2020 IEEE International Conference on Robotics and Automation (ICRA)
31 May - 31 August, 2020. Paris, France

Stereo Visual Inertial Odometry with Online Baseline Calibration

Yunfei Fan, Ruofu Wang, and Yinian Mao

Abstract— **Stereo-vision devices have rigorous requirements for extrinsic parameter calibration. In Stereo Visual Inertial Odometry (VIO), inaccuracy in or changes to camera extrinsic parameters may lead to serious degradation in estimation performance. In this manuscript, we propose an online calibration method for stereo VIO extrinsic parameters correction. In particular, we focus on Multi-State Constraint Kalman Filter (MSCKF [1]) framework to implement our method. The key component is to formulate stereo extrinsic parameters as part of the state variables and model the Jacobian of feature reprojection error with respect to stereo extrinsic parameters as sub-block of update Jacobian. Therefore we can estimate stereo extrinsic parameters simultaneously with inertial measurement unit (IMU) states and camera poses. Experiments on EuRoC dataset and real-world outdoor dataset demonstrate that the proposed algorithm produce higher positioning accuracy than the original S-MSCKF [2], and the noise of camera extrinsic parameters are self-corrected within the system.**

I. INTRODUCTION

In recent years, high-precision positioning technologies have progressed significantly, propelling the advancements in multiple application scenarios such as autonomous driving, robotics and unmanned aerial vehicles (UAVs), and augmented and virtual reality (AR and VR). In outdoor environments, GNSS such as GPS and RTK can be employed. In indoor and GPS-denied environments, Lidar and visual SLAM can be used. For applications that are limited by device size and weight requirements, the applicable positioning technology is rather limited in the absence of GPS. Since VIO only requires IMU and one or two camera modules to estimate ego-motion, it is naturally suitable for such scenarios. It has been reported that stereo-vision VIO system can improve the overall estimation accuracy over single-vision VIO system (S-MSCKF [2], VINS-Fusion [3,4]). A good stereo calibration ensures the epipolar lines of stereo images being parallel, which is the foundation for most stereo matching algorithms. However, in stereo VIO systems, the estimation accuracy heavily depends on camera extrinsic parameters calibration. With a poor calibration or slight changes in camera parameters during operation, stereo VIO positioning accuracy will drop sharply. Even with rigid and bulky frames, most stereo cameras cannot ensure that extrinsic parameters are unchanged during long course of operations. Within this context, an accurate calibration algorithm that is robust to changes in camera extrinsic parameters is highly desired.

In this paper, we propose a stereo VIO algorithm with online calibration to overcome the above issues. The core method is to formulate stereo camera extrinsic parameters (rotation and translation) into the set of state variables and model the relevance between feature reprojection error and stereo extrinsic parameters in update Jacobian, so that the stereo extrinsic parameters can be calibrated online as part of the state estimation. To accelerate the self-calibration process, the initial covariance of stereo extrinsic parameters is set to a large value. In addition, during the initial phase of the estimation, the threshold of the outlier rejection rule based on stereo extrinsic constraint on the algorithm frontend is relaxed to avoid too many inliers being mistakenly taken out.

Using EuRoC dataset and real-world outdoor dataset, we compare the proposed scheme with other state-of-the-art stereo VIO algorithm, specifically S-MSCKF. The experiments show that, without calibration errors, the proposed method performs similarly to S-MSCKF. Besides, when artificial noises are involved in the calibrated parameters, the proposed scheme can achieve rapid self-calibration and outperforms S-MSCKF in position estimation.

The rest of this paper is organized as follows: Section II introduces related works. Section III introduces system framework and derives analytical formulations. Section IV compares experimental results of the proposed scheme with those of VINS-Fusion [3,4] and S-MSCKF using EuRoC dataset and real-world outdoor datasets collected by UAVs as well as by a handheld device. Finally, the conclusions are summarized in section V.

II. RELATED WORK

The current scholarly works in VIO could be roughly divided into loosely-coupled [5,6] and tightly-coupled [1- 4,7] methods. Tightly-coupled methods put IMU information into state variables and optimize with vision information simultaneously, which is a mainstream direction currently. Tightly-coupled methods can be divided further into filter-based and optimization-based.

VIO methods based on non-linear optimization utilize all measurements, including IMU measurements and visual measurements, to find the optimal state variables to minimize the measurement error. Stereo VIO based on non-linear optimization includes OKVIS [7], VINS-Fusion [3,4], etc. Both OKVIS and VINS-Fusion perform online estimation of extrinsic parameters between the IMU and each camera, separately. However, due to the large number of state variables, even the current mainstream sliding window based VIO methods using non-linear optimization have a considerable demand for computational resource, and it is still difficult to run in real time on embedded platforms.

Filter based VIO methods are mainly based on Extended Kalman Filter (EKF) [1]. Generally, IMU is used for prediction, while visual information is used for update. They achieve almost the same level of accuracy as optimiza-

This work was supported by the Meituan-Dianping Group.

Yunfei Fan and Yinian Mao are with the Meituan-Dianping Group, Beijing, China (e-mail: {fanyunfei | maoyinian}@meituan.com).

Ruofu Wang is with University of Southern California, Los Angeles, CA 90007 USA (e-mail: ruofuwan@usc.edu)

978-1-7281-7395-5/20/$31.00 ©2020 IEEE

tion-based methods using relatively low computational resources. Thus they can run in real time on embedded platforms. S-MSCKF [2] is one of filter-based stereo VIO frameworks, which only estimates the extrinsic parameters between IMU and left camera online. In order to achieve real-time performance, our method is also based on MSCKF framework.

P Hansen et al. [8] and Yonggen Ling et al. [9] proposed approaches to estimate stereo extrinsic parameters online. They are all based on epipolar geometric constraints for online self-calibration of stereo extrinsic. However, because pure vision-based methods cannot self-calibrate the baseline fully, they can only achieve estimation of stereo extrinsic parameters with 5-DOF, while the length of the baseline cannot be estimated. Therefore, we use the IMU and the cameras jointly, to self-calibrate the 6-DOF stereo extrinsic parameters online.

III. MSCKF ALGORITHM FRAMEWORK

A. State definition

Following the definition of MSCKF in [1], IMU state is defined below:

$$X_I = \left[{}^{G}p_I{}^{T} \; {}^{G}v_I{}^{T} \; {}^{I}_{G}\bar{q}^{T} \; b_a{}^{T} \; b_g{}^{T} \right]^{T} \tag{1}$$

In this paper, different from [1, 2], both extrinsic parameters E_0 and E_1 of stereo VIO system shown in Fig. 1, are added into IMU states and calibrated online. The extended IMU states are defined:

$$X_I = \left[{}^{G}p_I{}^{T} \; {}^{G}v_I{}^{T} \; {}^{I}_{G}\bar{q}^{T} \; b_a{}^{T} \; b_g{}^{T} \; {}^{C^0}_{I}\bar{q}^{T} \; {}^{I}p_{C^0}{}^{T} \; {}^{C^1}_{C^0}\bar{q}^{T} \; {}^{C^0}p_{C^1}{}^{T} \right]^{T} \tag{2}$$

In these expressions, $\{G\}$ and $\{I\}$ are the global and inertial frame respectively, $\{C^0\}$ and $\{C^1\}$ are frame of C^0 and C^1 respectively. ${}^{G}p_I$ and ${}^{G}v_I$ are position and velocity of IMU expressed in $\{G\}$, respectively. 4×1 ${}^{I}_{G}\bar{q}$ represents the rotation from $\{G\}$ to $\{I\}$ (in this paper, quaternion obeys JPL rules). The vectors b_g and b_a are the biases of the measured angular velocity and linear acceleration from the IMU, separately. ${}^{C^0}_{I}\bar{q}$ represents the rotation from $\{I\}$ to $\{C^0\}$, and ${}^{I}p_{C^0}$ is the position of C^0 based on frame $\{I\}$ (${}^{C^0}_{I}\bar{q}$ and ${}^{I}p_{C^0}$ are the rotation and translation of extrinsic parameter E_0 respectively). Finally, ${}^{C^1}_{C^0}\bar{q}$ represents the rotation from frame $\{C^0\}$ to frame $\{C^1\}$, and ${}^{C^0}p_{C^1}$ is the position of C^1 based on frame $\{C^0\}$ (${}^{C^1}_{C^0}\bar{q}$ and ${}^{C^0}p_{C^1}$ are the rotation and translation of stereo extrinsic parameter E_1 respectively. We treat stereo extrinsic parameters as ${}^{C^1}_{C^0}\bar{q}$ and ${}^{C^0}p_{C^1}$ later).

The EKF error-state of X_I is defined accordingly:

$$\tilde{X}_I = \left[{}^{G}\tilde{p}_I{}^{T} \; {}^{G}\tilde{v}_I{}^{T} \; {}^{I}_{G}\tilde{\theta}^{T} \; \tilde{b}_a{}^{T} \; \tilde{b}_g{}^{T} \; {}^{C^0}_{I}\tilde{\theta}^{T} \; {}^{I}\tilde{p}_{C^0}{}^{T} \; {}^{C^1}_{C^0}\tilde{\theta}^{T} \; {}^{C^0}\tilde{p}_{C^1}{}^{T} \right]^{T} \tag{3}$$

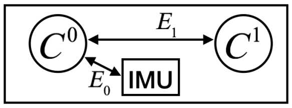


Figure 1. Structure diagram of sensor, and definition of extrinsic parameters

Except for quaternions, other states can be used with standard additive error (e.g. $x = \hat{x} + \tilde{x}$). the extended additive error of quaternion is defined in [10] (in this paper, quaternion error is defined in frame $\{I\}$, see details in [11])

$${}^{I}_{G}\bar{q} = \delta{}^{I}_{G}\bar{q} \otimes {}^{I}_{G}\hat{\bar{q}}, \;\; \delta{}^{I}_{G}\bar{q} = \left[1 \; \tfrac{1}{2}{}^{I}_{G}\tilde{\theta}\right]^{T} \tag{4}$$

similarly, the extended additive error of rotation matrix is defined:

$$R\left({}^{I}_{G}\bar{q}\right) = {}^{I}_{G}R, \;\; {}^{I}_{G}R = (1 - \left[{}^{I}_{G}\tilde{\theta}\right]_{\times}){}^{I}_{G}\hat{R} \tag{5}$$

B. State Propagation

Similar to EKF state propagation, MSCKF framework uses IMU data to propagate states. The difference is state augmentation at the moment of new image arrival. As can be seen from [1], The time evolution of IMU states are described below:

$$\begin{gathered} {}^{I}_{G}\dot{\bar{q}}(t) = \frac{1}{2}\Omega\big(\omega(t)\big){}^{I}_{G}\bar{q}(t) \\ \dot{b}_g(t) = n_{wg}(t) \\ {}^{G}\dot{v}_I(t) = {}^{G}a(t) \\ \dot{b}_a(t) = n_{wa}\,(t) \\ {}^{G}\dot{p}_I(t) = {}^{G}v_I(t) \end{gathered} \tag{6}$$

where ${}^{G}a$ represents the body acceleration in frame $\{G\}$. $\omega = \left[\omega_x \; \omega_y \; \omega_z\right]^{T}$ represents angular velocity of IMU expressed in frame $\{I\}$. And:

$$\Omega(\omega) = \begin{bmatrix} -[\omega]_\times & \omega \\ -\omega^T & 0 \end{bmatrix}, \; [\omega]_\times = \begin{bmatrix} 0 & -\omega_z & \omega_y \\ \omega_z & 0 & -\omega_x \\ -\omega_y & \omega_x & 0 \end{bmatrix} \tag{7}$$

ω_m and a_m are the gyroscope and accelerometer measurements separately. Ignored the effects of the planet's rotation, they are given by [1]:

$$\begin{gathered} \omega_m = \omega + b_g + n_g \\ a_m = R\left({}^{I}_{G}\bar{q}\right)\left({}^{G}a - {}^{G}g\right) + b_a + n_a \end{gathered} \tag{8}$$

where ${}^{G}g$ is gravitational acceleration, expressed in frame $\{G\}$. Applying Eq. (6) in Eq. (8), continuous dynamic model of IMU states can be obtained:

$$\begin{gathered} {}^{I}_{G}\dot{\hat{\bar{q}}} = \tfrac{1}{2}\Omega(\hat{\omega}){}^{I}_{G}\hat{\bar{q}}, \; \dot{\hat{b}}_g = 0_{3\times1}, \\ {}^{G}\dot{\hat{v}}_I = R\left({}^{I}_{G}\hat{\bar{q}}\right)^{T}\hat{a} + {}^{G}g \\ \dot{\hat{b}}_a(t) = 0_{3\times1}, \qquad {}^{G}\dot{\hat{p}}_I = {}^{G}\hat{v}_I \end{gathered} \tag{9}$$

moreover, $\hat{a} = a_m - \hat{b}_a$, $\hat{\omega} = \omega_m - \hat{b}_g$, continuous dynamic model of IMU error-state is defined by:

$$\dot{\tilde{X}}_I = F\tilde{X}_I + Gn_I \tag{10}$$

where $\mathrm{n_I} = \left[\ \mathrm{n_g^T}\ \mathrm{n_{\omega g}^T}\ \mathrm{n_a^T}\ \mathrm{n_{\omega a}^T}\right]^{\mathrm{T}}$ is the system noise. It depends on the IMU noise characteristics. Finally, the matrices F and G that appear in Eq. (10) are given by:

$$\mathrm{F} = \begin{bmatrix} 0_{3\times3} & \mathrm{I}_{3\times3} & 0_{3\times3} & 0_{3\times3} & 0_{3\times3} & 0_{3\times9} \\ 0_{3\times3} & 0_{3\times3} & -\mathrm{R}\left({}^{\mathrm{I}}_{\mathrm{G}}\hat{\bar{\mathrm{q}}}\right)^{\mathrm{T}}[\hat{\mathrm{a}}]_\times & -\mathrm{R}\left({}^{\mathrm{I}}_{\mathrm{G}}\hat{\bar{\mathrm{q}}}\right)^{\mathrm{T}} & 0_{3\times3} & 0_{3\times9} \\ 0_{3\times3} & 0_{3\times3} & -[\hat{\omega}]_\times & 0_{3\times3} & -\mathrm{I}_{3\times3} & 0_{3\times9} \\ 0_{18\times3} & 0_{18\times3} & 0_{18\times3} & 0_{18\times3} & 0_{18\times3} & 0_{18\times9} \end{bmatrix} \tag{11}$$

$$\mathrm{G} = \begin{bmatrix} 0_{3\times3} & 0_{3\times3} & 0_{3\times3} & \mathrm{I}_{3\times3} \\ 0_{3\times3} & 0_{3\times3} & -\mathrm{R}\left({}^{\mathrm{I}}_{\mathrm{G}}\hat{\bar{\mathrm{q}}}\right)^{\mathrm{T}} & 0_{3\times3} \\ -\mathrm{I}_{3\times3} & 0_{3\times3} & 0_{3\times3} & 0_{3\times3} \\ 0_{3\times3} & 0_{3\times3} & 0_{3\times3} & 0_{3\times3} \\ 0_{3\times3} & \mathrm{I}_{3\times3} & 0_{3\times3} & 0_{3\times3} \\ 0_{12\times3} & 0_{12\times3} & 0_{12\times3} & 0_{12\times3} \end{bmatrix} \tag{12}$$

Following Euler integration [4] of Eq. (10), discrete-time system matrix is given by:

$$\Phi = \mathrm{I} + \mathrm{F}\delta\mathrm{t} \tag{13}$$

Moreover, the propagation of covariance is given by:

$$\mathrm{P} = \Phi\mathrm{P}\Phi^{\mathrm{T}} + (\Phi\mathrm{G})\mathrm{Q}(\Phi\mathrm{G})^{\mathrm{T}}\delta\mathrm{t}$$

In this paper, covariance structure is defined as:

$$\mathrm{P}_{\mathrm{k|k}} = \begin{bmatrix} \mathrm{P}_{\mathrm{II}_{\mathrm{k|k}}} & \mathrm{P}_{\mathrm{IC}_{\mathrm{k|k}}} \\ \mathrm{P}_{\mathrm{IC}_{\mathrm{k|k}}}{}^{\mathrm{T}} & \mathrm{P}_{\mathrm{CC}_{\mathrm{k|k}}} \end{bmatrix} \tag{14}$$

Since the current state of IMU propagation doesn't change the pose of sliding window, we can formulate the covariance propagation method:

$$\mathrm{P}_{\mathrm{k+1|k}} = \begin{bmatrix} \mathrm{P}_{\mathrm{II}_{\mathrm{k+1|k}}} & \Phi\mathrm{P}_{\mathrm{IC}_{\mathrm{k|k}}} \\ \mathrm{P}_{\mathrm{IC}_{\mathrm{k|k}}}{}^{\mathrm{T}}\Phi^{\mathrm{T}} & \mathrm{P}_{\mathrm{CC}_{\mathrm{k|k}}} \end{bmatrix} \tag{15}$$

where, $\mathrm{P}_{\mathrm{II}_{\mathrm{k+1|k}}} = \Phi\mathrm{P}_{\mathrm{II}_{\mathrm{k|k}}}\Phi^{\mathrm{T}} + (\Phi\mathrm{G})\mathrm{Q}(\Phi\mathrm{G})^{\mathrm{T}}\delta\mathrm{t}$, and $\mathrm{P_{II}}$ represents covariance of IMU states. $\mathrm{P_{IC}}$ represents covariance of IMU states with respect to pose of cameras. $\mathrm{P_{CC}}$ represents covariance of pose of augmented cameras.

When a new image arrives, current state of system should be augmented (in this paper, we augment the left camera state similarly to [2]). Including augmented states, the extended states are defined as:

$$\hat{\mathrm{X}}_{\mathrm{k}} = \left[\hat{\mathrm{X}}_{\mathrm{I}}{}^{\mathrm{T}}\ \hat{\mathrm{X}}_{\mathrm{C}_0^0}{}^{\mathrm{T}}\ \hat{\mathrm{X}}_{\mathrm{C}_1^0}{}^{\mathrm{T}}\ \hat{\mathrm{X}}_{\mathrm{C}_2^0}{}^{\mathrm{T}} \cdots \hat{\mathrm{X}}_{\mathrm{C}_{\mathrm{N}}^0}{}^{\mathrm{T}}\right]^{\mathrm{T}} \tag{15}$$

where $\hat{\mathrm{X}}_{\mathrm{C}_{\mathrm{j}}^0} = \left[{}^{\mathrm{C}^0}_{\mathrm{G}}\hat{\bar{\mathrm{q}}}^{\mathrm{T}}\ {}^{\mathrm{G}}\hat{\mathrm{p}}_{\mathrm{C}_{\mathrm{j}}^0}{}^{\mathrm{T}}\right]^{\mathrm{T}}, \mathrm{j} = (0,1,\dots,\mathrm{N})$ represents the pose of augmented camera C^0. It is derived from extrinsic parameter E_0 and IMU states:

$$\begin{gathered} {}^{\mathrm{C}^0}_{\mathrm{G}}\hat{\bar{\mathrm{q}}} = {}^{\mathrm{C}^0}_{\mathrm{I}}\hat{\bar{\mathrm{q}}} \otimes {}^{\mathrm{I}}_{\mathrm{G}}\hat{\bar{\mathrm{q}}} \\ {}^{\mathrm{G}}\hat{\mathrm{p}}_{\mathrm{C}^0} = {}^{\mathrm{G}}\hat{\mathrm{p}}_{\mathrm{I}} + \mathrm{R}({}^{\mathrm{I}}_{\mathrm{G}}\hat{\bar{\mathrm{q}}})^{\mathrm{T}} \cdot {}^{\mathrm{I}}\hat{\mathrm{p}}_{\mathrm{C}^0} \end{gathered} \tag{16}$$

Hence, in Error State Kalman Filter (ESKF [12]) framework, error-state of system (including augmented cameras) is defined by:

$$\tilde{\mathrm{X}}_{\mathrm{k}} = \left[\tilde{\mathrm{X}}_{\mathrm{I}}{}^{\mathrm{T}}\ \tilde{\mathrm{X}}_{\mathrm{C}_0^0}{}^{\mathrm{T}}\ \tilde{\mathrm{X}}_{\mathrm{C}_1^0}{}^{\mathrm{T}}\ \tilde{\mathrm{X}}_{\mathrm{C}_2^0}{}^{\mathrm{T}} \cdots \tilde{\mathrm{X}}_{\mathrm{C}_{\mathrm{N}}^0}{}^{\mathrm{T}}\right]^{\mathrm{T}} \tag{17}$$

where, $\tilde{\mathrm{X}}_{\mathrm{C}_{\mathrm{j}}^0} = \left[{}^{\mathrm{C}_{\mathrm{j}}^0}_{\mathrm{G}}\tilde{\theta}^{\mathrm{T}}\ {}^{\mathrm{G}}\tilde{\mathrm{p}}_{\mathrm{C}_{\mathrm{j}}^0}{}^{\mathrm{T}}\right]^{\mathrm{T}}, \mathrm{j} = (0,\dots,\mathrm{N}-1)$ represents the error of jth augmented camera C^0. Moreover, augmented covariance is defined by:

$$\mathrm{P}'_{\mathrm{k|k}} = \begin{bmatrix} \mathrm{P}_{\mathrm{k|k}} & \mathrm{P}_{21}{}^{\mathrm{T}} \\ \mathrm{P}_{21} & \mathrm{P}_{22} \end{bmatrix} \tag{18}$$

Note that $\mathrm{P}_{21} = \mathrm{JP}_{\mathrm{k|k}}$, $\mathrm{P}_{22} = \mathrm{JP}_{\mathrm{k|k}}\mathrm{J}^{\mathrm{T}}$ are the augmented covariance with respect to jth augmented state, and J is the Jacobian of $\tilde{\mathrm{X}}_{\mathrm{C}_{\mathrm{j}}^0}$ with respect to the error-state vector.

$$\mathrm{J} = \begin{bmatrix} 0_{3\times3} & 0_{3\times3} & {}^{\mathrm{C}^0}_{\mathrm{G}}\hat{\mathrm{R}} & 0_{3\times6} & \mathrm{I}_{3\times3} & 0_{3\times3} & 0_{3\times(6\mathrm{N}+6)} \\ \mathrm{I}_{3\times3} & 0_{3\times3} & -{}^{\mathrm{I}}_{\mathrm{G}}\hat{\mathrm{R}}^{\mathrm{T}}\left[{}^{\mathrm{I}}\hat{\mathrm{p}}_{\mathrm{C}^0}\right]_\times & 0_{3\times6} & 0_{3\times3} & {}^{\mathrm{I}}_{\mathrm{G}}\hat{\mathrm{R}}^{\mathrm{T}} & 0_{3\times(6\mathrm{N}+6)} \end{bmatrix} \tag{19}$$

C. State Update

Similar to [2], we can formulate the reprojection of features from stereo. Different from [2], the extrinsic parameters E_1 employed in this paper is calibrated online. $\left({}^{\mathrm{C}_{\mathrm{i}}^0}_{\mathrm{G}}\bar{\mathrm{q}}\ \ {}^{\mathrm{G}}\mathrm{p}_{\mathrm{C}_{\mathrm{i}}^0}\right)$ and $\left({}^{\mathrm{C}_{\mathrm{i}}^1}_{\mathrm{G}}\bar{\mathrm{q}}\ \ {}^{\mathrm{G}}\mathrm{p}_{\mathrm{C}_{\mathrm{i}}^1}\right)$ are ith left and right camera pose at the same time instance respectively. Employing the stereo extrinsic, the pose of the right camera C^1 can be easily derived in terms of the left camera augmented(e.g. ${}^{\mathrm{C}^1}_{\mathrm{G}}\bar{\mathrm{q}} = {}^{\mathrm{C}^1}_{\mathrm{C}^0}\bar{\mathrm{q}} \otimes {}^{\mathrm{C}^0}_{\mathrm{G}}\bar{\mathrm{q}}$, ${}^{\mathrm{G}}\mathrm{p}_{\mathrm{C}^1} = {}^{\mathrm{G}}\mathrm{p}_{\mathrm{C}^0} + \mathrm{R}\left({}^{\mathrm{C}^0}_{\mathrm{G}}\bar{\mathrm{q}}\right)^{\mathrm{T}} \cdot {}^{\mathrm{C}^0}\mathrm{p}_{\mathrm{C}^1}$). The reprojection of stereo measurement, $\hat{\mathrm{z}}_{\mathrm{i}}^{\mathrm{j}}$ in ith pose is defined as:

$$\hat{\mathrm{z}}_{\mathrm{i}}^{\mathrm{j}} = \begin{pmatrix} \hat{\mathrm{u}}_{\mathrm{i},0}^{\mathrm{j}} \\ \hat{\mathrm{v}}_{\mathrm{i},0}^{\mathrm{j}} \\ \hat{\mathrm{u}}_{\mathrm{i},1}^{\mathrm{j}} \\ \hat{\mathrm{v}}_{\mathrm{i},1}^{\mathrm{j}} \end{pmatrix} = \begin{pmatrix} \frac{1}{{}^{\mathrm{C}_{\mathrm{i}}^0}\hat{\mathrm{Z}}_{\mathrm{j}}} & 0_{2\times2} \\ 0_{2\times2} & \frac{1}{{}^{\mathrm{C}_{\mathrm{i}}^1}\hat{\mathrm{Z}}_{\mathrm{j}}} \end{pmatrix} \begin{pmatrix} {}^{\mathrm{C}_{\mathrm{i}}^0}\hat{\mathrm{X}}_{\mathrm{j}} \\ {}^{\mathrm{C}_{\mathrm{i}}^0}\hat{\mathrm{Y}}_{\mathrm{j}} \\ {}^{\mathrm{C}_{\mathrm{i}}^1}\hat{\mathrm{X}}_{\mathrm{j}} \\ {}^{\mathrm{C}_{\mathrm{i}}^1}\hat{\mathrm{Y}}_{\mathrm{j}} \end{pmatrix} \tag{20}$$

Note that $\left[{}^{\mathrm{C}_{\mathrm{i}}^{\mathrm{k}}}\hat{\mathrm{X}}_{\mathrm{j}}\ {}^{\mathrm{C}_{\mathrm{i}}^{\mathrm{k}}}\hat{\mathrm{X}}_{\mathrm{j}}\ {}^{\mathrm{C}_{\mathrm{i}}^{\mathrm{k}}}\hat{\mathrm{X}}_{\mathrm{j}}\right]$ is the coordinate of jth feature in frame $\{\mathrm{C}^{\mathrm{k}}\}$ in ith camera pose of sliding window (k=0,1 represents left and right camera respectively). Measurement residual is defined as:

$$\mathrm{r}_{\mathrm{i}}^{\mathrm{j,k}} = \mathrm{z}_{\mathrm{i}}^{\mathrm{j,k}} - \hat{\mathrm{z}}_{\mathrm{i}}^{\mathrm{j,k}} \tag{21}$$

We can formulate least-squares system to optimize the coordinates of features. See details in [13]. Then, the reprojection error of jth feature observation in ith camera pose in sliding window is derived as:

$$\begin{gathered} \mathrm{r}_{\mathrm{i}}^{\mathrm{j,k}} = \mathrm{z}_{\mathrm{i}}^{\mathrm{j,k}} - \hat{\mathrm{z}}_{\mathrm{i}}^{\mathrm{j,k}} \approx \mathrm{H}_{\mathrm{X_i}}^{\mathrm{j,k}}\tilde{\mathrm{X}} + \mathrm{H}_{\mathrm{f_i}}^{\mathrm{j,k}}\,{}^{\mathrm{G}}\tilde{\mathrm{p}}_{\mathrm{f_j}} + \mathrm{n}_{\mathrm{i}}^{\mathrm{j,k}} \\ \mathrm{H}_{\mathrm{X_i}}^{\mathrm{j,0}} = \left[0_{2\times(27+6i)}\ H_1\ 0_{2\times6(N-i-1)}\right] \\ \mathrm{H}_{\mathrm{X_i}}^{\mathrm{j,1}} = \left[0_{2\times21}\ H_2\ 0_{2\times6i}\ H_3\ 0_{2\times6(N-i-1)}\right] \end{gathered} \tag{22}$$

where, $\mathrm{H}_{\mathrm{X_i}}^{\mathrm{j,0}}$ and $\mathrm{H}_{\mathrm{X_i}}^{\mathrm{j,1}}$ represents the Jacobian of $\mathrm{r}_{\mathrm{i}}^{\mathrm{j,0}}$ and $\mathrm{r}_{\mathrm{i}}^{\mathrm{j,1}}$ with respect to error-state. And H_1, H_2, H_3 are derived respectively by:

$$\begin{gathered} H_1 = \left[J_i^{j,0}\left[{}^{C_i^0}\hat{\mathrm{p}}_{f_j}\right]_\times\ \ -J_i^{j,0}\,{}^{C_i^0}_{\mathrm{G}}\hat{\mathrm{R}}\right] \\ H_2 = \left[J_i^{j,1}\left[{}^{C_i^1}\hat{\mathrm{p}}_{f_j}\right]_\times\ -J_i^{j,1}\,{}^{\mathrm{C}^1}_{\mathrm{C}^0}\hat{\mathrm{R}}\right] \end{gathered}$$

$$H_3 = \left[J_i^{j,1}\, {}^{C^1}_{C^0}\hat{R} \left[{}^{C_i^0}\hat{p}_{f_j} \right]_\times \; - J_i^{j,1}\, {}^{C_i^1}_{G}\hat{R} \right]$$

where $J_i^{j,0}$ and $J_i^{j,1}$ are defined as：

$$J_i^{j,k} = \frac{1}{\left({}^{C_i^k}\hat{Z}_j\right)^2} \begin{bmatrix} {}^{C_i^k}\hat{Z}_j & 0 & -\,{}^{C_i^k}\hat{X}_j \\ 0 & {}^{C_i^k}\hat{Z}_j & -\,{}^{C_i^k}\hat{Y}_j \end{bmatrix}, (k = 0,1)$$

Similar to original S-MSCKF [2], $H_{f_i}^j$ represents the Jacobian with respect to the error of feature coordinate. $H_{X_i}^j$ represents the Jacobian with respect to error-state. The core point in this paper is, different with S-MSCKF, in the Jacobian of reprojection error in right camera with respect to error-state, the sub-Jacobian of the reprojection error in right camera with respect to the error-state of stereo extrinsic E_1 is a non-zero block. It just models the reprojection error with respect to E_1. During state update, the E_1 will be calibrated online iteratively. n_i^j represents observation noise of jth feature in ith pose. We can stack Eq. (22) of all the observations with respect to the same feature:

$$r^j = z^j - \hat{z}^j \approx H_X^j \tilde{X} + H_f^j\, {}^G\tilde{p}_{f_j} + n^j \tag{23}$$

As EKF state variables are formulated regardless of feature coordinates, we can project Eq. (23) into the left null space of H_f^j, and marginalize the formula of feature error [14]:

$$r_o^j = V^T r^j = V^T(z^j - \hat{z}^j) \approx V^T H_X^j \tilde{X} + n_o^j \tag{24}$$

where, V represents the left null space of H_f^j, $n_o^j = V^T n^j$. Hence, Eq. (24) becomes the same as standard EKF update, and QR decomposition can be employed to accelerate the standard EKF update [1].

Similar to original S-MSCKF [2], the Observability Constrained EKF [15] is applied in our method for maintaining the consistency of the filter. And the strategy of feature update also comes from S-MSCKF.

D. Vision Frontend

In our implementation, for efficiency, FAST [16] corners are extracted as landmarks. Similar to [2-4], the KLT optical flow algorithm [17] is employed in feature matching of front and rear frames, as well as left and right frames. In stereo matching, essential matrix constraint is used to eliminate outliers. Different from [2-4], since stereo extrinsic parameters are calibrated online in this work, the stereo extrinsic parameters used in the frontend will also be time-varying. Since the initial extrinsic parameters may be inaccurate, the outlier rejection algorithm may incorrectly remove inliers during the initial phase of system start-up. Therefore, the constraint of outlier rejection using essential matrix relation should be weakened during the initial period of system startup to prevent serious errors. After the system runs for a period of time (i.e. 30 seconds in this paper), the essential matrix constraint could be set to the normal threshold.

IV. EXPERIMENTS

Two experiments are performed to evaluate the proposed algorithm. Firstly, we compare our method with state-of-the-art stereo VIO [2-4] on EuRoC dataset and a large scale dataset. Secondly, another experiment is performed with the stereo extrinsic containing initial noise to show the robustness and the validity of the proposed algorithm. All of the following algorithms run on Intel i9-9900k (3.6GHZ) desktop platform.

A. Dataset

EuRoC dataset is a visual-inertial dataset [18] produced by ASL team of ETH. Collected by UAV, the dataset includes stereo images of 20 FPS and IMU data of 200 Hz. It also provides ground truth trajectories from Leica MS50 lidar and Vicon motion capture system. The dataset consists of three scenarios and 11 sequences. Five of them are randomly selected for comparison.

Our large scale dataset includes 30 Hz stereo images and 500 Hz IMU collected by Mynteye S1030 camera shown in Fig. 2. The cameras is calibrated by Kalibr toolkit [19], and the ground truth is collected by 5Hz GPS of UBLOX NEO-M8N.

Our real-world dataset contains two scenes. The first dataset is the outdoor flight scene of UAV at 10m, 25m and 30m altitude. The horizontal trajectory distances are 1km, 1.1km and 2.1km separately, and the trajectories look like rectangles. The second dataset is an outdoor hand-held scene with a total distance of 1.5km. Therefore, all sequences are from large scale scenes. Fig. 3 shows sample images of the self-collected dataset.

B. RMSE comparison

RMSE, root mean square error, is a popular measurement to evaluate estimation accuracy. In the experiment, we compare proposed method with S-MSCKF and VINS-Fusion. The former is also based on MSCFK framework which cannot estimate stereo extrinsic online. The latter is an optimization based stereo VIO, which can estimate extrinsic between IMU and every camera. For fairness, we turn off the loop closure mode of VINS-Fusion.

In some cases, as the proposed algorithm has a relatively large initial threshold in frontend, we only compare trajectories after 30s.

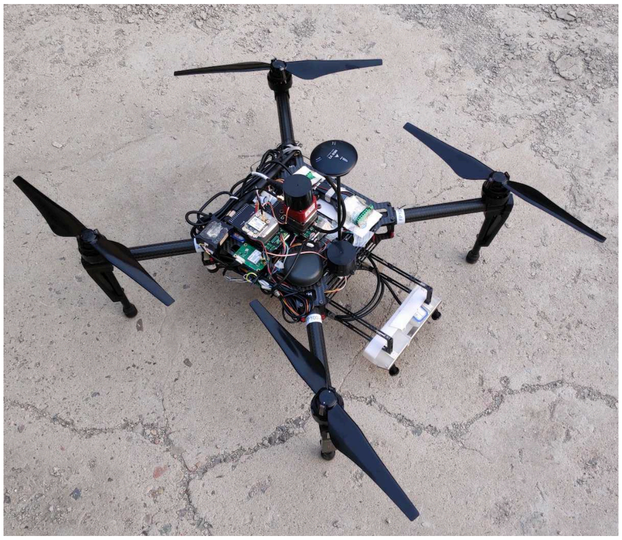


Figure 2. The device we used for our dataset. It contains oblique top-down global shutter stereo camera(AR0135, 30Hz) with 752×480 resolution and it contains a build-in IMU (ICM20602, 500Hz).

1) *In EuRoC dataset*

In Table 1, when the initial stereo extrinsic is normal, VINS-Fusion performs the best, and our method performs similarly to original S-MSCKF. Although VINS-Fusion has higher accuracy, it consumes more computational resource because of too many variables optimized at the same time (see the detail comparison in [2]), and the average CPU load on our machine is about 146%. On the contrary, the proposed method is filter-based. Thus, it has the advantage of both high efficiency and lightweight. The average CPU load of our method is only about 57%, which is less than 1/2 of VINS-Fusion. Our method is similar with S-MSCKF in terms of CPU load.

In Table 2, as expected, benefited from the online stereo extrinsic calibration, the estimation results with the initial noise in stereo extrinsic of our algorithm and VINS-Fusion are not degraded significantly compared with no noise situations. However, without stereo extrinsic estimation, S-MSCKF performs badly or even diverges.

2) *In large scale environment*

In large scale environment, the estimation accuracy of proposed method is similar with VINS-Fusion, both of which are superior to S-MSCKF. Especially in the handheld data (Fig. 4), because there is no stereo baseline estimation, the estimated scale of S-MSCKF has a large error. It indicates that in the large scale environment, online stereo extrinsic estimation is crucial for scale estimation.

Similarly, with stereo extrinsic containing initial noise, the proposed algorithm and VINS-Fusion still work well in most cases, but S-MSCKF diverges. Hence, the robustness of our method is validated.

C. *Stereo extrinsic estimation result*

As can be seen from Table 3, with different perturbations to X and Y direction of the initial stereo extrinsic parameters, our method can converge to be approximately the same as the off-line calibration results. Specifically, for errors in translation in X or Y axis, most of the final errors are limited to below 0.5 mm. For errors in any direction of rotation, the final error is controlled under 0.1 degree. It should be noted that in Z axis of translation, all final errors are around 5 mm, including the case with normal initial stereo extrinsic parameters. An intuitive explanation is that the Z axis of stereo camera device is aligned with UAV heading direction in EuRoC dataset. Almost all the time UAV moves towards the heading direction, which leads to a bigger error in the estimated offset along the depth direction.

Fig. 5 shows that, with different initial artificial perturbations, the estimated translation in X axis and rotation in yaw direction between two cameras change with time. The figure shows that in about 30 seconds the estimated translation in X axis converges, and in 5 seconds the estimated rotation in yaw converges. These results indicate that the proposed algorithm can effectively estimate the stereo extrinsic in a timely manner.

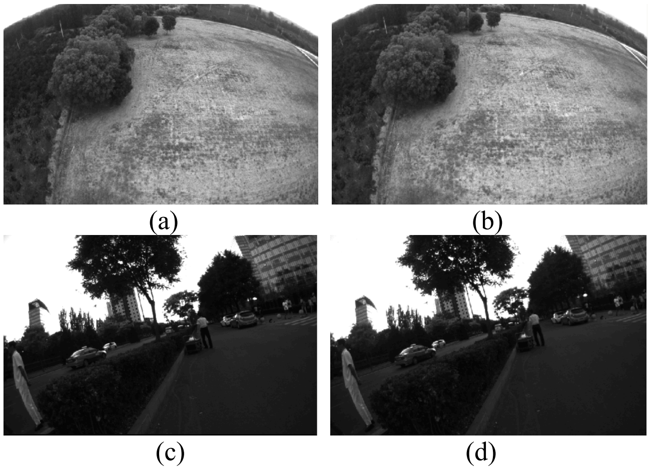


Figure 3. Sample images of large scale dataset. (a) and (b) are the images when the UAV is flying on 30 meters height. (c) and (d) are the images when the device is held by hand.

TABLE I. RMSE (m) comparison with normal initial stereo extrinsic. For EuRoC dataset, only trajectories after 30s were considered.

Data sequences	**VINS-Fusion**	**S-MSCKF**	**Our Method**
MH_03	0.080	0.211	0.223
MH_04	0.110	0.373	0.315
V1_03	0.129	0.260	0.195
V2_01	0.079	0.110	0.091
V2_02	0.035	0.139	0.163
UAV_10m	2.935	4.417	3.737
UAV_25m	4.068	4.674	4.768
UAV_30m	15.976	19.328	13.866
Hand_Held	14.223	41.969	9.480

TABLE II. RMSE (m) comparison with bad initial stereo extrinsic (added 2 deg. error in Z axis for rotation and 5mm error in baseline for translation). Only trajectories after 30s were considered, as it took time to estimate appropriate stereo extrinsic for filter-based method.

Data sequences	**VINS-Fusion**	**S-MSCKF**	**Our Method**
MH_03	0.087	-	0.302
MH_04	0.102	1.659	0.337
V1_03	0.195	0.585	0.235
V2_01	0.154	0.724	0.127
V2_02	0.067	0.454	0.165
UAV_10m	2.950	-	4.930
UAV_25m	4.076	-	11.213
UAV_30m	-	-	-
Hand_Held	18.610	-	24.861

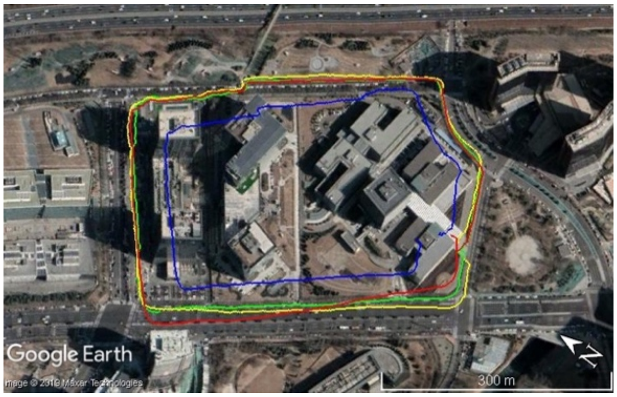
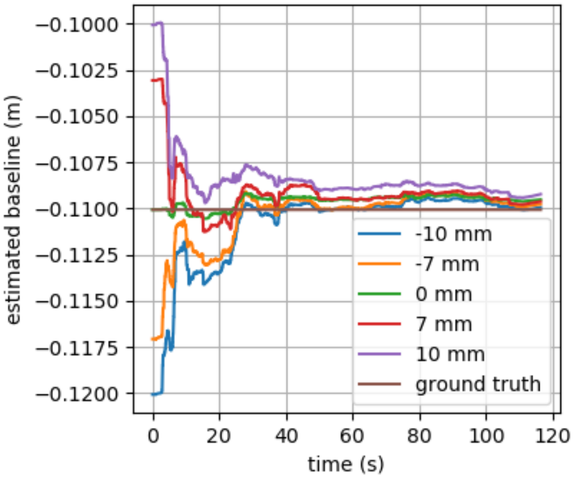


Figure 4. The estimated trajectories with good initial stereo extrinsic in hand held outdoor environment aligned to Google Map. VINS-Fusion (red), S-MSCKF (blue), our method (yellow) and GPS (green).

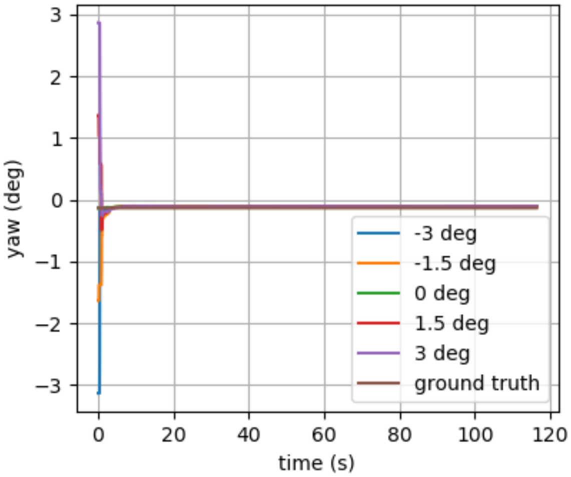
V. CONCLUSION

In this paper, we have presented an approach for online estimation of stereo extrinsic parameters based on S-MSCKF framework. The key component of our formulation is that the stereo extrinsic parameter E_1 is explicitly included in state variables, and the model between E_1 error and feature reprojection error is formulated. The resulting stereo VIO system significantly reduces the dependency on accurate offline stereo calibration. At the same time, the robustness and accuracy of the system are improved. Based on the experiments using EuRoC and real-world datasets, our scheme significantly outperforms the original S-MSCKF when there are perturbations to camera parameters. Especially, given inaccurate extrinsic parameters, our method can converge to an accurate estimation of extrinsic parameters over a few dozens of seconds. Since our method is filter-based, the computational requirement is much lower than those of optimization-based methods (e.g. VINS-Fusion), without significantly degrading the accuracy and robustness of the algorithm.

In future work, we will focus on real-time evaluation of the certainty of stereo extrinsic parameters.



(a)



(b)

Figure 5. With different initial artifical perturbations, estimated baseline (translation in X axis) and rotation in yaw between two cameras changing with time compared with offline calibration results using V2_02_medium data of EuRoC. (a) shows the translation and (b) shows the rotation.

TABLE III. Given different artificial initial perturbations, the final estimation errors of translation (mm) and rotation (Euler Angles in degree) between two cameras compared to the offine calibration ground truth. V2_02_medium data sequence of EuRoC is used in this experiment.

Errors in translation	**-10 mm**	**-7 mm**	**0 mm**	**+7 mm**	**+10 mm**
X	0.127	0.295	0.547	0.442	0.838
Y	0.248	-0.026	-0.204	-0.020	-0.040
Z	5.425	5.720	5.253	5.547	5.554
Errors in rotation	**-3 deg**	**-1.5 deg**	**0 deg**	**+1.5 deg**	**+3 deg**
Roll	0.096	0.097	0.093	0.096	0.087
Pitch	-0.078	-0.077	-0.080	-0.080	-0.086
Yaw	0.027	0.026	0.027	0.025	0.026

REFERENCES

[1] A. I. Mourikis and S. I. Roumeliotis, "A multi-state constraint kalman filter for vision-aided inertial navigation," inProceedings 2007 IEEE International Conference on Robotics and Automation, pp. 3565 – 3572,2007.

[2] K. Sun, K. Mohta, B. Pfrommer, M. Watterson, S. Liu, Y. Mulgaonkar,C. J. Taylor, and V. Kumar, "Robust stereo visual inertial odometry forfast autonomous flight," IEEE Robotics and Automation Letters, vol. 3,pp. 965 – 972, April 2018.[3] M. Brossard, S. Bonnabel, and J.

[3] T. Qin, J. Pan, S. Cao, and S. Shen, "A general optimization-based framework for local odometry estimation with multiple sensors", arXiv preprint arXiv:1901.03638, 2019.

[4] T. Qin and S. Shen, "Online temporal calibration for monocular visual-inertial systems", in 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, 2018, pp. 3662–3669.

[5] S. Weiss, M. W. Achtelik, S. Lynen, M. Chli, and R. Siegwart, "Real-time onboard visual-inertial state estimation and self-calibration of MAVs in unknown environments," Robotics and Automation (ICRA), 2012 IEEE International Conference on, pp. 957-964, 2012.

[6] S. Shen, Y. Mulgaonkar, N. Michael, and V. Kumar, "Multi-sensor fusion for robust autonomous flight in indoor and outdoor environments with a rotorcraft MAV," Robotics and Automation (ICRA), 2014 IEEE International Conference on, pp. 4974-4981, 2014.

[7] Stefan Leutenegger, Simon Lynen, Michael Bosse, Roland Siegwart and Paul Timothy Furgale. Keyframe-based visual-inertial odometry using nonlinear optimization. The International Journal of Robotics Research, 2015.

[8] Hansen P, Alismail H, Rander P, et al. Online continuous stereo extrinsic parameter estimation[C]//2012 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2012: 1059-1066.

[9] Ling Y, Shen S. High-precision online markerless stereo extrinsic calibration[C]//2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2016: 1771-1778.

[10] N. Trawny and S. Roumeliotis, "Indirect Kalman filter for 6D pose estimation," University of Minnesota, Dept. of Comp. Sci. & Eng., Tech. Rep, vol. 2, 2005.

[11] Li, M. and Mourikis, A. I. (2011). Consistency of EKF-based visual-inertial odometry. Technical report, University of California Riverside. www.ee.ucr.edu/~mourikis/tech reports/VIO.pdf.

[12] J. Sola`, "Quaternion kinematics for the error-state kalman filter," CoRR, vol. abs/1711.02508, 2017.

[13] L. Clement, V. Peretroukhin, J. Lambert, and J. Kelly. The battle for filter supremacy: A comparative study of the multi-state constraint kalman filter and the sliding window filter. In Computer and Robot Vision (CRV), 2015 12th Conference on, pages 23–30, 2015.

[14] Y. Yang, J. Maley, and G. Huang, "Null-space-based marginalization: Analysis and algorithm," in Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems, Vancouver, Canada, Sep. 24-28, 2017, pp. 6749-6755.

[15] J. A. Hesch, D. G. Kottas, S. L. Bowman, and S. I. Roumeliotis, "Observability-constrained vision-aided inertial navigation," University of Minnesota, Dept. of Comp. Sci. & Eng., MARS Lab, Tech. Rep, vol. 1, 2012.

[16] M. Trajkovic´ and M. Hedley, "Fast corner detection," Image and vision computing, vol. 16, no. 2, pp. 75–87, 1998.

[17] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision (ijcai)," Proceedings of the 7th International Joint Conference on Artificial Intelligence (IJCAI '81), pp. 674–679, April 1981.

[18] Burri M, Nikolic J, Gohl P, et al. The EuRoC micro aerial vehicle datasets[J]. The International Journal of Robotics Research, 2016, 35(10): 1157-1163.

[19] Rehder J, Nikolic J, Schneider T, et al. Extending kalibr: Calibrating the extrinsics of multiple IMUs and of individual axes[C]//2016 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2016: 4304-4311.