Stereo Visual-Inertial Odometry with Multiple Kalman Filters Ensemble

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

Abstract—In this paper, we present a stereo visual-inertial odometry (VIO) 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 (VO) 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 6DOF (Degree of Freedom) 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 3-5 times of the camera’s sampling rate. This means that the IMU will output 3-5 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
We design a cascading fusion architecture, which fuses the orientation and position filters working on different frequencies to estimate the 6DOF odometry of the system. In order to support low-precision IMUs working at hostile environments, we come up with the second challenge to design a robust VIO fusion algorithm which can overcome the conflict mentioned in the first challenge. Here we use separated algorithm with multiple Kalman filters ensemble. To overcome challenging environments and present a novel stereo VIO to support low-precision IMUs working at hostile environments in low-computational resources [10]. Here we decompose the classical integrated VIO filter into three sub-filters, Attitude Filter (AF), Orientation Filter (OF) and Position Filter (PF). This separated fusion framework can support 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.

- Separated fusion for sensors with varied sampling rate
- Cascading and multiple-level-fusion architecture
- Ensemble with low-cost linear sub-filters

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 unified nonlinear filter, those three linear sub-filters in our VIO only need to estimate a few state vectors, they also do not need to compute the Jacobians in the EKF or sigma point in the UKF.
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, 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 orientation filter 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 Kurt Konolige et al [6], which only uses the orientation filter 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 above 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, orientation filter and position filter. In our position filter, the propagation and observation of the position states come from the IMU and the VO respectively, which is same as the first category. In our orientation filter, the propagation and observation of the orientation states come from the VO and the IMU respectively, which is 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-precision IMUs.

III. STEREO VISUAL-INERTIAL ODOMETRY

Before introducing the detailed filters, the coordinates used in our stereo VIO are introduced first, shown in figure 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 \( \{ G \} \). 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 \( \{ G \} \).
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIE.2016.2573765, IEEE Transactions on Industrial Electronics

Fig. 3. Fusion intervals for three sub-filters. The Orientation Filter has the largest fusion interval $\Delta \chi$; the fusion interval of Position Filter is $\Delta T$; and the Attitude Filter has the smallest fusion interval $\Delta t$.

C. 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 (ZNCC) score [22]. Using those feature pairs from left-right cameras, we can reconstruct the sparse 3D points. By tracking the feature pairs between sequential images with the SURF descriptors, the motion estimation can be described as a 3D-to-2D problem [23], which refers to estimating the motion from the sparse 3D features calculated by the stereo vision system in the earlier frames and the corresponding matched 2D features in current frames. In our 3D-to-2D motion estimation, RANSAC is also used to remove those outliers. The 3D-to-2D feature pairs from both cameras are all considered in the same optimization function, which tries to minimize the re-projection 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 0.3m or the norm of the Rodrigues representation of the rotation from successive candidate frame to the current key-frame is larger than 0.25radian.

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 only integrated from the measured angular speed may suffer from drift. Thus we apply an attitude Kalman filter [21], i.e. filter 1 in figure 2, to output long-term stable attitude, i.e. the angles of roll and pitch.

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

$$
\mathbf{a} = \mathbf{R}(\mathbf{u} + (-\mathbf{g})) = -\mathbf{g} + \mathbf{R}\mathbf{u} = \mathbf{R}[0 \ 0 \ g]^T + \mathbf{R}\mathbf{u} = gx + \mathbf{R}\mathbf{u}
$$

(1)

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

$$
\mathbf{x} = [\sin(pitch) \ \cos(pitch) \ \sin(roll) \ \cos(pitch) \ \cos(roll)]^T
$$

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

Based on the kinematics, we have $\mathbf{R} = [\hat{\theta} \times]\mathbf{R}$, that is

$$
\dot{\mathbf{x}} = [\hat{\theta} \times]\mathbf{x}
$$

(2)

Here $\hat{\theta}$ is the 3D angular speed of the IMU. $[\hat{\theta} \times]$ is the skew symmetric matrix, it is also denoted as $S(\hat{\theta})$. Discretizing formula (2), we can get

$$
\mathbf{x}_n = \mathbf{A}_{n-1}\mathbf{x}_{n-1}
$$

(3)

Here $\mathbf{A}_n = I + \frac{S(\hat{\theta})\sin(|\hat{\theta}|\tau)}{|\hat{\theta}|\tau} + \frac{S(\hat{\theta})(1-\cos(|\hat{\theta}|\tau))}{|\hat{\theta}|^2\tau^2}$. $\tau$ is the sampling cycle of IMU. $\mathbf{A}_n$ is an orthogonal rotation matrix. $\mathbf{I}$ is an identity matrix of $3 \times 3$.

Based on formula (1) and (3), we can get the system model:

$$
\begin{align*}
\{ & \mathbf{x}_n = \mathbf{A}_{n-1}\mathbf{x}_{n-1} \\
& \mathbf{y}_n = \mathbf{x}_n + \mathbf{R}_n \mathbf{u}_n/g
\end{align*}
$$

(4)

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

$$
\begin{align*}
\{ & \mathbf{x}_n = \mathbf{A}_{n-1}\mathbf{x}_{n-1} + \omega_{n-1} \\
& \mathbf{y}_n = \mathbf{H}_n \mathbf{x}_n + \nu_n
\end{align*}
$$

(5)

$\mathbf{H}_n = \sigma_n \mathbf{I}$, $\sigma_n$ is a binary variable which equals 1 or 0 determined by the acceleration. $\mathbf{H}_n$ 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

\footnote{According to [21], attitude refers to the robot’s orientation relative to the gravity vector, usually described by pitch and roll.}
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

**Propagaton:** 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 \(\omega\) is a tuning parameter in this filter and it is assumed to be a diagonal matrix with non-zero entries. The diagonal elements of \(Q\) can be estimated through the noise of angular speed measurements. The covariance matrix \(W\) of \(\nu\) 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 which determines the value of \(\sigma_n\).

**C. Orientation Estimation with Indirect Kalman Filter**

The orientation filter is used to estimate the three orientation angles of the system, i.e. yaw, pitch and roll. In our orientation filter, we introduce the indirect Kalman filter [24], which uses the orientation error space represented by the error quaternion instead of the orientation represented by the unit quaternion. According to the indirect Kalman filter, we introduce the indirect Kalman filter [24], which uses the orientation error space represented by the error quaternion instead of the orientation represented by the unit quaternion. The state vector of this filter state \(Q\) instead of the orientation represented by the unit quaternion can be estimated mainly by the acceleration threshold which determines the value of \(\sigma_n\).

We use the unit quaternion, \(q\), 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 \(\bar{q}\) to denote the estimate of \(q\), and \(\delta q\) to denote the error between \(q\) and \(q\) as following:

\[
\dot{q} = \bar{q} \otimes \delta q \quad \text{(6)}
\]

\(\otimes\) denotes the multiplication operation of the quaternion.

Assuming the rotation estimated by VO between two adjacent key-frames can be denoted with \(\Delta q\). Then we have

\[
\dot{q}_k = \bar{q}_k \otimes \Delta q_k \quad \text{(7)}
\]

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

\[
\delta q_k \approx \bar{q}_k \otimes \Delta q_k \approx \bar{q}_k \otimes \delta q_k \otimes \Delta q_k \quad \text{(8)}
\]

Then we have the following deriviation,

\[
\delta q_k = (\bar{q}_k \otimes \Delta q_k)^T \otimes (\bar{q}_k \otimes \delta q_k \otimes \Delta q_k) = \delta q_{k-1} \otimes \Delta q_{k-1} \otimes \Delta q_k \quad \text{(9)}
\]

Here, \(\delta q_{k-1}\) can be denoted as

\[
\delta q_{k-1} = [\delta q_0 \quad \delta q_1 \quad \delta q_2 \quad \delta q_3] \quad \text{(10)}
\]

Then

\[
\delta q_k = \Delta q_k^T \otimes [\delta q_0 \quad 0 \quad 0 \quad 0] \otimes \Delta q_{k-1} + \\
\Delta q_{k-1} \otimes [0 \quad \delta q_1 \quad \delta q_2 \quad \delta q_3] \otimes \Delta q_{k-1}
\]

\[
= [\delta q_0 \quad 0 \quad 0 \quad 0] \otimes [0 \quad 0 \quad 0 \quad 0] + [\begin{bmatrix} 0 \\ \Delta R_{k-1}^T \\ \Delta q_{k-1} \end{bmatrix} \begin{bmatrix} \delta q_1 \\ \delta q_2 \\ \delta q_3 \end{bmatrix}] \quad \text{(11)}
\]

\(\Delta R\) is the rotation matrix corresponding to \(\delta q\). 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 \(\Delta R^T\).

With the transformation of Modified Rodrigues Parameters (MRPs) [18]:

\[
\delta e = [\delta q_1/(1 + \delta q_0) \quad \delta q_2/(1 + \delta q_0) \quad \delta q_3/(1 + \delta q_0)]^T \quad \text{(12)}
\]

According to formula (9), \(\delta q_0\) is invariant between state \(k-1\) to \(k\). Then we have:

\[
\begin{bmatrix} \delta q_1/(1 + \delta q_0) \\ \delta q_2/(1 + \delta q_0) \\ \delta q_3/(1 + \delta q_0) \end{bmatrix}_k = \Delta R_{k-1}^T \begin{bmatrix} \delta q_1/(1 + \delta q_0) \\ \delta q_2/(1 + \delta q_0) \\ \delta q_3/(1 + \delta q_0) \end{bmatrix}_{k-1} \quad \text{(13)}
\]

And can be abbreviated as:

\[
\delta e_k = \Delta R_{k-1}^T \delta e_{k-1} \quad \text{(14)}
\]

MRPs will bring additional convenience during the calculation of the transformation. The inverse transformation, from \(\delta e\) to \(\delta q\), is

\[
\delta q_0 = \frac{1-||\delta e||^2}{1+||\delta e||^2} \quad \delta q = (1 + \delta q_0)\delta e \quad \text{(15)}
\]

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

\[
x_k = \Delta R_{k-1} x_{k-1} + \omega_{k-1} \quad \text{(16)}
\]

If we use \(\delta e\) as the measurement vector, the observation equation of the system is

\[
y_k = x_k + \nu_k \quad \text{(17)}
\]

\(\omega\) in formula (14) denotes process noise and \(\nu\) in formula (17) represents observation noise of the system.

Then the orientation filter is given in the following:
Filter 2 Orientation Filter

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

Observation:
(I) Get $\textit{yaw}$ from IMU measurement, $\textit{pitch}$ and $\textit{roll}$ from Filter 1 at the time that the last key-frame is acquired.
(II) Get $\textit{yaw}$ from IMU measurement, $\textit{pitch}$ and $\textit{roll}$ 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 $x_k$ and $y_k$ 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 $\omega$ and the covariance matrix $W$ of measurement noise $v$ are set to be diagonal matrices. 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 $\textit{roll}$, $\textit{pitch}$, $\textit{yaw}$ 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

$$\hat{x}_k = \hat{x}_{k-1} + K_k (y_k - \hat{x}_k) = \delta e_{vO}^k + K_k (\delta e_{IMU}^k - \delta e_{vO}^k)$$

$$\delta e_{vO}^k$$ in formula (16) denotes the MRPs of $\delta q_{vO}^k$. And,$$
\delta q_{IMU}^k = \Delta q_{vO} \otimes \Delta q_{IMU}$$

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

After filtering, $\hat{x}_k$ 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 orientation filter; the first one is that our filter occurs in current frame, 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 position filter uses the measurements of IMU as the forecast and the observation comes from the VO, which is opposite to our orientation filter. In additional, the fusion intervals of these two filters are designed to be different based on the following considerations.

The orientation filter 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. Firstly, the estimation of the position and velocity from the measurements of IMU will suffer from the long fusion interval due to the Abbe error 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; Secondly, 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 constrain 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 figure 9 demonstrates the rough estimation of the velocity when using large fusion interval.

In order to solve the above problems, we set the fusion interval of our position filter 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 position filter can be reduced to minimum even when the key-frame technique is used.

In our position filter, the state vector of the position filter is $[P \ V]^T$, i.e. position vector $P$ and linear velocity vector $V$ in the coordinate $\{G\}$. The cycles of the camera and IMU are denoted as $\Delta T$ and $\Delta t$ 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:

$$[P]_m = [\begin{bmatrix} 1 & \Delta t \end{bmatrix}] [P]_{m-1} + \sum_{j=1}^{i-1} A^{-j}Bu_{n-j}$$

Where $A = [\begin{bmatrix} 1 & \Delta t \end{bmatrix}]$, $B = [\begin{bmatrix} \Delta t^2/2 \end{bmatrix}]$ and $u_n$ 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 figure 3.

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

$$\begin{bmatrix} x_m \\ y_m \end{bmatrix} = \begin{bmatrix} 1 & \Delta t \end{bmatrix} [\begin{bmatrix} x_{m-1} \\ y_{m-1} \end{bmatrix} + \sum_{j=1}^{i-1} A^{-j}Bu_{n-j} + \omega_{n-j}]$$

Then the position filter is given as follows:

Filter 3 Position Filter

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

Propagation: For each camera cycle, $\Delta T$
(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 $\omega$ and the covariance matrix $W$ of measurement noise $\nu$ are taken as hyper-parameters in this filter and are set to be diagonal matrices with non-zero 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 position filter, 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 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 Kurt Konolige et al [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 Jean-Philippe Tardif et al [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 Stefan Leutenegger et al [11]. In this method, Graph Optimization (G2O) [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. Their algorithm is denoted as optimization in the following experiments.

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 53cm and can capture high-quality images with a resolution of $1226 \times 370$ pixels. We use the synchronized data set, i.e. “2011_09_30_drive_0018”, which is calibrated and synchronized at a frame rate of 10HZ. 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.

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:

$$AEX = \frac{\sum_{i=1}^{N} |X_i^{GPS} - X_i^{VO}|}{N}, \quad AEY = \frac{\sum_{i=1}^{N} |Y_i^{GPS} - Y_i^{VO}|}{N}$$

$$AED = \frac{\sum_{i=1}^{N} \sqrt{(X_i^{GPS} - X_i^{VO})^2 + (Y_i^{GPS} - Y_i^{VO})^2}}{N}$$

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 $i$th frame, $X_i$ and $Y_i$ are the output results in X direction and Y direction obtained by the corresponding odometry method at the $i$th 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 figure 4. Figure 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 $(X_i^{GPS}, Y_i^{GPS})$ point are also plotted in this figure.

During the experiments of KITTI datasets, pure VO estimation fails 5 times in dataset “2011 09 30_drive_0018”. The positions that VO failed are also plotted, point $a$ to point $e$ in figure 4, obviously, those positions are almost at turning
with uphill and downhill. The images captured during the experiments are shown in figure 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 12cm, 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 than ours; The most important challenge is that images captured from our system are often blurred, shown in figure 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.

We also present the error curves of all the methods compared with the ground truths from GPS. Figure 5 shows the distance error curve in each frame and the corresponding 3-sigma error boundary 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.

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

![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.](image)

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 figure 6. Our IMU can output measurements at a rate of 100Hz, and the stereo camera works at a frame rate of 15Hz, and the resolution for each camera is $640 \times 480$. In this experiment, the toy vehicle drove on our campus, which has a rough terrain environment.
**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.

![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).](Image)

In the first experiment, our toy vehicle drove around a lawn for four circles and the length of each circle is about 150 meters. As the roads around the lawn are quite rough, there are many frames where the images are blurred. Figure 7 shows the trajectory estimated by our approach. The white point in figure 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 error in the starting point to evaluate the performances. We set 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.

![Image](Image)

In the second experiment, the toy vehicle drove around a large lawn for about 200 meters and went back to the starting point. Trajectories of four approaches are plotted in figure 8. The closed-loop errors of all approaches are given in table III. In this experiment, the vehicle drove fast on the rough terrain which causes a lot of blurred images between point \(A\) and point \(B\) as marked in figure 8. According to the results shown in figure 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 figure 8). As we know, the blurred images between point \(A\) and point \(B\) will lead to many failures on the VO estimation. Fusion1 does not provide position filter, 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.

![Image](Image)

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.

**TABLE II**

<table>
<thead>
<tr>
<th></th>
<th>1st circle error (m)</th>
<th>2nd circle error (m)</th>
<th>3rd circle error (m)</th>
<th>4th circle error (m)</th>
<th>Average error (m)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fusion1</td>
<td>1.27</td>
<td>1.91</td>
<td>2.24</td>
<td>3.90</td>
<td>2.33±1.26</td>
</tr>
<tr>
<td>Fusion2</td>
<td>1.76</td>
<td>3.78</td>
<td>6.49</td>
<td>6.53</td>
<td>4.64±2.37</td>
</tr>
<tr>
<td>Stereo VIO</td>
<td>0.90</td>
<td>1.17</td>
<td>1.66</td>
<td>1.78</td>
<td>1.38±0.17</td>
</tr>
<tr>
<td>VIO-Key</td>
<td>0.91</td>
<td>1.92</td>
<td>2.08</td>
<td>2.18</td>
<td>1.99±0.17</td>
</tr>
<tr>
<td>VIO-All</td>
<td>0.38</td>
<td>3.70</td>
<td>4.12</td>
<td>5.87</td>
<td>3.77±3.42</td>
</tr>
<tr>
<td>Optimization</td>
<td>2.51</td>
<td>4.99</td>
<td>5.69</td>
<td>4.26</td>
<td>4.36±1.86</td>
</tr>
</tbody>
</table>

![Image](Image)

We also plot the velocity curves estimated by the PF of Stereo VIO, Fusion 2 and VIO-Key in figure 9. In figure 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 seconds the same as the real experiment. While 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 position filter fed with key frame-pairs will output less accurate velocity compared with the position filter fed with all camera frame-pairs. From figure 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.

In the third experiment, the toy vehicle equipped with a GPS sensor drove around a long distance about 1000 meters in the same rough terrain environment as the experiment 2. The estimated trajectories and error curves by various methods are given in figure 10 and figure 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.
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 present a novel multiple Kalman filters ensemble algorithm for VIO in challenging environments. To adapt the mismatch of the fusion interval, our algorithm uses separated orientation filter and position filter 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 orientation filter 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.

![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.](image)

![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.](image)

![Fig. 9. Velocity of x direction estimated by Stereo VIO, Fusion 2, and VIO-Key.](image)

### TABLE III

<table>
<thead>
<tr>
<th>Closed-loop error (m)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fusion1</td>
</tr>
<tr>
<td>Fusion2</td>
</tr>
<tr>
<td>Stereo VIO</td>
</tr>
<tr>
<td>VIO-Key</td>
</tr>
<tr>
<td>VIO-All</td>
</tr>
<tr>
<td>Optimization</td>
</tr>
</tbody>
</table>
TABLE IV
AVERAGE TRAJECTORY ERRORS OF FOUR ALGORITHMS IN THE DATASET CAPTURED BY SMALL MOBILE PLATFORM

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>AEX (m)</th>
<th>ASE (m)</th>
<th>AED (m)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stereo VIO</td>
<td>20.00±2.13</td>
<td>6.92±1.10</td>
<td>22.60±1.15</td>
</tr>
<tr>
<td>Optimization</td>
<td>27.17±2.99</td>
<td>7.18±1.97</td>
<td>29.86±2.03</td>
</tr>
<tr>
<td>Fusion1</td>
<td>32.49±2.74</td>
<td>5.76±1.64</td>
<td>34.71±2.11</td>
</tr>
<tr>
<td>Fusion2</td>
<td>30.44±3.07</td>
<td>7.93±1.53</td>
<td>40.43±2.94</td>
</tr>
</tbody>
</table>


Yong Liu (M’11) received the B.S. degree in computer science and engineering and the Ph.D. degree in computer science, both 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 in 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.

Yue Wang (SM’10) received his BSc degree in Communication Engineering from Zhejiang University of Technology, Hangzhou, P.R. China in 2011. Currently, he is pursuing his PhD supervised by Professor Rong Xiong in Control Science and Engineering, Zhejiang University, Hangzhou, China. He is a joint PhD student with Stanford University, California, United States, funded by China Scholarship Council. His research interests include mobile robot, machine learning and big data analysis.

Hong Huang received the B.S. degree in School of Power and Mechanical Engineering, Wuhan University, Wuhan, in 2011, and the M.S. degree supervised by Associate Professor Yong Liu in Control Science and Engineering, Zhejiang University, Hangzhou, China in 2014. He is currently a positioning and navigation algorithm engineer of Guozi Robot Company. His research interests include multiple sensor fusion, positioning and navigation, Computer vision, and SLAM.

Xiaofeng Liu received the B.S. degree from Xi’an Jiaotong University, Xi’an, China in 2013, and M.S. Degree from Zhejiang University, Hangzhou, China in 2016. He is currently working in China Securities Depository and CLeaning Corporation Limited. 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 pursuing 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.

Xiaojia Xie received the B.S. degree in Mechatronics Engineering from University of Electronic Science and Technology of China, Chengdu, China in 2014. He is currently pursuing 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.