2018 IEEE International Conference on Robotics and Automation (ICRA)
May 21-25, 2018, Brisbane, Australia

Detection and Resolution of Motion Conflict in Visual Inertial Odometry

VIO中的运动冲突的检测与解决

Benzun Pious Wisely Babu[1], David Cyganski[1], James Duckworth[1], Soohwan Kim[3]

新

***Abstract*— In this paper, we present a novel method to detect and resolve *motion conflicts* in visual-inertial odometry. Recently, it has been common to integrate an IMU sensor with visual odometry in order to improve localization accuracy and robustness. However, when a disagreement between the two sensor modalities occurs, the localization accuracy reduces drastically and leads to irreversible errors. In such conditions, multiple motion estimates based on the set of observations used are possible. This creates a conflict (*motion conflict*) in determining which observations to use for accurate ego-motion estimation. Therefore, we present a method to detect motion conflicts based on per-frame positional estimate discrepancy and per-landmark reprojection errors. Additionally, we also present a method to resolve motion conflicts by eliminating inconsistent IMU and landmark measurements. Finally, we implement Motion Conflict aware Visual Inertial Odometry (MC-VIO) by combining both detection and resolution of motion conflicts. We perform quantitative and qualitative evaluation of MC-VIO on visually and inertially challenging datasets. Experimental results indicate that the MC-VIO algorithm reduces the increase in absolute trajectory error by 80% and the relative pose error by 60% for scenes with motion conflict, in comparison to the state-of-the-art reference VIO algorithm.**

I. INTRODUCTION

Ego-motion estimation is a fundamental problem in mobile devices such as autonomous cars, humanoids and even augmented reality. Recent work on Visual Odometry (VO) [1] has achieved accurate pose tracking in real-time with visual features [2], [3], [4], [5] and direct pixel information [6], [7], [8], [9]. However, due to the single sensor modality, VO is prone to fail in challenging situations such as texture-less visual environments and changing-light conditions.

To overcome this limitation, IMU measurements are applied to improve the robustness of VO, leading to Visual-Inertial Odometry (VIO). The IMU has a higher sampling rate but its accuracy decreases with drift over time. On the other hand, VO has a lower frame rate but its accuracy increases with repeated observations over time. Thanks to this complementary nature, VIO produces more accurate and robust estimation in highly dynamic environments that lack photometric and geometric variations. The linear acceleration and angular velocity obtained from IMU sensors are fused using an Extended Kalman Filter or nonlinear optimization in loosely coupled [10] or tightly coupled integration [11], [12], [13].

[1]Benzun Pious Wisely Babu, David Cyganski and James Duckworth are with Worcester Polytechnic Institute, 100 Institute Road, Worcester, MA, USA {bpwiselybabu, cyganski, rjduck}@wpi.edu

[2]Soohwan Kim is with the Commonwealth Scientific and Industrial Research Organisation (CSIRO). This work was done when he was at Bosch Research and Technology Center, Palo Alto, USA. soohwan.kim@csiro.au

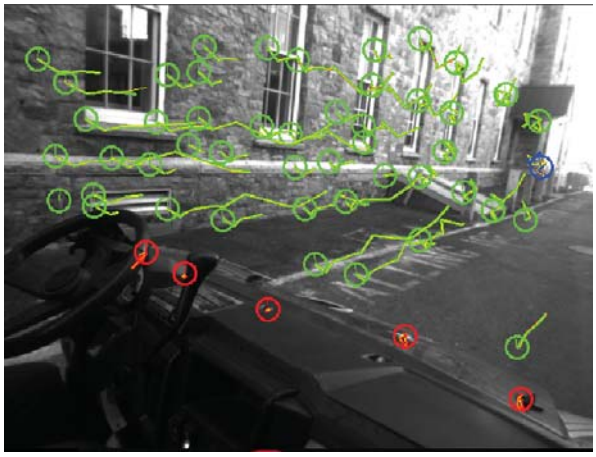

Fig. 1. Example of motion conflict: A moving camera in car sees static landmarks outside (green) with large optical flow and moving landmarks on the dashboard (red) with small optical flow. Each group of landmarks produce a different ego-motion estimate. Landmarks with noisy optical flow (blue) give incorrect ego-motion estimates.

Often, the underlying assumption of VIO that the visual and inertial measurements are in agreement with each other is violated in real-world applications. For example, consider a robot driving a vehicle [14] or a passenger carrying a VIO device on a mobile platform (Fig. 1). In order to estimate the motion of the VIO device with respect to the inertial frame, the IMU measurements and the landmarks outside the vehicle can be used, as they are in agreement with respect to the targeted motion. However, the landmarks inside the vehicle do not convey information regarding this motion. This condition is reversed if we need to estimate the motion of VIO with respect to the vehicle frame. A fundamental disagreement, as opposed to that due to noise, between measurements in a multi-sensor device which reduces the accuracy and robustness of the estimated motion is termed as *motion conflict*.

In this paper, we present novel methods to detect and resolve motion conflicts. Particularly, to detect motion conflicts we implement (1) a per-frame conflict detector that combines the difference in the positional estimates between the IMU-only and visual-only system and (2) a per-landmark conflict detector based on the marginal reprojection error from multiple views. To resolve the motion conflicts, we implement (1) an IMU dominated approach that retains only the inertial constraints during motion conflict and (2) a selective fusion approach that additionally retains the visual constraints that are in agreement with inertial motion during motion conflict.

One might suggest RANSAC [15] or M-estimators [16] as

不能使用RANSAC的原因：

an outlier detection and rejection approach to solve motion conflict. However, these approaches find the best fit using maximal support, and cannot determine which ego-motion is consistent with the inertial frame when maximal support is not available for it. Additionally, in scenes where two motions with similar support exist, RANSAC might oscillate between them, as it has no way to determine which motion to consistently follow.

The main contributions of this paper are:

- methods for motion conflict detection
- methods for motion conflict resolution
- evaluation of Motion Conflict aware VIO (MC-VIO) in visually and inertially challenging scenes.

Section II summarizes the existing visual localization approaches for the dynamic world. Next, section III provides a brief mathematical foundation for VIO. It is followed by section IV which extends VIO to situations with motion conflicts. Section V and section VI describe our approach for detecting motion conflicts and resolving motion conflicts respectively. Next, section VII provides a description of the implementation of Motion Conflict aware VIO (MC-VIO) using a triple window optimization strategy. Section VIII presents qualitative and quantitative evaluation of the MC-VIO algorithm. Finally, Section IX presents conclusions and direction for future work.

II. BACKGROUND

With the seminal paper by Smith et al. [17], the fundamentals of visual localization using Simultaneous Localization and Mapping (SLAM) were formulated. This led to the probabilistic formulation of SLAM using EKF [2] and FastSLAM [18]. The current focus on SLAM research is to improve its robustness to real-life applications [19].

To extend SLAM from static world to dynamic world cases, outlier rejection schemes to improve matching, such as Joint Compatibility Branch and Bound [20] or RANSAC [21] have been suggested. In contrast to this approach which rejects matches on dynamic objects, SLAMMOT [22], SLAMIDE [23] use matches from dynamic objects to improve the pose estimation. Recently, Reddy et al. [24] suggested to use a factor graph based approach to track moving cars with ego-motion estimation for autonomous driving assistance. However, none of the existing approaches handle multi-sensor localization in visually and inertially challenging environments.

Mourikis and Roumeliotis [25] performed tight coupling of visual and inertial sensors using an augmented state Kalman filter. Careful initialization and calibration of the IMU was essential to prevent the divergence of the EKF. Pre-integration of the IMU measurements was suggested by [26] to quickly estimate the initial conditions for a loosely coupled visual inertial odometry system. OKVIS [11] introduced nonlinear optimization based approach for visual inertial odometry. ORB SLAM [5] extended bias initialization to algorithms with loop closure.It is worth mentioning that all existing visual inertial algorithms reject measurements that do not agree with the inertial measurements as outliers.

In contrast to existing approaches, our work aims at developing a visual inertial localization algorithm that selects measurements based on the frame of reference in which the motion is to be estimated. The selection is performed by analyzing the conflicts that exist between the sensors and the estimator.

III. Visual Inertial Odometry

The objective of localization in robotics is to estimate the trajectory of the system with respect to the world frame W based on observation made on the sensor frame S. The trajectory is a part of the state of the VIO system ($\mathbf{X}^W_{0:N}$) consisting of the pose ${}^W\mathbf{p}_{WS}$, orientation $\mathbf{q}_{WS}$ and velocity ${}^S\mathbf{v}$. Additionally, the IMU linear acceleration and rotational velocity biases $\mathbf{b}_a, \mathbf{b}_g$ and the position of landmarks ${}^W\mathbf{l}_j$ are also added to the state vector. We estimate the trajectory based on the Maximum à Posteriori (MAP) criterion optimization of the state,

$$\mathbf{X}_k := \left[{}^W\mathbf{p}_{WS}{}^\top,\quad \mathbf{q}_{WS}{}^\top,\quad {}^S\mathbf{v}_{WS}{}^\top,\quad \mathbf{b}_g{}^\top,\quad \mathbf{b}_a{}^\top,\quad \mathbf{l}_1^{W\top},\quad \ldots,\quad \mathbf{l}_n^{W\top}\right]_k^\top \in \mathbb{R}^3 \times S^3 \times \mathbb{R}^9 \times \mathbb{R}^{4n}$$

The observations of the VIO state are made using a synchronized IMU-stereo camera pair. Our observations at each time step k consists of feature matches $\mathbf{z}_{1:k-1}$, and raw IMU measurements $\mathbf{u}_k = \{{}^S\tilde{\mathbf{a}},\ {}^S\tilde{\boldsymbol{\omega}}_{WS}\}$.

The MAP estimation of the state is represented as

$$\hat{\mathbf{X}}_k = \underset{X}{\operatorname{argmax}}\, \mathrm{P}\left(\mathbf{X}_k \mid \mathbf{z}_{1:k-1}, \mathbf{u}_{1:k}\right) \tag{1}$$

The error state $\delta\mathbf{X}_k$ is described using local parameterization $\delta\boldsymbol{\chi}_k$ in the tangent space of the state manifold.

$$\begin{aligned} \hat{\mathbf{X}}_k &= \mathbf{X}_k \oplus \delta\boldsymbol{\chi}_k \\ \delta\chi_k &= \Phi^{-1}(\log(\mathbf{X}_k)) \\ \mathbf{X}_k &= \exp\left(\Phi\left(\delta\boldsymbol{\chi}_k\right)\right) \end{aligned} \tag{2}$$

$$\delta\boldsymbol{\chi}_k := \left[\delta\mathbf{p}^\top,\quad \delta\boldsymbol{\alpha}^\top,\quad \delta\mathbf{v}^\top,\quad \delta\mathbf{b}_g^\top,\quad \delta\mathbf{b}_a^\top,\quad \delta\mathbf{l}_j^\top\right]^\top \in \mathbb{R}^{15+3N}$$

The previous estimate state $\hat{\mathbf{X}}_{k-1}$ is propagated ($\mathbf{X}_k = f(\hat{\mathbf{X}}_{k-1}, \mathbf{u}_k)$) using $\mathbf{u}_k$ according to the IMU kinematics $f(\cdot)$ described in Equation (3), similar to [11].

$$\begin{aligned} {}^W\dot{\mathbf{p}} &= \mathbf{C}_{\mathbf{WS}}\,{}^S\mathbf{v} \\ \dot{\boldsymbol{q}}_{WS} &= \frac{1}{2}\boldsymbol{\Omega}\left({}^S\tilde{\boldsymbol{\omega}}_{\boldsymbol{WS}} - \mathbf{b}_g\right)\mathbf{q}_{WS} \\ {}^S\dot{\mathbf{v}}_{WS} &= \left({}^S\tilde{\mathbf{a}}_{WS} - \mathbf{b}_a\right) + {}^W\mathbf{g} \\ \dot{\mathbf{b}}_g &= \mathbf{n}_{bg} \\ \dot{\mathbf{b}}_a &= -\frac{1}{\tau}\mathbf{b}_a + \mathbf{n}_{ba} \end{aligned} \tag{3}$$

where $\boldsymbol{\Omega}(\cdot)$ defines the cross product matrix operator for rotation rates, $\mathbf{g}$ represents gravity and $\mathbf{n}_{bg}, \mathbf{n}_{ba}$ are random walk noises obtained from the IMU manufacturer. The nonlinear continuous time IMU kinematics $f(\cdot)$ can be linearized and represented as a difference equation:

$$\delta\dot{\boldsymbol{\chi}} \approx \mathbf{F}_\mathbf{d}(\mathbf{X}_k)\,\delta\boldsymbol{\chi}_k + \mathbf{Q}(\mathbf{X}_k) \tag{4}$$

where $\mathbf{Q}$ is the process noise and $\mathbf{F_d}$ is the first order Taylor series approximation of $f(\cdot)$.

With the higher rate inertial measurements, the state of the system is propagated and when the camera measurements are available, an update based on the error state of the system is formulated. The prediction error is the difference between the prior state $\mathbf{X}_k$ and the posterior state $\hat{\mathbf{X}}_k$,

$$\mathbf{e}_s^k(\mathbf{X}_k, \mathbf{X}_{k+1}, \mathbf{z}_k, \mathbf{u}_{k-1:k}) = \begin{bmatrix} {}^W\hat{\mathbf{p}}_k - {}^W\mathbf{p}_k \\ 2(\hat{\mathbf{q}}_k \oplus \mathbf{q}_k^{-1}) \\ {}^S\hat{\mathbf{v}}_k - {}^S\mathbf{v}_k \\ \hat{\mathbf{b}}_{g_k} - \mathbf{b}_{g_k} \\ \hat{\mathbf{b}}_{a_k} - \mathbf{b}_{a_k} \end{bmatrix} \tag{5}$$

A non-linear camera measurement model is used to convert the measurement $\mathbf{z}_k$ into a landmark state ${}^W\mathbf{l}$. The transformation $\mathbf{T}_{SC}$ converts the measurement from the camera coordinate frame C to the system coordinate frame S. $\pi(\cdot)$ is the projection function defined by the camera sensor model.

$${}^W\mathbf{l} = \mathbf{T}_{WS}\mathbf{T}_{SC}(\pi^{-1}(\mathbf{z}_k)) \tag{6}$$

The reprojection error of the landmark ${}^W\mathbf{l}_j$ observed by the camera i after the state propagation is given by

$$\mathbf{e}_r^{i,j,k} := \mathbf{z}^{i,j,k} - \pi_i(\mathbf{T}_{CS}\hat{\mathbf{T}}_{SW}{}^W\mathbf{l}_j) \tag{7}$$

Finally, a joint optimization combining both the prediction error and the reprojection error at the image time step k is formulated as a weighted sum [11]. The weight matrix $\mathbf{W}_r$ and $\mathbf{W}_s$ are determined from the inverse of the covariance in the measurements.

$$\mathcal{J}(\delta\chi_k) := \underbrace{\sum_{k=1}^{K}\sum_{i}\sum_{j\in J(k,i)} \mathbf{e}_r^{i,k,j\top}\mathbf{W}_r\mathbf{e}_r^{i,k,j}}_{\text{reprojection error}} + \underbrace{\sum_{k=2}^{K}\mathbf{e}_s^{k\top}\mathbf{W}_s\mathbf{e}_s^k}_{\text{prediction error}} \tag{8}$$

IV. Motion conflict

A robust multi-sensor localization system will have to seamlessly track states not only in a simple static world but also in a complex dynamic world. Some examples of challenging conditions encountered in the real world are:

1) Visually challenging where
   - secondary moving objects not fixed to the world are visible.
   - reflective objects with non lambertian reflection model are visible.
2) Visually and inertially challenging where
   - a VIO system is on a moving vehicle/elevator.

The estimation of ego-motion in such conditions are prone to failure. This is because, based on the set of points selected, (Fig. 1) the estimation can yield different state estimates. The observability of the states differs based on the set of selected visual matches as not all matches are consistent with the estimator assumptions such as static world or lambertian surface.

In a visually and inertially challenging scene, ego-motion is observable by both external (camera) and internal (IMU) sensors. When secondary motion such as when in a vehicle or in an elevator is only observable by one of these sensors, multiple consistent state estimates are possible. Hence, when multiple consistent motions with the physical world are observed by the internal or external sensors, it is important to determine which of these motions are consistent with the robots motion estimator and which describe secondary motions. Failure to determine the correct motion as ego-motion will lead to irrecoverable errors in the VIO system. We term this as *motion conflict*. A generalized Hidden Markov Model

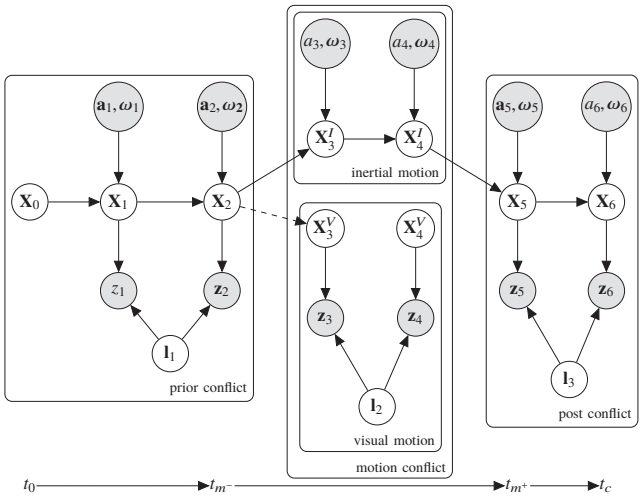

Fig. 2. A generalized Hidden Markov Model (HMM) for VIO in scenes with *motion conflict*. The observations are represented with gray circles and states with white circles. During motion conflict interval $[t_m^-, t_m^+]$ the state of the system is forked.

with time varying states is used to model the VIO (Fig. 2). When there is only one consistent motion observable by the VIO estimator, the state estimation is similar to the existing VIO system [11] but when there is *motion conflict*, the state is forked into two $(\mathbf{X}^I, \mathbf{X}^V)$ independent states. A local visual frame V is used to describe motion non-consistent with the ego-motion of the VIO.

$$\mathbf{X}_k^V := \left[\mathbf{p}_V^{VS\top}, \quad \mathbf{q}_{VS}{}^\top, \quad \mathbf{l}_0^{V\top}, \quad \ldots, \quad \mathbf{l}_n^{V\top}\right]_k^\top \in \mathbb{R}^3 \times S^3 \times \mathbb{R}^{4n}$$

$$\mathbf{X}_k^I := \left[\mathbf{p}_W^{WS\top}, \quad \mathbf{q}_{WS}{}^\top, \quad {}^S\mathbf{v}_{WS}{}^\top, \quad \mathbf{b}_g^\top, \quad \mathbf{b}_a^\top\right]_k^\top \in \mathbb{R}^3 \times S^3 \times \mathbb{R}^9$$

By maintaining two independent states during *motion conflict*, corruption of bias states due to inconsistent visual updates is prevented. Assuming the last state before conflict as $\mathbf{X}_{m^-}$, the state estimated by inertial observations is described by:

$$\hat{\mathbf{X}}^I = \underset{\mathbf{X}_k^I}{\operatorname{argmax}} \, \mathrm{P}(\mathbf{X}_{m^-})\,\mathrm{P}(\mathbf{X}_{k-1}^I \mid \mathbf{X}_{m^-})\,\mathrm{P}(\mathbf{X}^I{}_k \mid \mathbf{X}_{k-1}^I, \mathbf{u}_k) \tag{9}$$

and the state estimated by the visual observations is described by

$$\hat{\mathbf{X}}_k^V = \underset{\mathbf{X}^V}{\operatorname{argmax}} \, \mathrm{P}(\mathbf{X}_{m^-})\,\mathrm{P}(\mathbf{X}_{k-1}^V \mid \mathbf{X}_{m^-})\,\mathrm{P}(\mathbf{X}^V{}_k \mid \mathbf{X}_{k-1}^V, \mathbf{z}^{i,j,k}) \tag{10}$$

In Equation (9) and (10) we use the last estimated state before the motion conflict as a prior $P(\mathbf{X}_{m^-})$. The transition probability $P(\mathbf{X}_{k-1}^I \mid \mathbf{X}_{m^-})$ describes the last IMU sensor

在改变检测器阈值时，false positive rate和true positive rate。超过一个阈值，位置差异无法区分运动冲突和噪声。因此，δMC检测器true positive rate是limited。当匹配比低于阈值Mr且位置差异超过阈值δMC。根据运动冲突序列中采集的样本数据对阈值进行调整。

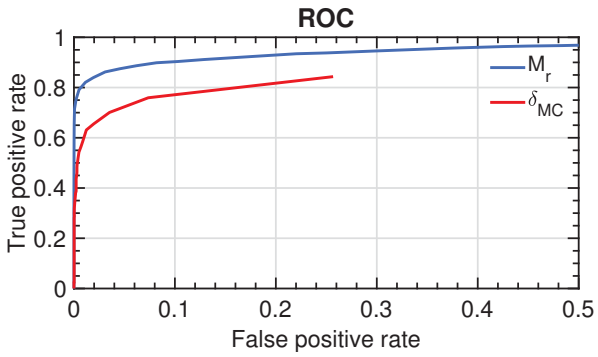

Fig. 3. The Receiver Operating Characteristic (ROC) for the landmark based M_r and pose based δ_{MC} *motion conflict* detectors.

states given the last estimated full VIO state. We have assumed, in this example, the VIO state to be aligned to the inertial frame. Hence we have assumed the transition probability to be identity. On the other hand, The transition probability $P(\mathbf{X}^V \mid \mathbf{X}_{m^-})$ is not directly observable.

V. Detecting motion conflicts

In order to determine which state corresponds to $\mathbf{X}_{m^-}$ and when to apply motion conflict resolution, the motion conflict interval $[t_m, t_m^+]$ needs to be estimated. To detect this interval, first the previous frame is assumed to be the last frame before motion conflict. Then, $\mathbf{X}_{k-1}$ is used to determine $P(\mathbf{X}_k^I)$ and $P(\mathbf{X}_k^V)$. Using these estimated states, tests are performed to determine if the state $\mathbf{X}_k$ corresponds to the scene with motion conflict.

If the state $\mathbf{X}_k$ describes a condition with motion conflict, the error residuals associated with the landmark state estimation will be large since the static world assumption is violated. This is our intuition for developing a per-landmark motion conflict detector.

A per-landmark motion conflict detector response is defined for each landmark $\mathbf{l}_j$ based on the residual error associated with all its previous S observations $\mathbf{z}_{ij}$ and the corresponding projection of the landmark estimate based on IMU-only estimated state $\hat{\mathbf{X}}_i^I$.

$$\delta_{l_j} := \sum_{i \in S} \left(\mathbf{z}_{ij} - h(\hat{\mathbf{X}}_i^I, \mathbf{l}_j) \right) \tag{11}$$

To convert the per-landmark motion conflict detector to a per-frame landmark based detector we used a matching ratio M_r. M_r was calculated as the ratio of the number of landmarks with response greater than threshold $\delta_{l_j}^*$ in the current frame to the total number of landmarks in the current frame.

$$M_r := \frac{\#\text{ landmarks without conflict}}{\#\text{ landmarks}} \tag{12}$$

On the other hand, based on the discrepancy of the estimated poses in the state $\hat{\mathbf{X}}_k^V$ and $\hat{\mathbf{X}}_k^I$, a per-frame motion conflict detector based on the positional discrepancy was defined as follows.

$$\delta_{MC} = \|\hat{\mathbf{p}}_k^V - \hat{\mathbf{p}}_k^I\|_\Sigma \tag{13}$$

The discrepancy was weighted by the relative uncertainty Σ obtained from the state estimate of $\mathrm{P}(\hat{X}_k^V)$ and $\mathrm{P}(\hat{X}_k^I)$. Fig. 3 presents the trade off between the false positive and the true positive rates on varying the detector thresholds. Beyond a threshold, the positional discrepancy cannot distinguish motion conflicts from noise, hence the true positive rate of the δ_{MC} detector was limited. A frame was declared to have motion conflict if the matching ratio was below the threshold M_r^* and the positional discrepancy exceeded the threshold δ_{MC}^*. The thresholds were tuned based on sample data collected in sequences with motion conflict.

VI. Resolving motion conflicts

Since localization is performed with respect to the inertial frame, the ego-motion of the VIO device is consistent with the inertial motion $\mathbf{X}^I$ during motion conflict. Hence, once motion conflict is over at t_{m^+}, the inertial state $\mathbf{X}^I$ is used as an initialization point to estimate a post motion conflict state $\mathbf{X}_{m^+}$. However, since the biases in $\mathbf{X}^I$ after conflict are not updated with visual measurements, there is a drift in the trajectory during motion conflict. Therefore, we corrected the biases when $\mathbf{X}_{m^+}$ was estimated with visual updates. The post motion conflict bias $\mathbf{b}_{a_{m^+}}$ was back-propagated into the $\mathbf{X}^I$ states during motion conflict based on linear interpolation.

$$\mathbf{b}_a{}^I(t) = \frac{t - t_{m^-}}{t_{m^+} - t_{m^-}} (\mathbf{b}_{a_{m^+}} - \mathbf{b}_{a_{m^-}}) + \mathbf{b}_{a_{m^-}} \tag{14}$$

During motion conflict, two approaches based on total and partial independence of visual and inertial measurements were used to maintain the states.

只关注IMU

A. IMU dominated motion conflict resolution

In this approach, we have assumed complete independence of visual and inertial measurements. We rejected all visual measurements as possible disagreements with the inertial measurements. The state $\mathbf{X}^V$ had all the landmarks observed during the motion conflict duration, that were discarded once motion conflict was over.

当IMU与VO一致时使用VO，来提高精度

B. Selective motion conflict resolution

In this approach, we have assumed that during motion conflict there were some visible landmarks in agreement with the inertial measurements. Hence, landmarks that were in agreement with the inertial measurements were moved from the state $\mathbf{X}^V$ and added to state $\mathbf{X}^I$. This provided partial observability of biases during motion conflict and had better localization accuracy.

VII. MC-VIO

We combined the detection and resolution techniques discussed in the previous section to implement Motion Conflict aware Visual Inertial Odometry (MC-VIO). A sliding window based optimization [27] approach was used for the state estimation. The delayed linearization of the problem during motion conflict was essential to maintain consistency. Keyframes similar to [11] were used to improve accuracy of the estimation problem while keeping the computation bounded. The sliding window was divided into three sections (Fig. 4). In the keyframe window, marginalized states [11] and the associated landmarks were maintained. In the IMU

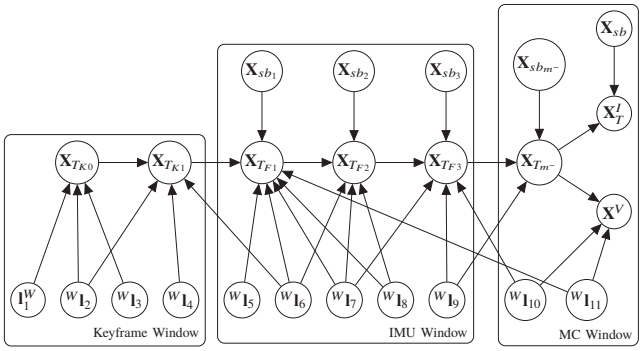

Fig. 4. A triple window optimization problem is constructed when motion conflict is detected. $\mathbf{X}_T = \left[{}^W\mathbf{p}, {}^W\mathbf{q}_{WS}\right]$, $\mathbf{X}_{sb} = \left[{}^S\mathbf{v}, \mathbf{b}_a, \mathbf{b}_g\right]$

window, consecutive frames without marginalization were maintained. Finally, the MC window was only maintained when motion conflict was detected.

With every new image frame received, first the combined per-frame motion conflict detector was applied. Based on the result the estimate was constructed as a triple window or double window optimization. If motion conflict was detected, the MC window was added. Additionally, the state $\mathbf{X}_{m-}$ was converted to a keyframe to prevent marginalization errors. If the end of motion conflict sequence t_{m+} was detected, any existing MC-window was resolved based on the resolution approaches described above and post motion conflict state $\mathbf{X}_{m+}$ was converted back to double window optimization. The marginalizations were performed using a Schur-Complement operation.

VIII. Evaluation

We present the quantitative and qualitative evaluation of our implementation of a reference VIO algorithm similar to [11] and two different configurations of MC-VIO described in Section VI-A and VI-B (denoted as Mode 1 and Mode 2 in the following). The Absolute Tracking Error (ATE) and the Relative Position Error (RPE) [28] were used as metrics to evaluate the localization accuracy of the algorithms.

A. Quantitative evaluation

1) Accuracy: Since there are no standard publicly available datasets containing scenes with motion conflict, artificial motion conflicts were simulated into the EuROC dataset [29]. A noise image representing secondary motion was used to corrupt the dataset at randomized n intervals of m duration. The evaluation for $n = 3$, $m = 5$ for a total duration of 15 seconds motion conflict is presented in Table I. The best performance in each dataset is highlighted. The results on unaltered EuROC dataset without simulated motion conflict are presented in Table II for comparison. We observed an average increase of 0.71m in ATE from 0.218m to 0.934m (≈328%) with the introduction of motion conflict. By using IMU dominated motion conflict resolution approach (Mode 1), the average increase in ATE was reduced to 0.13m (0.349 – 0.218). Thus the increase in ATE error reduced from 0.71m to 0.13m (≈ 80%). A similar increase of 0.135m/s in RPE from 0.199m/s to 0.334m/s (≈ 67%) was observed.

TABLE I

Evaluation of MC-VIO on motion conflict simulated EuROC dataset

Dataset	ATE $[m]$			RPE $[m/s]$		
	VIO	Mode1	Mode2	VIO	Mode1	Mode2
MH_01_easy	0.667	**0.375**	0.388	0.078	0.085	**0.070**
MH_02_easy	0.496	**0.227**	0.255	0.086	**0.065**	0.085
MH_03_medium	0.346	**0.261**	0.299	**0.092**	0.112	0.097
MH_04_difficult	2.152	**0.358**	0.423	0.473	0.405	**0.402**
MH_05_difficult	0.556	0.412	**0.341**	0.399	0.390	**0.387**
V1_01_easy	0.349	0.191	**0.176**	0.245	0.250	**0.210**
V1_02_medium	1.259	**0.340**	0.399	**0.470**	**0.470**	**0.470**
V1_03_difficult	2.470	0.820	**0.760**	0.444	**0.409**	0.420
V2_01_easy	0.397	0.247	**0.130**	0.163	0.182	**0.128**
V2_02_medium	0.611	**0.285**	0.491	0.191	**0.172**	0.176
mean	0.934	**0.349**	0.365	0.334	0.254	**0.244**
std.	0.778	0.179	0.178	0.218	0.152	0.157

TABLE II

Evaluation of MC-VIO on unaltered EuROC dataset as baseline.

EuROC Dataset	ATE $[m]$			RPE $[m/s]$		
	VIO	Mode1	Mode2	VIO	Mode1	Mode2
mean	0.218	0.247	**0.202**	0.199	0.200	**0.198**
std.	0.105	0.153	0.117	0.164	0.165	0.166

In the IMU dominated motion conflict resolution approach (Mode 1), the increase in RPE was reduced to 0.055m/s (0.254 – 0.199). Thus the increase in RPE error, reduced from 0.135 to 0.055m/s (≈ 60%). We also note the selective resolution approach (Mode 2) had better performance to Mode 1 and outperformed Mode 1 in datasets with high dynamic motions.

2) Accuracy with motion conflict duration: The duration of motion conflict was increased to observe its impact on tracking accuracy. We compared the performance of the reference VIO [11] algorithm against the MC-VIO algorithm on the EuROC dataset with simulated motion conflicts as we increased motion conflict duration. Fig. 5 shows that when the motion conflict duration increases, ATE error grew much slower for MC-VIO algorithm than for the reference VIO algorithm.

B. Qualitative evaluation

A number of challenging datasets were collected using a custom built in-house synchronized IMU-stereo camera pair [30] to evaluate the robustness and accuracy of MC-VIO in real world conditions. We present the evaluation on two

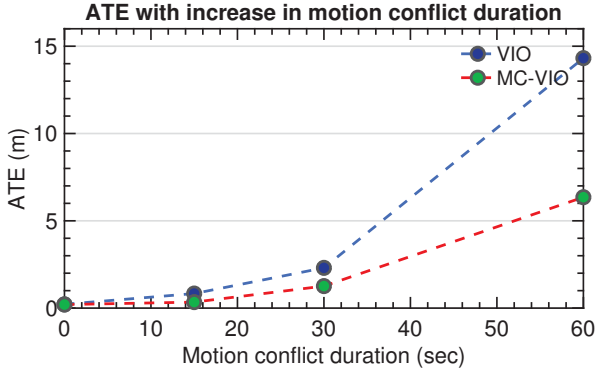

Fig. 5. The ATE error increases with increase in the motion conflict duration. The ATE in MC-VIO grows slower than VIO.

datasets which exemplify visually and inertially challenging environments.

1) Visually challenging environment: A dataset was collected in an indoor environment where an operator with the VIO system walked three loops around a rectangular corridor. The first loop did not contain any motion conflict and was used as reference. In the next two loops, three motion conflict intervals were deliberately introduced by observing an object not attached to the world frame (hand carried laptop computer) while walking. Fig. 6 presents several image frames taken during a period where motion conflict was introduced.

Fig. 7 presents the resultant trajectory plot overlaid on the floor plan of the building. With the reference VIO algorithm, there was a large drift in the position within a motion conflict interval that led to a total shift in the trajectory. This led to a total shift in the final resultant trajectory. With MC-VIO Mode 1, consistent trajectory paths were generated during the motion conflict intervals. Furthermore, with MC-VIO Mode 2, we observed that the second and third loop trajectories had more similarities to the first loop trajectory compared to MC-VIO Mode 1.

2) Visually and inertially challenging environment: A dataset was collected in an outdoor environment in a vehicle that took two loops around a parking lot with the VIO system carried by a passenger. The first loop did not contain any motion conflict and was used as a reference. In the next loop, motion conflict was introduced deliberately by observing the scene inside the vehicle. Fig. 8 presents several image frames taken during the period where motion conflict was introduced.

Fig. 9 presents the resultant trajectory plot overlaid on the map of the parking lot. With the reference VIO there was a very large drift in the trajectory of the path generated. The MC-VIO algorithm with IMU dominated motion conflict resolution (Mode 1) generated a trajectory consistent with the reference first loop trajectory. However, the MC-VIO algorithm with selective motion conflict resolution (Mode 2) generated a trajectory that was consistent with the reference loop trajectory and successfully terminated near the starting point.

IX. CONCLUSIONS

In visually and inertially challenging environments, *motion conflicts* occur. Motion conflict, if not handled correctly, can lead to large irreversible errors in VIO systems. A generalized HMM has been proposed to model motion conflict. Novel motions for detection and resolution techniques were combined in our Motion Conflict aware Visual Inertial Odometry (MC-VIO) algorithm. Quantitative results indicated that MC-VIO was successful in reducing ATE and RPE in scenes with motion conflict. Furthermore, qualitative results showed the robustness of MC-VIO in real world conditions. We believe that motion conflict detection and resolution is essential in a multi-sensor localization algorithm for robust localization in real world conditions. Future work will explore detection and resolution using learning based approaches.

ACKNOWLEDGMENT

The authors would like to thank the RTC-HMI1 team at Bosch Research and Technology Center, Sunnyvale for supporting this research.

REFERENCES

[1] D. Scaramuzza and F. Fraundorfer, "Visual odometry part I: The first 30 years and fundamentals," *IEEE Robotics and Automation Magazine*, vol. 18, no. 4, pp. 80–92, 2011.

[2] A. J. Davison, I. D. Reid, N. D. Molton, and O. Stasse, "MonoSLAM: real-time single camera slam," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 6, pp. 1051–1067, 2007.

[3] G. Klein and D. Murray, "Parallel Tracking and Mapping for Small AR Workspaces," in *Proc. IEEE and ACM International Symposium on Mixed and Augumented Reality*, Nara, Japan, Nov 2007, pp. 225–234.

[4] A. Geiger, J. Ziegler, and C. Stiller, "StereoScan: Dense 3D Reconstruction in Real-time," in *Intelligent Vehicles Symposium (IV)*, 2011.

[5] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardos, "ORB-SLAM: A Versatile and Accurate Monocular SLAM System," *IEEE Transactions on Robotics*, vol. 31, no. 5, pp. 1147 – 1163, 2015.

[6] C. Kerl, J. Sturm, and D. Cremers, "Dense visual slam for RGB-D cameras," in *Proc. of the Int. Conf. on Intelligent Robot Systems (IROS)*, Tokyo, Japan, Sep 2013, pp. 2100–2106.

[7] C. Forster, M. Pizzoli, and D. Scaramuzza, "SVO: Fast semi-direct monocular visual odometry," in *IEEE Intl. Conf. on Robotics and Automation, ICRA*, Hong Kong, China, May 2014.

[8] J. Engel, J. Stückler, and D. Cremers, "Large-scale direct SLAM with stereo cameras," in *Proc. IEEE Int. Conf. Intelligent Robots and Systems*, Hamburg, Germany, Sep 2015, pp. 1935–1942.

[9] J. Engel, V. Koltun, and D. Cremers, "Direct sparse odometry," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 3, pp. 611–625, March 2018.

[10] S. Sirtkaya, B. Seymen, and A. A. Alatan, "Loosely coupled Kalman filtering for fusion of Visual Odometry and inertial navigation," in *Proceedings of the 16th International Conference on Information Fusion*, July 2013, pp. 219–226.

[11] S. Leutenegger, S. Lynen, M. Bosse, R. Siegwart, and P. Furgale, "Keyframe-based visual-inertial odometry using nonlinear optimization," *The International Journal of Robotics Research*, vol. 34, no. 3, pp. 314–334, 2015.

[12] C. Forster, L. Carlone, F. Dellaert, and D. Scaramuzza, "On-manifold preintegration for real-time visual–inertial odometry," *IEEE Transactions on Robotics*, vol. PP, no. 99, 2016.

[13] T. Laidlow, M. Bloesch, W. Li, and S. Leutenegger, "Dense rgb-d-inertial slam with map deformations," in *Proceeding of IEEE International Conference on Intelligent Robots and Systems*, 2017.

[14] C. Liu, C. G. Atkeson, S. Feng, and X. Xinjilefu, "Full-body motion planning and control for the car egress task of the DARPA robotics challenge," in *2015 IEEE-RAS 15th International Conference on Humanoid Robots (Humanoids)*, Nov 2015, pp. 527–532.

[15] A. Z. Zhu, N. Atanasov, and K. Daniilidis, "Event-based Visual Inertial Odometry," in *Proceeding of IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[16] B. W. Babu, S. Kim, Z. Yan, and L. Ren, "σ-DVO: Sensor Noise Model Meets Dense Visual Odometry," in *Proceeding of IEEE International Symposium on Mixed and Augmented Reality*, 2016, pp. 18–26.

[17] R. C. Smith and P. Cheeseman, "On the representation and estimation of spatial uncertainty," *The International Journal of Robotics Research*, vol. 5, no. 4, pp. 56 – 68, 1986.

[18] M. Montemerlo, S. Thrun, D. Koller, and B. Wegbreit, "Fastslam: A factored solution to the simultaneous localization and mapping problem," in *Eighteenth National Conference on Artificial Intelligence*, 2002, pp. 593–598.

[19] C. Cadena, L. Carlone, H. Carrillo, Y. Latif, D. Scaramuzza, J. Neira, I. D. Reid, and J. J. Leonard, "Simultaneous localization and mapping: Present, future, and the robust-perception age," *IEEE Transactions on Robotics*, vol. 32, no. 6, pp. 1309–1332, Dec 2016.

[20] J. Neira and J. D. Tardos, "Data association in stochastic mapping using the joint compatibility test," *IEEE Transactions on Robotics and Automation*, vol. 17, no. 6, pp. 890–897, Dec 2001.

Fig. 6. The image frames captured when deliberate motion conflict was introduced by observing a hand carried laptop computer not attached to the world frame while walking.

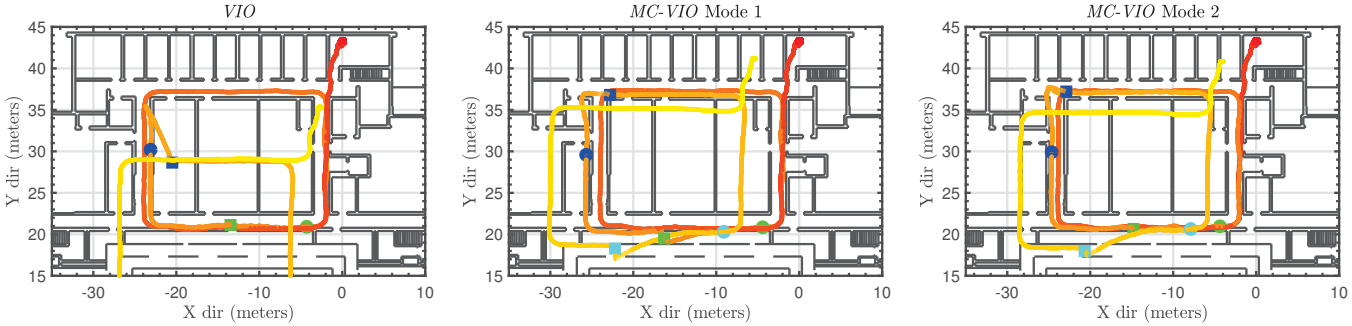

Fig. 7. Circle - start of motion conflict. Square - end of motion conflict. To denote passage of time, loops are colored progressively from red (start) to yellow (end). With MC-VIO Mode 2, we have a trajectory which is consistent with the reference first loop trajectory.

Fig. 8. The image frames captured when deliberate motion conflict was introduced by looking inside the vehicle while it was moving.

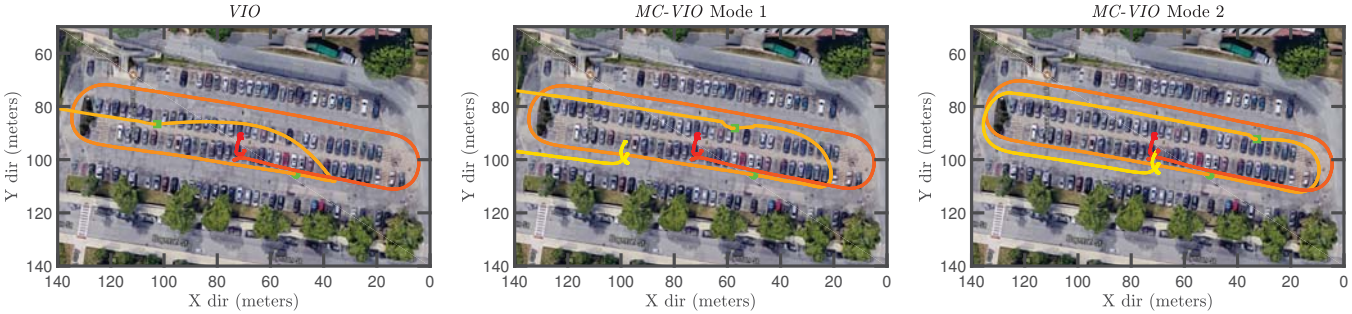

Fig. 9. Circle - start of motion conflict. Square - end of motion conflict. To denote passage of time, loops are colored progressively from red (start) to yellow (end). With MC-VIO Mode 2, we have a trajectory which is consistent with the reference first loop trajectory.

[21] J. Civera, O. G. Grasa, A. J. Davison, and J. M. M. Montiel, "1-Point RANSAC for extended kalman filtering: Application to real-time structure from motion and visual odometry," *Journal of Field Robotics*, vol. 27, no. 5, p. 609âĂŞ631, 2010.

[22] C.-C. Wang and C. Thorpe, "Simultaneous localization and mapping with detection and tracking of moving objects," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, Washington, DC, May 2002.

[23] C. Bibby and I. Reid, "Simultaneous Localisation and Mapping in Dynamic Environments (SLAMIDE) with Reversible Data Association," in *Proceedings of Robotics: Science and Systems*, Atlanta, GA, USA, June 2007.

[24] N. D. Reddy, I. Abbasnejad, S. Reddy, A. K. Mondal, and V. Devalla, "Incremental real-time multibody VSLAM with trajectory optimization using stereo camera," in *Proc. IEEE Int. Conf. Intelligent Robots and Systems*, Daejeon, Korea, Oct 2016.

[25] A. I. Mourikis and S. I. Roumeliotis, "A multi-state constraint kalman filter for vision-aided inertial navigation," in *Proc. IEEE Int. Conf. International Conference on Robotics and Automation*, Hamburg, Germany, April 2007, pp. 3565–3572.

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

[27] H. Strasdat, A. J. Davison, J. M. M. Montiel, and K. Konolige, "Double window optimisation for constant time visual slam," in *2011 International Conference on Computer Vision*, Nov 2011, pp. 2352–2359.

[28] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers, "A benchmark for the evaluation of RGB-D SLAM systems," in *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2012, pp. 573–580.

[29] M. Burri, J. Nikolic, P. Gohl, T. Schneider, J. Rehder, S. Omari, M. W. Achtelik, and R. Siegwart, "The euroc micro aerial vehicle datasets," *The International Journal of Robotics Research*, vol. 35, no. 10, pp. 1157–1163, 2016.

[30] B. P. W. Babu, D. Cyganski, and J. Duckworth, "Gyroscope assisted scalable visual simultaneous localization and mapping," in *2014 Ubiquitous Positioning Indoor Navigation and Location Based Service (UPINLBS)*, Nov 2014, pp. 220–227.