Stereo Visual-Inertial Odometry With Multiple Kalman Filters Ensemble

Yong Liu, *Member, IEEE*, Rong Xiong, *Member, IEEE*, Yue Wang, *Student Member, IEEE*, Hong Huang, Xiaojia Xie, Xiaofeng Liu, and Gaoming Zhang

***Abstract*—In this paper, we present a stereo visual-inertial odometry algorithm assembled with three separated Kalman filters, i.e., attitude filter, orientation filter, and position filter. Our algorithm carries out the orientation and position estimation with three filters working on different fusion intervals, which can provide more robustness even when the visual odometry estimation fails. In our orientation estimation, we propose an improved indirect Kalman filter, which uses the orientation error space represented by unit quaternion as the state of the filter. The performance of the algorithm is demonstrated through extensive experimental results, including the benchmark KITTI datasets and some challenging datasets captured in a rough terrain campus.**

***Index Terms*—Kalman filter, multi-sensor fusion, pose estimation, robot vision, visual-inertial odometry.**

I. Introduction

VISUAL-INERTIAL odometry (VIO) is a comprehensive technique, which fuses the information from both the visual odometry (VO) and the inertial measurement unit (IMU) in order to estimate the six degrees of freedom (6DOF) pose. Therefore, the VIO can combine the advantages of the visual sensors and the inertial sensors, and can provide more accurate long-term 6DOF odometry estimation. In fact, the VIO has become an essential technique for mobile robots, especially in environments without GPS.

With recent advances in the robotics applications, more and more mobile robots are deployed in complicated and hostile environments, such as the field rescue robots [1], legged robots [2], [3], service robot [4], [5], etc. As a result, the VIO faces new challenges in order to work under these environments. There are two main challenges for the VIO systems working in hostile environments.

A. Mismatch of the Fusion Interval With VO and IMU

There is an intrinsic conflict in VIO due to the different measurement principles of VO and IMU. As we all know, the sampling rate of the IMU is normally three to five times of the camera's sampling rate. This means that the IMU will output three to five measurements between two adjacent images. Classical Kalman filters combine visual and inertial cues in two nonoptimal ways. The first way is to ignore the abundant measurements from the IMU in order to synchronize the measurements from both the IMU and the VO [6]. This method is obviously unsatisfactory because it loses dynamic information from the IMU, which can trace better than an assumed model if with a higher sampling rate [7]; the second way is to propagate the IMU's state model for several times before the VO updating the state estimates [8]. However, this method equals to integrating measurements of the IMU between the adjacent updates of the VO in essence. In this situation, the drift of the estimation from the IMU will be raised significantly with the increasing of the fusion interval, as the IMU requires the VO to rectify its drift as soon as possible, especially for the translation drift during the fusion of the VIO. On the other hand, the VO always suffers from the image pairs with small motion, which will lead to failure or low estimation accuracy. So the VO tends to implement a key-frame technique, which tries to use the image pairs with a relative large motion (equal to increasing the temporal fusion interval), to improve the estimation accuracy of the 6D motion. Then, the accuracy of the estimation from the VO can be improved with the increasing fusion interval, which is in conflict with the propagation of the IMU. We call this conflict as *mismatch of the fusion interval*, which is more severe in challenging environments, such as rough roads, bumping, illumination variation, occlusions, etc., and will lead to frequent VO failures. So the first challenge is how to design the VIO filter that can balance the requirements of VO and IMU on the fusion interval simultaneously.

B. Maintain robust VIO Estimation With Low-Precision IMU

IMUs which can offer high accuracy for extended periods of time are both bulky and expensive [9]. However, small mobile platforms require the VIO system to be small volume and low price and, thus, highly precise IMUs have to be replaced with noised low-cost ones [10]. Currently, there are a lot of VIO fusion algorithms and benchmark datasets, but most of them need to deploy a highly precise IMU to measure the vehicle movements [11]–[13]. Provided with noised IMU measurement, their algorithms are proved to be not always effective in challenging environments. In fact, it is much more difficult to get precision estimation from low-cost sensors with low-computational resources [10]. Here, we come up the second challenge to design a robust VIO fusion algorithm that can support low-precision

Manuscript received June 27, 2015; revised December 21, 2015, January 28, 2016, and March 21, 2016; accepted May 1, 2016. Date of publication May 27, 2016; date of current version September 9, 2016. This work was supported in part by the National Natural Science Foundation of China under Grant U1509210 and Grant 61473258 and in part by the Natural Science Foundation of Zhejiang Province under Grant LR13F030003. *(Corresponding authors: Yong Liu and Rong Xiong.)*

The authors are with the Institute of Cyber-Systems and Control, Zhejiang University, Hangzhou 310027, China (e-mail: yongliu@iipc.zju.eud.cn; rxiong@iipc.zju.edu.cn; yuewang@iipc.zju.edu.cn; honghuang@iipc.zju.edu.cn; xjxie@iipc.zju.edu.cn; xfliu@iipc.zju.edu.cn; gmzhang@iipc.zju.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org

Digital Object Identifier 10.1109/TIE.2016.2573765

IMU working at hostile environments in order to make the platforms portable and economic.

In this paper, we address the above two challenges in challenging environments and present a novel stereo VIO algorithm with multiple Kalman filters ensemble. To overcome the conflict mentioned in the first challenge, we use separated orientation and position filters (PFs) working on different frequencies to estimate the 6DOF odometry of the system. In order to support low-precision IMU addressing on the second challenge, we design a cascading fusion architecture, which fuses the *pitch*, *roll* twice with the attitude filter and orientation filter (OF) to obtain long-term stable and accurate orientation. The main three contributions of this paper are given as follow:

1) Separated Fusion for Sensors With Varied Sampling Rate: We decompose the classical integrated VIO filter into three subfilters, attitude filter (AF), OF, and PF. This separated fusion framework can support better multiple sensors working on different sampling rates and fuse them with varied updating intervals. The PF in our VIO fuses at each sequenced image frame, and can estimate a precision velocity with the minimal fusion interval in the VIO to reduce the error caused by the drift of the IMU. The fusion cycle of the OF is set to the time interval of two adjacent key frames in the VO, thus the OF can employ more precise orientation results estimated by the VO from the image pairs with a larger motion during the longer interval. The AF can provide accurate fusion results on the pitch and roll based on the gravity with the fusion interval same as the output cycle of the IMU. By this way, our VIO fusion algorithm can take advantages of all the useful information in order to give better pose estimation.

2) Cascading and Multiple-Level-Fusion Architecture: We use a cascading fusion architecture to estimate orientation, which enables better support on low-precision IMUs. Our VIO uses multiple-level-fusion, which combines the first level of AF and the second level of OF, to output robust and accurate orientation. A further PF is used to estimate the position and velocity by fusing the information from the IMU and the VO. This can be more robust against the VO failure and large drifts in low-precision IMUs.

3) Ensemble With Low-Cost Linear Subfilters: We provide a novel low-cost implementation of the VIO estimation, which may be more competitive when concerning deploying the VIO into the embedded computation system. Compared with the uniformed nonlinear filter, those three linear subfilters in our VIO only need to estimate a few state vectors, they also do not need to compute the Jacobi in the EKF or sigma point in the UKF.

The following section provides an overview of the related works on the VIO. The stereo VIO with multiple Kalman filters ensemble is described in Section III. Finally, experiments are described in Section IV followed by concluding remarks and future work in Section V.

II. Related Work

The VIO is an extension of the research on visual odometry [14]. Generally, there are two kinds of VIO systems based on the number of cameras, i.e., stereo VIO [6], [7] and monocular VIO [15]–[19]. As the monocular vision system cannot recover its scale in the estimation, most monocular VIO systems [15]–[18] employ tightly coupled approach, which combines the disparate raw data of vision and inertial sensors in a single, optimum filter, rather than cascading filters, one for each sensor [20]. While almost all the stereo VIO systems employ loosely coupled approach, they use separate inertial navigation and VO based structure-from-motion blocks running at different rates and exchanging information [20].

Recent tightly coupled work [11] introduces the nonlinear optimization into the VIO and treats the visual-inertial fusion as an optimization problem. This work has a clear fusion framework and can achieve superior performance when both kinds of sensors can provide high-quality data. The weaknesses are also obviously, the results are sensitive to the data quality and it is also easy to converge to the suboptimal solutions once there are not enough constraints.

When considering the loosely coupled fusion, Kalman filter is the most popular one among various solutions. There are two categories based on their data flows of the prediction and observation in the filter.

The first category [8] uses the measurements of the IMU to predict the states by the kinematics model and the observations come from the estimated results of the VO. As the measurements of the IMU include the linear acceleration, those methods of the first category are able to provide accurate estimation of the linear velocity with the kinematics model in a short time interval, those methods are suitable to estimate motions with variable velocities. However, those methods are also sensitive to the bias and drift of the IMU, as they only use the IMU to forecast future motion. The process to integrate the measurements of the IMU with the kinematics model can be regarded as an open-loop system, any small bias and drift on the measurements will be amplified by the integration operation and cause large estimation errors in the VO. Especially, when the key-frame technique is employed in the VO, the time interval that needs to predict by the IMU will increase, this will bring severer error accumulation. A typical example of the first category is proposed by Tardif, George, Laverne, Kelly and Stentz [8]. They used the position, velocity, orientation, bias of the linear acceleration, and the bias of the angular speed of the system as the model state vector. The position, velocity, and orientation of the system in the states are calculated from the measurements of the IMU. They then use the position and orientation estimated by the key-frame based VO as the observations and fuse with an EKF. Obviously, the accuracy of their prediction model will be significantly reduced as the fusion interval is increasing. This happens especially with low-precision IMUs. Furthermore, the high dimension and the correlation of the state vector will also increase the complexity and difficulty of implementation.

On the other hand, the second category [6] uses the estimated results of the VO to predict the model state and the observations come from the measurements of the IMU. Those methods in the second category are able to attach an additional low-level attitude filter [21], which is applied in the measurements of the IMU directly and can provide long-term, stable, drift-free attitude of the system. The advantage of this attitude filter is obviously, it can provide long-term accurate attitude angles,

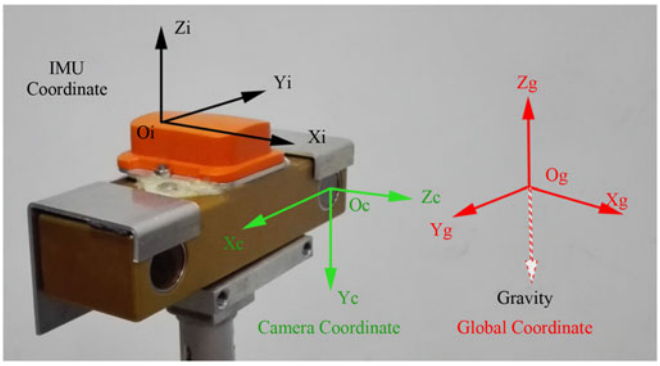

Fig. 1. Coordinate used in our stereo VIO system, including IMU(orange sensor in figure) and stereo camera (yellow sensor in figure). The IMU coordinate is represented by (Xi, Yi, Zi) and its origin *Oi*; the camera coordinate is (Xc, Yc, Zc) and *Oc*; the global coordinate is (Xg, Yg, Zg) and *Og*. The direction of the gravity is also shown in the global coordinate.

which will bring significant improvements on the accuracy of the odometry estimation. As the estimated results of the VO cannot provide the linear velocity, there will be some difficulties to include the linear velocity in the state equation. That is why all the methods of the second category only consider the OF and those methods of the second category can also be called inertial-aided visual odometry. The main drawback of these methods is that the position estimation will be unavailable once the VO fails. A typical example of the second category is presented by Konolige, Agrawal, and Sol [6], which only uses the OF between the IMU and the VO and achieves a dramatically improvement on the long-term VO accuracy. However their method cannot predict the position and velocity once the VO fails. To solve these problems, we present our cascading fusion architecture and introduce the PF in our approach.

In short, both categories of the aforesaid approaches cannot sufficiently address on the challenge of *mismatch of the fusion interval*. They are not able to take full use of the information from both the camera and the IMU. In this paper, we propose a new VIO algorithm, which uses separated attitude filter, OF, and PF. In our PF, the propagation and observation of the position states come from the IMU and the VO, respectively, which is the same as the first category. In our OF, the propagation and observation of the orientation states come from the VO and the IMU, respectively, which is the same as the second category. Thus, our approach has the potential to combine the advantages of both categories and suppress their drawbacks to achieve accurate estimation when using low-precise IMUs.

III. Stereo Visual-Inertial Odometry

Before introducing the detailed filters, the coordinates used in our stereo VIO are introduced first, shown in Fig. 1. There are three coordinates used in our stereo VIO, the original global coordinate $\{G\}$, IMU coordinate $\{I\}$, and stereo camera coordinate $\{C\}$. We parallel *X*-*O*-*Y* plane of $\{G\}$ to the horizontal plane. The *Z* axis points opposite to gravity. The *X*-axis points forward of the mobile platform, and the *Y*-axis is determined by the right-hand rule. Then, the task of VIO is to estimate the 6DOF pose of the IMU-affixed coordinate $\{I\}$ with respect to

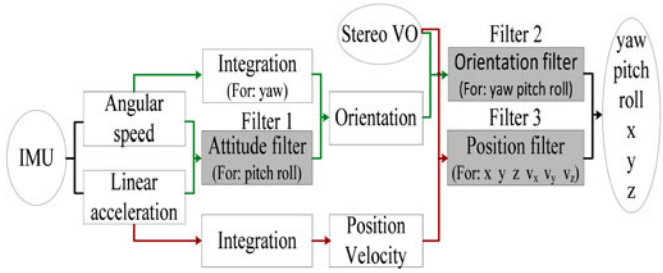

Fig. 2. Framework of our stereo VIO with three filters. The first filter fuses angular speed and linear acceleration of the IMU to get drift-free attitude estimation. The second filter is an indirect Kalman filter designed for orientation fusion of the VO and the IMU. The third filter is a PF of the VO and the IMU.

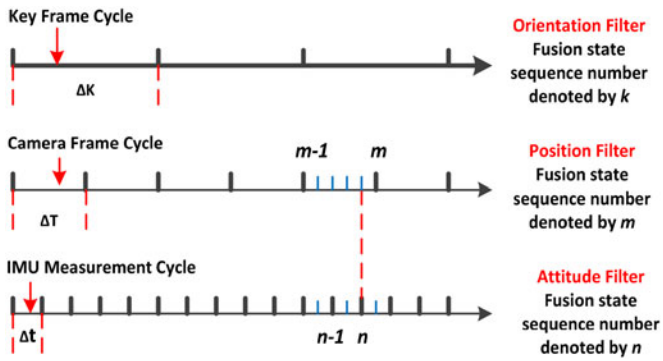

Fig. 3. Fusion intervals for three sub-filters. The OF has the largest fusion interval ΔK; the fusion interval of PF is ΔT; and the Attitude Filter has the smallest fusion interval Δt.

$\{G\}$. In our VIO, $\{C\}$ is set at the coordinate of the left camera. As the relative pose between $\{C\}$ and $\{I\}$ is rigid, we can calibrate their relative pose in advance. For simplicity, in the following filters, the 6DOF pose calculated by the VO is transformed to $\{I\}$ with the known rigid transformation between $\{I\}$ and $\{C\}$. Then, our filters only need to predict the relative 6DOF pose between $\{I\}$ and $\{G\}$.

The detailed fusion processing of stereo VO and the IMU is shown in Fig. 2. Fig. 2 shows that our stereo VIO runs separate inertial sensor fusion and vision based structure from motion fusion. It belongs to the loosely coupled approach. This choice is based on the following two reasons. The first is that the stereo VO is more precise and complete compared to the monocular VO, it is also able to avoid the problem of scale ambiguity in the monocular VO. Using loosely coupled approach will not break the natural integrity of the stereo VO modular; the second reason is that the covariances of the stereo VO and the IMU are varied, sometimes their covariances may lie on different scales and change with the time, thus it will lead to intractable estimation for their coupled variance if using tightly coupled approach.

Fig. 3 shows three different fusion intervals used in our approach for those three subfilters mentioned in Fig. 2.

A. Stereo Visual Odometry

The basic idea of the stereo VO is to estimate the motion between two adjacent frames by tracking the same feature points projected on these frames. In our VIO, the stereo VO is treated

as an independent module. CenSure detector is widely used due to its robustness and low-computation complexity. We use CenSure detector to obtain interest points in the left camera image and find their corresponding feature points in the right camera by searching the matched points along the baseline with a minimal zero-mean normalized cross-correlation score [22]. Using those feature pairs from left-right cameras, we can reconstruct the sparse 3-D points. By tracking the feature pairs between sequential images with the SURF descriptors, the motion estimation can be described as a 3-D-to-2-D problem [23], which refers to estimating the motion from the sparse 3-D features calculated by the stereo vision system in the earlier frames and the corresponding matched 2-D features in current frames. In our 3-D-to-2-D motion estimation, RANSAC is also used to remove those outliers. The 3-D-to-2-D feature pairs from both cameras are all considered in the same optimization function, which tries to minimize the reprojection errors of the images and will concern the rigid constraint of the stereo vision system.

We also implement the key-frame selection and the Levenberg–Marquardt optimization into the estimation to improve the accuracy. Similar to [11], the key-frame in our approach is selected based on the motion between frames. Given the current key-frame, the next key-frame is selected when either the norm of the translation from successive candidate frame to the current key-frame is larger than 0.3 m or the norm of the Rodrigues representation of the rotation from successive candidate frame to the current key-frame is larger than 0.25 rad.

Although the stereo VO is quite precise in most of conditions, its performance is also sensitive to many factors, such as complexity of the motions, illumination changing, environmental features and quality of images etc. Especially in those high speed small autonomous mobile platforms, the images are easy to be blurred, which will lead to failure on feature matching and thus the stereo VO will fail to estimate the motions. In this condition, the inertial sensor is helpful to recover the motions. This is also the reason that the stereo VIO will be more robust than the stereo VO.

B. Drift-Free Attitude Estimation

Although the IMU can output relatively reliable measurements of the angular speed and the linear acceleration, the attitude[1] only integrated from the measured angular speed may suffer from drift. Thus, we apply an attitude Kalman filter [21], i.e., filter 1 in Fig. 2, to output long-term stable attitude, i.e. the angles of *roll* and *pitch*.

Assuming $\boldsymbol{u}$ is the 3D linear acceleration of the IMU in coordinate $\{G\}$, $\boldsymbol{g} = [0 \quad 0 \quad -g]^T$ is the gravity acceleration in coordinate $\{G\}$, and R is the rotation matrix from $\{G\}$ to $\{I\}$. We use $\boldsymbol{a}$ to denote the linear acceleration measured by IMU in coordinate $\{I\}$. So the measured a will include both the gravity acceleration and its linear acceleration

$$\begin{aligned} \boldsymbol{a} &= R(\boldsymbol{u} + (-\boldsymbol{g})) = -R\boldsymbol{g} + R\boldsymbol{u} \\ &= R[0\ 0\ g]^T + R\boldsymbol{u} = g\boldsymbol{x} + R\boldsymbol{u}. \end{aligned} \tag{1}$$

[1]According to [21], attitude refers to the robot's orientation relative to the gravity vector, usually described by pitch and roll.

In formula (1), $\boldsymbol{x}$ is the third column of R, it is only related to the attitude of IMU, and can be represented with *pitch* and *roll*. That is,

$$\begin{aligned} \boldsymbol{x} = [&-sin(\text{pitch}) \quad cos(\text{pitch})\, sin(\text{roll}) \\ &\times cos(\text{pitch})\, cos(\text{roll})]^T. \end{aligned}$$

In this equation, *yaw* is not correlated. If the acceleration $\boldsymbol{u}$ is treated as the disturbance, then we can observe $\boldsymbol{x}$ by $\boldsymbol{a}$ in our filter. Because $\boldsymbol{a}$ is not accumulated with time, so attitude estimation is free from drift.

Based on the kinematics, we have $\dot{R} = \lfloor \dot{\theta} \times \rfloor R$, that is

$$\dot{\boldsymbol{x}} = \dot{\theta} \times \boldsymbol{x}. \tag{2}$$

Here, $\dot{\theta}$ is the 3-D angular speed of the IMU. $\lfloor \dot{\theta} \times \rfloor$ is the skew symmetric matrix, it is also denoted as $S(\dot{\theta})$. Discretizing formula (2), we can obtain

$$\boldsymbol{x}_n = A_{n-1}\boldsymbol{x}_{n-1}. \tag{3}$$

Here, $A_n = \mathbf{I} + \frac{S(\dot{\theta}_n)\sin(\|\dot{\theta}_n\|t)}{\|\dot{\theta}_n\|} + \frac{S^2(\dot{\theta}_n)(1-\cos(\|\dot{\theta}_n\|T))}{\|\dot{\theta}_n\|^2}$. t is the sampling cycle of IMU. A_n is an orthogonal rotation matrix. $\mathbf{I}$ is an identity matrix of 3×3.

Based on formulas (1) and (3), we can obtain the system model:

$$\begin{cases} \boldsymbol{x}_n = A_{n-1}\boldsymbol{x}_{n-1} \\ \boldsymbol{y}_n = \boldsymbol{x}_n + R_n\boldsymbol{u}_n/g. \end{cases} \tag{4}$$

In (4), $\boldsymbol{y} = \boldsymbol{a}/g\, R_n\boldsymbol{u}_n/g$ is treated as disturbance. When acceleration $\boldsymbol{u}_n$ is small, we can introduce a measurement noise $\boldsymbol{v}$ to model accelerations. Also, we can introduce a process noise $\boldsymbol{\omega}$ to measure inaccuracies in modeling and gyro noise. Then, we can obtain the model used in the filtering equations

$$\begin{cases} \boldsymbol{x}_n = A_{n-1}\boldsymbol{x}_{n-1} + \boldsymbol{\omega}_{n-1} \\ \boldsymbol{y}_n = H_n\boldsymbol{x}_n + \boldsymbol{v}_n. \end{cases} \tag{5}$$

$H_n = \sigma_n\mathbf{I}, \sigma_n$ is a binary variable that equals 1 or 0 determined by the acceleration. H_n is designed to reduce error from large acceleration. When acceleration is larger than the given threshold, $\sigma_n = 0$. In this circumstance, the system has no observation and estimates the attitude only by formula (3) until the next measurement of the IMU is obtained. Even though the system switches between model with observation and model without observation, this Kalman filter can be proved to be stable by both theory and real world experiments [21].

Filter 1: Attitude Filter.

Propagation: For each measurement of the IMU, propagate the filter state x and covariance Q with angular speed obtained from the IMU.

Update: If the acceleration of the IMU is less than the given threshold $(2m/s^2)$, $\sigma_n = 1$ otherwise $\sigma_n = 0$.

The covariance matrix Q of $\boldsymbol{\omega}$ is a tuning parameter in this filter and it is assumed to be a diagonal matrix with nonzero entries. The diagonal elements of Q can be estimated through the noise of angular speed measurements. The covariance matrix W of $\boldsymbol{v}$ is also a diagonal matrix, assuming the measurement of

each acceleration axis is independent. The diagonal elements of W can be estimated mainly by the acceleration threshold that determines the value of σ_n.

C. Orientation Estimation With Indirect Kalman Filter

The OF is used to estimate the three orientation angles of the system, i.e., *yaw*, *pitch,* and *roll*. In our OF, we introduce the indirect Kalman filter [24], which uses the orientation error space represented by the error unit quaternion instead of the orientation represented by the unit quaternion as the state of the filter. The state vector of this indirect filter does not need to be positive or unit and it has only three elements. Its state propagation model and measurement model are also much simpler. Moreover, the processing of the data fusion occurred in the error space is represented by the error quaternion, which could be closer to a linear space and, thus, more suitable to the Kalman filter.

We use the unit quaternion, ${}^G_I q$ to represent the relative orientation from $\{I\}$ to $\{G\}$. All the orientation values estimated by VO are also transformed to the coordinate of $\{I\}$. We use ${}^G_I \hat{q}$ to denote the estimate of ${}^G_I q$, and δq to denote the error between ${}^G_I \hat{q}$ and ${}^G_I q$ as following:

$$ {}^G_I q = {}^G_I \hat{q} \otimes \delta q. \tag{6}$$

$\otimes$ denotes the multiplication operation of the quaternion.

Assuming the rotation estimated by VO between two adjacent key-frames can be denoted with Δq. Then, we have

$$ {}^G_I \hat{q}_k = {}^G_I \hat{q}_{k-1} \otimes \Delta q_{k-1}. \tag{7}$$

In our approach, we assume the values of δq is normal distribution with mean zero and tiny variances, then based on formula (6),

$$ {}^G_I q_k \approx {}^G_I q_{k-1} \otimes \Delta q_{k-1} = {}^G_I \hat{q}_{k-1} \otimes \delta q_{k-1} \otimes \Delta q_{k-1}. \tag{8}$$

Then, we have the following derivation:

$$\begin{aligned} \delta q_k &= {}^G_I \hat{q}_k^T \otimes {}^G_I q_k \\ &= \left({}^G_I \hat{q}_{k-1} \otimes \Delta q_{k-1}\right)^T \otimes \left({}^G_I \hat{q}_{k-1} \otimes \delta q_{k-1} \otimes \Delta q_{k-1}\right) \\ &= \Delta q_{k-1}^T \otimes \delta q_{k-1} \otimes \Delta q_{k-1}. \end{aligned}$$

Here, δq_{k-1} can be denoted as

$$\begin{aligned} \delta q_{k-1} &= \begin{bmatrix} \delta q_0 & \delta q_1 & \delta q_2 & \delta q_3 \end{bmatrix}^T_{k-1} \\ &= \begin{bmatrix} \delta q_0 & 0 & 0 & 0 \end{bmatrix}^T_{k-1} + \begin{bmatrix} 0 & \delta q_1 & \delta q_2 & \delta q_3 \end{bmatrix}^T_{k-1}. \end{aligned}$$

Then,

$$\begin{aligned} \delta q_k &= \Delta q_{k-1}^T \otimes \begin{bmatrix} \delta q_0 & 0 & 0 & 0 \end{bmatrix}^T_{k-1} \otimes \Delta q_{k-1} \\ &+ \Delta q_{k-1}^T \otimes \begin{bmatrix} 0 & \delta q_1 & \delta q_2 & \delta q_3 \end{bmatrix}^T_{k-1} \otimes \Delta q_{k-1} \\ &= \begin{bmatrix} \delta q_0 & 0 & 0 & 0 \end{bmatrix}^T_{k-1} + \begin{bmatrix} 0 & \mathbf{0} \\ \mathbf{0} & \Delta R^T_{k-1} \end{bmatrix} \begin{bmatrix} 0 & \delta q_1 & \delta q_2 & \delta q_3 \end{bmatrix}^T_{k-1} \\ &= \begin{bmatrix} (\delta q_0)_{k-1} \\ \Delta R^T_{k-1} \begin{bmatrix} \delta q_1 \\ \delta q_2 \\ \delta q_3 \end{bmatrix}_{k-1} \end{bmatrix}. \end{aligned} \tag{9}$$

ΔR is the rotation matrix corresponding to Δq. According to formula (9), we can find the scalar of the error quaternion remaining constant, while the vector of the error quaternion is transformed by ΔR^T.

With the transformation of modified rodrigues parameters (MRPs) [18]:

$$\delta \boldsymbol{e} = \begin{bmatrix} \delta q_1 / (1 + \delta q_0) & \delta q_2 / (1 + \delta q_0) & \delta q_3 / (1 + \delta q_0) \end{bmatrix}^T. \tag{10}$$

According to formula (9), δq_0 is invariant between state *k-1* to *k*. We then have

$$\begin{bmatrix} \delta q_1 / (1 + \delta q_0) \\ \delta q_2 / (1 + \delta q_0) \\ \delta q_3 / (1 + \delta q_0) \end{bmatrix}_k = \Delta R^T_{k-1} \begin{bmatrix} \delta q_1 / (1 + \delta q_0) \\ \delta q_2 / (1 + \delta q_0) \\ \delta q_3 / (1 + \delta q_0) \end{bmatrix}_{k-1} \tag{11}$$

which can be abbreviated as

$$\boldsymbol{\delta e_k = \Delta R^T_{k-1} \delta e_{k-1}}. \tag{12}$$

MRPs will bring additional convenience during the calculation of the transformation. The inverse transformation, from $\delta \boldsymbol{e}$ to δq, is

$$\delta q_0 = \frac{1 - ||\delta \boldsymbol{e}||^2}{1 + ||\delta \boldsymbol{e}||^2} \qquad \boldsymbol{\delta q} = (1 + \delta q_0)\, \boldsymbol{\delta e}. \tag{13}$$

Assuming the noise of the system is additive, the state model of the system is

$$\boldsymbol{x}_k = \Delta R^T_{k-1} \boldsymbol{x_{k-1}} + \boldsymbol{\omega_{k-1}}|_{\boldsymbol{x}=\delta \boldsymbol{e}}. \tag{14}$$

If we use $\delta \boldsymbol{e}$ as the measurement vector, the observation equation of the system is

$$\boldsymbol{y}_k = \boldsymbol{x}_k + \boldsymbol{v}_k. \tag{15}$$

$\boldsymbol{\omega}$ in (14) denotes process noise and $\boldsymbol{v}$ in formula (15) represents observation noise of the system.

Then the OF is given in the following:

Filter 2: OF.

Propagation: For each key-frame captured by stereo cameras, propagate the filter state $\delta \boldsymbol{e}$ and covariance Q with relative rotation calculated by VO.

Observation:

(I) Get *yaw* from IMU measurement, *pitch* and *roll* from *Filter 1* at the time that the last key-frame is acquired.

(II) Get *yaw* from IMU measurement, *pitch* and *roll* from *Filter 1* at the time that the current key-frame is acquired.

(III) Compute the relative rotation between two key-frames with the angles above.

Update: Each time a key-frame is obtained, perform a Kalman update.

Because $\boldsymbol{x}_k$ and $\boldsymbol{y}_k$ are both vectors in the approximate-linear orientation error space, elements of them can be regarded as decoupled with each other. So the covariance matrix Q of process noise $\boldsymbol{\omega}$ and the covariance matrix W of measurement noise $\boldsymbol{v}$ are set to be diagonal matrixes. The sensors (camera and

IMU) are installed close to horizontally on the vehicle and their attitudes have small variation amplitude, diagonal elements of Q and W can be determined by the uncertainty of *roll*, *pitch*, *yaw* estimation from different sensors, respectively. Diagonal elements of Q can be determined by the uncertainty of VO's motion estimation. Diagonal elements of W can be determined by the uncertainty of IMU's orientation estimation.

Finally, update equation of the filter turns out to be

$$\begin{aligned}\hat{\boldsymbol{x}}_k &= \hat{\boldsymbol{x}}_k^- + K_k(\boldsymbol{y}_k - \hat{\boldsymbol{x}}_k^-) = \delta\hat{\boldsymbol{e}}_k^- + K_k(\delta\hat{\boldsymbol{e}}_{\mathrm{IMU_k}}^- \delta\hat{\boldsymbol{e}}_{\mathrm{VO_k}}^-) \\ &\approx \delta\hat{\boldsymbol{e}}_{\mathrm{VO_k}}^- + K_k\delta\boldsymbol{e}_{\mathrm{IMU_k}}^{VO}. \end{aligned} \tag{16}$$

$\delta\boldsymbol{e}_{\mathrm{IMU}}^{\mathrm{VO}}$ in (16) denotes the MRPs of $\delta q_{\mathrm{IMU}}^{VO}$, and

$$\delta q_{\mathrm{IMU}}^{VO} = \Delta q_{VO}^T \otimes \Delta q_{\mathrm{IMU}}.$$

Here, Δq_{IMU} denotes the rotation between two adjacent key-frames measured by the IMU, and Δq_{VO} denotes the rotation between two adjacent key-frames measured by the VO. $\delta q_{\mathrm{IMU}}^{VO}$ denotes the relative orientation between Δq_{VO} and Δq_{IMU}.

After filtering, $\hat{\boldsymbol{x}}_k$ is transformed into unit quaternion by formula (13) and combined with the result of VO to obtain the final estimates. Obviously, there are two advantages in our OF; the first one is that our filter occurs in current orientation's error space, which meets the linear requirement of Kalman filter. The second advantage is that our filter does not require addition operation on the unit quaternion, this will maintain the unit constraint of the unit quaternion.

D. Position Filtering

As mentioned in Section II, our PF uses the measurements of IMU as the forecast and the observation comes from the VO, which is opposite to our OF. In additional, the fusion intervals of these two filters are designed to be different based on the following considerations.

The OF working on the interval of the key-frame can improve the estimation accuracy of the orientation. However, the increment on the fusion interval also brings two drawbacks. First, the estimation of the position and velocity from the measurements of IMU will suffer from the long fusion interval due to the Abbe error[2] of the large acceleration measurement (most of the time caused by large gravity acceleration), even tiny bias on the linear acceleration will lead to dramatic position error accumulation over time. Thus, the accurate estimation of the position and velocity with IMU is only limited to the short time interval; second, the large fusion interval will also lead to rough estimations on the position and velocity during the interval, because of the random drift of IMU and lack of feedback. This will constraint the implementation of the VIO in the cases of visual servo, which need smooth position and velocity output in a short cycle. The experiment in Fig. 9 demonstrates the rough estimation of the velocity when using large fusion interval.

In order to solve the aforesaid problems, we set the fusion interval of our PF as the camera cycle. And we calculate each frame's motion relative to its previous key-frame, thus we can obtain the VO output at each frame. Then, the fusion interval in PF can be reduced to minimum even when the key-frame technique is used.

In our PF, the state vector of the PF is $[\boldsymbol{P}\ \boldsymbol{V}]^T$, i.e., position vector $\boldsymbol{P}$ and linear velocity vector $\boldsymbol{V}$ in the coordinate $\{G\}$. The cycles of the camera and IMU are denoted as ΔT and Δt, respectively, and $i = \Delta T/\Delta t$. According to the integral of the IMU's measurements, the motion between two adjacent camera frames, i.e., frame m and frame m-1, can be estimated with the following discretized model:

$$\begin{bmatrix}\boldsymbol{P}\\ \boldsymbol{V}\end{bmatrix}_m = \begin{bmatrix}\mathbf{I} & \Delta T * \mathbf{I}\\ 0 & \mathbf{I}\end{bmatrix}\begin{bmatrix}\boldsymbol{P}\\ \boldsymbol{V}\end{bmatrix}_{m-1} + \sum_{j=1}^{i} A^{j-1}B\boldsymbol{u}_{\boldsymbol{n-j}} \tag{17}$$

where

$$A = \begin{bmatrix}\mathbf{I} & \Delta t * \mathbf{I}\\ 0 & \mathbf{I}\end{bmatrix}, \quad B = \begin{bmatrix}\Delta t^2/2 * \mathbf{I}\\ \Delta t * \mathbf{I}\end{bmatrix}$$

and $\boldsymbol{u_n}$ is the linear acceleration in coordinate $\{G\}$ measured by the IMU and whose timestamp is closest to and less than (or equal to) the camera's capture moment of frame m, shown in Fig. 3.

Let $\boldsymbol{x} = [\boldsymbol{P} \quad \boldsymbol{V}]^T$, there are also two noise sources, the process noise $\boldsymbol{\omega}$ which incorporate linear acceleration noise and the VO measurement noise $\boldsymbol{v}$. Then, the filter equations are

$$\begin{cases}\boldsymbol{x}_m = \begin{bmatrix}\mathbf{I} & \Delta T * \mathbf{I}\\ 0 & \mathbf{I}\end{bmatrix}\boldsymbol{x}_{m-1} + \sum_{j=1}^{i} A^{j-1}B(\boldsymbol{u_{n-j}} + \boldsymbol{\omega}_{n-j})\\ \boldsymbol{y}_m = \begin{bmatrix}\mathbf{I} & 0\end{bmatrix}\boldsymbol{x}_m + v_m.\end{cases} \tag{18}$$

Then, the PF is given as follow:

Filter 3: PF.

Initialization: set initial values for the filter state $[\boldsymbol{P}\ \boldsymbol{V}]^T$.

Propagation: For each camera cycle, ΔT

(I) use formula (17) to estimate the values of position and velocity with the linear acceleration obtained from IMU.

(II) Propagate covariance with the state transformation matrix.

Update: For each camera cycle, perform a Kalman update with the measurement of the VO between two camera frames.

The covariance matrix Q of process noise ω and the covariance matrix W of measurement noise $\boldsymbol{v}$ are taken as hyperparameters in this filter and are set to be diagonal matrices with nonzero entries. As the uncertainty of the acceleration mainly comes from the Abbe error of large acceleration measurement, the diagonal elements of Q can be estimated from the direction and the norm value of the acceleration measurement in the global coordinate $\{G\}$. The diagonal elements of W can be determined by the uncertainty of motion estimation from VO.

With the PF, the linear velocity can be computed accurately. When VO fails, we still can use the maintained linear velocity and the acceleration provided by the IMU to estimate the translation motion, the orientation can also be estimated by the

[2] Abbe error, also called sine error, describes the magnification of angular error over distance. The error of orientation brings large translational acceleration errors from the large acceleration measurement.

angular speed measured by the IMU. Then, our stereo VIO can provide robust estimation under various conditions.

IV. Experiments

In this section, we carry out two kinds of experiments to evaluate the proposed method. In our comparative experiments, we compare with three state-of-the-art approaches as follows:

The first method is an IMU aided stereo visual odometry algorithm proposed by, Agrawal, and Sol [6]. This algorithm only fuses orientation information of the IMU and the VO by an EKF. In their approach, the results of VO are used to forecast and the measurements of the IMU are used to give observation. At the same time, the measurements of the IMU are used to make drift-free attitude estimation. Their algorithm is denoted as *Fusion1* in the following experiments.

The second method is a visual-inertial fusion algorithm proposed by Tardif, George, Laverne, Kelly, and Stentz [8]. This method uses a delayed Kalman filter to fuse position and orientation information from both the IMU and the VO. As their filter puts the last values of the position and the Euler angle estimated by the filter into the state vector, it is called delayed Kalman filter. In that filter, the measurements of the IMU are used to forecast and the results of VO are used to observe. Their algorithm is denoted as *Fusion2* in the following experiments.

The third one is the key-frame-based nonlinear optimization algorithm presented by Leutenegger *et al.* [11]. In this method, Graph Optimization [25] is used as the optimization framework. The measurements of IMU and visual estimation are tightly coupled. As the outlier rejection is performed by applying a chi-square test with IMU-based pose predictions, there is no RANSAC involved in their approach. And the IMU measurement is integrated with landmark reprojection errors in the probabilistic manner. This method is denoted as *optimization* in the following experiments.[3]

A. Experiments on KITTI Datasets

In the first experiment, we use the KITTI datasets [26] to evaluate our approach and compare with the other three state-of-the-art methods. The KITTI datasets are captured on an autonomous vehicle platform named Annieway. It is equipped with an inertial navigation system (GPS/IMU, OXTS RT 3003), measurements from the GPS can be used as ground truths. There are also two grayscale cameras, two color cameras and other sensors mounted on top of the vehicle.

In our experiment, we use the data captured by the IMU and the image pairs of the color cameras. These two color cameras are mounted horizontally with a baseline of 53 cm and can capture high-quality images with a resolution of 1226×370 pixels. We use the synchronized dataset, i.e., "2011_09_30_drive_0018," which is calibrated and synchronized at a frame rate of 10 Hz. The images in KITTI datasets have high-quality features. Since the velocity of the vehicle is almost constant, all the images in these datasets can be regarded as key-frames of VO and almost will not lead to any VO failure. So the problem of *mismatch of fusion interval* is not major in this condition.

[3]For a fair comparison, the window size of the *optimization* method is set to 2 in our experiments.

TABLE I
Average Trajectory Errors of Five Algorithms in the Kitti Datasets

	AEX (m)	AEY (m)	AED (m)
Stereo VIO	2.12 ± 2.52	2.82 ± 2.23	$\mathbf{3.53 \pm 3.37}$
Optimization	3.26 ± 4.43	1.81 ± 2.34	3.72 ± 5.02
Fusion1	7.95 ± 10.07	4.35 ± 4.68	9.06 ± 11.10
Fusion2	7.09 ± 8.85	4.38 ± 5.04	8.33 ± 10.19
VO only	18.68 ± 15.54	5.98 ± 7.58	19.61 ± 17.29

As the KITTI datasets can provide ground-truth GPS positions in the plane of X and Y direction for each image frame. In the following experiments, we introduce three quantitative metrics to evaluate the performance of varied methods. These three metrics are *Average Error in X direction* (AEX), *Average Error in Y direction* (AEY) and *Average Error in Distance* (AED) respectively, calculated as follows:

$$\mathrm{AEX} = \frac{\sum_{i=1}^{N} \left| X_i - X_i^{\mathrm{GPS}} \right|}{N}, \quad \mathrm{AEY} = \frac{\sum_{i=1}^{N} \left| Y_i - Y_i^{\mathrm{GPS}} \right|}{N}$$
$$\mathrm{AED} = \frac{\sum_{i=1}^{N} \sqrt{\left(X_i - X_i^{\mathrm{GPS}}\right)^2 + \left(Y_i - Y_i^{\mathrm{GPS}}\right)^2}}{N}. \tag{19}$$

Here, N is the number of frames, X_i^{GPS} and Y_i^{GPS} are the ground-truth values in X direction and Y direction obtained by the GPS at the ith frame, X_i and Y_i are the output results in X direction and Y direction obtained by the corresponding odometry method at the ith frame. Both AEX and AEY depend on the choice of the coordinate system, slight rotation of the coordinate system might cause very different error, while AED is invariant to the choice of coordinate systems. So AED is the most important and discriminated evaluation metric.

The results of AEX, AEY, and AED and their corresponding variances for different methods are given in Table I. In the experiments, we also compare with the odometry results estimated by the original stereo VO method (denoted as VO *only* in figures), which is introduced in Section III A. The results show that our stereo VIO method can achieve the best performance on AED among all the five methods, the variance of stereo VIO on AED is also smallest among all the methods, which indicates the output results of the stereo VIO are most stable compared with other methods.

The visual results on the odometry estimations of different methods are shown in Fig. 4. Fig. 4 shows the trajectories consisting of each (X_i, Y_i) point estimated by odometry methods in all the frames, the ground-truth trajectories for each $(X_i^{\mathrm{GPS}}, Y_i^{\mathrm{GPS}})$ point are also plotted in this figure.

During the experiments of KITTI datasets, pure VO estimation fails five times in dataset "2011_09_30_drive_0018." The positions that VO failed are also plotted, point *a* to point *e* in Fig. 4, obviously, those positions are almost at turning corners. The huge biases of the pure stereo VO's trajectories in Fig. 4 also indicate that pure stereo VO may not provide satisfied odometry

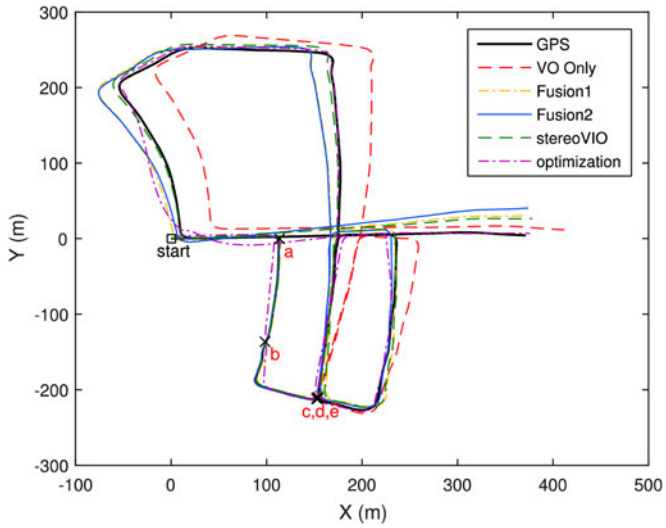

Fig. 4. Results on dataset "2011_09_30_drive_0018." Comparison on ground truth GPS, VIO, VO, Optimization, Fusion 1, and Fusion2. The mark *a-e* are the positions that VO failed.

Fig. 6. Our toy vehicle VIO platform and the sample images captured by our platform (including different scenes, variation of illumination, moving objects, and blur).

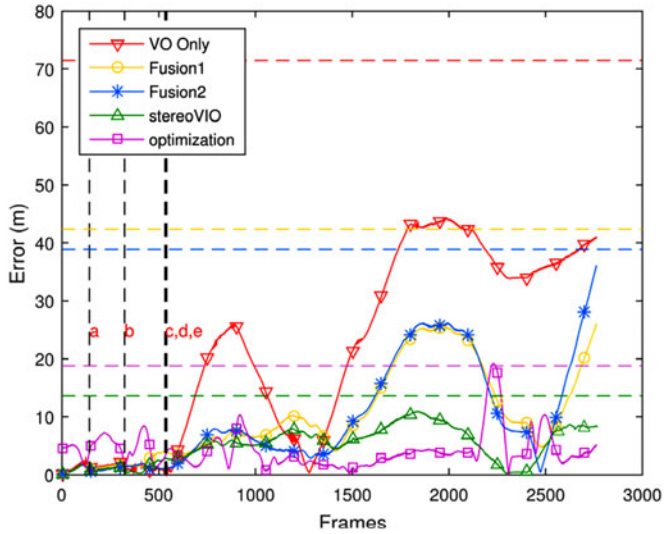

Fig. 5. Error curves of varied methods comparing with the ground truths from GPS on dataset "2011_09_30_drive_0018". The horizontal coordinate of the figure is the frame sequence of the camera. The vertical coordinate is the bias value between the ground truth and the estimation in the corresponding frame $(\sqrt{(X_i - X_i^{GPS})^2 + (Y_i - Y_i^{GPS})^2})$. The curve close to the zero line means the error is small. The mark *a-e* are the positions that VO failed. The horizontal dash line represents the 3-sigma error boundary of the corresponding method with the same color.

estimation when the platform drives a long distance and contains some VO failures. The results in Fig. 4 show that all the VIO algorithms can achieve much better odometry trajectories compared with the pure stereo VO algorithm.

We also present the error curves of all the methods compared with the ground truths from GPS. Fig. 5 shows the distance error curve in each frame and the corresponding 3-sigma error boundary[4] of the error curve for each method. The results show that our VIO can have smaller overall errors and error boundary compared with other state-of-the art methods. So our VIO is capable to navigate in real world for kilometers and performs better than other state-of-the-art algorithms.

B. Experiments on Small Mobile Platform

In order to evaluate the performance of our approach under universal and challenging environments, we also build navigation system, equipped with an IMU (Xsens MTi28A53G35) and a PointGrey Bumblebee2 stereo camera, mounted on a small toy vehicle, shown in Fig. 6. Our IMU can output measurements at a rate of 100 Hz, and the stereo camera works at a frame rate of 15 Hz, and the resolution for each camera is 640×480. In this experiment, the toy vehicle drove on our campus, which has a rough terrain environment with uphill and downhill. The images captured during the experiments are shown in Fig. 6.

The IMU (MTi28A53G35) used in our platform is very noisy and low-cost, the bias and noise of our IMU are significant larger than other IMUs used in the state-of-the-art VIO methods [11]–[13].

Comparing to the KITTI datasets, the datasets captured by our small mobile platform have more challenges and irregular motion styles. The challenges include:

The baseline of the stereo camera used in our system is 12 cm, which is much less than that in KITTI datasets. It is well known that shorter baseline will lead to larger triangulation errors in the stereo VO system; the view of our system is also much narrower than the view in KITTI, which will lead to less feature points in stereo matching and, thus, reduce the accuracy; the IMU used in KITTI is much more accurate than ours. Its drift is also much less

[4]Here the 3-sigma error boundary is calculated as 3σ, and σ is the standard deviation of the error curve on frames for the corresponding method.

Fig. 7. Experiment 1: trajectory obtained by our VIO in the circle experiment. As there are uphill and downhill in circles, the trajectories of every circles are arbitrary due to the rough terrain.

than ours; The most important challenge is that images captured from our system are often blurred, shown in Fig. 6, due to jolt when the toy vehicle bumping on rough terrain. Furthermore, the rough terrain will also lead to unstable stochastic motions which will lead to many failures to the estimation of the VO.

In our experiments, the velocity of the toy vehicle is random, and the intervals of the key-frames are usually much larger than the intervals of the camera images. As the pure VO is easy to fail, the *mismatch of the fusion interval* is especially severe to the accuracy of VIO in this condition. To evaluate the effectiveness that the separated filters working on different fusion intervals can accommodate this situation well, we also introduce another two comparable VIO methods.

VIO-Key, which applies the fusion of the OF and PF presented in this paper between every two key-frames. Then, those two filters' fusion intervals are manually set as the cycle of the key-frames;

VIO-All, which applies the fusion of the OF and PF presented in this paper between every two image frames. Those two filters' fusion intervals are manually set as the capture cycle of the vision sensor.

In the first experiment, our toy vehicle drove around a lawn for four circles and the length of each circle is about 150 m. As the roads around the lawn are quite rough, there are many frames where the images are blurred. Fig. 7 shows the trajectory estimated by our approach. The white point in Fig. 7 is the starting point that our platform will drive through in each circle. In this experiment, we will use the closed-loop error in the starting point to evaluate the performances. We set the initial global zero point at the starting point, and then the closed-loop error is calculated as $\sqrt{x^2+y^2+z^2}$, where (x, y, z) is the algorithm's output of the position when the platform returns to the starting point. The results of all six approaches are given in Table II. The results show that our stereo VIO has the smallest closed-loop error in each circle.

TABLE II
CLOSED-LOOP ERRORS OF SIX ALGORITHMS IN THE FIRST EXPERIMENT WITH THE SMALL MOBILE PLATFORM

	1^{st} circle error (m)	2^{nd} circle error (m)	3^{th} circle error (m)	4^{th} circle error (m)	Average error (m)
Fusion1	1.27	1.91	2.24	3.90	2.33 ± 1.26
Fusion2	1.76	3.78	6.49	6.53	4.64 ± 5.37
Stereo VIO	**0.90**	**1.17**	**1.66**	**1.78**	$\mathbf{1.38 \pm 0.17}$
VIO-Key	1.01	1.93	2.08	3.35	2.09 ± 0.92
VIO-All	1.38	3.70	4.12	5.87	3.77 ± 3.42
Optimization	2.51	4.99	5.69	4.26	4.36 ± 1.86

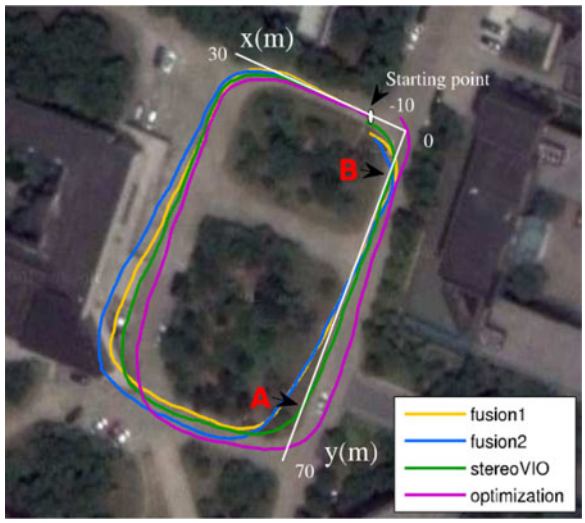

Fig. 8. Experiment 2: results of our VIO and other three state-of-the-art algorithms, the mobile platform drives anticlockwise, and the white point is the start point.

TABLE III
CLOSED-LOOP ERROR IN THE SECOND EXPERIMENT DRIVING AROUND A LARGE LAWN CIRCLE

	Closed-loop error (m)
Fusion1	4.43
Fusion2	4.16
Stereo VIO	1.91
VIO-Key	4.36
VIO-All	4.87
Optimization	6.57

In the second experiment, the toy vehicle drove around a large lawn for about 200 m and went back to the starting position. Trajectories of four approaches are plotted in Fig. 8. The closed-loop errors of all approaches are given in Table III. In this experiment, the vehicle drove fast on the rough terrain that causes a lot of blurred images between point A and point B as marked in Fig. 8. According to the results shown in Fig. 8, we can find that *fusion1*, *fusion2, and optimization* have obvious bias on their estimations, and our stereo VIO can estimate the trajectory that is closest to the real one (at least, it is closest to the starting point in Fig. 8). As we know, the blurred images between point A and point B will lead to many failures on

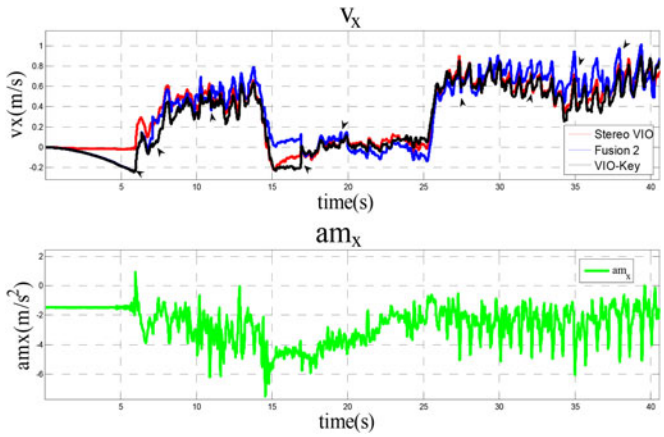

Fig. 9. Velocity of *x* direction estimated by Stereo VIO, Fusion 2, and VIO-Key.

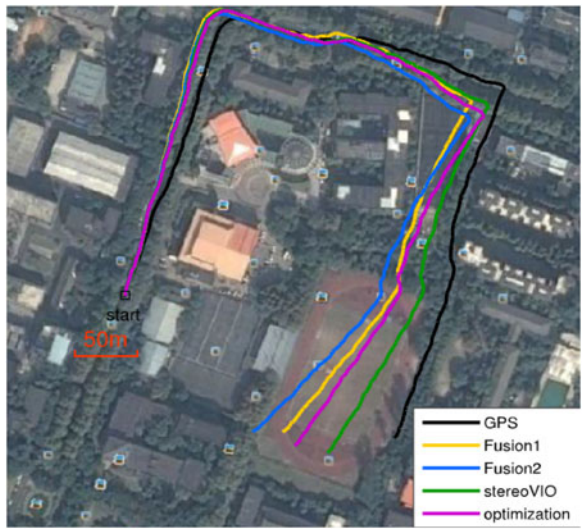

Fig. 10. Experiment 3: trajectories of four algorithms comparing with the ground truths from GPS.

the VO estimation. *Fusion1* does not provide PF, thus it will lose translation information when VO fails. *Fusion2* does not use attitude filter to obtain stable and drift-free attitude for the system, so the noised attitude estimation from our low-precise IMU will impair the fusion results especially when the VO fails. The large bias on *optimization* method may be caused by the failure of nonlinear optimization, as the highly noised IMU and VO failures in this dataset cannot give reliable constraints for the optimization.

In both experiments, the results show that the proposed Stereo VIO, which fuse at varied OF and PF fusion intervals, will always outperform VIO-Key and VIO-All, especially in experiment 2, where the VO failure occurs frequently. The VIO-All fuses at every two image frames and the effect of IMU is excessively suppressed, while VIO-Key fuses at every two key-frames, its estimation accuracy will be greatly decreased once the VO failure occurs.

We also plot the velocity curves estimated by the PF of *Stereo VIO*, *Fusion 2,* and *VIO-Key* in Fig. 9. In Fig. 9, the positions of *Stereo VIO* fuse with all the sequential camera frame pairs. The positions of *Fusion 2* and *VIO-Key* fuse with selected key frame pairs. The green line shows the corresponding acceleration measurements by the IMU along its *x*-axis. From the green line, we can observe that the vehicle keeps stationary at the first 6 s the same as the real experiment, hile there is a significant linear velocity drift on the key-frames based fusion (blue line, black line). The black arrows show some timestamps where the key-frames are selected, it is also notable that the linear velocity undergoes sharp jump at these timestamps. Thus, the PF fed with key frame pairs will output less accurate velocity compared with the PF fed with all camera frame pairs. From Fig. 9, we can observe that *Stereo VIO*'s minimal position fusion interval has better velocity estimation, which is crucially important when VO fails.

Thus, our method can successfully avoid the conflict between minimizing the fusion interval to reduce the error of the IMU drift and the use of a large interval to improve the VO estimation accuracy.

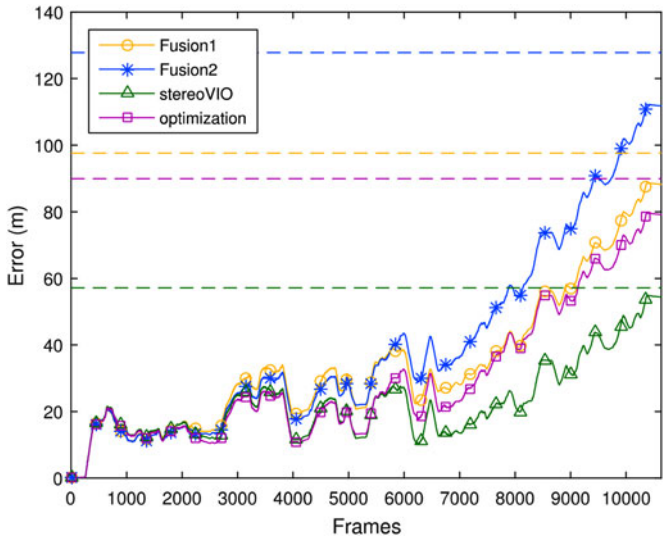

Fig. 11. Error curves of varied methods comparing with the ground truths from GPS on experiment 3. The horizontal coordinate of the figure is the frame sequence of the camera. The vertical coordinate is the bias value between the ground truth and the estimation in the corresponding frame ($\sqrt{(X_i - X_i^{GPS})^2 + (Y_i - Y_i^{GPS})^2}$). The curve close to the zero line means the error is small. The horizontal dash line represents the 3-sigma error boundary of the corresponding method with the same color.

In the third experiment, the toy vehicle equipped with a GPS sensor drove around a long distance about 1000 m in the same rough terrain environment as the experiment 2. The estimated trajectories and error curves by various methods are given in Figs. 10 and 11. The AEX, AEY, and AED results are also shown in Table IV. The results show that our algorithm can also achieve the best AED and corresponding variance among all the methods.

TABLE IV
AVERAGE TRAJECTORY ERRORS OF FOUR ALGORITHMS IN THE DATASET CAPTURED BY SMALL MOBILE PLATFORM

	AEX (m)	AEY (m)	AED (m)
Stereo VIO	20.00 ± 13.03	6.92 ± 5.10	$\mathbf{22.60 \pm 11.52}$
Optimization	27.17 ± 21.84	7.18 ± 5.17	29.88 ± 20.03
Fusion1	32.49 ± 22.74	5.76 ± 4.44	34.37 ± 21.08
Fusion2	38.04 ± 30.71	7.93 ± 5.53	40.43 ± 29.14

The experiments in our small mobile platform show our approach can achieve superior performance comparing with other state-of-the-art methods.

V. CONCLUSION AND FUTURE WORK

We have presented a novel multiple Kalman filters ensemble algorithm for VIO in challenging environments. To adapt the *mismatch of the fusion interval*, our algorithm uses separated OF and PF working on different cycles to estimate the 6DOF odometry of the system. In order to obtain high accuracy in long-term estimations, we introduce an attitude filter fusing with the input accelerations and angular speed of the IMU to obtain long-term stable attitude. In our approach, the OF is built on the orientation error space, which is a local linear space and more suitable for the Kalman filter. The experiments carried out in this paper have proved that our algorithm is superior to other state-of-the-art algorithms.

Compared with traditional fusion methods which need to consider complex nonlinear coupling of the states, the main advantage of our approach is to employ a simple hierarchical fusion architecture assembling with multiple simple Kalman filters, which are easy to be implemented on those hardware platforms with limited resources. Experimental results show the performance of our approach is also robust to the low-precision IMU, thus our method may have broad application prospects on low-cost hardware systems such as ARM, FPGA, etc.

In our future work, we will further optimize our VIO algorithm, reduce its computation complexity, and try to implement it in the compact embedded platform. Integration of other low-cost sensors, such as conventional low-cost GPS, is also a possible future task. As our current attitude filter sets fixed biases for the gyroscope and accelerometer, there remains an unknown bias in the output of our attitude filter. In addition, our attitude filter is not tuned to handle large accelerations and, thus, the filter is sensitive to attitude errors caused by them. Therefore, we will try to introduce the DCM-based attitude estimation algorithm [27] to improve the performance of our method by estimating the gyroscope biases online and adapting attitude filter to handle large accelerations. In addition of this future improvement, we will also explore the method to adjust accelerometer biases using the output of the VO, as the velocity measured by VO should be equal to the velocity integrated from bias corrected accelerations.

REFERENCES

[1] W. Wang, W. Dong, Y.U. Su, D. Wu and Z. Du, "Development of search-and-rescue robots for underground coal mine applications," *J. Field Robot.*, vol. 31, no. 3, pp. 386–407, Feb. 2014.

[2] B. Ugurlu and A. Kawamura, "On the backward hopping problem of legged robots," *IEEE Trans. Ind. Electron.*, vol. 61, no. 3, pp. 1632–1634, Mar. 2014.

[3] A. Suzumura and Y. Fujimoto, "Real-time motion generation and control systems for high wheel-legged robot mobility," *IEEE Trans. Ind. Electron.*, vol. 61, no. 7, pp. 3648–3659, Jul. 2014.

[4] R. C. Luo and C. C. Lai, "Multisensor fusion-based concurrent environment mapping and moving object detection for intelligent service robotics," *IEEE Trans. Ind. Electron.*, vol. 61, no. 8, pp. 4043–4051, Aug. 2014.

[5] R. C. Luo and C. C. Lai, "Enriched indoor map construction based on multisensor fusion approach for intelligent service robot," *IEEE Trans. Ind. Electron.*, vol. 59, no. 8, pp. 3135–3145, Aug. 2012.

[6] K. Konolige, M. Agrawal and J. Sol, "Large-scale visual odometry for rough terrain," in *Robotics Research, Springer Tracts in Advanced Robotics*, vol. 66. Berlin, Germany: Springer, 2011, pp. 201–212.

[7] S. Sirtkaya, B. Seymen and A. Alatan, "Loosely coupled kalman filtering for fusion of visual odometry and inertial navigation", in *Proc. Int. Conf. FUSION*, Istanbul, Turkish, Jul. 2013, pp. 219–226.

[8] J. P. Tardif, M. George, M. Laverne, A. Kelly and A. Stentz, "A new approach to vision-aided inertial navigation," in *Proc. IEEE/RSJ Int. Conf. IROS*, Taipei, Taiwan, Oct. 18–22, 2010, pp. 4161–4168.

[9] H. Chao, C. Coopmans, L. Di and Y. Chen, "A comparative evaluation of low-cost IMUs for unmanned autonomous systems," in *Proc. IEEE Conf. Multisensor Fusion Integr. Intell. Syst.*, Salt Lake City, UT, USA, Sep. 2010, pp. 211–216.

[10] W. Wang and G. Xie, "Online high-precision probabilistic localization of robotic fish using visual and inertial cues," *IEEE Trans. Ind. Electron.*, vol. 62, no. 2, pp. 1113–1124, Feb. 2015.

[11] S. Leutenegger, P. Furgale, V. Rabaud, M. Chli, K. Konolige, R. Siegwart and W. Garage, "Keyframe-based visual-inertial slam using nonlinear optimization," in *Proc. Robot. Sci. Syst.*, Berlin, Germany, Jun. 2013. pp. 1–7.

[12] T. Lupton and S. Sukkarieh, "Visual-inertial-aided navigation for high-dynamic motion in built environments without initial conditions," *IEEE Trans. Robot.*, vol. 28, no. 1, pp. 61–76, Feb. 2012.

[13] J. A. Hesch, D. G. Kottas, S. L. Bowman and S. I. Roumeliotis, "Camera-IMU-based localization: Observability analysis and consistency improvement," *Int. J. Robot. Res.*, vol. 33, no. 1, pp. 182–201, Jan. 2014.

[14] D. Scaramuzza and F. Fraundorfer, "Visual odometry Part I: The first 30 years and fundamentals," *IEEE Robot. Autom. Mag.*, vol. 18, no. 4, pp. 80–92, Dec. 2011.

[15] M. Li and A. I. Mourikis, "Improving the accuracy of EKF-based visual-inertial odometry," in *Proc. IEEE Int. Conf. Robot. Autom.*, St Paul, MN, USA, May 14–18, 2012, pp. 828–835.

[16] E. Jones and S. Soatto, "Visual-inertial navigation, mapping and localization: A scalable real-time causal approach," *Int. J. Robot. Res.*, vol. 30, no. 4, pp. 407–430, Apr. 2011.

[17] S. Hilsenbeck, A. Moller, R. Huitl, G. Schroth, M. Kranz and E. Steinbach, "Scale-preserving long-term visual odometry for indoor navigation," in *Proc. Int. Conf. Indoor Positioning Indoor Navigation*, Univ. New South Wales, Sydney, Australia, Nov. 13–15, 2012, pp. 1–10.

[18] J. Kelly and G. Sukhatme, "Visual-inertial sensor fusion: Localization, mapping and sensor-to-sensor self-calibration," *Int. J. Robot. Res.*, vol. 30, no. 1, pp. 56–79, Jan. 2011.

[19] S. Weiss and R. Siegwart, "Real-time metric state estimation for modular vision-inertial systems," in *Proc. IEEE Int. Conf. Robot. Autom.*, Shanghai, China, May 9–13, 2011, pp. 4531–4537.

[20] P. Corke, J. Lobo, and J. Dias, "An introduction to inertial and visual sensing," *Int. J. Robot. Res.*, vol. 26, no. 6, pp. 519–535, Jun. 2007.

[21] H. Rehbinder and H. Xiaoming, "Drift-free attitude estimation for accelerated rigid bodies," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 21–26, 2001, pp. 4244–4249.

[22] D. Luigi, M. Stefano, and T. Federico, "ZNCC-based template matching using bounded partial correlation", *Pattern Recognit. Lett.*, vol. 26, no. 14, pp. 2129–2134, Oct. 2005.

[23] H. Alismail, B. Browning, and M. B. Dias. "Evaluating pose estimation method for stereo visual odometry on robots," in *Proc. 11th Int. Conf. Intell. Auton. Syst.*, 2010, pp. 101–110.

[24] N. Trawny and S. Roumeliotis, "Indirect kalman filter for 3d attitude estimation," Dept. Comput. Sci. Eng., University of Minnesota, Minneapolis, MN, USA, *Tech. rep.*, 2005.
[25] R. Kummerle, G. Grisetti, H. Strasdat, K. Konolige and W. Buigard, "g2o: A general framework for graph optimization," in *Proc. IEEE Int. Conf. Robot. Autom.*, Shanghai, China, May 2011, pp. 3607–3613.
[26] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," in *Proc. IEEE Conf. CVPR*, Providence, RI, USA, Jun. 16–21, 2012, pp. 3354–3361.
[27] H. Hyyti and A. Visala, "A DCM based attitude estimation algorithm for low-cost MEMS IMUs," *Int. J. Navig. Obs.*, vol. 2015, Article ID 503814, p. 18, 2015.

Yong Liu (M'11) received the B.S. degree in computer science and engineering and the Ph.D. degree in computer science from Zhejiang University, Zhejiang, China, in 2001 and 2007, respectively.

He is currently an Associate Professor with the Institute of Cyber-Systems and Control, College of Control Science and Engineering, Zhejiang University. His current research interests include machine learning, robotics vision, and information fusion. He has published over 30 research papers on machine learning, computer vision, information fusion, and robotics.

Rong Xiong (M'10) received the B.Sc. and M.Sc. degrees in computer science and engineering and the Ph.D. degree in control science and engineering from Zhejiang University, Hangzhou, China, in 1994, 1997, and 2009, respectively.

She has been with the State Key Laboratory of Industrial Control Technology, Zhejiang University, since 1997, where she is currently a Professor and directs the Robotics Laboratory. Her current research interests include machine vision, simultaneous localization and mapping, motion planning, and control for humanoid robots. She is the common corresponding author of this paper.

Yue Wang (S'10) received the B.Sc. degree in communication engineering from Zhejiang University of Technology, Hangzhou, China, in 2011. He is currently working toward the Ph.D. degree in control science and engineering at Zhejiang University, Hangzhou, China. He is a joint Ph.D. student with Stanford University, Stanford, CA, USA, funded by the China Scholarship Council.

His research interests include mobile robots, machine learning, and big data analysis.

Hong Huang received the B.S. degree from the School of Power and Mechanical Engineering, Wuhan University, Wuhan, China, in 2011, and the M.S. degree in control science and engineering, from Zhejiang University, Hangzhou, China, in 2014.

He is currently a Positioning and Navigation Algorithm Engineer with Guozi Robot Company, Hangzhou. His research interests include multiple sensor fusion, positioning and navigation, computer vision, and SLAM.

Xiaojia Xie received the B.S. degree in mechatronics engineering from the University of Electronic Science and Technology of China, Chengdu, China, in 2014. He is currently working toward the M.S. degree in control science and engineering at Zhejiang University, Hangzhou, China.

His research interests include visual inertial odometry and multiple sensor fusion.

Xiaofeng Liu received the B.S. degree from Xi'an Jiaotong University, Xi'an, China in 2013, and the M.S. degree from Zhejiang University, Hangzhou, China, in 2016.

He is currently with the China Securities Depository and Clearing Corporation Limited, Beijing, China. His research interests include computer vision, machine learning, and data mining.

Gaoming Zhang received the B.S. degree in control science and engineering from Harbin Institute of Technology, Harbin, China, in 2014. He is currently working toward the M.S. degree in control science and engineering at Zhejiang University, Hangzhou, China.

His research interests include mobile robot path planning and visual navigation.