sensors

MDPI

Article

A Multi-Sensor Fusion MAV State Estimation from Long-Range Stereo, IMU, GPS and Barometric Sensors

Yu Song [1,2,*], Stephen Nuske [1] and Sebastian Scherer [1]

[1] Robotics Institute, Carnegie Mellon University, Pittsburgh, PA 15213, USA; nuske@cmu.edu (S.N.); basti@andrew.cmu.edu (S.S.)

[2] School of Electronic and Information Engineering, Beijing Jiaotong University, Beijing 100044, China

* Correspondence: songyu@bjtu.edu.cn; Tel.: +86-10-158-1101-4311

Academic Editors: Gabriel Oliver-Codina, Nuno Gracias and Antonio M. López

Received: 1 October 2016 ; Accepted: 16 December 2016; Published: 22 December 2016

Abstract: State estimation is the most critical capability for MAV (Micro-Aerial Vehicle) localization, autonomous obstacle avoidance, robust flight control and 3D environmental mapping. There are three main challenges for MAV state estimation: (1) it can deal with aggressive 6 DOF (Degree Of Freedom) motion; (2) it should be robust to intermittent GPS (Global Positioning System) (even GPS-denied) situations; (3) it should work well both for low- and high-altitude flight. In this paper, we present a state estimation technique by fusing long-range stereo visual odometry, GPS, barometric and IMU (Inertial Measurement Unit) measurements. The new estimation system has two main parts, a stochastic cloning EKF (Extended Kalman Filter) estimator that loosely fuses both absolute state measurements (GPS, barometer) and the relative state measurements (IMU, visual odometry), and is derived and discussed in detail. A long-range stereo visual odometry is proposed for high-altitude MAV odometry calculation by using both multi-view stereo triangulation and a multi-view stereo inverse depth filter. The odometry takes the EKF information (IMU integral) for robust camera pose tracking and image feature matching, and the stereo odometry output serves as the relative measurements for the update of the state estimation. Experimental results on a benchmark dataset and our real flight dataset show the effectiveness of the proposed state estimation system, especially for the aggressive, intermittent GPS and high-altitude MAV flight.

Keywords: multi-sensor fusion; GPS-denied state estimation; long-range stereo visual odometry; absolute and relative state measurements; stochastic cloning EKF

1. Introduction

Light weight Micro-Aerial Vehicles (MAVs) equipped with sensors can autonomously access environments that are difficult to access for ground robots. Due to this capability, MAVs have become popular in many robot missions, e.g., structure inspection, environment mapping, reconnaissance and large-scale data gathering. Compared with ground robots, there are two main challenges for MAV autonomous navigation: (1) limited payload, power and onboard computing resources, so only light-weight compact sensors (like cameras) can be integrated for MAV applications; and (2) MAVs usually move with fast and aggressive six DOF (Degrees Of Freedom) motions. Accordingly, their state estimation, environment perception and obstacle avoidance are more difficult than ground robots.

Robust, accurate and smooth high-rate state estimation is the most critical capability to realize truly autonomous flight of MAVs. The state estimator reports the six DOF MAV pose and the velocity, so the output of the estimator serves as the input for environment mapping, motion planning and trajectory-following control. GPS (Global Positioning System) combined with the IMU (Inertial

Measurement Unit) state estimation technique has been widely utilized for providing MAV high-rate state information. Applications of low-rate GPS are limited to open environments, and also, GPS cannot provide accurate positioning information for MAV, especially in terms of altitude. As a complimentary sensor for GPS, the IMU measures the tri-axis accelerations and rotation rates in the IMU body frame, and the velocity and orientation are calculated as the integral of accelerations and rotation rates over time. For low-cost commercial IMUs, the inertia integral will drift very fast without global rectification information. As a result, the integration of additional sensing is a possible way to further improve state estimation redundancy, accuracy and robustness.

Because of the low cost, low energy consumption and satisfactory accuracy, camera-based Visual Odometry (VO) is an ideal choice for providing additional measurements. Stereo visual sensors reconstruct the environment features with the metric scale from the stereo baseline, so stereo-based VO easily generates six DOF pose measurements. The performance of stereo VO highly depends on the ratio between the stereo baseline and environmental depth, namely the baseline-depth ratio. The depth standard deviation from stereo is proportional to the quadratic of depth; thus, stereo VO is limited to a short range. As with the results reported in reference [1], at stereo disparities lower than 10 pixels, the depth triangulation from a single stereo rig tends to follow a non-Gaussian curve with a long tail. For cases with a large baseline-depth ratio (e.g., MAV high-altitude flights), stereo almost degenerates to a monocular system, thus losing the capability of pose measurements.

In this paper, we present a state estimation system that utilizes long-range stereo odometry that can degrade to a monocular system at high altitude and integrates GPS, barometric and IMU measurements. The estimation system has two main parts: an EKF (Extended Kalman Filter) estimator that loosely fuses both absolute state measurements (GPS, barometer) and the relative state measurements (IMU, VO) is derived and discussed in detail; a long-range stereo VO is designed both for low- and high-altitude odometry calculation. The odometry takes the EKF prediction information for robust camera pose tracking and feature matching, and the stereo VO outputs serve as the relative measurements for the update of the EKF state. There are three main highlights for the system:

(1) The state estimation system utilizes both absolute state measurement sensors (GPS, barometer), the relative six DOF pose state measurement provided by VO. To deal with both absolute and relative state measurements effectively, we derive a new stochastic cloning EKF state estimator to generate accurate and smooth state estimation both for GPS-available and GPS-denied environments.

(2) We developed a robust long-range stereo VO that works well both for low- and high-altitude cases. At low altitude, the VO utilizes stereo images; that means the features are directly triangulated by stereo pairs with a fixed static stereo baseline. At high altitude, the ratio between the scene depth and stereo baseline becomes large, and the stereo pair almost degenerates to a monocular system. In this situation, the additional stereo observations over time are fused by both multi-stereo triangulation and a multi-view stereo inverse depth filter for long-range feature depth generation.

(3) The EKF estimator and long-range VO coordinate to improve the robustness of the method. The IMU integral prediction information from the EKF estimator is used both for guiding image-feature matching and long-range VO optimization. Additionally, the VO is utilized as the relative measurement for the update of the EKF state.

The rest of the paper is organized as follows: Section 2 reviews the related work. Section 3 presents the proposed long-range stereo VO. In Section 4, a new EKF state estimator that combines both absolute and relative state measurements (GPS, barometer, IMU and long-range stereo VO) is derived and discussed. Finally, the experimental results will be reported and analyzed, followed by conclusions in Section 5.

2. Related Work

Since the VO concept was first derived, many VO algorithms have been proposed. Monocular VOs were developed on the basis of the "structure from motion" idea in computer vision. The first monocular VO uses a fundamental matrix between the two images to recover the camera motion [2]. As a milestone for modern VO or vSLAM (visual Simultaneous Localization And Mapping), PTAMwas proposed by Klein et al. in 2009 [3]. PTAM utilizes two threads for odometry: one is for sparse-features local mapping optimization, and the other one is for online camera pose tracking on the built local map. The original PTAM can only be used for small environments, and its modified versions have been applied for MAV state estimation with a down-looking camera system [4]. With a similar idea to PTAM, some other monocular VO or SLAM algorithms were proposed. For example, SVO [5], which was proposed by Foster et al. in 2014, utilizes image patches for camera direct tracking, because the features are only detected for selected key-frames, and also, no descriptors are extracted; therefore, SVO is faster than PTAM. At the same time, SVO was limited for high-rate cameras since direct tracking is employed. SDVO [6] and LSD-SLAM [7] are the first semi-dense monocular VO and SLAM algorithms. The idea for LSD-SLAM is also from the PTAM framework, but using high-gradient pixels for camera pose tracking (frame to key-frame tracking) and using the pose tracking result to refine the semi-dense depth of key-frames. As a state-of-the-art sparse feature VO or SLAM approach, ORBSLAM [8] has an additional thread for loop-closure detection and global batch optimization. For MAV state estimation using a monocular camera, there are three main limitations: (1) it is difficult for initial map generation since a single image cannot provide depth information; (2) the translation is up to scale, and the scale easily drifts over time; therefore, it usually should be combined with other sensors like an IMU or laser to recover the absolute scale; and (3) the monocular VO is not robust enough for MAVs' fast motions, unless high-rate cameras are utilized.

As a result, the stereo and RGBD sensors are more suitable for MAV application. Due to the limited range for depth perception (<4 m), most RGBD VOs are designed for indoor environments. As the first direct VO technique, DVO [9] utilizes the photometric projection consistency to track the current camera pose with respect to (w.r.t.) the last pose by using both depth and color images. Since RGBD cameras have the ability of dense depth perception, some point-cloud matching techniques have also been used [10]. A monocular-RGBD combined approach was proposed in 2015 [11], in which the key idea is to triangulate features that lack effective depth perception (>4 m). Libviso2 [12] and FOVIS [13] are the two well-known algorithms for stereo VO. Both algorithms utilize "reprojection error minimization" to calculate six DOF odometry. In stereo PTAM [14], the correspondence between the map points and current features is matched with a constant velocity motion model. Stereo version ORBSLAM follows the same idea as the monocular ORBSLAM and uses the static stereo baseline to reconstruct the map points both for map initialization and odometry calculation. As a result, the stereo version ORBSLAM is more robust than the monocular one. Stereo LSDSLAM is derived from monocular LSDSLAM, and the main difference lies in the direct utilization of dense stereo disparity for key-frame depth generation [15]. The SOFTalgorithm [16] was designed on the basis of Libviso2. The main contribution lies in the decoupling of the rotation and translation calculations: the five-point algorithm is utilized to calculate the rotation motion first, then with the reprojection error minimization for translation optimization. Recently, a long-range stereo VO algorithm was proposed with a 0.006 baseline-depth ratio. The stereo baseline is 2 m, and the maximum altitude is 120 m for a fixed-wing MAV. In this approach, an initial map is estimated with a monocular technique using only stereo left images, and then, the rejection errors of the reconstructed map points on the stereo right images are considered to recover scale. In this way, even the stereo baseline is very weak compared with environmental depth; it still provides useful metric scale information for batch optimization. Furthermore, the multi-view stereo bundle adjustment technique is used by taking the stereo right images as additional views for map point triangulation. While the algorithm cannot operate online, as reported in the experimental results, the long-range VO runs 1.35 Hz on an offline desktop PC [1].

For the fusion state estimation, the approaches can be classified into loosely coupled and tightly coupled in terms of sensor information utilization. For loosely-coupled algorithms, various sensors generate state measurements independently, and fuse the state measurements using either filters (EKF, UKF, PF) or smoothers (G2O [17], iSAM [18]). For loosely-coupled methods, the state estimation rate is equal to the highest state measurement rate. As a result, loosely-coupled approaches are suitable for high-rate requirements, e.g., for MAV control. Furthermore, loosely-coupled methods are more robust to sensor failures. Weiss et al. [4,19] proposed a multi-sensor fusing approach with IMU and monocular SLAM for MAV. The technique utilizes modified PTAM to provide an absolute pose measurement, so other absolute sensors are limited for integration. Furthermore, the system state will slowly drift due to the SLAM error accumulation over time. A UKF loosely-coupled state estimator was proposed for fusing stereo VO, GPS, IMU and barometer [20]. The estimator utilizes the unscented transform to calculate the transition probability densities. Accordingly, no Jacobians are required, and the transition probability densities can be computed up to the third non-linearity. Because the state vector has 21 elements, 43 sigma points will be propagated both for prediction and updates. The computational cost is much higher than an EKF. In reference [21], a stochastic cloning Kalman filter was proposed for fusing relative state measurements by augmenting state with the delayed pose. With this idea, the IMU and VO are fused in an EKF framework [22]. For tightly-coupled approaches, the sensors do not report state estimates on their own. Instead, all of the sensing information is combined to calculate a final output. It has been proven that tightly-coupled approaches outperform loosely-coupled algorithms in terms of accuracy. The EKFSLAM and MSCKFare two filter-based tightly-coupled VO/IMU fusion approaches [23]. In EKF-SLAM, the positions of visual landmarks in a sliding window are integrated. The IMU integral is used for pose and velocity prediction, and the sliding window SLAM system will be updated by visual landmark re-observations. In MSCKF, the system state is augmented by N delayed IMU poses; the delayed IMU poses are connected by visual landmarks. This is similar to the local bundle adjustment without iterated relinearization. Leutenegger et al. proposed a key-frame-based visual-inertial odometry with local bundle adjustment, with special focus on the marginalization of measurement terms outside the bundle adjustment window [24]. For smoother-based visual-inertia odometry, the IMU integral has to be recalculated because of the change of the linearization point in the batch optimization iteration steps. To reduce the re-integral computational cost, an IMU pre-integral approach in the IMU body frame has been discussed in reference [25].

3. Long-Range Stereo Odometry

The sparse features-based stereo VO algorithms are popular for robotics navigation applications. A key aspect of the stereo VO is to minimize a nonlinear error cost function by projecting the local map 3D points or 3D points generated from the reference stereo frame to the current stereo image pair. Current stereo VO algorithms have two main limits for MAV applications: (1) a lack of robustness in fast motion, especially for rotation; and (2) no correct estimation at long-ranges. In this section, we address the two main limitations of the stereo VO implementation to make stereo VO robust for MAV long-range high-altitude applications.

3.1. Long-Range Stereo Odometry Pipeline

Stereo depth reconstruction with a fixed static baseline is limited to a short range. For static stereo triangulation, the feature depth z is associated with the stereo matching disparity d as: $z = f_x \frac{B}{d}$ (where f_x is the focus length in pixels and B is the length of static stereo baseline in meters). Suppose the stereo matching disparity has variance σ_d^2; the triangulated depth variance σ_z^2 by stereo is as Equation (1). It is clear that the stereo depth standard deviation σ_z is proportional

to a quadratic of depth z. The depth error increases very quickly for the small disparity, long-range stereo measurements and, thus, cannot be utilized for VO optimization.

$$\sigma_z^2 = (\frac{\partial z}{\partial d})^2 \sigma_d^2 = \frac{f_x^2 B^2}{d^4} \sigma_d^2 = \frac{z^4}{f_x^2 B^2} \sigma_d^2 \quad (1)$$

Long-range stereo depth error (bias) can be effectively reduced by introducing additional stereo observation over time, namely multi-view stereo with a dynamic pseudo baseline. The pseudo baseline between the stereo frames can be used for the triangulation of the long-range stereo points. The fixed stereo baseline can provide an absolute scale constraint. Based on this idea, we developed a sparse feature-based stereo VO both for short- and long-range cases. The pipeline of the proposed long-range stereo VO is shown in Figure 1. It is a key-frame-based VO technique. The local map consists of a set of 3D sparse map points that is generated by selected key-frames. Furthermore, IMU information is integrated to further improve the robustness for aggressive camera motion and repetitive texture environments. Based on the current stereo baseline-depth ratio, the VO system switches both key-frame and new map point generation strategies between stereo and monocular modes:

(1) For a short range (e.g., MAV low-altitude flight, as shown in Figure 2a, the VO works with a stereo mode. For each new selected key-frame, most of the new features are directly triangulated by the stereo camera pair with the static stereo baseline. For some long-range points, they are triangulated using both the pseudo-baseline formed by the key-frame's poses and the static stereo baseline. In stereo mode, the environment structure is close to the camera; the image context easily changes especially for camera rotation. Therefore, the key-frames and its features are inserted into the local map relatively densely.

(2) For a long range (e.g., high-altitude flight, as shown in Figure 2b, the VO switches to monocular mode. The key-frames are inserted sparsely to provide enough relative motion between the key-frames for long-range triangulation. When VO is in a long-range mode, no features will be directly triangulated by static stereo. Because most of the "short-range points" will be outliers due to an incorrect matching from a low or repetitive texture area, such as sky, cloud and trees, instead, the new features will first be triangulated using both a dynamic pseudo baseline and a static stereo baseline. For the new features that cannot be triangulated by the pseudo baseline, we insert them into a "candidate queue". The feature depth will be iteratively refined by subsequently tracking stereo information with a multi-view inverse depth filter. If the inverse depth converges, the candidate feature will be added into the map and then used for camera pose tracking.

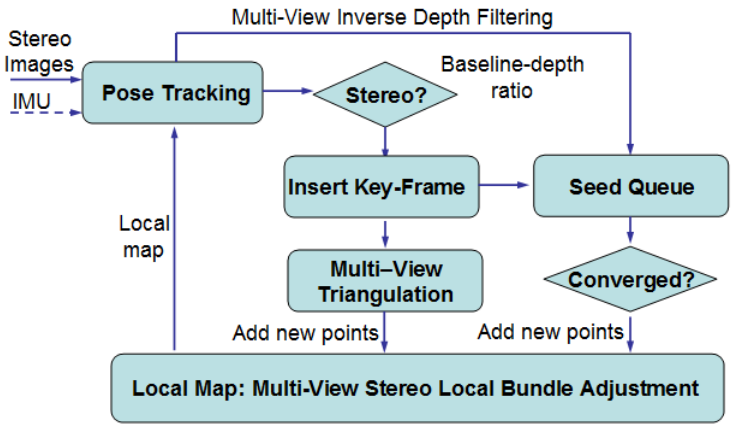

Figure 1. Long-range visual odometry pipeline.

Figure 2. Short-range and long-range environments. (**a**) Short-range environment. In this situation, most of the features are directly triangulated by static stereo baseline. (**b**) The long-range environment can also detect some large disparities of "short-rang" features (red dot) due to incorrect matching for the repetitive texture area. For this case, most of the "short-range" features are outliers; thus, they cannot be directly triangulated by static stereo baseline.

3.2. Long-Range Point Generation Using Multi-View Stereo Triangulation

The most critical aspect for long-range stereo is feature depth generation. For each new key-frame, its features can be classified into three groups:

(1) the features have been matched with the map.
(2) new features with an effective stereo depth (i.e, short-range points, with enough stereo disparity).
(3) new features with small disparities (long-range points).

The new long-range points without depth will first be triangulated using both the pseudo-baseline and the static stereo baseline from multi-view stereo measurements. The pseudo baseline is formed by the "relative pose" between the neighboring key-frames. As shown in Figure 3, the current left image feature is searched in the previous key-frame's left image feature set on the basis of an epipolar constraint, and for each key-frame, the matched feature pairs also have their own corresponding features in the right image. To make the matching more robust, the epipolar constraint between right image features is also checked. As a result, for each new map point, four matched features can be obtained between two key-frames, and the map point is triangulated as the intersection point of the four rays in the sense of least-squares.

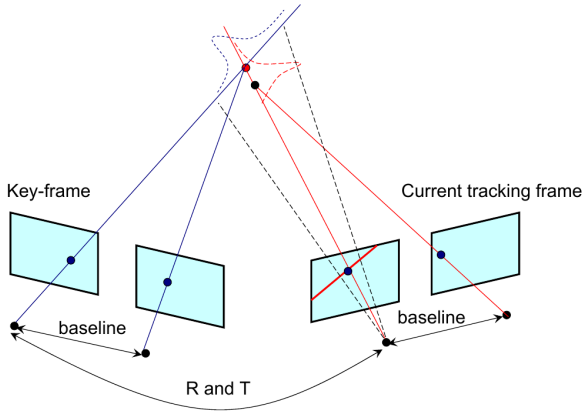

Figure 3. Multi-view observations by stereo. Between the two frames, the camera motions R and T provide a dynamic pseudo baseline, and for each stereo frame, the feature position is constrained by the static stereo baseline.

3.3. Long-Range Point Generation by Multi-View Stereo Inverse Depth Filtering

The inter-key-frames' triangulation is a kind of delayed depth generation approach because only features that can be viewed by at least two key-frames can be triangulated. For the exploration mode (e.g., the stereo moves forward), there are some new features that belong to the current key-frame itself; thus, they cannot be triangulated in time. An illustrative example is shown in Figure 4. To also apply these kinds of new features for subsequent camera pose tracking, we designed an inverse depth filter for each new candidate. For stereo, the inverse depth $\rho = \frac{1}{z} = \frac{d}{f_x B}$ is proportional to disparity d; as a result, the inverse depth uncertainty is easily modeled by a Gaussian distribution:

$$\sigma_\rho^2 = \frac{1}{f_x^2 B^2}\sigma_d^2 \tag{2}$$

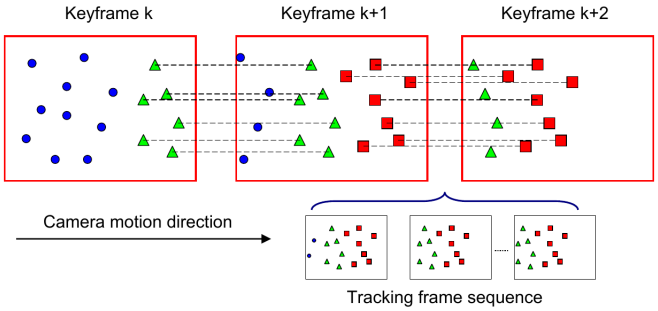

Figure 4. An example for the camera exploration mode. For the k-th key-frame; the blue points indicate the "old" features that have been matched with the map; and green triangles are the new features that await triangulation. For the $(k + 1)$-th key-frame, green triangle features can be triangulated, and some new features (red rectangle) wait for the next key-frame for triangulation. Between the $(k + 1)$-th key-frame and the $(k + 2)$-th key-frame, there is a set of tracking frames that also can provide useful measurements for the new features (red rectangle); we integrate all of the multi-view observations for the new feature using an inverse depth filter.

For each long-range candidate feature that belongs to the new inserted key-frame, its initial inverse depth prior is directly obtained from noisy static stereo depth triangulation, denoted as $\mathcal{N}(\rho_0, \frac{1}{f_x^2 B^2}\sigma_d^2)$. During the subsequent pose tracking, each new tracking frame is utilized to filter the initial distribution $\mathcal{N}(\rho_0, \frac{1}{f_x^2 B^2}\sigma_d^2)$, and the new feature candidate will be added to the map until its inverse depth variance is smaller than a given threshold. Ideally, for each new tracking frame, we can obtain two new observations for the candidate feature: (1) the inverse depth observation distribution for the candidate is calculated from the tracking frame static stereo matching; and (2) the inverse depth observation distribution can also be obtained by the dynamic pseudo baseline formed by the motion between the current tracking frame and its reference key-frame. Therefore, the filtered inverse depth distribution can be updated by the two new observations.

Denote as the 3D coordinate of a candidate feature with $z_0 = 1$ as $P_0 = (x_0, y_0, 1)^T$ in the key-frame coordinate and its corresponding matching point in the current tracking frame with $z_1 = 1$ is $P_1 = (x_1, y_1, 1)^T$. The motion from the key-frame to the current tracking frame is R_{10}, $t_{10} = (t_x, t_y, t_z)^T$, so the relationship of the two points is:

$$\frac{1}{\rho_1}P_1 = \frac{1}{\rho_0}R_{10}P_0 + t_{10} \tag{3}$$

where ρ_1 and ρ_0 represent the inverse depth measurements in the current tracking frame and key-frame, respectively.

For the current tracking frame, we observe the inverse depth stereo ρ_1 with its variance $\frac{1}{f_x^2 B^2}\sigma_d^2$. Therefore, the new measured inverse depth and its variance in the key-frame coordinate

are calculated by projecting the new measurement $\mathcal{N}(\rho_1, \frac{1}{f_x^2 B^2}\sigma_d^2)$ to the key-frame coordinate based on the last row of Equation (3):

$$\begin{aligned} \rho_0^s &= \frac{\frac{1}{\rho_1} - t_z}{R_{10}(3)P_0} \\ \sigma_{\rho_0^s}^2 &= (\frac{\rho_0^s}{\rho_1})^4 (\frac{1}{R_{10}(3)P_0 f_x B})^2 \sigma_d^2 \end{aligned} \tag{4}$$

where $R_{10}(3)$ represents the third row of rotation matrix R_{10} and σ_d^2 is the new stereo disparity variance in the current tracking frame (we set $\sigma_d^2 = 1$).

The inverse depth triangulation distribution using the motion from the key-frame to the current tracking frame is also derived from Equation (3) (with the first row and the last row). We have:

$$\begin{aligned} \rho_0^e &= \frac{R_{10}(1)P_0 - R_{10}(3)P_0 x_1}{t_z x_1 - t_x} \\ \sigma_{\rho_0^e}^2 &= (\frac{R_{10}(3)P_0 t_x - R_{10}(1)P_0 t_z}{(t_z x_1 - t_x)^2 f_x})^2 \sigma_{u1}^2 \end{aligned} \tag{5}$$

where $R_{10}(1)$ represents the first row of rotation matrix R_{10} and σ_{u1}^2 describes the matching error variance along the epipolar line in the current tracking frame; we set $\sigma_{u1}^2 = 4$ in our experiments.

To remove the outlier inverse depth measurements, both of the two new inverse depth hypotheses are further tested with prior $\mathcal{N}(\rho_0, \sigma_{\rho_0}^2)$ using χ^2 compatibility testing at 0.95. After passing the test, the posterior of the inverse depth distribution for the candidate feature is updated by multiplying the prior with the new measurements $\mathcal{N}(\rho_0^s, \sigma_{\rho_0^s}^2)$ and $\mathcal{N}(\rho_0^e, \sigma_{\rho_0^e}^2)$, that is:

$$\mathcal{N}(\rho_0^+, \sigma_{\rho_0^+}^2) = \mathcal{N}(\rho_0, \sigma_{\rho_0}^2)\mathcal{N}(\rho_0^s, \sigma_{\rho_0^s}^2)\mathcal{N}(\rho_0^e, \sigma_{\rho_0^e}^2) \tag{6}$$

3.4. Local Bundle Adjustment for Multi-View Stereo Optimization

The long-range stereo points generated by either triangulation or inverse depth filtering may still be noisy. An effective approach to further improve the feature 3D reconstruction accuracy is multi-view stereo local Bundle Adjustment (BA). During the local BA, the re-projection errors for both left and right images are considered. If the map points are reconstructed with an incorrect scale, the re-projection error on the right images will be large. Accordingly, the "weak" static stereo baseline can provide an absolute scale constraint for local BA optimization. The Jacobian J_{pi} of the rejection residual $\epsilon_{reproj}(i)$ w.r.t. the map point $P_i = (X_i, Y_i, Z_i)^T$ is:

$$J_{Pi} = \begin{bmatrix} \frac{\partial \epsilon_{reproj}(i)}{\partial u_i^l} \frac{\partial u_i^l}{\partial P_i} \\ \frac{\partial \epsilon_{reproj}(i)}{\partial v_i^l} \frac{\partial v_i^l}{\partial P_i} \\ \frac{\partial \epsilon_{reproj}(i)}{\partial u_i^r} \frac{\partial u_i^r}{\partial P_i} \end{bmatrix} = -\frac{1}{Z_c} \begin{bmatrix} f_x & 0 & -f_x \frac{X_c}{Z_c} \\ 0 & f_y & -f_y \frac{Y_c}{Z_c} \\ f_x & 0 & -f_x \frac{X_c - B}{Z_c} \end{bmatrix} R \tag{7}$$

where $P_c = (X_c, Y_c, Z_c)^T$ is the map point 3D coordinate in the left camera frame system. The first two rows are the residual Jacobian w.r.t. the left image and the last row is for right image. R is the camera rotation matrix.

The factor graph for the long-range stereo is shown in Figure 5. We add a unary edge $I_{4\times4}$ to each key-frame pose vertex. Consequently, the local BA will mainly focus on the map point optimization, and the key-frame's pose can only be changed in a small range. The factor graph is more like a structure-only bundle adjustment since the camera pose tracking has been fused with the IMU motion information (the IMU coupled odometry will be discussed in Section 3.5).

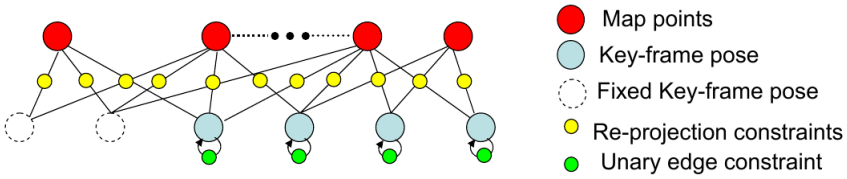

Figure 5. Factor graph for long-range stereo local bundle adjustment.

3.5. IMU Tightly-Coupled Odometry Calculation

The integration of an IMU motion prior to stereo VO has two advantages: (1) it provides a good initial motion guess for feature guided matching; (2) it gives a motion prior constraint for odometry optimization. We designed a tightly-coupled stereo VO by adding an IMU integral constraint into the 3D-2D re-projection cost non-linear optimization framework. Figure 6 shows the factor graph for the stereo VO; the camera pose tracking w.r.t. the local map can also be seen as a motion-only bundle adjustment. In this graph, map points and reference frame pose are fixed; only the current pose is set free for optimization. The cost function is:

$$\begin{aligned}(R,t) = argmin(w(\sum_{i=1}^{N} \|l_i - \pi^l(P_i;R,t)\|^2 \\ + \|r_i - \pi^r(P_i;R,t)\|^2) + (1-w)\|I_{imu} - (R,t)^T\|^2)\end{aligned} \tag{8}$$

where the current camera pose (R,t) is calculated by minimizing a non-linear re-projection error cost function. The 3D point in the local map is $P_i = (x_i, y_i, z_i)$; its matched 2D features in the current stereo rig are $l_i = (u_i^l, v_i^l)$ and $r_i = (u_i^r, v_i^r)$ for left and right images; π^l and π^r are the 3D-2D re-projection model for left and right cameras, respectively. N indicates the number of matched features. I_{imu} denotes the IMU motion integral between the current stereo frame and the reference stereo frame. The term $\|I_{imu} - (R,t)^T\|^2$ represents the IMU integral residual. $w \subset [0,1)$ is the weight for the IMU integral constraint.

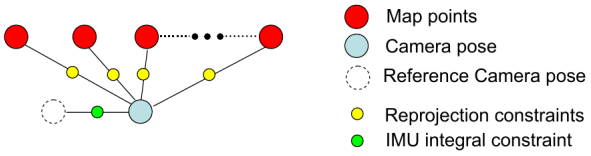

Figure 6. Factor graph for stereo IMU tightly-coupled odometry.

The optimal solution for the camera pose tracking is obtained by Levenberg–Marquardt iteration:

$$(J_x^T J_x + \lambda I)\Delta X = -J_x \epsilon_x \tag{9}$$

where J_x and ϵ_x are the Jacobian and residual at current pose x for the stereo pose tracking system. It has the form:

$$J_x = \begin{bmatrix} w(J_{reproj}) \\ (1-w)(I_{6\times 6}) \end{bmatrix} \tag{10}$$

$$\epsilon_x = \begin{bmatrix} w(\epsilon_{reproj}) \\ (1-w)(\epsilon_{imu}) \end{bmatrix} \tag{11}$$

where $I_{6\times 6}$ is a 6×6 unit matrix. J_{reproj} is the Jacobian for feature re-projection error. ϵ_{reproj} is feature re-projection error. ϵ_{imu} indicates the IMU integral residual.

For each map point $P_i = (Xi, Yi, Zi)^T$, its 3D-2D reprojection error $\epsilon_{reproj}(i)$ is calculated as:

$$\begin{aligned}\epsilon_{reproj}(i) &= m_i - \pi(P_i; R, t) \\ &= m_i - \frac{1}{Z_c}\begin{bmatrix} f_x & 0 & u_0 \\ 0 & f_y & v_0 \end{bmatrix}\left[R\begin{pmatrix} X_i \\ Y_i \\ Z_i \end{pmatrix} + t - \begin{pmatrix} B \\ 0 \\ 0 \end{pmatrix}\right]\end{aligned} \tag{12}$$

where $m_i \in l_i, r_i$ indicates the measured feature coordinate for the left or the right images. Z_c is the feature depth by transforming the map point to the left camera coordinate frame. $B = 0$ for the left camera, and $B = -baseline$ for right camera. f_x, f_y, u_0, v_0 are the stereo intrinsic parameters.

For the optimization, we utilize the minimal parametrization for the camera pose R, t in Lie manifold $\mathbf{SE}(3)$ denoted as: $X = (\theta_x, \theta_y, \theta_z, t_x, t_y, t_z)^T$. The Jacobian $J_{reproj}(i)$ for the 3D-2D re-projection error $\epsilon_{reproj}(i)$ w.r.t. the camera pose X is:

$$\begin{aligned} J_{reproj}(i) &= \begin{bmatrix} \frac{\partial \epsilon_{reproj}(i)}{\partial u_i^l}\frac{\partial u_i^l}{\partial X} \\ \frac{\partial \epsilon_{reproj}(i)}{\partial v_i^l}\frac{\partial v_i^l}{\partial X} \\ \frac{\partial \epsilon_{reproj}(i)}{\partial u_i^r}\frac{\partial u_i^r}{\partial X} \end{bmatrix} \\ &= \begin{bmatrix} f_x\frac{X_cY_c}{Z_c^2} & -f_x\frac{X_c^2+Z_c^2}{Z_c^2} & -f_x\frac{Y_c}{Z_c} & -f_x\frac{1}{Z_c} & 0 & f_x\frac{X_c}{Z_c^2} \\ f_y\frac{Y_c^2+Z_c^2}{Z_c^2} & -f_y\frac{X_cY_c}{Z_c^2} & -f_y\frac{X_c}{Z_c} & 0 & -f_y\frac{1}{Z_c} & f_y\frac{Y_c}{Z_c^2} \\ f_x\frac{X_cY_c}{Z_c^2} - B\frac{Y_c}{Z_c^2} & -f_x\frac{X_c^2+Z_c^2}{Z_c^2} + B\frac{X_c}{Z_c^2} & -f_x\frac{Y_c}{Z_c} & -f_x\frac{1}{Z_c} & 0 & f_x\frac{X_c}{Z_c^2} - B\frac{1}{Z_c^2} \end{bmatrix} \end{aligned} \tag{13}$$

where $P_c = (X_c, Y_c, Z_c)^T$ is the map point 3D coordinate in the left camera frame, i.e., $P_c = RP_i + t$. The first two rows are the residual Jacobian w.r.t. the left image, and the last row is for right image.

As a result, for N stereo features, the final system Jacobian J_x has $3N + 6$ rows. Additionally, based on the incremental solution $\Delta X = (\Delta\theta, \Delta t)$ from Equation (9), the update of the current camera pose is expressed as:

$$\begin{aligned} R &= \exp\left([\Delta\theta]_\times\right)R \\ t &= \exp\left([\Delta\theta]_\times\right)t + \Delta t \end{aligned} \tag{14}$$

where $[\Delta\theta]_\times$ is the skew-symmetric matrix of the incremental rotation vector $\Delta\theta$ and $\exp\left([\Delta\theta]_\times\right)$ is an exponential map.

4. Robust Multi-Sensor Fusion Based on a Stochastic Cloning EKF

This section presents an EKF state estimator for the multi-sensor loosely-coupled state estimation. In the EKF, IMU measurements are utilized to propagate the system state and covariance. For the update of the EKF state, both absolute measurements (GPS and barometer) and relative state measurements (stereo VO) are fused. The coordinate systems for the EKF estimator are shown in Figure 7. The navigation frame is a local NED (North-East-Down) frame, and the initial position is determined by the first GPS measurement. The EKF estimates the IMU body frame pose w.r.t. the navigation frame. The transformation from the camera frame to the IMU body frame is denoted as T_{is}, and the GPS receiver coordinate in the IMU body frame is t_{ig}.

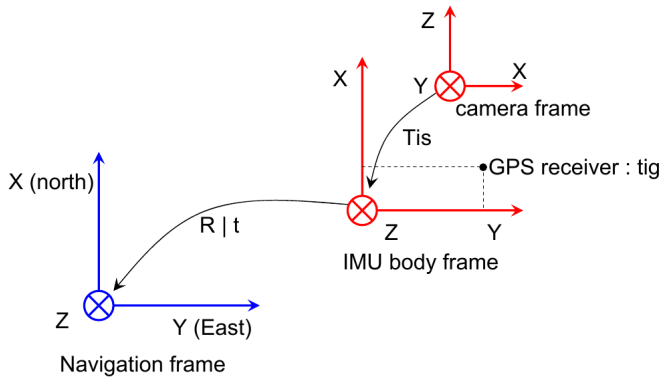

Figure 7. Definition of coordinate frames for EKF state estimation. The navigation frame (or world frame) is a local NED (North-East-Down) frame. The transforms R and t from the IMU body frame to the navigation frame will be estimated by the EKF. The parameters T_{is} (from the camera frame to the IMU body frame) and t_{ig} (GPS receiver coordinate in IMU body frame) are obtained by calibration.

4.1. IMU Integration

The IMU sensor measures the tri-axis accelerations and tri-axis angular rates w.r.t. the IMU body frame. The measurements given by the IMU are corrupted by Gaussian noise and a slowly varying bias terms, which must be removed before state estimation processing. Furthermore, the IMU accelerometers measure the force, which must be compensated by gravity. The following continuous-time model expresses the relationship between the IMU measured signals and true ones:

$$\begin{aligned} \omega_m &= \omega + b_g + n_g \\ a_m &= a + R^T g + b_a + n_a \end{aligned} \tag{15}$$

where $\omega_m \in \mathbb{R}^3$ and $a_m \in \mathbb{R}^3$ are the measured acceleration and angular rate, respectively. $\omega \in \mathbb{R}^3$ and $a \in \mathbb{R}^3$ indicate the true signals. n_g and n_a are zero-mean Gaussian $\mathcal{N}(0, \sigma_g^2)$ and $\mathcal{N}(0, \sigma_a^2)$; $b_g \in \mathbb{R}^3$ and $b_a \in \mathbb{R}^3$ are slowly varying bias terms for the accelerometer and gyroscope, respectively. Additionally, $g \in \mathbb{R}^3$ is gravity acceleration; the rotation matrix $R \in \mathbf{SO}(3)$ indicates the current IMU pose w.r.t. the navigation frame.

The estimated angular rate and acceleration rate are denoted as $\hat{\omega} \in \mathbb{R}^3$, $\hat{a} \in \mathbb{R}^3$, respectively. Additionally, the estimated bias terms for angular rate and acceleration are $\hat{b}_g$ and $\hat{b}_a$; we have:

$$\hat{\omega} = \omega_m - \hat{b}_g, \hat{a} = a_m - \hat{b}_a \tag{16}$$

Denote $\delta b_g = b_g - \hat{b}_g$, $\delta b_a = b_a - \hat{b}_a$ as the bias errors between the true bias b_g, b_a and the estimated bias $\hat{b}_g$, $\hat{b}_a$, and the slowly varying motion for bias errors are modeled as:

$$\delta \dot{b}_g = r_g, \delta \dot{b}_a = r_a \tag{17}$$

where $r_g \sim \mathcal{N}(0, \sigma_{rg}^2)$ and $r_a \sim \mathcal{N}(0, \sigma_{ra}^2)$ are zero-mean Gaussian.

Based on the above IMU kinematic model, the discrete IMU integral equations are:

$$\begin{aligned} p(k+1) &= p(k) + v(k)dt + \frac{1}{2}\hat{a}dt^2 \\ v_b(k+1) &= v_b(k) + (\hat{a} - [\hat{\omega}]_\times v_b(k))dt \\ R(k+1) &= R(k)\exp\left([\hat{\omega}dt]_\times\right) \end{aligned} \tag{18}$$

where $p(k) \in \mathbb{R}^3$ indicates the three D.O.F position w.r.t. the navigation frame at instant k. $v_b(k)$ is the velocity defined in the IMU body frame, and $R(k) \in \mathbf{SO}(3)$ is the rotation matrix w.r.t. the navigation frame. $[\hat{\omega}dt]_\times$ is a skew-symmetric matrix of the angular rate integral rotation vector $\hat{\omega}dt$; $\exp\left([\hat{\omega}dt]_\times\right)$ is an exponential map in the Lie manifold $\mathbf{SO}(3)$. dt is the IMU sampling time.

4.2. EKF State Definition and Jacobians

Based on the IMU integral equations and bias error model, the EKF system state S is defined as:

$$S = (p, \delta\theta, v_b, \delta b_g, \delta b_a)^T \in \mathbb{R}^{15} \tag{19}$$

where $p \in \mathbb{R}^3$ indicates position w.r.t. the navigation frame, $\delta\theta \in \mathbb{R}^3$ is the error rotation vector w.r.t. the IMU body frame, $v_b \in \mathbb{R}^3$ is the velocity w.r.t. the IMU body frame and $\delta b_g \in \mathbb{R}^3$, $\delta b_a \in \mathbb{R}^3$ are the current bias error terms.

The estimated rotation matrix is defined as $\hat{R} \in \mathbf{SO}(3)$, so the true rotation matrix $R \in \mathbf{SO}(3)$ after the rotation error compensation is calculated by matrix right multiplication:

$$R = \hat{R}\exp\left([\delta\theta]_\times\right) \tag{20}$$

where $[\delta\theta]_\times$ is skew-symmetric matrix of error rotation vector $\delta\theta$.

Based on the above system state definition, the system state dynamics $\dot{S}$ is derived as:

$$\begin{aligned} \dot{p} &= \hat{R}\exp\left([\delta\theta]_\times\right)v_b \\ \dot{\delta\theta} &= \exp\left([\delta\theta]_\times\right)(\hat{\omega} - \delta b_g - n_g) \\ \dot{v_b} &= -[\hat{\omega} - \delta b_g - n_g]_\times v_b + (\hat{R}\exp\left([\delta\theta]_\times\right))^T g + \hat{a} - \delta b_a - n_a \\ \dot{\delta b_g} &= r_g \\ \dot{\delta b_a} &= r_a \end{aligned} \tag{21}$$

Therefore, the Jacobian matrix $\frac{d\dot{S}}{dS} \in \mathbb{R}^{15\times15}$ for the system dynamics is obtained as:

$$\frac{\partial\dot{S}}{\partial S} = \begin{pmatrix} 0_{3\times3} & -\hat{R}[v_b]_\times & \hat{R}\exp\left([\delta\theta]_\times\right) & 0_{3\times3} & 0_{3\times3} \\ 0_{3\times3} & -[\hat{\omega} - \delta b_g - n_g]_\times & 0_{3\times3} & -\exp\left([\delta\theta]_\times\right) & 0_{3\times3} \\ 0_{3\times3} & [\hat{R}^T g]_\times & -[\hat{\omega} - \delta b_g - n_g]_\times & -[v_b]_\times & -I_{3\times3} \\ 0_{3\times3} & 0_{3\times3} & 0_{3\times3} & 0_{3\times3} & 0_{3\times3} \\ 0_{3\times3} & 0_{3\times3} & 0_{3\times3} & 0_{3\times3} & 0_{3\times3} \end{pmatrix} \tag{22}$$

where $I_{3\times3}$ denotes the 3×3 identity matrix and $0_{3\times3}$ denotes the 3×3 zero matrix.

The system state noise input consists of IMU measurement noise and bias error noise, that is:

$$W = (n_g, n_a, r_g, r_a)^T \in \mathbb{R}^{12} \tag{23}$$

As a result, the Jacobian matrix $\frac{d\dot{S}}{dW} \in \mathbb{R}^{15\times 12}$ w.r.t. the system noise is:

$$\frac{\partial \dot{S}}{\partial W} = \begin{pmatrix} 0_{3\times 3} & 0_{3\times 3} & 0_{3\times 3} & 0_{3\times 3} \\ -I_{3\times 3} & 0_{3\times 3} & 0_{3\times 3} & 0_{3\times 3} \\ -[v_b]_\times & -I_{3\times 3} & 0_{3\times 3} & 0_{3\times 3} \\ 0_{3\times 3} & 0_{3\times 3} & I_{3\times 3} & 0_{3\times 3} \\ 0_{3\times 3} & 0_{3\times 3} & 0_{3\times 3} & I_{3\times 3} \end{pmatrix} \tag{24}$$

Based on the relationship between the continuous-time and discrete-time systems, the final Jacobians for state covariance propagation are:

$$J_S = \frac{\partial \dot{S}}{\partial S} dt + I_{15\times 15}, J_W = \frac{\partial \dot{S}}{\partial W} dt \tag{25}$$

Remark 1. *In this section, the rotational error $\delta\theta$ is defined in a local coordinate system (current IMU body frame). As a result, the covariance $\Sigma^2_{\delta\theta}$ is also w.r.t. the current local frame. On the basis of the adjoint map of $\boldsymbol{SO}(3)$, the error rotate vector $\delta\theta$ can be expressed in global navigation frame $\hat{R}\exp([\delta\theta]_\times) = \exp([\hat{R}\delta\theta]_\times)\hat{R}$. The covariance for error rotation w.r.t. the navigation frame is $\hat{R}\Sigma^2_{\delta\theta}\hat{R}^T$.*

4.3. Treatment of VO Relative State Measurement Using Delayed State Stochastic Cloning

Our state estimation system utilizes both absolute state measurements (GPS provides absolute position and velocity measurement in the NED coordinate system; the barometer provides absolute state measurement for altitude) and the relative six D.O.F pose measurement (between the two stereo frames) provided by long-range stereo VO. To deal-with both absolute and relative state measurements, the system state defined in Equation (19) is further augmented by stochastic cloning of a delayed pose $p_l, \delta\theta_l$, which is updated with the previous VO measurement, namely:

$$\tilde{S} = (S^T, p_l, \delta\theta_l)^T \in \mathbb{R}^{21} \tag{26}$$

During the system state propagation, the delayed pose $p_l, \delta\theta_l$ is kept as constant; that means $\dot{p}_l = 0$ and $\delta\dot{\theta}_l = 0$. Therefore, the Jacobians for the augmented state $\tilde{S}$ are:

$$\tilde{J_S} = \begin{pmatrix} J_S & 0_{15\times 6} \\ 0_{6\times 15} & I_{6\times 6} \end{pmatrix} \in \mathbb{R}^{21\times 21} \tag{27}$$

$$\tilde{J_W} = \begin{pmatrix} J_W \\ 0_{6\times 12} \end{pmatrix} \in \mathbb{R}^{21\times 12} \tag{28}$$

The augmented state covariance is denoted as $\tilde{P}(k) \in \mathbb{R}^{21\times 21}$. Accordingly, the covariance propagation for the state augmented system is given as:

$$\tilde{P}(k+1 \mid k) = \tilde{J_S}\tilde{P}(k)\tilde{J_S}^T + \tilde{J_W}Q(K)\tilde{J_W}^T \tag{29}$$

For the system initialization, the initial system state covariance is of the form:

$$\tilde{P}(0) = \begin{pmatrix} \Sigma^2_p & 0 & 0 & 0 & 0 & \Sigma^2_p & 0 \\ 0 & \Sigma^2_\theta & 0 & 0 & 0 & 0 & \Sigma^2_\theta \\ 0 & 0 & \Sigma^2_{vb} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \Sigma^2_{bg} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \Sigma^2_{ba} & 0 & 0 \\ \Sigma^2_p & 0 & 0 & 0 & 0 & \Sigma^2_p & 0 \\ 0 & \Sigma^2_\theta & 0 & 0 & 0 & 0 & \Sigma^2_\theta \end{pmatrix} \tag{30}$$

Long-range stereo VO generates the relative six D.O.F motion measurement between the two visual frames. The relative measurement model is defined as:

$$\begin{aligned} \Delta p &= \exp\left(-[\delta\theta_l]_\times\right) R_l^T (p - p_l) \\ \Delta\theta &= \log(\exp\left(-[\delta\theta_l]_\times\right) R_l^T \hat{R} \exp\left([\delta\theta]_\times\right)) \end{aligned} \tag{31}$$

where $\Delta p \in \mathbb{R}^3$ is a position increment from the current pose $p, \hat{R}$ to the delay pose p_l, R_l and $\Delta\theta \in \mathbb{R}^3$ is the rotation increment. $R_l \in \mathbf{SO}(3)$ is the rotation matrix for previous visual updated orientation (i.e., the delayed state orientation), and $\delta\theta_l$ indicates the error rotation vector for the delayed state. $\hat{R} \in \mathbf{SO}(3)$ is the rotation matrix for the current orientation, and $\delta\theta$ is the current error rotation vector. The matrix logarithm $\log(R_l^T \hat{R})$ maps the rotation matrix $R_l^T \hat{R}$ to a rotation vector.

For Jacobians with a relative translation Δp w.r.t. system state S, we have:

$$\begin{aligned} \frac{\partial \Delta p}{\partial p} &= \exp\left(-[\delta\theta_l]_\times\right) \mid_{\delta\theta_l=0} R_l^T = R_l^T \\ \frac{\partial \Delta p}{\partial p_l} &= -\exp\left(-[\delta\theta_l]_\times\right) \mid_{\delta\theta_l=0} R_l^T = -R_l^T \\ \frac{\partial \Delta p}{\partial \delta\theta_l} &= \frac{\partial \exp\left(-[\delta\theta_l]_\times\right)}{\partial \delta\theta_l} \mid_{\delta\theta_l=0} R_l^T (p - p_l) = [R_l^T (p - p_l)]_\times \end{aligned} \tag{32}$$

where the derivative $\frac{\partial \Delta p}{\partial \delta\theta_l}$ is derived based on the first-order Taylor expansion for the exponential map at $\delta\theta_l = 0$, that is $\exp\left(-[\delta\theta_l]_\times\right) \mid_{\delta\theta_l=0} \approx 1 - [\delta\theta_l]_\times$. Additionally, the anti-commutativity rule for skew-symmetric matrix, namely: $[\delta\theta_l]_\times R_l^T (p - p_l) = -[R_l^T (p - p_l)]_\times \delta\theta_l$.

The Jacobians for the $\Delta\theta$ are computed as:

$$\begin{aligned} \frac{\partial \Delta\theta}{\partial \delta\theta} &= \frac{\partial \log(R_l^T \hat{R} \exp\left([\delta\theta]_\times\right))}{\partial \delta\theta} \mid_{\delta\theta=0} = Adj(R_l^T \hat{R}) = R_l^T \hat{R} \\ \frac{\partial \Delta\theta}{\partial \delta\theta_l} &= \frac{\partial \log(\exp\left(-[\delta\theta_l]_\times\right) R_l^T \hat{R})}{\partial \delta\theta_l} \mid_{\delta\theta_l=0} = -Adj(I_{3\times3}) = -I_{3\times3} \end{aligned} \tag{33}$$

where $Adj(R)$ is the adjoint map in $R \in \mathbf{SO}(3)$, and it has the property of $Adj(R) = R$. The derivative for the matrix logarithm is derived by the first-order approximation of Campbell–Baker–Hausdorff formula. For the logarithm map derivative with a unified form like $\frac{\partial \log(A \exp\left([\delta\theta]_\times\right)) B}{\partial \delta\theta} \mid_{\delta\theta=0}$, its derivative can be estimated by $Adj(A)$ under the condition that AB approximates the identity. More details for the logarithm derivation can be found in reference [26].

As a result, the VO relative measurement Jacobian is expressed as:

$$H_{vo} = \begin{pmatrix} R_l^T & 0_{3\times12} & -R_l^T & [R_l^T (p - p_l)]_\times \\ 0_{3\times3} & R_l^T \hat{R} & 0_{3\times12} & -I_{3\times6} \end{pmatrix} \tag{34}$$

Denote the VO relative measurement as $(\Delta p_{vo}, \Delta\theta_{vo})^T$; the measurement residual is given by:

$$\tilde{r} = \begin{pmatrix} \Delta p_{vo} - \Delta p \\ \Delta\theta_{vo} \ominus \Delta\theta \end{pmatrix} \tag{35}$$

where the rotational vector residual $\Delta\theta_{vo} \ominus \Delta\theta$ is defined as: $log(\Delta R^{-1} \Delta R_{vo})$. $\Delta R = \exp\left([\Delta\theta]_\times\right)$ is the predicted rotation matrix from the current state to the delayed state. Additionally, the $\Delta R_{vo} = \exp\left([\Delta\theta_{vo}]_\times\right)$ is the VO measured one.

It is worthwhile to note that, after each VO relative measurement update, the delayed portion vector of the state $p_l, \delta\theta_l$ is set equal to the current updated pose $p(k+1), \delta\theta(k+1)$, and the state covariance matrix is updated by "cloning" the corresponding covariance blocks from the current

state covariance to delayed pose covariance. To update the EKF state, we should first transform the VO measurement from the visual frame to the IMU body frame using the visual-IMU relative pose calibration T_{is}; suppose the VO measurement in visual frame is Z_s; its corresponding measurement in the IMU body frame is:

$$Z_i = T_{is} Z_s {T_{is}}^{-1} \tag{36}$$

The update of the EKF state is standard, that is:

$$\begin{aligned} K &= \tilde{P}(k+1 \mid k) H^T (\tilde{P}(k+1 \mid k) H^T + R)^{-1} \\ \tilde{S}(k+1) &= \tilde{S}(k) + K\tilde{r} \end{aligned} \tag{37}$$

The EKF covariance update uses Joseph's form to avoid the negative definition, that is:

$$\tilde{P}(k+1) = (I - KH)\tilde{P}(k+1 \mid k)(I - KH)^T + KRK^T \tag{38}$$

Remark 2. *After the VO relative measurement update, the updated covariance $\tilde{P}(k+1)$ should keep two important properties: (1) it should be lower than IMU state propagation covariance $\tilde{P}(k+1 \mid k)$ since VO information is available to the system; and (2) it must be increased compared with previous error covariance $\tilde{P}(k)$. Otherwise, the absolute measurement (GPS and barometer) will lose the ability of updating the state estimation. Compared with pseudo absolute measurement VO update approaches, the covariance using delayed state cloning EKF can meet the two properties. This will be verified in Sections 5.3 and 5.4.*

4.4. Update of EKF State Using Absolute State Measurements

GPS provides absolute position and velocity measurement in the NED frame system; suppose the heading of the initial EKF navigation frame is aligned with the NED frame; the GPS measurement model is:

$$Z_{gps} = \begin{bmatrix} p + \hat{R} \exp([\delta\theta]_\times) t_{ig} \\ \hat{R} \exp([\delta\theta]_\times)(v_b + [\hat{\omega} - \delta b_g]_\times t_{ig}) \end{bmatrix} \tag{39}$$

where $t_{ig} \in \mathbb{R}^3$ is the translation from the GPS receiver to the IMU body frame, as explained in Figure 7. The GPS measurement Jacobian is derived as:

$$H_{gps} = \begin{pmatrix} I_{3\times3} & -\hat{R}[t_{ig}]_\times & 0_{3\times3} & 0_{3\times3} & 0_{3\times3} \\ 0_{3\times3} & -\hat{R}[v_b]_\times & \hat{R} & [t_{ig}]_\times & 0_{3\times3} \end{pmatrix} \tag{40}$$

Since GPS measurement in altitude has a large uncertainty, the GPS height and velocity in altitude are not utilized to update the EKF state. Only the position and velocity for north and east are kept as GPS measurements, namely $Z_{gps} = (p_n, p_e, v_n, v_e)^T \in \mathbb{R}^4$. Consequently, the third and the sixth rows for the GPS Jacobian H_{gps} are also removed.

We utilized the "GPS health status", which reports how many satellites can be seen by the receiver, to determine the current GPS measurement covariance. For bad "GPS health status", GPS will report a large covariance. It is worth mentioning that the χ^2 test at 0.95 is utilized to verify the compatibility between current GPS measurement and the system predicted state. If GPS measurement "jumps" due to perturbation (e.g., multipath), the system will reject the GPS measurement automatically. In fact, all of the sensor measurements are firstly checked by the χ^2 test before they are utilized for state estimation. As a result, the EKF state estimator is robust to any sensor failures.

The barometer provides absolute altitude measurements w.r.t. the navigation frame. The navigation frame is a local NED frame, so the barometer measures the negative altitude w.r.t. the NED coordinate. As a result, the barometer measurement model is:

$$Z_{baro} = -p_d \tag{41}$$

where p_d denotes the z component for current position. Its Jacobian is:

$$H_{baro} = \begin{pmatrix} 0 & 0 & -1 & 0_{1\times 18} \end{pmatrix} \tag{42}$$

For the EKF implementation, a ring buffer with a 2-s time is kept to save all of the incoming sensor data. As shown in Figure 8, when a new VO measurement arrives, its time stamp is usually not the most up to date due to the image transmission and the stereo VO calculation delay. For this case, after the update of the EKF state on the VO time stamp, all of the subsequent IMU integral should be re-integrated to re-predict the current state. The same processing is also carried out for GPS and barometric measurements. To further decrease the computational cost of IMU re-integration, the IMU pre-integral technique in the IMU body frame can be utilized.

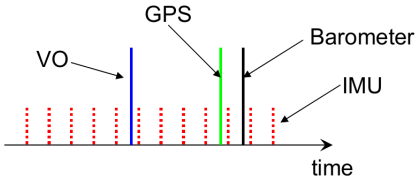

Figure 8. A ring buffer is utilized to keep 2 s of incoming sensor information. Due to the image transmission and Visual Odometry (VO) calculation delay, the newly-arriving VO measurements do not have the newest time stamp. The VO will update the EKF state on the VO time, and the subsequent IMU integral is re-integrated to predict the current state.

Remark 3. *After an absolute measurement update of the EKF state by the GPS or barometer, both the current pose covariance* $\Sigma^2_p, \Sigma^2_\theta$ *and the delayed pose covariance* $\Sigma^2_{p_l}, \Sigma^2_{\theta_l}$ *are decreased. Furthermore the current pose covariance* $\Sigma^2_p, \Sigma^2_\theta$ *should be higher than the delayed pose covariance* $\Sigma^2_{p_l}, \Sigma^2_{\theta_l}$*. Otherwise, the VO relative measurement will lose the ability to update of the EKF state.*

5. Results

5.1. Experimental System

In this section, we present the experimental results for the proposed technique. Figure 9a shows the MAV developed by our team for the state estimation experiments; some additional sensors, including a commercial GPS, stereo camera, an IMU (Microstrain 3DM-GX3-35 [27]) a barometer and a spinning laser scanner, were carried onboard by the MAV for dataset gathering. All the sensors are hardware triggered, so the timestamps for different onboard sensors have the same time reference. The IMU was recorded at 100 Hz; the barometer rate was 7 Hz; GPS was 4 Hz; and stereo was 10 Hz. For the forward facing stereo (about -15° pitch) with a 0.41 m static baseline, its effective short-range measurement is 13.41 m with a 10-pixel stereo disparity. The datasets are gathered in Fort Indiantown Gap, Pennsylvania. Figure 9b shows the experimental scenario from Google Earth and the schematic flight trajectory for the MAV. For the dataset, the MAV flies 12 min with aggressive six D.O.F motions. Some topical scenarios for this experiments are shown in Figure 10.

Figure 9. Our MAV for the dataset recording and experimental environment. (**a**) MAV developed by our team for the state estimation experiments; (**b**) experimental scenario from Google Earth and the schematic flight trajectory for the MAV.

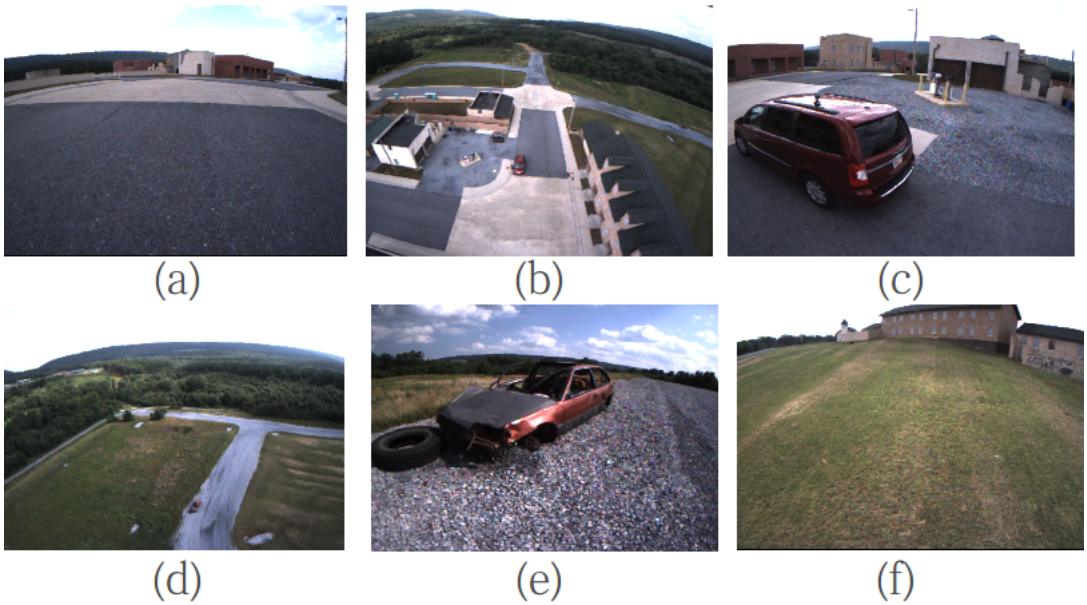

Figure 10. Some typical images from the experiment: (**a**) the MAV taking off; (**b**,**d**) the high-altitude flight; (**c**,**e**) low-altitude flight to capture the images for the two cars' stereo reconstruction; and (**f**) the similar texture environment.

5.2. Performance of IMU Tightly-Coupled Long-Range Stereo Odometry

We first utilized the EuRoC (European Robotics Challenge) dataset [28] to evaluate the proposed long-range stereo VO. Then, we compared the performance of long-range VO with that of the state-of-the-art sparse feature ORBSLAM 2.0. For fair comparison, the loop-closing detection thread and global bundle adjustment for ORBSLAM2.0 are deactivated. The EuRoc dataset contains nine stereo-IMU datasets recorded by a quadrocopter in three different indoor environments. We utilize the six datasets in VICON [29] room due to the six D.O.F ground truth being provided for the evaluation. The proposed long-range VO and long-range VO tightly coupled with IMU and ORBSLAM 2.0 were tested and compared. In the experiments, we recorded the odometry outputs, including timestamp, position and orientation quaternion. Furthermore, the VICON data are used as the corresponding ground truth. Relative Pose Error (RPE) is used as the evaluation measure. For the IMU integral, the first one second of the IMU dataset is filtered and used for calculating the

initial roll, pitch and gravity acceleration with respect to the inertial frame. Figure 11 shows the 3D environments reconstructed by our stereo VO for the "room-1" dataset.

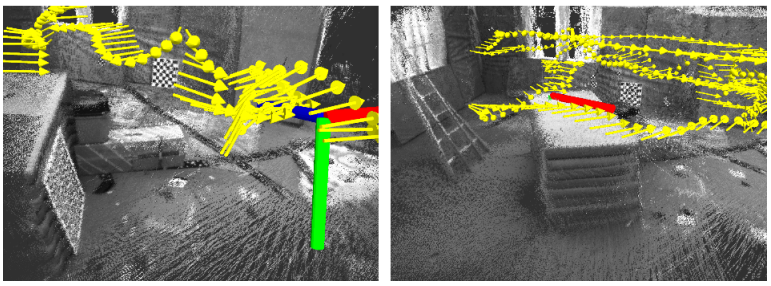

Figure 11. 3D environments reconstructed by our stereo VO for the EuRoC (European Robotics Challenge) "room-1" dataset.

Figure 12 gives the RPE evaluation results. From the results, the proposed VO shows similar performance as the ORBSLAM 2.0. For the last dataset (room 2-difficult), all three approaches failed to track the pose at 67 s due to the serious image motion blur and low texture (so, we do not report the result for the last dataset). The VICON room datasets were recorded in relatively small indoor environments; the proposed VO works with pure stereo mode for most situations. Furthermore, the vehicle flies around the room several times for each dataset, so ORBSLAM 2.0 utilizes the built map to localize the vehicle, namely with the localization mode. In comparison, our VO is a sliding-window VO for the onboard memory and computing resource consideration; only some key-frames and their corresponding map points are kept for pose tracking and local BA. As a result, our VO works with exploration mode, which is more difficult than localization mode.

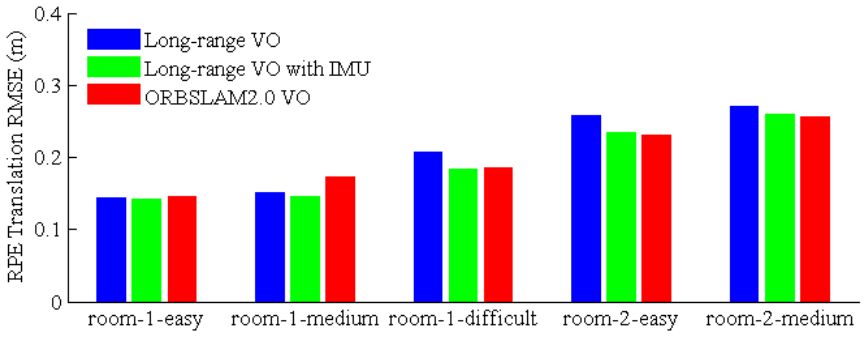

Figure 12. Relative Pose Error (RPE) evaluation results.

To test the long-range performance, the dataset recorded by our MAV is utilized. For this experiment, GPS and the barometer were utilized as the ground truth. The results are shown in Figure 13. For ORBSLAM2.0 VO, it easily fails to track aggressive MAV motion, so we also integrate the same IMU tightly-coupled strategy (Section 3.5) in ORBSLAM 2.0 VO. The altitude estimation results are shown in Figure 14; ORBSLAM 2.0 VO fails at 400 s of the dataset. At this time, the MAV altitude is sharply increasing, so ORBSLAM 2.0 VO cannot deal with this long-range high-altitude case. The path RMSE is listed in Table 1. The first row is long-range VO RMSE before the ORBSLAM 2.0 VO fails; the last row is the path RMSE for the entire MAV 12-min flights.

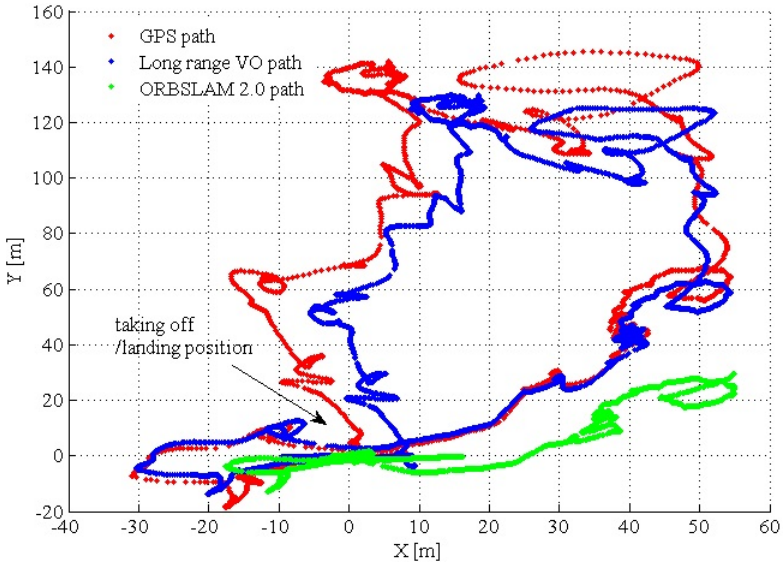

Figure 13. Comparison of the GPS path, proposed long-range VO path and ORBSLAM 2.0 path. The MAV takes off at position (0,0). At this position, we recorded the initial GPS and barometric measurements as the offsets for the subsequence ground truth measurements. The initial MAV heading is aligned with the NED coordinate. ORBSLAM 2.0 (also with IMU tightly coupled) fails at 400 s of the dataset.

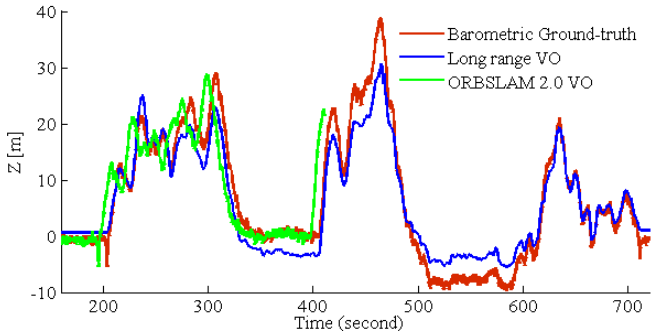

Figure 14. Comparison of long-range VO path and ORBSLAM 2.0 path in the altitude direction. ORBSLAM 2.0 fails at 400 seconds of the dataset. At this time, the MAV altitude increases sharply, so ORBSLAM 2.0 cannot deal with this long-range high-altitude case.

Table 1. Comparisons of path RMSE (m).

Method	RSME x	RMSE y	RMSE z
Long-range VO	1.4936	3.0465	2.2860
ORBSLAM 2.0 VO	5.0012	21.1514	3.3277
Long-range VO	5.8547	7.6728	4.5409

5.3. Performance of Multi-Sensor Fusion State Estimation

With the same dataset, we tested the multi-sensor fusion state estimation performance. The result of MAV path estimation is shown in Figure 15. In this experiment, we utilized the VO pseudo absolute measurement update of EKF state [4,19] for the first 220 s. Then, the VO measurement is switched to relative mode, as discussed in Section 4.3. The purpose of using pseudo absolute measurement for the first 220 s is two-fold: (1) the proposed long-range VO can provide absolute pose measurement w.r.t. local map; also, the VO absolute pose drifts slowly over time,

as the result shown in Section 5.1; it can be used for short-term absolute update of EKF state for the initial phase; (2) since the VO absolute update is used for the first 220 s, the system estimation will be overconfident. Other absolute measurement sensors with larger uncertainty than that of VO absolute measurement will be prohibited for system state estimation. This experiment will show this effect by comparing the MAV path estimation consistency.

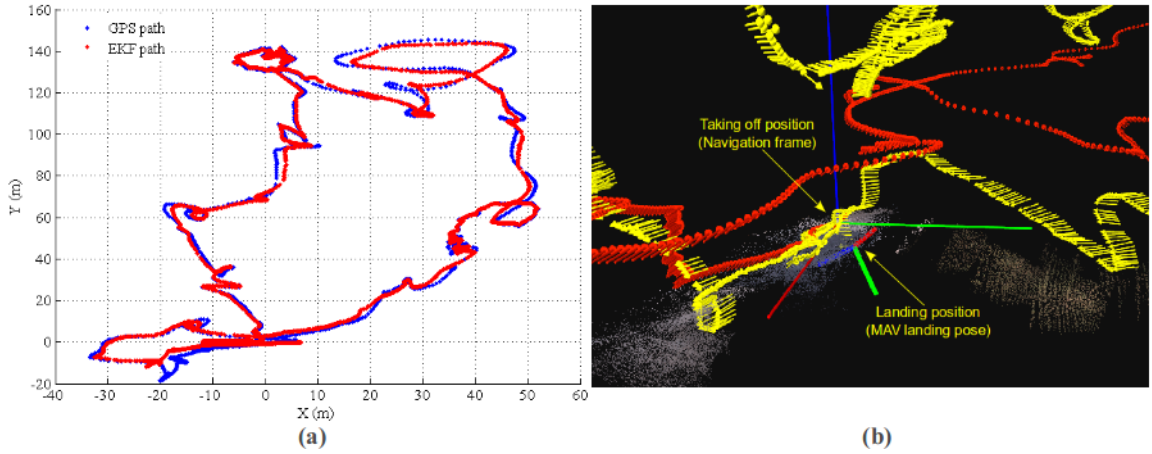

Figure 15. EKF state estimation result. (**a**) EKF estimation path in the X-Y plane; the MAV takes off at (0,0) and comes back after 12 min of flight; (**b**) local details for EKF state estimation. The yellow arrow sequence is the 3D path from EKF, and the red sequence is the GPS path. Clearly, GPS reports bad positioning information in the altitude direction.

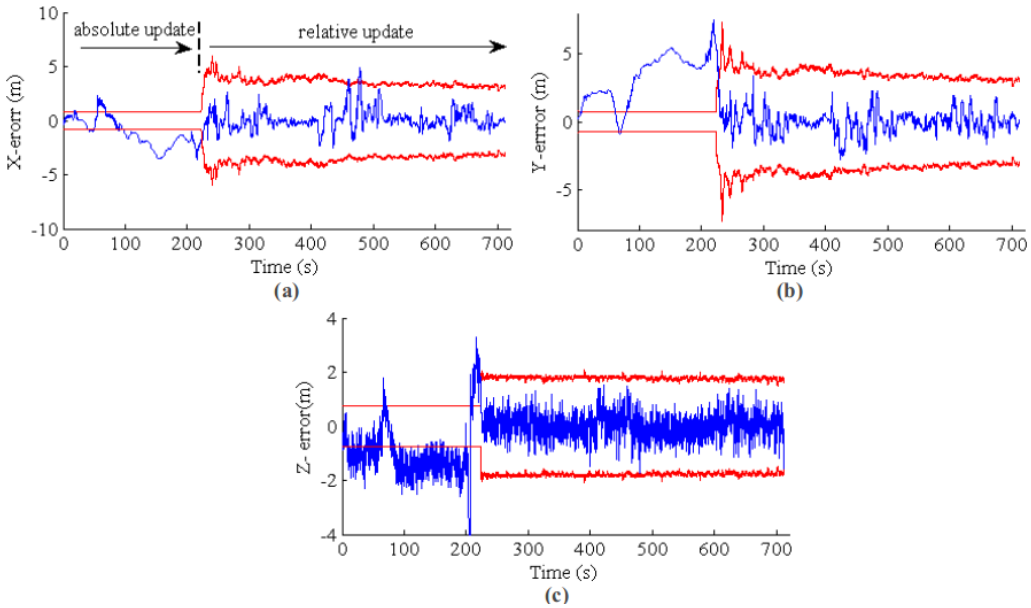

Figure 16. EKF state estimation error and consistency. (**a**–**c**) The results for X, Y and Z, respectively. The blue line is the position error, and the red line is the 3σ error band from the state covariance. Because VO absolute measurement is fused in the initial phase, the state estimation is overconfident and inconsistent. Furthermore, the GPS and barometric measurements are prohibited from updating the EKF state. After the VO is switched to the relative measurement model, the EKF state becomes consistent, and the error is bounded in the 3σ band.

Figure 16 plots the state estimation results for MAV position. The blue line is the EKF positioning error, and the red line is the 3σ error band from EKF state covariance. Because VO absolute

measurement is fused in the initial phase, the system covariance from EKF almost kept constant (or a little bit decreasing), and the state estimation is inconsistent compared with the GPS (and barometric) ground truth. In fact, the position error in the X and Ydirections are mainly from the inaccurate initial MAV heading w.r.t. the NED coordinate. However, the initial heading cannot be corrected because GPS is prohibited from updating the EKF state. After the VO is switched to the relative model, the EKF becomes consistent, and the estimation error is bounded in the 3σ error band.

5.4. Performance of GPS Outage Situations

A natural question for multi-sensor fusing is "why do we fuse VO for state estimation?". In this section, the VO information fusion performance for GPS outage situations is investigated. To simulate GPS outage, we do not use GPS for the update of the EKF state in some time periods. Furthermore, the state estimation performance with and without VO fusing are compared. In the first group of experiments, GPS is deactivated for 1 min (GPS lost at 300 s and recovered at 360 s). For the second group of experiments, the GPS outage is relatively long term (about 120 s); we turned off the GPS update at 500 s and never activated it again. The results for the first group of experiments are shown in Figures 17 and 18 for GPS outage with/without VO fusing, respectively. In Figure 17, when the GPS was lost at 300 s, the system fuses the IMU, barometer and VO relative measurement. As expected, the error covariance is slowly increased by fusing VO relative motion information. Furthermore, the state estimation is still consistent and accurate. By comparison, both of error and error covariance are sharply increased without VO fusing. The main reason for the fast drift without VO lies in the noisy IMU measurement. The IMU signal is corrupted by vibration when the MAV motors are powered, especially for the acceleration. Accordingly, the system velocity smooth estimation can only be kept for a short time and will drift quickly by the accumulating IMU acceleration noise. For the altitude direction, both with and without VO can report reasonable results since barometric absolute measurement is provided.

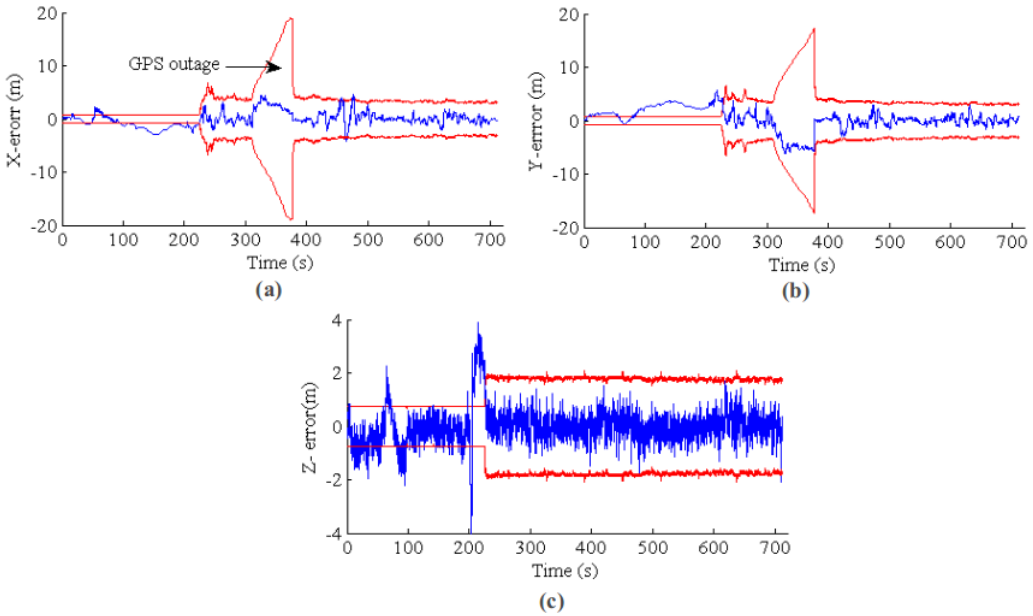

Figure 17. EKF state estimation for GPS outage from 300 s to 360 s. The blue line is the estimation error, and the red line is the 3σ error band. As expected, the error covariance is slowly increased by fusing VO relative information. Furthermore, the state estimation is consistent and accurate. (**a**–**c**) The results for X, Y and Z, respectively.

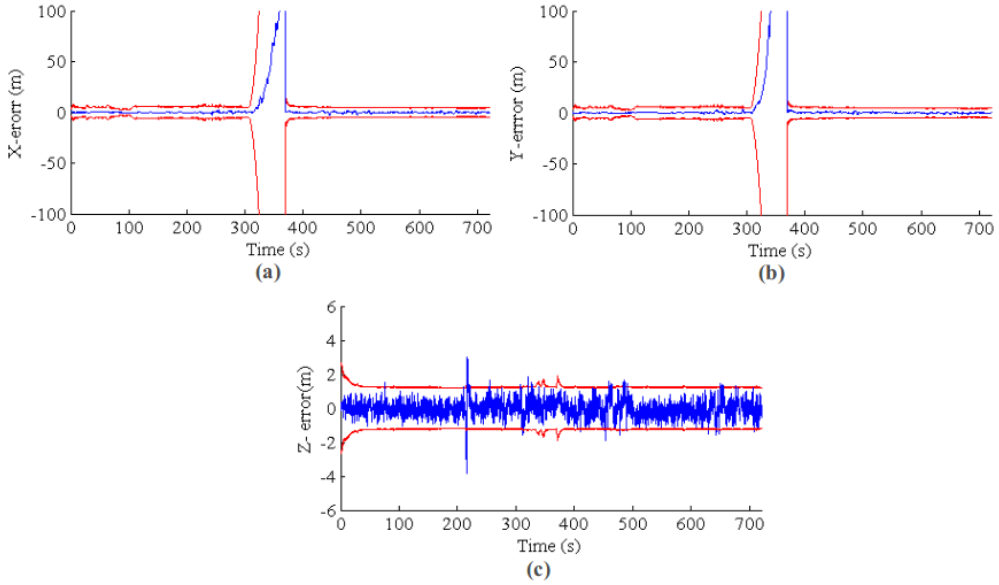

Figure 18. EKF state estimation without the VO update for GPS outage from 300 s to 360 s. Both the error and error covariance are sharply increasing due to the noisy IMU integral. (**a**–**c**) The results for X, Y and Z, respectively. The blue line is the position error, and the red line is the 3σ error band from the state covariance.

The results for the second group of experiments are shown in Figures 19 and 20 for GPS outage with and without VO fusion, respectively. The results are similar to that of the first experiments. For the state estimation with VO measurements, the state error and covariance are increasing smoothly; while the state estimation information becomes unusable without a VO relative update. Figure 21 shows the estimated paths for with and without VO measurements update, respectively. It is clear that the EKF state estimator with VO update can generate a smooth, low drift result after the GPS is lost. For quantitative comparison, the RMSEs for the two group experiments are listed in Table 2.

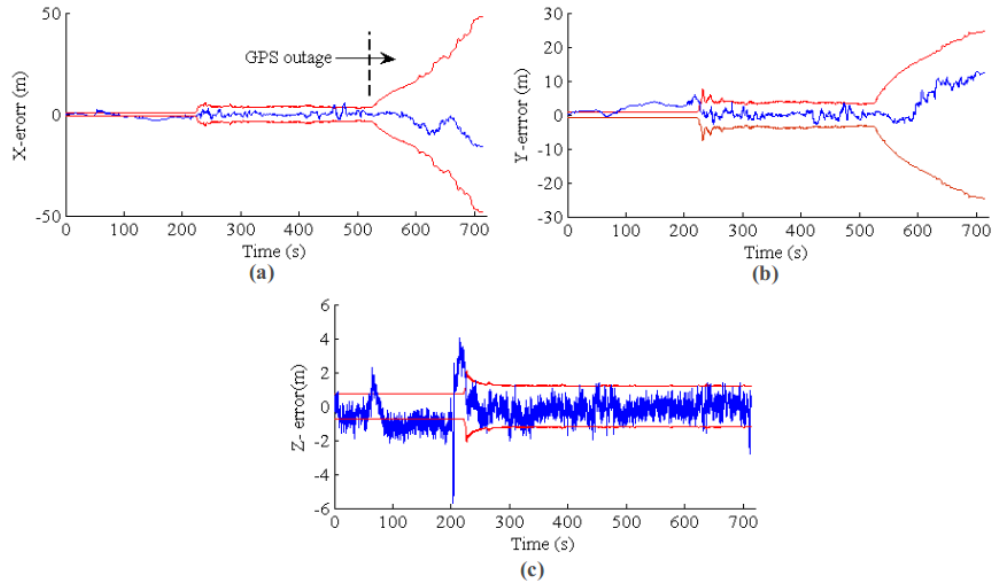

Figure 19. EKF state estimation for a long-term GPS outage from 500 s to the end. The blue line is the estimation error, and the red line is the 3σ error band. (**a**–**c**) The results for X, Y and Z, respectively.

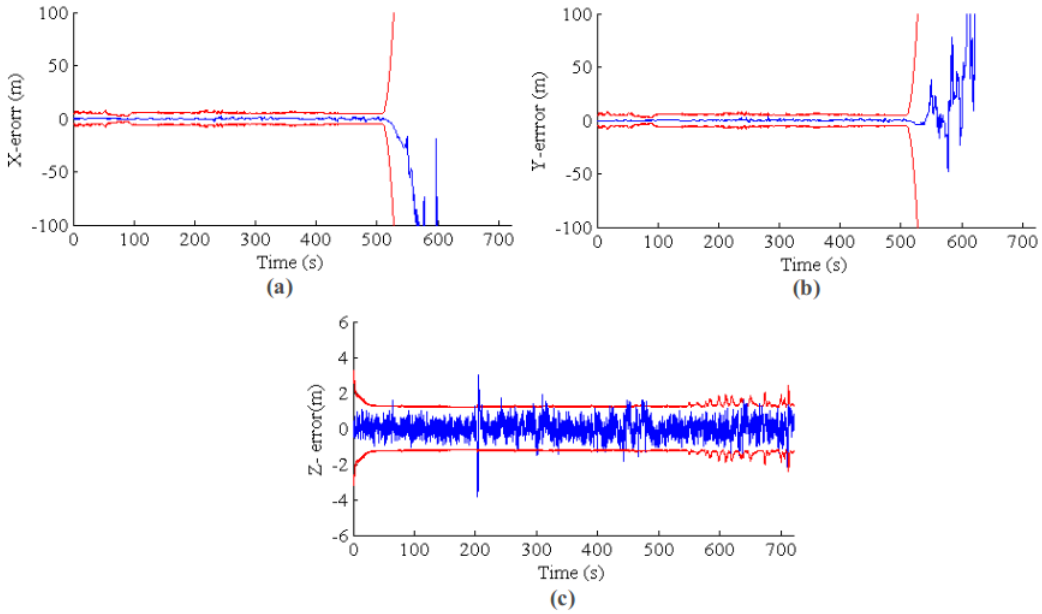

Figure 20. EKF state estimation without VO update for GPS outage from 500 s to the end. Both of the error and error covariance are sharply increased, and state estimation becomes unusable. (**a**–**c**) The results for X, Y and Z, respectively. The blue line is the position error, and the red line is the 3σ error band from the state covariance.

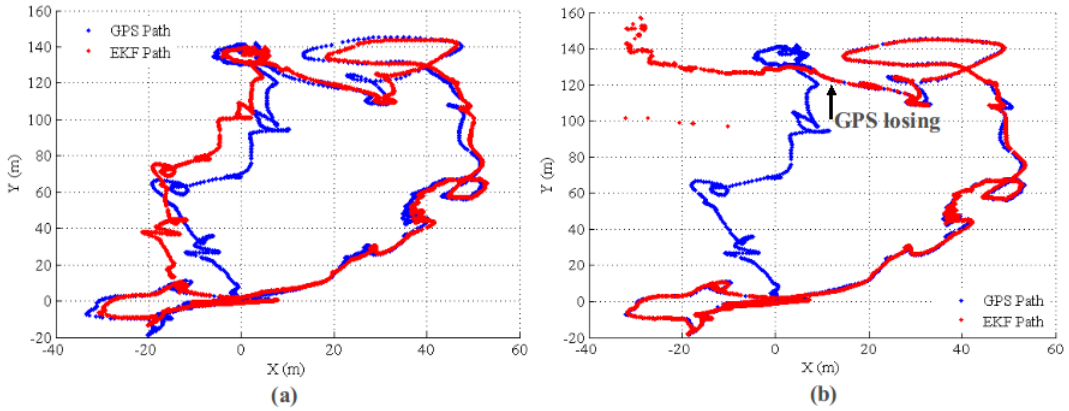

Figure 21. Estimated paths with/without VO update. The GPS outage starts at 500 s. (**a**) Path with VO relative update; (**b**) path without VO relative update.

Table 2. Comparisons of path RMSE (m) for GPS outage.

Method	RSME x	RMSE y	RMSE z
EKF (300 s to 360 s, GPS lost)	1.3782	2.2670	0.5859
EKF without VO (300 s to 360 s, GPS lost)	19.8595	66.4899	0.6047
EKF (500 s to 720 s, GPS lost)	3.5654	3.8767	0.5535
EKF without VO (500 s to 720 s, GPS lost)	595.9539	141.5476	0.5973

5.5. Performance of Timing

All of the experiments were performed offline using the recorded dataset for detailed performance analysis. We have tested the "timing performance" for the state estimation system both on a desktop PC (Intel i7, 2.8 GHZ) and small-sized onboard ARM-based computers (NVIDIA Jetson TX1 [30] and low-cost Odroid XU4 [31]).

The EKF can run in real time both for a desktop PC and two kinds of onboard ARM-based computers. The EKF algorithm is currently developed for ROS (Robot Operation System). It only requires 2 to 3 ms (including state propagation, update, rolling-buffer recalculation and state publication) even for the slowest low-cost Odroid XU4. The long-range stereo VO can run 20 Hz on the desktop PC (so, the proposed VO can run in real time if the Intel i7 onboard computer is utilized for MAV), 7 Hz on the NVIDIA Jetson TX1 onboard computer and 5 Hz on the slowest low-cost Odroid XU4. The most time-consuming part for stereo VO is feature detection, descriptor extraction and stereo matching. For the Intel i7 computer, it requires about 28 ms. In comparison, for the slowest Odroid XU4, it requires about 144 ms (five-times slower), because the current stereo VO implementation uses some functions from OpenCV, which is much slower than that running on the Intel i7 computer. Next, we will focus on code optimization to reduce stereo VO time consumption on the low-cost Odroid XU4.

6. Conclusions

In this paper, we present an EKF multi-sensor state estimator by fusing long-range VO, GPS, IMU and a barometer for MAV navigation in both GPS-available and GPS-denied environments.

Firstly, we derived a new long-range stereo VO. The main reason for using the stereo rather than monocular lies in the absolute scale that can be directly provided, while the performance of stereo VO highly depends on the ratio between the stereo baseline and the environmental depth. For high-altitude flights, stereo generally degenerates to monocular, making it ineffective for new feature depth generation. To explore this problem, we discussed a long-range stereo VO technique. On the basis of the current baseline-depth ratio, the odometry switches the working mode between short range and long range. For long range, the stereo almost degenerates to monocular, but the stereo "weak static baseline" can still provide useful physical scale information both for new map point generation and VO calculation. The new feature depth is estimated by introducing additional stereo observations through time. The performance of long-range VO was evaluated using both EuRoC datasets and our own dataset, results showing that the proposed VO improves the performance for long-range environments.

Secondly, a new state estimation system is derived to fuse both absolute state measurement sensors (GPS, barometer) and the relative 6 D.O.F pose state measurement provided by long-range VO. The EKF estimator and long-range visual estimation help each other to improve the robustness of the method. The IMU integral prediction from the EKF estimator is used both for guiding image-feature matching and long-range VO optimization. Additionally, the VO is utilized as the relative measurements for the update of the EKF state, especially for the GPS outage situations. To our best knowledge, the proposed system is the first EKF estimator for fusing both relative measurements (VO, IMU) and absolute measurements (GPS, barometer) with different time stamps. The performance of the proposed EKF estimator is investigated and compared. Results verified the effectiveness of the state estimation system.

Supplementary Materials: The experimental video for EKF fusion estimation is available online at https://www.youtube.com/watch?v=LYszVoEboWY.

Acknowledgments: This research is supported by the Office of Naval Research (Grant No. N00014-14-1-0693), the National Science Foundation (Grant No. IIS-1328930), the National National Science Foundation of China (Grant No. 61573053), the Beijing Natural Science Foundation (Grant No. 4162501) and the State Key Laboratory of Robotics and System (Grant No. SKLRS-2016-KF-02). We acknowledge Geetesh Dubey, Daniel Maturana, Sezal Jain, Sankalp Arora and Shichao Yang for the MAV system design and suggestion. We acknowledge the anonymous reviewers for their helpful suggestions and comments.

Author Contributions: Yu Song contributed to the theory research, the design of experiments and writing. Sebastian Scherer and Stephen Nuske contributed to scientific advising and proof reading.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Warren, M.; Corke, P.; Upcroft, B. Long-range stereo visual odometry for extended altitude flight of unmanned aerial vehicles. *Int. J. Robot Res.* **2015**, *35*, 381–403.
2. Nister, D.; Naroditsky, O.; Bergen, J. Visual odometry. In Proceedings of the 2004 IEEE International Conference on Computer Vision and Pattern Recognition, Washington, DC, USA, 27 June–2 July 2004; pp. 652–659.
3. Klein, G.; Murray, D. Parallel tracking and mapping on a camera phone. In Proceedings of the 2009 IEEE International Symposium on Mixed and Augmented Reality, Orlando, FL, USA, 19–22 October 2009; pp. 83–86.
4. Weiss, S.; Achtelik, M.W.; Lynen, S. Monocular vision for long-term micro aerial vehicle state estimation: A compendium. *J. Field Robot* **2013**, *30*, 803–831.
5. Forster, C.; Pizzoli, M.; Scaramuzza, D. SVO: Fast semi-direct monocular visual odometry. In Proceedings of the 2014 IEEE International Conference on Robotics and Automation, Hong Kong, China, 31 May–7 June 2014; pp. 15–22.
6. Engel, J.; Sturm, J.; Cremers, D. Semi-dense visual odometry for a monocular camera. In Proceedings of the 2013 IEEE International Conference on Computer Vision, Sydney, Australia, 1–8 December 2013; pp. 1449–1456.
7. Engel, J.; Schops, T.; Cremers, D. LSD-SLAM: Large-scale direct monocular SLAM. In Proceedings of the 2014 European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; pp. 834–849.
8. Artal, R.M.; Montiel, J.M.M.; Tardos, J.D. ORB-SLAM: A versatile and accurate monocular SLAM system. *IEEE Trans. Robot* **2015**, *31*, 1147–1163.
9. Kerl, C.; Sturm, J.; Cremers, D. Robust odometry estimation for RGB-D cameras. In Proceedings of the 2013 IEEE International Conference on Robotics and Automation, Karlsruhe, Germany, 6–10 May 2013; pp. 3748–3754.
10. Fang, Z.; Scherer, S. Experimental study of odometry estimation methods using RGB-D cameras. In Proceedings of the 2014 IEEE/RSI International Conference on Intelligent Robots and Systems, Chicago, IL, USA, 14–18 September 2014; pp. 680–687.
11. Zhang, J.; Kaess, M.; Singh, S. A real-time method for depth enhanced visual odometry. *Auton Robot*. **2015**, 1–13, doi:10.1007/s10514-015-9525-1.
12. Geiger, A.; Ziegler, J.; Stiller, C. StereoScan: Dense 3d reconstruction in real-time. In Proceedings of the 2011 IEEE International Symposium on Intelligent Vehicles, Baden, Germany, 5–9 Jun 2011; pp. 963–968.
13. Huang, A.S.; Henry, P.; Fox, D. Visual odometry and mapping for autonomous flight using an RGB-D camera. In Proceedings of the 2011 International Symposium of Robotics Research, Flagstaff, AZ, USA, 28 August– 1 September 2011; pp. 235–252.
14. Pire, T.; Fischer, T.; Civera, J. Stereo parallel tracking and mapping for robot localization. In Proceedings of the 2015 IEEE/RSI International Conference on Intelligent Robots and Systems, Hamburg, Germany, 28 September–2 October 2015; pp. 1373–1378.
15. Engel, J.; Stueckler, J.; Cremers, D. Large-scale direct SLAM with stereo cameras. In Proceedings of the 2015 IEEE/RSI International Conference on Intelligent Robots and Systems, Hamburg, Germany, 28 September–2 October 2015; pp. 1935–1942.
16. Cvisic, I.; Petrovic, I. Stereo odometry based on careful feature selection and tracking. In Proceedings of the 2015 European Conference on Mobile Robots, Lincoln, UK, 2–4 September 2015; pp. 1–5.
17. Kuemmerle, R.; Grisetti, G.; Strasdat, H.; Konolige, K.; Burgard, W. G2o: A general framework for graph optimization. In Proceedings of the 2011 IEEE International Conference on Robotics and Automation, Shanghai, China, 9–13 May 2011; pp. 3607–3613.
18. Kaess, M.; Johannsson, H.; Roberts, R. iSAM2: Incremental smoothing and mapping using the Bayes tree. *Int. J. Robot Res.* **2012**, *31*, 216–235.
19. Weiss, S. Vision Based Navigation for Micro Helicopters. Ph.D. Thesis, Eidgenössische Technische Hochschule (ETH), Zurich, Switzerland, 2012.
20. Chambers, A.; Scherer, S.; Yoder, L.; Jain, S.; Nuske, S.; Singh, S. Robust multi-Sensor Fusion for micro aerial vehicle navigation in GPS-degraded/denied environments. In Proceedings of the 2014 American Control Conference, Portland, OR, USA, 4–6 June 2014; pp. 1892–1899.

21. Mourikis, A.; Roumeliotis, S. On the treatment of relative-pose measurements for mobile robot localization. In Proceedings of the 2006 IEEE International Conference on Robotics and Automation, Orlando, FL, USA, 15–19 May 2006; pp. 2277–2284.
22. Tardif, J.; George, M.; Laverne, M.; Kelly, A. A new approach to vision-aided inertial navigation. In Proceedings of the 2010 IEEE/RSI International Conference on Intelligent Robots and Systems, Taipei, Taiwan, 18–22 October 2010; pp. 4161–4168.
23. Li, M.Y. High-precision, consistent EKF-based visual-inertial odometry. *Int. J. Robot Res.* **2012**, *32*, 690–711.
24. Leutenegger, S.; Lynen, S.; Bosse, M.; Siegwart, R.; Furgale, P. Keyframe-based visual-inertial slam using nonlinear optimization. *Int. J. Robot Res.* **2014**, *34*, 314–334.
25. Forster, C.; Carloney, L.; Dellaerty, F.; Scaramuzza, D. IMU preintegration on manifold for efficient visual-inertial maximum-a-posteriori estimation. In Proceedings of the 2015 Robotics: Science and Systems, Rome, Italy, 13–17 July 2015; pp. 1–20.
26. Strasdat, H. Local Accuracy and Global Consistency for Efficient Visual SLAM. Ph.D. Thesis, Imperial College London, London, UK, 2012.
27. Microstrain 3DM-GX3-35. LORD MicroStrain Ltd. Available online: http://www.microstrain.com/inertia/3dm-gx3-35 (accessed on 20 December 2016).
28. VICON. Vicon Motion Systems Ltd. Available online: http://www.vicon.com (accessed on 20 December 2016).
29. Burri, M.; Nikolic, J.; Gohl, P.; Schneider, T.; Rehder, J.; Omari, S.; Achtelik, M.; Siegwart, R. The EuRoC micro aerial vehicle datasets. *Int. J. Robot Res.* **2016**, *35*, 1157–1163.
30. NVIDIA Jetson TX1. Nvidia Ltd. Available online: http://www.nvidia.com/object/jetson-tx1-module.html (accessed on 20 December 2016).
31. Odroid XU4. Hardkernel Ltd. Available online: http://www.hardkernel.com/main/main.php (accessed on 20 December 2016).