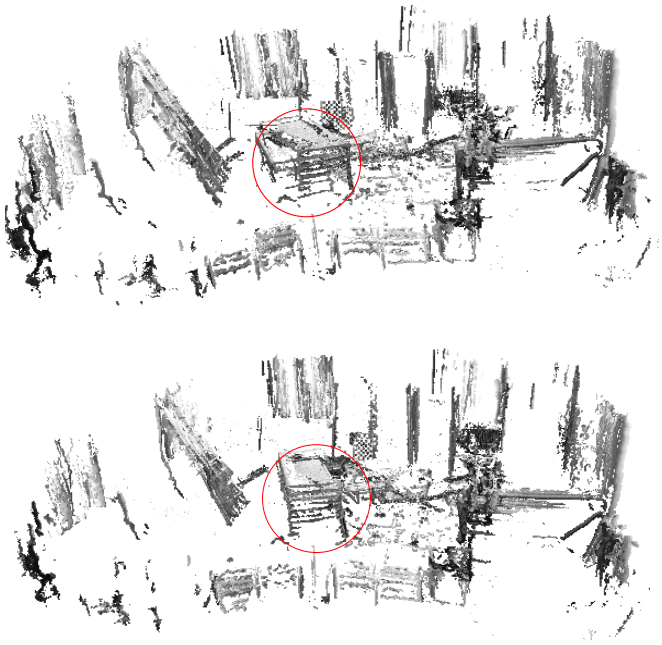

Fig. 1: Tight fusion of the IMU measurements with direct image alignment results in more accurate position tracking (bottom) compared to the odometry system that only relies on image alignment (top). The reconstructed pointclouds come from pure odometry, no loop closures were enforced.

In experiments we demonstrate the benefits of tight IMU integration with our Stereo LSD-SLAM approach towards loose integration or vision-only approaches. Our method performs very well on challenging sequences with strong illumination changes and rapid motion. We also compare our method with state-of-the-art keypoint-based methods and demonstrate that our method can achieve better accuracy on challenging sequences.

This work has been partially supported by grant CR 250/9-2 (Mapping on Demand) of German Research Foundation (DFG) and grant 16SV6394 (AuRoRoll) of BMBF.

V. Usenko, J. Engel, J. Stückler and D. Cremers are with the Department of Computer Science, Technical University of Munich, Germany {usenko, engelj,stueckle,cremers}@in.tum.de

II. Related Work

There exists a vast amount of research towards monocular and stereo visual odometry, 3D reconstruction and visual-inertial integration. In this section we will give an overview over the most relevant related publications, in particular focussing on direct vs. keypoint-based approaches, as well as tight vs. loosely coupled IMU integration.

While direct methods have a long history – first works including the work of Irani et al. [15] for monocular structure-and-motion – the first complete, real-time capable direct stereo visual-odometry was the work of Comport et al. [3]. Since then, direct methods have been omni-present in the domain of RGB-D cameras [18] [27], as they directly provide the required pixel-wise depth as sensor measurement. More recently, direct methods have become popular also in a monocular environment, prominent examples include DTAM [26], SVO [10] and LSD-SLAM [4].

At the same time, much progress has been made in the domain of IMU integration: due to their complementary nature and abundant presence in all modern hardware set-ups, IMUs are well-suited to complement vision based systems – providing valuable information about short-term motion and rendering global roll, pitch, and scale observable. In early works, visual-inertial fusion has been approached as a pure sensor-fusion problem: Vision is treated as an independent, black-box 6DoF sensor which is fused with inertial measurements in a filtering framework [30] [23] [6]. This so-called *loosely coupled* approach allows to use existing vision-only methods – such as PTAM [19], or LSD-SLAM [4] – without modifications; and the chosen method can easily be substituted for another one. On the other hand, in this approach, the vision part does not benefit from the availability of IMU data. More recent works therefore follow a tightly coupled approach, treating visiual-inertial odometry as one integrated estimation problem, optimally exploiting both sensor modalities.

Two main categories can be identified: Filtering-based approaches [21] [2] [29] operate on a probabilistic representation – mean and covariance – in a Kalman-filtering framework. One of the filtering approaches [13] claims to combine IMU measurements with direct image tracking, but does not provide a systematic evaluation and comparison to the state of the art methods. Optimization-based approaches on the other hand operate on an energy-function based representation in a non-linear optimization framework. While the complementary nature of these two approaches has long been known [8], the energy-based approach [20] [16] [17], – which we employ in this paper – allows for easily and adaptively re-linearizing energy terms if required, thereby avoiding systematic error integration from linearization. Another example of energy-based approach is presented in [9], which combines IMU measurements with direct tracking of a sparse subset of points in the image. In contrary to our method the old states are not marginalized out which on one hand allows for loop closures, but on the other hand does not guarantee the bounded update time in the worst case.

III. Contribution.

The main novelty of this paper is the formulation of *tight* IMU integration into *direct* image alignment within a non-linear energy-minimization framework. We show that including this sensor modality which in most practical cases is abundantly available helps to overcome the non-convexity of the photometric error, thereby eliminating one of the main weaknesses of direct approaches over keypoint-based methods. We evaluate our approach on different datasets and compare it to alternative stereo visual-inertial odometry systems, out-performing state-of-the-art keypoint-based methods in terms of accuracy in many cases. In addition, our method estimates accurate, metrically scaled, semi-dense 3D reconstructions of the environment, while running in real-time on a modern CPU.

IV. Notation

Throughout the paper, we will write matrices as bold capital letters ($\mathbf{R}$) and vectors as bold lower case letters ($\boldsymbol{\xi}$). We will represent rigid-body poses directly as elements of $\mathfrak{se}(3)$, which – with a slight abuse of notation – we write directly as vectors, i.e., $\boldsymbol{\xi} \in \mathbb{R}^6$. We then define the pose concatenation operator $\circ\colon \mathfrak{se}(3) \times \mathfrak{se}(3) \to \mathfrak{se}(3)$ directly on this notation as $\boldsymbol{\xi} \circ \boldsymbol{\xi}' := \log\left(\exp(\boldsymbol{\xi})\exp(\boldsymbol{\xi}')\right)$.

For each time-step i, our method estimates the camera's rigid-body pose $\boldsymbol{\xi}_i \in \mathfrak{se}(3)$, its linear velocity $\boldsymbol{v}_i \in \mathbb{R}^3$ expressed in the world coordinate system, and the IMU bias terms $\boldsymbol{b}_i \in \mathbb{R}^6$ for the 3D acceleration and 3D rotational velocity measurements of the IMU. A full state is hence given by $\boldsymbol{s}_i := \left[\boldsymbol{\xi}_i^T \boldsymbol{v}_i^T \boldsymbol{b}_i^T\right]^T \in \mathbb{R}^{15}$. For ease of notation, we further define pose concatenation and subtraction directly on this state-space as

$$\boldsymbol{s}_i \oplus s_i' := \begin{bmatrix} \boldsymbol{\xi}_i \circ \boldsymbol{\xi}_i' \\ \boldsymbol{v}_i + \boldsymbol{v}_i' \\ \boldsymbol{b} + \boldsymbol{b}_i' \end{bmatrix} \tag{1}$$

and

$$\boldsymbol{s}_i \ominus s_i' := \begin{bmatrix} \boldsymbol{\xi}_i \circ \boldsymbol{\xi}_i'^{-1} \\ \boldsymbol{v}_i - \boldsymbol{v}_i' \\ \boldsymbol{b} - \boldsymbol{b}_i' \end{bmatrix}. \tag{2}$$

The full state vector $\boldsymbol{s} := \left[\boldsymbol{s}_1^T \dots \boldsymbol{s}_N^T\right]^T$ includes the states of all frames.

V. Direct Visual-Inertial Stereo Odometry

We tightly couple direct image alignment – minimization of the photometric error – with non-linear error terms arising from inertial integration. In contrast to a loosely coupled approach, where the vision system runs independently of the IMU and is only fused afterwards, such tight integration maintains correlations between all state variables and thereby arbitrates directly between visual and IMU measurements.

Our photometric error formulation is directly based on the formulation proposed in LSD-SLAM [4], including robust Huber weights and normalization by the propagated depth variances. Recently, we extended this approach to stereo cameras [5], and augmented it with affine lighting correction.

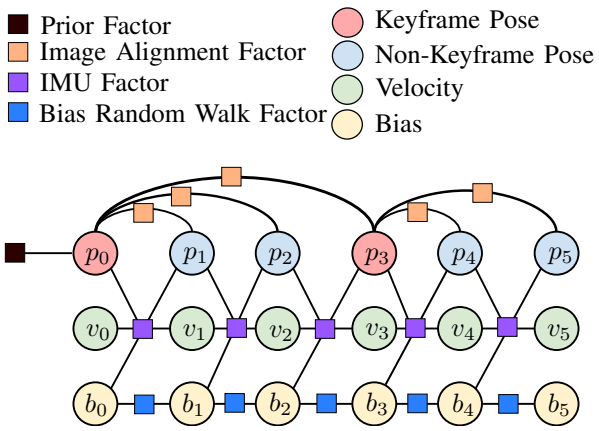

Fig. 2: Factor graph representing the visual-inertial odometry optimization problem. Poses of the keyframes are shown in red, poses of other frames in blue, velocities in green and biases in yellow. Poses and velocities are connected to the pose, velocity and biases of the previous frame by an IMU factor. The pose of each frame is connected to the pose of the keyframe by a VO factor, and factors between biases constrain their random walk.

We then formulate a joint optimization problem to recover the full state containing camera pose, translational velocity and IMU biases of all frames i. The overall energy that we want to minimize is given by

$$E(\boldsymbol{s}) := \frac{1}{2}\sum_{i=1}^{N} E^{I}_{i\to\text{ref}(i)}(\boldsymbol{\xi}_i, \boldsymbol{\xi}_{\text{ref}(i)}) + \frac{1}{2}\sum_{i=2}^{N} E^{\text{IMU}}(\boldsymbol{s}_{i-1}, \boldsymbol{s}_i), \tag{3}$$

where $E^{I}_{i\to\text{ref}(i)}$ and E^{IMU}_{i} are image and IMU error function terms, respectively. This optimization problem can be interpreted as maximum a-posterior estimation in a probabilistic graphical model (s. Fig. 2).

To achieve real-time performance, we do not optimize over an unboundedly growing number of state variables. Instead, we marginalize out all state variables other than the current image, its predecessor, and its reference keyframe. Through marginalization, all prior estimates and measurements are included with their uncertainty in the optimization.

Note that both modalities complement each other very well in a joint optimization framework – beyond the level of simple averaging of their motion estimates: Images can provide rich information for robust visual tracking. Depending on the observed scene, full 6-DoF relative motions can be observable. Degenerate cases, however, can occur in which the observed scene does not provide sufficient information for fully constrained tracking (e.g. pointing the camera at a texture-less wall). In this case, IMU measurements provide complementary measurements that bridge the gaps in observability.

IMUs typically operate at a much higher frequency than the frame rate of the camera and make measuring gravity direction and eliminating drift in roll and pitch angles possible. The downside of IMUs is, however, that they measure relative poses only indirectly through rotational velocities and linear accelerations. They are noisy and need to be integrated and compensated for gravity which strongly depends on the accuracy of the pose estimate. The measurements come with unknown, drifting biases that need to be estimated using an external reference such as vision. While IMU information is incremental, any images can be aligned towards each other that have sufficient overlap. This allows for incorporating relative pose measurements between images that are not in direct temporal sequence–enabling more consistent trajectory estimates.

A. Direct Semi-Dense Stereo Odometry

We base visual tracking on Stereo LSD-SLAM [5]:

- We track the motion of the camera towards a reference keyframe in the map. We create new keyframes, if the camera moved too far from existing keyframes.
- We estimate a semi-dense depth map in the current reference keyframe from static and temporal stereo cues. For static stereo we exploit the fixed baseline between the pair of cameras in the stereo configuration. Temporal stereo is estimated from pixel correspondences in the reference keyframe towards subsequent images based on the tracked motion.

There are several benefits of complementing static with temporal stereo in a tracking and mapping framework. Static stereo makes reconstruction scale observable. It is also independent of camera movement, but is constrained to a constant baseline, which limits static stereo to an effective operating range. Temporal stereo requires non-degenerate camera movement, but is not bound to a specific range as demonstrated in [4]. The method can reconstruct very small and very large environments at the same time. Finally, through the combination of static with temporal stereo, multiple baseline directions are available: while static stereo typically has a horizontal baseline – which does not allow for estimating depth along horizontal edges, temporal stereo allows for completing the depth map by providing other motion directions.

1) Direct Image Alignment: The pose between two images I_1 and I_2 is estimated by minimizing the photometric residuals

$$r^{I}_{\boldsymbol{u}}(\boldsymbol{\xi}) := aI_1(\boldsymbol{u}) + b - I_2(\boldsymbol{p}'). \tag{4}$$

where $\boldsymbol{p}' := \mathbf{T}_{\boldsymbol{\xi}}\pi^{-1}(\boldsymbol{u}, D_1(\boldsymbol{u}))$ and $\boldsymbol{\xi}$ transforms from image frame I_2 to I_1. The parameters a and b correct for affine lighting changes between the images and are optimized alongside the pose $\boldsymbol{\xi}$ in an alternating fashion, as described in [5]. We also determine the uncertainty $\sigma^{I}_{r,\boldsymbol{u}}$ of this residual [4]. The optimization objective for tracking a current frame towards a keyframe is thus given by

$$E^{I}_{\text{cur}\to\text{ref}}(\boldsymbol{\xi}_{\text{cur}}, \boldsymbol{\xi}_{\text{ref}}) := \sum_{u\in\Omega_{D_1}} \rho\left(\left[\frac{r^{I}_{\boldsymbol{u}}(\boldsymbol{\xi}^{-1}_{\text{ref}} \circ \boldsymbol{\xi}_{\text{cur}})}{\sigma^{I}_{r,\boldsymbol{u}}}\right]^2\right), \tag{5}$$

where ρ is the Huber norm. This objective is minimized using the iteratively re-weighted Levenberg-Marquardt algorithm.

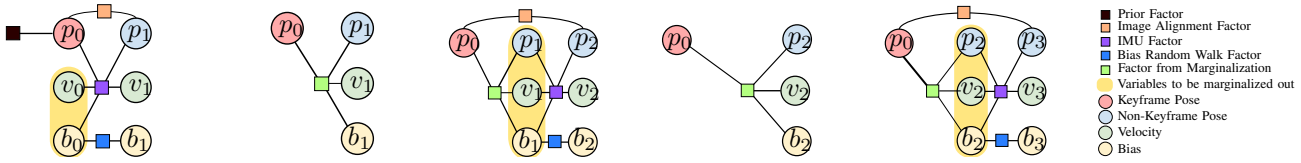

Fig. 3: Evolution of the factor graph with marginalization of old states. After adding new frame and optimizing the estimates of the variables in graph, all variables except the keyframe pose and the current frame pose, velocity and biases are marginalized out.

2) Depth Estimation: Scene geometry is estimated for pixels of the key frame with high image gradient, since they provide stable disparity estimates. Fig. 1 shows an example of such a semi-dense depth reconstruction. We estimate depth both from static stereo (i.e., using images from different physical cameras, but taken at the same point in time) as well as from temporal stereo (i.e., using images from the same physical camera, taken at different points in time).

We initialize the depth map with the propagated depth from the previous keyframe. The depth map is subsequently updated with new observations in a pixel-wise depth-filtering framework. We also regularize the depth maps spatially and remove outliers [4].

a) Static Stereo: We determine the static stereo disparity at a pixel by a correspondence search along its epipolar line in the other stereo image. In our case of stereo-rectified images, this search can be performed very efficiently along horizontal lines. We use the SSD photometric error over five pixels along the scanline as a correspondence measure. If a depth estimate with uncertainty is available, the search range along the epipolar line can be significantly reduced. Due to the fixed baseline, we limit disparity estimation to pixels with significant gradient along the epipolar line, making the depth reconstruction semi-dense.

Static stereo is integrated in two ways. If a new stereo keyframe is created, the static stereo in this keyframe stereo pair is used to initialize the depth map. During tracking, static stereo in the current frame is propagated to the reference frame and fused with its depth map.

b) Temporal Stereo: For temporal stereo we estimate disparity between the current frame and the reference keyframe using the pose estimate obtained through tracking. These estimates are fused in the keyframe. Only pixels are updated with temporal stereo, whose expected inverse depth error is sufficiently small. This also constrains depth estimates to pixels with high image gradient along the epipolar line, producing a semi-dense depth map.

B. IMU Integration

Underlying our IMU error function terms is the following nonlinear dynamical model. Let the pose $\boldsymbol{\xi}$ consist of the position $\boldsymbol{p}$ and rotation $\mathbf{R}$ of the IMU expressed in the world frame. Note that the velocity estimate $\boldsymbol{v}$ also is in the world frame. According to the IMU measurements of rotational velocities $\boldsymbol{\omega}_z$ and linear accelerations $\boldsymbol{a}_z$ the pose of the IMU evolves as

$$\dot{\boldsymbol{p}} = \boldsymbol{v} \tag{6}$$

$$\dot{\boldsymbol{v}} = \mathbf{R}\left(\boldsymbol{a}_z + \boldsymbol{\epsilon}_a - \boldsymbol{b}_a\right) + \boldsymbol{g} \tag{7}$$

$$\dot{\mathbf{R}} = \mathbf{R}\left[\boldsymbol{\omega}_z + \boldsymbol{\epsilon}_\omega - \boldsymbol{b}_\omega\right]_\times \tag{8}$$

where $[\cdot]_\times$ is the skew-symmetric matrix such that for vectors a, b, $[a]_\times b = a \times b$. The process noise $\boldsymbol{\epsilon}_a$, $\boldsymbol{\epsilon}_\omega$, $\boldsymbol{\epsilon}_{b,a}$, and $\boldsymbol{\epsilon}_{b,\omega}$ affect the measurements and their biases $\boldsymbol{b}_a$ and $\boldsymbol{b}_\omega$ with Gaussian white noise. Hence, for the biases $\dot{\boldsymbol{b}}_a = \boldsymbol{\epsilon}_{b,a}$ and $\dot{\boldsymbol{b}}_\omega = \boldsymbol{\epsilon}_{b,\omega}$. Note that we neglect the effect of Coriolis force in this model.

IMU measurements typically arrive at a much higher frequency than camera frames. We do not add independent residuals for each individual IMU measurement, but integrate the measurements into a condensed IMU measurement between the image frames. In order to avoid frequent reintegration if the pose or bias estimates change during optimization, we follow the pre-integration approach proposed in [22] and [14]. We integrate the IMU measurements between timestamps i and j in the IMU coordinate frame and obtain pseudo-measurements $\Delta\boldsymbol{p}_{i\to j}$, $\Delta\boldsymbol{v}_{i\to j}$, and $\mathbf{R}_{i\to j}$.

We initialize pseudo-measurements with $\Delta\boldsymbol{p}_{i\to i} = 0$, $\Delta\boldsymbol{v}_{i\to i} = 0$, $\mathbf{R}_{i\to i} = \mathbf{I}$, and assuming the time between IMU measurements is Δt we integrate the raw measurements:

$$\Delta\boldsymbol{p}_{i\to k+1} = \Delta\boldsymbol{p}_{i\to k} + \Delta\boldsymbol{v}_{i\to k}\Delta t \tag{9}$$

$$\Delta\boldsymbol{v}_{i\to k+1} = \Delta\boldsymbol{v}_{i\to k} + \mathbf{R}_{i\to k}\left(\boldsymbol{a}_z - \boldsymbol{b}_a\right)\Delta t \tag{10}$$

$$\mathbf{R}_{i\to k+1} = \mathbf{R}_{i\to k}\exp([\boldsymbol{\omega}_z - \boldsymbol{b}_\omega]_\times\,\Delta t). \tag{11}$$

Given the initial state and integrated measurements the state at the next time-step can be predicted:

$$\boldsymbol{p}_j = \boldsymbol{p}_i + (t_j - t_i)\boldsymbol{v}_i + \frac{1}{2}(t_j - t_i)^2\boldsymbol{g} + \mathbf{R}_i\Delta\boldsymbol{p}_{i\to j} \tag{12}$$

$$\boldsymbol{v}_j = \boldsymbol{v}_i + (t_j - t_i)\boldsymbol{g} + \mathbf{R}_i\Delta\boldsymbol{v}_{i\to j} \tag{13}$$

$$\mathbf{R}_j = \mathbf{R}_i\mathbf{R}_{i\to j}. \tag{14}$$

For the previous state $\boldsymbol{s}_{i-1}$ and IMU measurements $\boldsymbol{a}_{i-1}$, $\boldsymbol{\omega}_{i-1}$ between frames i and $i-1$, the method yields a prediction

$$\widehat{\boldsymbol{s}}_i := h\left(\boldsymbol{\xi}_{i-1}, \boldsymbol{v}_{i-1}, \boldsymbol{b}_{i-1}, \boldsymbol{a}_{i-1}, \boldsymbol{\omega}_{i-1}\right) \tag{15}$$

of the pose, velocity, and biases in frame i with associated covariance estimate $\widehat{\Sigma}_{s,i}$. Hence, the IMU error function

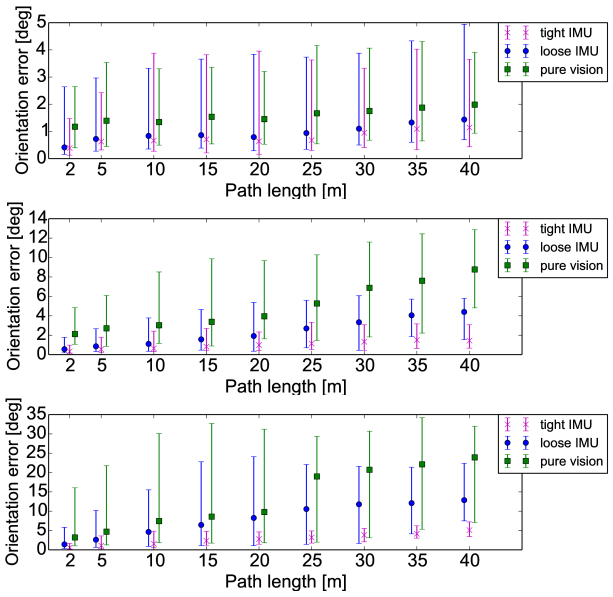

Fig. 4: Orientation error evaluated over different segment lengths for the three on the EuRoC dataset sequences. While both loosely-coupled and tightly-coupled IMU integration significantly decrease the error as global roll and pitch become observable, the tightly coupled approach is clearly superior. In particular in the last sequence – which includes strong motion blur and illumination changes – direct tracking directly benefits from tight IMU integration. See also Fig. 5

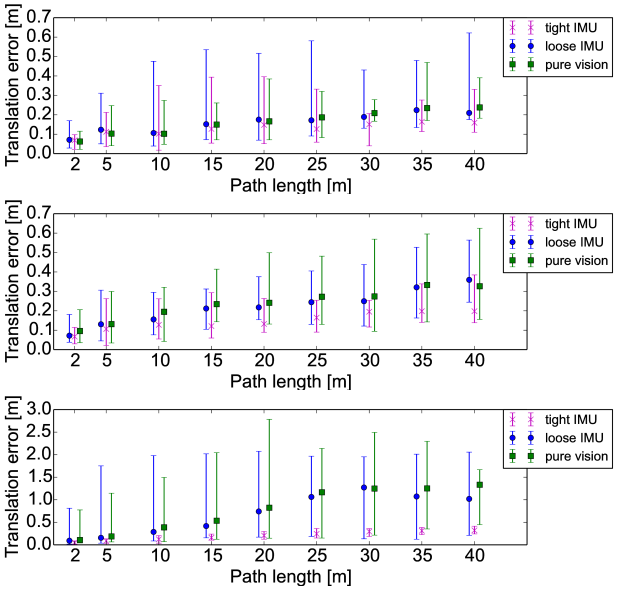

Fig. 5: Translational drift evaluated over different segment lengths for the three EuRoC sequences. As for rotation, the tightly coupled approach clearly performs best, see also Fig. 4.

terms are

$$E^{\text{IMU}}(\boldsymbol{s}_{i-1}, \boldsymbol{s}_i) := (\boldsymbol{s}_i \ominus \widehat{\boldsymbol{s}}_i)^T \, \widehat{\boldsymbol{\Sigma}}_{s,i}^{-1} \, (\boldsymbol{s}_i \ominus \widehat{\boldsymbol{s}}_i) \,. \tag{16}$$

C. Optimization

The error function in eq. (3) can be written as

$$E(\boldsymbol{s}) = \frac{1}{2} \boldsymbol{r}^T \mathbf{W} \boldsymbol{r} \tag{17}$$

$$= \frac{1}{2} \begin{bmatrix} \boldsymbol{r}_I^T & \boldsymbol{r}_{\text{IMU}}^T \end{bmatrix} \begin{bmatrix} \mathbf{W}_I & 0 \\ 0 & \mathbf{W}_{\text{IMU}} \end{bmatrix} \begin{bmatrix} \boldsymbol{r}_I \\ \boldsymbol{r}_{\text{IMU}} \end{bmatrix} . \tag{18}$$

The weights either implement the Huber norm on the photometric residuals $\boldsymbol{r}_I$ using iteratively re-weighted least-squares, or correspond to the inverse covariances of the IMU residuals $\boldsymbol{r}_{\text{IMU}}$ (eq.(16)). We optimize this objective using the Levenberg-Marquardt method. Linearizing the residual around the current state

$$\boldsymbol{r}(\boldsymbol{s} \oplus \boldsymbol{\delta s}) = \boldsymbol{r}(\boldsymbol{s}) + \mathbf{J}_s \boldsymbol{\delta s} \tag{19}$$

where

$$\mathbf{J}_s = \left. \frac{d\boldsymbol{r}(\boldsymbol{s} \oplus \boldsymbol{\delta s})}{d\boldsymbol{\delta s}} \right|_{\boldsymbol{\delta s} = \mathbf{0}}, \tag{20}$$

the error function $E(\boldsymbol{s})$ can be approximated around current state $\boldsymbol{s}$ with a quadratic function

$$E(\boldsymbol{s} \oplus \boldsymbol{\delta s}) = \mathbf{E}_{\boldsymbol{s}} + \boldsymbol{\delta s}^T \boldsymbol{b}_{\boldsymbol{s}} + \frac{1}{2} \boldsymbol{\delta s}^T \mathbf{H}_{\boldsymbol{s}} \boldsymbol{\delta s} \tag{21}$$

$$\boldsymbol{b}_{\boldsymbol{s}} = \mathbf{J}_{\boldsymbol{s}}^T \mathbf{W} \boldsymbol{r}(\boldsymbol{s}) \tag{22}$$

$$\mathbf{H}_{\boldsymbol{s}} = \mathbf{J}_{\boldsymbol{s}}^T \mathbf{W} \mathbf{J}_{\boldsymbol{s}} \tag{23}$$

where $\boldsymbol{b}_{\boldsymbol{s}}$ is the Jacobian and $\mathbf{H}_{\boldsymbol{s}}$ is the Hessian approximation of $E(\boldsymbol{s})$ and $\boldsymbol{\delta s}$ is a right-multiplied increment to the current state.

This function is minimized through $\boldsymbol{\delta s} = -\mathbf{H}_{\boldsymbol{s}}^{-1} \boldsymbol{b}_{\boldsymbol{s}}$, yielding the state update $\boldsymbol{s} \leftarrow \boldsymbol{s} \oplus \boldsymbol{\delta s}$. This update and relinearization process is repeated until convergence.

D. Partial Marginalization

To constrain the size of optimization problem, we perform partial marginalization and keep the set of optimized states at a small constant size. Specifically, we only optimize for the current frame state $\boldsymbol{s}_i$, the state of the previous frame $\boldsymbol{s}_{i-1}$, and the state $\boldsymbol{s}_{\text{ref}}$ of the reference frame used for tracking. If we split our state space $\boldsymbol{s}$ into $\boldsymbol{s}_\lambda$ and $\boldsymbol{s}_\mu$, where $\boldsymbol{s}_\lambda$ are the state variables we want to keep in the optimization, and $\boldsymbol{s}_\mu$ are the parts of the state that we want to marginalize out, we can rewrite the update step as follows

$$\begin{bmatrix} \mathbf{H}_{\mu\mu} & \mathbf{H}_{\mu\lambda} \\ \mathbf{H}_{\lambda\mu} & \mathbf{H}_{\lambda\lambda} \end{bmatrix} \begin{bmatrix} \boldsymbol{\delta s}_\mu \\ \boldsymbol{\delta s}_\lambda \end{bmatrix} = \begin{bmatrix} \boldsymbol{b}_\mu \\ \boldsymbol{b}_\lambda \end{bmatrix} . \tag{24}$$

Applying the schur complement to the upper part of the system we find

$$\boldsymbol{\delta s}_\lambda = -\left(\mathbf{H}_{\lambda\lambda}^*\right)^{-1} \boldsymbol{b}_\lambda^*, \tag{25}$$

$$\mathbf{H}_{\lambda\lambda}^* = \mathbf{H}_{\lambda\lambda} - \mathbf{H}_{\lambda\mu} \mathbf{H}_{\mu\mu}^{-1} \mathbf{H}_{\mu\lambda}, \tag{26}$$

$$\boldsymbol{b}_\lambda^* = \boldsymbol{b}_\lambda - \mathbf{H}_{\lambda\mu} \mathbf{H}_{\mu\mu}^{-1} \boldsymbol{b}_\mu. \tag{27}$$

which represents a system for $E^*(\boldsymbol{s}_\lambda)$ with states $\boldsymbol{s}_\mu$ marginalized out. Figure 3 shows the evolution of the graph with the marginalization procedure applied after adding every new frame to the graph.

E. Changing the Linearization Point

Partial marginalization fixes the linearization point of $\boldsymbol{s}_\lambda$ for the quantities involving both $\boldsymbol{s}_\mu$ and $\boldsymbol{s}_\lambda$ in eq. (25). Further optimization, however, changes the linearization point such that a relinearization would be required. We avoid the tedious explicit relinearization using a first-order approximation. If we represent the new linearization point $\boldsymbol{s}'_\lambda$ by the old linearization point $\boldsymbol{s}_\lambda$ and an increment $\boldsymbol{\Delta s}_\lambda$,

$$\boldsymbol{s}'_\lambda = \boldsymbol{s}_\lambda \oplus \underbrace{\boldsymbol{s}_\lambda^{-1} \oplus \boldsymbol{s}'_\lambda}_{=:\boldsymbol{\Delta s}_\lambda}, \tag{28}$$

we can change the linearization point of the current quadratic approximation of E^* through

$$E^*(\boldsymbol{s}'_\lambda \oplus \boldsymbol{\delta s}_\lambda) = E^*(\boldsymbol{s}_\lambda \oplus \boldsymbol{\Delta s}_\lambda \oplus \boldsymbol{\delta s}_\lambda) \tag{29}$$

$$\approx E^*(\boldsymbol{s}_\lambda \oplus (\boldsymbol{\Delta s}_\lambda + \boldsymbol{\delta s}_\lambda)). \tag{30}$$

The approximation made holds only if both $\boldsymbol{\Delta s}_\lambda$ and $\boldsymbol{\delta s}_\lambda$ are small – as both represent updates to the state, this is a valid assumption. We can then approximate the error function linearized around $\boldsymbol{s}'_\lambda$:

$$E^*(\boldsymbol{s}'_\lambda \oplus \boldsymbol{\delta s}_\lambda) = {E^*}_{\lambda'} + \boldsymbol{\delta s}_\lambda^T \boldsymbol{b}^*_{\lambda'} + \frac{1}{2}\boldsymbol{\delta s}_\lambda^T \mathbf{H}^*_{\lambda'\lambda'}\boldsymbol{\delta s}_\lambda, \tag{31}$$

$${E^*}_{\lambda'} = {E^*}_{\lambda} + \frac{1}{2}{\boldsymbol{\Delta s}_\lambda}^T \mathbf{H}^*_{\lambda\lambda}\boldsymbol{\Delta s}_\lambda + \boldsymbol{\Delta s}_\lambda^T \boldsymbol{b}^*_\lambda, \tag{32}$$

$$\boldsymbol{b}^*_{\lambda'} = \boldsymbol{b}^*_\lambda + \mathbf{H}^*_{\lambda\lambda}\boldsymbol{\Delta s}_\lambda, \tag{33}$$

$$\mathbf{H}^*_{\lambda'\lambda'} = \mathbf{H}^*_{\lambda\lambda}. \tag{34}$$

F. Statistical Consistency

Our framework accumulates information from many sources, in particular it uses (1) IMU observations, (2) static stereo, (3) temporal stereo / direct tracking and (4) a smoothness-prior on the depth. While old camera poses are correctly marginalized, we discard all pose-depth and depth-depth correlations: For each image alignment factor, depth values are treated as independent (noisy) input (Eq. 5). In turn, the effect of noisy poses is approximated during depth estimation [7]. While from a statistical perspective this is clearly incorrect, it allows our system to use hundreds of thousands of residuals per direct image alignment factor in real-time. Furthermore, it becomes unnecessary to drop past observations in order to preserve depth-depth independence, as done in [19]. Also note that – in contrast to monocular LSD-SLAM – much of the depth information originates from static stereo, which in fact is independent of the tracked camera poses.

VI. Results

We evaluate our approach both qualitatively and quantitatively on three different datasets, including a direct comparison to state-of-the-art feature-based VI odometry methods.

We have selected the datasets as they are especially challenging for direct visual odometry methods. They contain contrast changes, pure rotations, aggressive motions and relatively few frames per second, and for all of them the pure monocular algorithm [4] fails to track the sequence until the end. For Malaga dataset we used the default calibration parameters for evaluation. For all other datasets an offline calibration was performed using the *Kalibr* framework [11].

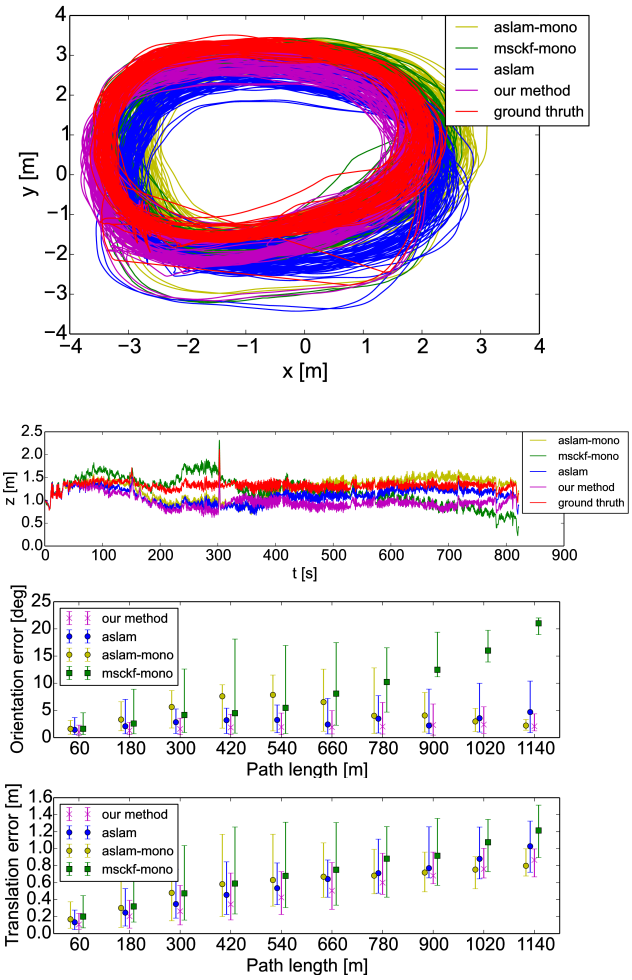

Fig. 6: Long-run comparison with state-of-the-art keypoint-based VI odometry methods, both filtering-based (msckf) and optimization-based (aslam). Dataset and results reported in [20]. Top: horizontal trajectory plot. Middle: height estimate. Bottom: Translational/rotational drift evaluated over different segment lengths.

We compare loosely- with tightly-coupled IMU integration with Stereo LSD-SLAM. The loosely coupled version runs the direct image alignment process separately, and only the final pose estimation result of the alignment is included into the optimization as a relative pose constraint between reference and current frame. With regards to the reconstruction accuracy, ground truth is not available on the datasets, such that it can only be judged qualitatively. Nevertheless, as tracking is based on the reconstruction, its accuracy implicitly depends on the trajectory estimate.

A. EuRoC Dataset

This dataset is obtained from the European Robotics Challenge (EuRoC), and contains three calibrated stereo video

Fig. 7: Qualitative results on a subset of Malaga Urban Dataset. Semi-dense reconstructions of selected parts of the map are shown on the left, and the overall trajectory with semi-dense reconstruction is shown in the middle. On the right a map of the city with overlayed trajectory measured with GPS is presented.

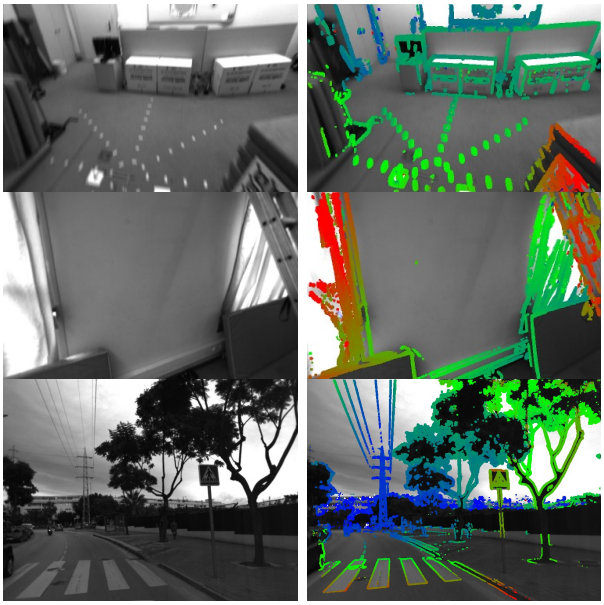

Fig. 8: Images from the EuRoC (upper row: motion blur, middle row: textureless) and Malaga datasets (bottom row) with semi-dense depth estimates. Semi-dense depth maps, with color-coded depth estimates are shown on the left.

sequences with corresponding IMU measurements, recorded with a Skybotix VI sensor. They were obtained from a quadrocopter flying indoors, and are in increasing difficulty: The third and most challenging sequence includes fast and aggressive motion, strong illumination changes as well as motion-blur and poorly textured views; some example images are shown in Fig. 8, as well as in the attached video.The images are provided with all required calibration parameters and motion-capture based ground truth, at WVGA resolution.

On this dataset, we evaluate the difference between tight IMU integration, loose IMU integration and purely vision-based LSD-SLAM. With the two upper plots in Fig. 6, we give a qualitative impression of the absolute trajectory estimate as in [20]. Since visual odometry does not correct for drift like a SLAM or full bundle adjustment method, the quantitative performance of the algorithm can be judged from the relative pose error (RPE) measure in the two bottom plots. The results in Fig. 5 and Fig. 4 demonstrate that our tightly-integrated, direct visual-inertial odometry method outperforms loose IMU integration both in translation and orientation drift. Both IMU methods in turn are better than the purely vision-based approach. The differences become particularly obvious for the last sequence, as here the tight IMU integration greatly helps to overcome non-convexities in the photometric error, allowing to seamlessly track through parts with strong motion blur.

Qualitatively, the reconstruction in Fig. 1 demonstrates a significant reduction in drift through tight IMU integration. The improved performance becomes apparent through the well-aligned, highlighted reconstructions which are viewed at the beginning and the end of the trajectory.

B. Long-Term Drift Evaluation

The second dataset contains a 14 minutes long sequence designed to evaluate long-term drift, captured with the same hardware setup as the EuRoC dataset. It is described and evaluated in [20] facilitating direct numeric comparison.

On this dataset, we compare our method with stereo depth estimation to the keypoint-based nonlinear optimization methods presented in [20] (aslam, aslam-mono) and the filtering-based approach in [25] (msckf). The aslam methods come in a stereo (aslam) and a monocular (aslam-mono) version. From Fig. 5 we can observe that the proposed method outperforms the filtering-based approach and the state-of-the-art keypoint-based optimization methods. Note,

that our method at the same time provides a semi-dense 3D reconstruction of the environment.

C. Malaga Dataset: Autonomous Driving

Third, we provide qualitative results on the Malaga dataset [1], obtained from a car-mounted stereo camera. For this dataset only raw GPS position without orientation is available as ground-truth such that we cannot provide a quantitative evaluation. Figure 8 shows a resulting trajectory, a semi-dense reconstruction of the environment, and a city map overlaid with GPS signal obtained on the Malaga dataset. These qualitative results demonstrate our algorithm in a challenging outdoor application scenario. Repetitive texture, moving cars and pedestrians, and direct sunlight pose gross challenges to vision-based approaches.

VII. Conclusion

We have presented a novel approach to direct, tightly integrated visual-inertial odometry. It combines a fully direct structure and motion approach – operating on per-pixel depth instead of individual keypoint observations – with tight, minimization-based IMU integration. We show that the two sensor sources ideally complement each other: stereo vision allows the system to compensate for long-term IMU bias drift, while short-term IMU constraints help to overcome non-convexities in the photometric tracking formulation, allowing to track through large inter-frame motion or intervals without visual information. Our method can out-perform existing feature-based approaches in terms of tracking accuracy, and simultaneously provides accurate semi-dense 3D reconstructions of the environment, while running in real-time on a standard laptop CPU.

In future work, we will investigate tight IMU integration with monocular LSD-SLAM. We also plan to employ this technology for localization and mapping with flying and mobile robots as well as handheld devices.

References

[1] J.-L. Blanco, F.-A. Moreno, and J. González-Jiménez. The málaga urban dataset: High-rate stereo and lidars in a realistic urban scenario. *International Journal of Robotics Research*, 2014.

[2] M. Bloesch, S. Omari, M. Hutter, and R. Siegwart. Robust visual inertial odometry using a direct ekf-based approach. In *Intelligent Robots and Systems (IROS), 2015 IEEE/RSJ International Conference on*, pages 298–304, Sept 2015.

[3] A. Comport, E. Malis, and P. Rives. Accurate quadri-focal tracking for robust 3d visual odometry. In *Intl. Conf. on Robotics and Automation (ICRA)*, 2007.

[4] J. Engel, T. Schöps, and D. Cremers. LSD-SLAM: Large-scale direct monocular SLAM. In *European Conference on Computer Vision (ECCV)*, September 2014.

[5] J. Engel, J. Stückler, and D. Cremers. Large-scale direct SLAM with stereo cameras. In *Intl. Conf. on Intelligent Robot Systems (IROS)*, 2015.

[6] J. Engel, J. Sturm, and D. Cremers. Camera-based navigation of a low-cost quadrocopter. In *Intl. Conf. on Intelligent Robot Systems (IROS)*, 2012.

[7] J. Engel, J. Sturm, and D. Cremers. Semi-dense visual odometry for a monocular camera. In *Intl. Conf. on Computer Vision (ICCV)*, 2013.

[8] R. M. Eustice, H. Singh, and J. J. Leonard. Exactly sparse delayed-state filters for view-based SLAM. *IEEE Transactions on Robotics*, 22(6):1100–1114, December 2006.

[9] C. Forster, L. Carlone, F. Dellaert, and D. Scaramuzza. Imu preintegration on manifold for efficient visual-inertial maximum-a-posteriori estimation. In *Proceedings of Robotics: Science and Systems*, Rome, Italy, July 2015.

[10] C. Forster, M. Pizzoli, and D. Scaramuzza. SVO: fast semi-direct monocular visual odometry. In *Intl. Conf. on Robotics and Automation (ICRA)*, 2014.

[11] P. Furgale, J. Rehder, and R. Siegwart. Unified temporal and spatial calibration for multi-sensor systems. In *Intelligent Robots and Systems (IROS), 2013 IEEE/RSJ International Conference on*, pages 1280–1286, Nov 2013.

[12] A. Geiger. Are we ready for autonomous driving? the KITTI vision benchmark suite. In *Intl. Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[13] J. Gui, D. Gu, and H. Hu. Robust direct visual inertial odometry via entropy-based relative pose estimation. In *Mechatronics and Automation (ICMA), 2015 IEEE International Conference on*, pages 887–892, Aug 2015.

[14] V. Indelman, S. Williams, M. Kaess, and F. Dellaert. Information fusion in navigation systems via factor graph based incremental smoothing. *Journal of Robotics and Autonomous Systems, RAS*, 61(8):721–738, August 2013.

[15] M. Irani and P. Anandan. All about direct methods, 1999.

[16] E. S. Jones and S. Soatto. Visual-inertial navigation, mapping and localization: A scalable real-time causal approach. *Int. J. Rob. Res.*, 2011.

[17] N. Keivan, A. Patron-Perez, and G. Sibley. Asynchronous adaptive conditioning for visual-inertial slam. In *International Symposium on Experimental Robotics*, 2014.

[18] C. Kerl, J. Sturm, and D. Cremers. Robust odometry estimation for RGB-D cameras. In *Intl. Conf. on Robotics and Automation (ICRA)*, 2013.

[19] G. Klein and D. Murray. Parallel tracking and mapping for small AR workspaces. In *Intl. Symp. on Mixed and Augmented Reality (ISMAR)*, 2007.

[20] S. Leutenegger, S. Lynen, M. Bosse, R. Siegwart, and P. Furgale. Keyframe-based visual–inertial odometry using nonlinear optimization. *The International Journal of Robotics Research*, page 0278364914554813, 2014.

[21] M. Li and A. Mourikis. High-precision, consistent EKF-based visual-inertial odometry. *International Journal of Robotics Research*, 32:690–711, 2013.

[22] T. Lupton and S. Sukkarieh. Visual-inertial-aided navigation for high-dynamic motion in built environments without initial conditions. *IEEE Transactions on Robotics*, 28(1):61–76, 2012.

[23] L. Meier, P. Tanskanen, F. Fraundorfer, and M. Pollefeys. Pixhawk: A system for autonomous flight using onboard computer vision. In *Intl. Conf. on Robotics and Automation (ICRA)*, 2011.

[24] O. Miksik, V. Vineet, M. Lidegaard, R. Prasaath, M. Nießner, S. Golodetz, S. L. Hicks, P. Perez, S. Izadi, and P. H. S. Torr. The semantic paintbrush: Interactive 3d mapping and recognition in large outdoor spaces. In *Proceedings of the 33nd annual ACM conference on Human factors in computing systems*. ACM, 2015.

[25] A. I. Mourikis and S. I. Roumeliotis. A multi-state constraint Kalman filter for vision-aided inertial navigation. In *Proceedings of the IEEE International Conference on Robotics and Automation*, pages 3565–3572, Rome, Italy, April 10-14 2007.

[26] R. Newcombe, S. Lovegrove, and A. Davison. DTAM: Dense tracking and mapping in real-time. In *Intl. Conf. on Computer Vision (ICCV)*, 2011.

[27] R. A. Newcombe, S. Izadi, O. Hilliges, D. Molyneaux, D. Kim, A. J. Davison, P. Kohli, J. Shotton, S. Hodges, and A. Fitzgibbon. KinectFusion: Real-time dense surface mapping and tracking. In *Intl. Symp. on Mixed and Augmented Reality (ISMAR)*, 2011.

[28] A. Stelzer, H. Hirschmüller, and M. Görner. Stereo-vision-based navigation of a six-legged walking robot in unknown rough terrain. *Int. J. Rob. Res.*, 2012.

[29] P. Tanskanen, T. Naegeli, M. Pollefeys, and O. Hilliges. Semi-direct ekf-based monocular visual-inertial odometry. In *Intelligent Robots and Systems (IROS), 2015 IEEE/RSJ International Conference on*, pages 6073–6078, Sept 2015.

[30] S. Weiss, M. Achtelik, S. Lynen, M. Chli, and R. Siegwart. Real-time onboard visual-inertial state estimation and self-calibration of mavs in unknown environments. In *Proc. of the IEEE International Conference on Robotics and Automation (ICRA)*, 2012.