2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
Macau, China, November 4-8, 2019

Robust High Accuracy Visual-Inertial-Laser SLAM System

Zengyuan Wang[1], Jianhua Zhang[1], Shengyong Chen[2], Conger Yuan[1], Jingqian Zhang[1] and Jianwei Zhang[3]

Abstract— **In recent years, many excellent works on visual-inertial SLAM and laser-based SLAM have been proposed. Although inertial measurement unit (IMU) significantly improve the motion estimate performance by reducing the impact of illumination variation or texture-less region on visual tracking, tracking failures occur when in such an environment for a long time. Similarly, when in structure-less environments, laser module will fail since lack of sufficient geometric features. Besides, motion estimation by moving lidar has the problem of distortion since range measurements are received continuously. To solve these problems, we propose a robust and high-accuracy visual-inertial-laser SLAM system. The system starts with a visual-inertial tightly-coupled method for motion estimation, followed by scan matching to further optimize the estimation and register point cloud on the map. Furthermore, we enable modules to be adjusted automatically and flexibly. That is, when one of these modules fails, the remaining modules will undertake the motion-tracking task. For further improving the accuracy, loop closure and proximity detection are implemented to eliminate drift accumulation. When loop or proximity is detected, we perform six degree-of-freedom (6-DOF) pose graph optimization to achieve the global consistency. The performance of our system is verified on public dataset, and the experimental results show that the proposed method achieves superior accuracy against other state-of-the-art algorithms.**

I. INTRODUCTION

In recent decades, many efforts have been devoted to improve the accuracy and robustness of Simultaneous Localization and Mapping (SLAM) in different environments. Since SLAM uses sensors mounted on the robot to sense the environment and its own movement, many researchers try to improve the performance of motion estimation by integrating multi-modality sensors. With the development of technology, common sensors used in SLAM are getting smaller and cheaper. This makes it possible to carry multiple sensors on a single device for better localization and mapping performance at an acceptable price and volume.

We present a robust SLAM system fusing measurements from a camera, an IMU and a 3D lidar. It is a complete SLAM system that includes visual-inertial-laser odometry, loop closure, proximity detection, and global pose optimization. The proposed odometry method starts with a visual-inertial tightly-coupled odometry to output a motion approximation, then a scan matching method is used to optimize the approximation. Experiments on public dataset show that the presented odometry method achieves better performance than LOAM [1], which is the state-of-the-art laser SLAM. And appearance-based loop closure is used to eliminate the drift accumulation. However, appearance-based loop detection is limited by camera's field of view. Therefore, we use laser measurements for proximity detection, which can determine whether the robot visits the same position regardless of the robot's horizontal orientation. When loop or proximity is detected, we carry out the pose graph optimization to achieve the global consistency of the system.

Furthermore, the system is able to handle sensor degradation. For instance, when robots stay in an illumination variance or texture-less area for a long time, visual-inertial odometry (VIO) will fail due to the loss of visual tracking. When VIO failure is detected, system will automatically adjust workable modules, using laser-only method to estimate motion. Likewise, in structure-less environment, no sufficient geometric features can be extracted from point cloud, and the scan matching can not be performed, therefore, only VIO is used for motion tracking. The VIO used in our system relies on the component of VINS-Mono [2]. Although our approach partly relies on the components of LOAM and VINS-Mono, there are several innovations that make our system significantly different from them. Compared with VINS-Mono, we further optimize the visual-inertial odometry using scan matching. And compared with LOAM, we provide a motion priori for scan matching, and maintain a keyframe database for storing poses, images, and point clouds to achieve global optimization. In addition, the ability to handle sensor degradation is one of our advantages. The main contributions of this paper are as follows:

- High-accuracy visual-inertial-laser odometry. Starting with a visual-inertial tightly-coupled odometry, followed by a laser scan matching to refine the estimation.
- The system can be flexibly and automatically adjusted to deal with sensor degradation.
- Combining the visual-based loop closure and laser-based proximity detection to add global constraints.
- The 6-DOF global pose optimization to maintain global consistency

The rest of this paper is organized as follows: In Section II, we review the related work. Assumptions and notations are described in Section III. In Section IV, we introduce visual-inertial-laser odometry in detail. In Section V, we explain the method to improve robustness in severe environments. Loop closure, proximity detection and 6-DOF global pose graph optimization are presented in Section VI. Experimental

*This work was supported by National Key R&D Program of China (2018YFB1305200), and National Natural Science Foundation of China (U1509207, 61876167, 61802348)

[1] The authors are with Computer Science and Technology, Zhejiang University of Technology. (E-mail: zengyuanwang@outlook.com; zjh@zjut.edu.cn). Jianhua Zhang is the corresponding author.

[2] Shengyong Chen is with School of Computer Science and Engineering, Tianjin University of Technology. E-mail: sy@ieee.org

[3] Jianwei Zhang is with TAMS group, department of informatics, University of Hamburg.

evaluations of the proposed method are presented in Section VII and conclusion is made in Section VIII.

II. RELATED WORK

In this section, we give a summary of multi-modality fusion SLAM. Visual-inertial state estimation has been an active research field over the years since IMU is a useful complement to visual input. According to the way how visual measurements and inertial measurements combine with each other, the visual-inertial odometry can be divided into loosely-coupled method and tightly-coupled method [2]–[6]. Loosely-coupled method [7] [8] treats IMU as an independent module to assist vision-only motion estimation obtained from the visual mesurements. Tightly-coupled method jointly optimizes visual and inertial measurements from the raw measurements level. VINS-Mono [2] is a tightly-coupled visual-inertial state estimator, using a nonlinear optimization-based approach to acquire accurate visual-inertial odometry by fusing pre-integrated inertial measurements and visual observations. Since the odometry can only use a few adjacent keyframes for motion estimation, errors from one moment will inevitably accumulate to the next. Therefore, some visual-inertial and visual-only SLAM systems [2] [6] integrate visual-based loop closure to maintain global consistency. The visual-based loop closure is usually implemented through the bag of words [9] approach.

Lidar has been attached great importance in autonomous driving. However, moving lidar has the problem of distortion which can be removed by a laser scanner itself or other sensors. Zhang and Singh [1] presented a laser odometry method that model the motion as constant velocity during a scan. There are some algorithms combine lidar and IMU [10] [11], where IMU can provide a motion priori to help account for high frequency motion estimation. And Zhang and Singh [12] proposed an algorithm that tightly couples camera and lidar, model the visual odometry drift as linear motion. Scherer et al. [13] presented a method that loosely couples a stereo camera pair with an IMU to estimate motion, and uses the estimated motion to register point cloud into map. Some methods use visual odometry output as motion approximation, and further match laser scans to optimize the motion [14] [15]. Inspired by the same concept, Zhang and Singh [16] proposed a system that fusing range, visual, and inertial measurements. All of these sensors are loosely-coupled, and the visual-inertial module is filter-based. Our method is inspired by this, however we use a tightly-coupled optimization-based visual-inertial odometry method that simultaneously estimates pose, IMU bias and features, which makes motion approximation more accurate. In addition, visual-based loop closure, laser-based proximity detection, and global pose optimization are implemented in our system to maintain global consistency.

III. ASSUMPTIONS AND NOTATIONS

We assume that the intrinsic parameters of the camera are known. The extrinsic parameters between the three sensors are calibrated, and they are time synchronized.

There are four coordinate systems in our system. Lidar coordinate system $\{L\}$, camera coordinate system $\{C\}$, IMU coordinate system $\{I\}$, and world coordinate system $\{W\}$. We define the lidar coordinate system after initialization as world coordinate system. Sensor coordinate system changes as the sensor moves, so we denote S_i as the coordinate system of sensor S at time i. And T_b^a represents the transformation between coordinate system a and b. Denote X_j^S as the homogeneous coordinate of point j in coordinate system $\{S\}$, $X_i^S = [x \quad y \quad z \quad 1]^T$.

IV. ODOMETRY

In this section, we explain in detail how to perform online motion estimation that combines visual, inertial and laser measurements. Our approach starts with a tightly-coupled visual-inertial method to estimate the motion, followed by a scan matching to further optimize the estimated motion and register point cloud on the map. The diagram illustrating the pipeline of the proposed method is shown in Fig. 1.

A. Visual-Inertial Odometry

The visual-inertial odometry in our system relies on the component of VINS-Mono [2]. VINS-Mono proposed a sliding window estimator based on nonlinear optimization, which tightly fuses pre-integrated IMU measurements with visual observations. It minimizes the sum of the Mahalanobis norm of IMU measurement residual and visual measurement residual to obtain a maximum posteriori estimation. By solving the nonlinear problem, the poses and bias of IMU can be obtained. For the sake of brevity, the detailed procedure of visual-inertial odometry can be found in [2].

B. Scan Matching

The structure of scan matching module is shown in Fig. 2. Laser points are received continuously and registered in the coordinate system L_{t_i}, where t_i is the time that point i is received. Define t_k as the start time of laser scan k, $k \in Z^+$, and t_{k+1} the end time of scan k. Let $\tilde{P}_k$ be the set of points received during scan k. Since sensors are synchronous, we can acquire the motion approximation $\tilde{T}^{I_k}_{I_{k+1}}$ of IMU from t_k to t_{k+1} through VIO module, and using the extrinsic parameters to calculate the motion approximation $\tilde{T}^{L_k}_{L_{k+1}}$ of lidar, the time t_i is abbreviated as i when used as a superscript or subscript for readability :

$$\tilde{T}^{L_k}_{L_{k+1}} = T^L_I \tilde{T}^{I_k}_{I_{k+1}} (T^L_I)^{-1} \tag{1}$$

where T^L_I is the extrinsic parameters between the lidar and the IMU.

Based on the assumption that the device moves at a constant speed during a scan, we can use linear interpolation to acquire the motion from the start time of scan to the time when point i is received:

$$\tilde{T}^{L_k}_{L_i} = (\frac{t_i - t_k}{t_{k+1} - t_k})\tilde{T}^{L_k}_{L_{k+1}} \tag{2}$$

Then we can register these points to a common coordinate L_{k+1} by projecting the points to the end of the scan:

$$X_i^{L_{k+1}} = (\tilde{T}^{L_k}_{L_{k+1}})^{-1}(\tilde{T}^{L_k}_{L_i})X_i^{L_i} \tag{3}$$

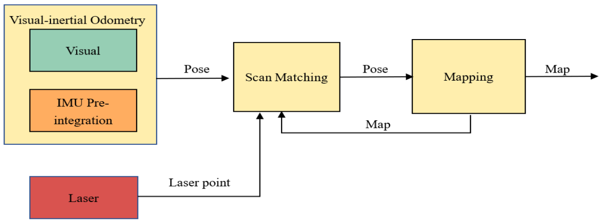


Fig. 1. The pipeline of the proposed localization and mapping system.

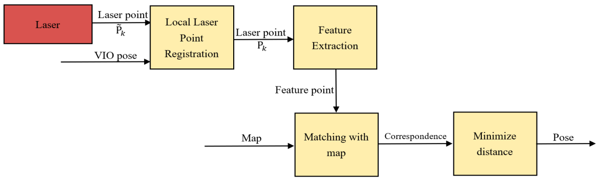


Fig. 2. A block diagram illustrating of scan matching module.

where $X_i^{L_i}$ is the coordinate of point i in L_i, and $X_i^{L_{k+1}}$ is the coordinate of point i in L_{k+1}. All points in the set $\tilde{P}_k$ are projected to L_{k+1} to obtain points set P_k.

Then extracting edge points and planar points from P_k according the smoothness of local surfaces around the points. Define extracted features set as F_k. The predicted transformation of lidar at time t_{k+1} in world coordinate is :

$$\tilde{T}_{L_{k+1}}^{W} = T_{L_k}^{W} \tilde{T}_{L_{k+1}}^{L_k} \tag{4}$$

where $T_{L_k}^{W}$ is the pose of lidar at time t_k that has been optimized by scan matching. Using the estimated transformation $\tilde{T}_{L_{k+1}}^{W}$ to project features F_k to the world coordinate system. Matching projected features with map, we can find an edge line as the correspondence for an edge point, and a planar patch as the correspondence for a planar point. The distance of edge point to edge line and the distance of planar point to planar patch can be calculated, the function of distance of point i to its correspondence can be written as:

$$d_i = f(X_i^{L_{k+1}}, \tilde{T}_{L_{k+1}}^{W}), \quad i \in F_k \tag{5}$$

The detailed process of finding correspondence of the feature and calculating their distance can be found in [1]. And the sum of distances from all points in F_k to their matching edge lines or planar patches can be described as the following nonlinear function:

$$\boldsymbol{d} = \boldsymbol{f}(\tilde{T}_{L_{k+1}}^{W}) \tag{6}$$

Then scan matching can be formulated into a nonlinear optimization problem, and solved by minimizing the $\boldsymbol{d}$ towards zero. We solve this problem through Newton gradient-descent method. If it converges, we can get a optimized motion estimation $T_{L_{k+1}}^{W}$ of lidar. Finally, we register the point cloud received during scan k to world coordinate system using the optimized motion estimation $T_{L_{k+1}}^{W}$.

V. ROBUSTNESS ENHANCEMENT

Although the IMU improves the robustness of visual motion tracking, tracking loss of visual-inertial odometry occurs when the device walks in the illumination changing or texture-less environment for a long time. Similarly, in structure-less environment, such as long corridors, scan matching can not work properly. To tackle these problems, we enable the system modules to be automatically and flexibly reorganized. When the visual-inertial odometry module or scan matching module fails due to sensor degradation, the remaining working modules will be adjusted automatically to take on the major motion estimation task.

In the case of the number of features tracked in the current frame is below a threshold, the estimated position and orientation in the sliding window saltates compared with the previous estimation, and large change in IMU bias estimation, the system determines that the visual-inertial odometry module is invalid. In this case, the visual-inertial odometry module will no longer output the motion approximation, instead it will send a signal that enables the scan to scan matching module, as shown in Fig. 3. This module is based on the component of LOAM, in a nutshell, firstly extract geometric features from $\tilde{P}_k$, and find edge line or planar patch in P_{k-1} as their correspondences, then compute the distance from a feature point to its correspondence, finally recover the lidar's motion between two consecutive scan by minimizing the overall distances of the feature points to their correspondences. Complete and detailed scan to scan matching method description can be found in [1].

When in structure-less environment, scan matching is unable to refine motion estimates since no sufficient geometric features can be extracted. In this case, the visual-inertial odometry output skips the scan matching module and is used for registering points on the map directly as shown in Fig.4. When the sensor degradation problem is solved, all modules resume normal operation.

VI. LOOP CLOSURE AND PROXIMITY DETECTION

Loop closure and proximity detection are used to eliminate drift accumulation. Loop detection is appearance-based and proximity detection is laser-based. When loop or proximity is detected, we conduct 6-DOF pose graph optimization to maintain global consistency. A keyframe database is maintained for global optimization, and keyframes in the database are numbered starting from 1 based on the creation time. A keyframe contains the pose of the device , the image features, and geometric features extracted from laser points.

A. Loop closure and Proximity Detection

DBoW2 [9] is used for loop detection. When the appearance similarity of two frames reaches the threshold, we consider a loop is detected. The transformation between current keyframe and its loop closure keyframe is solved by perspective-n-point (PnP) method [17]. However, loop detection is limited by the camera's field of view. To deal with such situations, we implement laser-based proximity

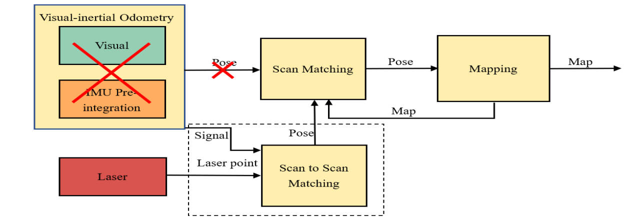


Fig. 3. A block diagram of the odometry pipeline when failure occurs on visual-inertial odometry module. The part in the dotted box replaces the work of VIO and output motion approximation.

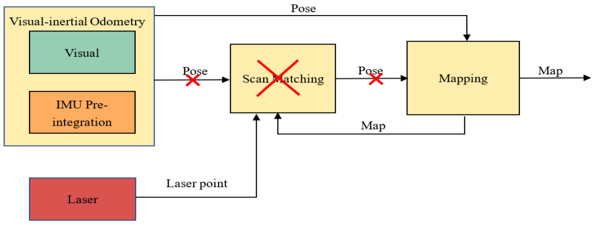


Fig. 4. A block diagram of the odometry pipeline when failure occurs on scan matching module.

detection. Since 3D lidar has 360° horizontal field of view, proximity detection is able to find whether the device is revisiting a place without considering the horizontal orientation. To avoid redundant constraints, the keyframes that have been detected loop will not be used for proximity detection.

Fig. 5 illustrates how proximity detection works. Firstly, iterate through the keyframe database, and calculate the relative distance between the traversed keyframe and current keyframe using the poses in the world coordinate system. If the relative distance is less than R_1, using relative transformation calculated before as prediction, and using scan to scan matching to update the relative transformation. If the updated distance is smaller than R_2, $R_2 < R_1$, a proximity is considered being detected. If there are more than one nodes that meet the requirements, we select the earlier one as the proximity node, in order to eliminate the drift as much as possible. Furthermore, assuming that the current keyframe is k, the keyframes $k-e, ..., k-1, k$ are not used for the current proximity detection, e is a empirical value.

B. Global Pose Graph Optimization

When the loop or proximity is detected, we add a loop closure edge or a proximity edge to pose graph, and then a 6-DOF pose graph optimization is performed. The keyframe corresponds to a vertex in the pose graph, and there are three types of edges that connect the vertices.

1) Sequential Edge: a vertex will establish several sequential edges to its previous vertices. The value of the sequential edge corresponds to the relative pose transformation between two vertices, which is acquired directly from visual-inertial-laser odometry. The value of the sequential edge that connects vertex i and j is:

$$\begin{aligned} \tilde{t}_j^i &= (\tilde{q}_i^w)^{-1}(\tilde{t}_j^w - \tilde{t}_i^w) \\ \tilde{q}_j^i &= (\tilde{q}_i^w)^{-1}(\tilde{q}_j^w) \end{aligned} \tag{7}$$

where $\tilde{t}_i^w$ and $\tilde{q}_i^w$ are the orientation and position of keyframe i in the world coordinate system taken directly from odometry respectively.

2) Loop Closure Edge: if a vertex has a loop closure connection, it will be connected with loop closure vertex by a loop closure edge. Likewise, the loop closure edge contains 6-DOF relative pose transformation $\tilde{t}_j^i$, $\tilde{q}_j^i$. The value of the loop closure edge is solved by PnP.

3) Proximity Edge: if the vertex has a proximity connection, a proximity edge will be added to link it and its proximity vertex. The value of proximity edge is obtained from scan to scan matching.

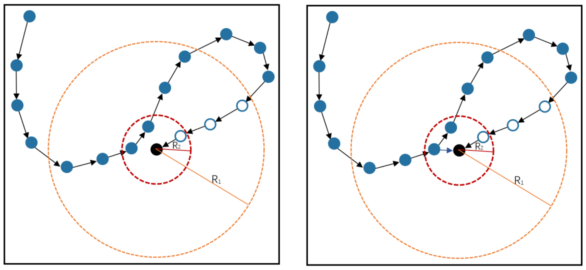


Fig. 5. A diagram illustrating the proximity detection procedure. The black node is current node, the nodes in orange dashed circle are used for the proximity detection. The e blue hollow nodes are not used for detection. The radius of red dashed circle limit the length the link to be created. In left image, there are two nodes fit the length limitation, only the early one is selected as proximity. In the right image, a new proximity link is added.

The residual of the edge between vertex i and vertex j is defined as:

$$r_{i,j}(q_i^w, t_i^w, q_j^w, t_j^w) = \begin{bmatrix} (q_i^w)^{-1}(t_j^w - t_i^w) - \tilde{t}_j^i \\ (q_i^w)^{-1}(q_j^w)(\tilde{q}_j^i)^{-1} \end{bmatrix} \tag{8}$$

The whole pose graph with three types edges are optimized by minimizing the following cost function:

$$\begin{aligned} \arg\min_{q,t}\{ \sum_{(i,j)\in A} \|r_{i,j}\|^2 + \sum_{(i,j)\in B} \rho_1(\|r_{i,j}\|^2) \\ + \sum_{(i,j)\in C} \rho_2(\|r_{i,j}\|^2)\} \end{aligned} \tag{9}$$

In the expression, A is the set of all sequential edges, B is the set of all loop closure edges, and C is the set of all proximity edges. $\rho_1(\cdot)$ and $\rho_2(\cdot)$ are Huber norm to reduce the influence of any possible incorrect loop or proximity. The nonlinear optimization problem is solved by Ceres Solver [18]. With the increase of moving distance, the size of the pose graph becomes unlimited. Since the computation cost of graph optimization is positively correlated with the size of the graph, a downsample process is implemented to keep the keyframe database at a limited size. The keyframe with proximity or loop closure constrain will be retained, while other keyframes which are too close to their neighbors will be removed.

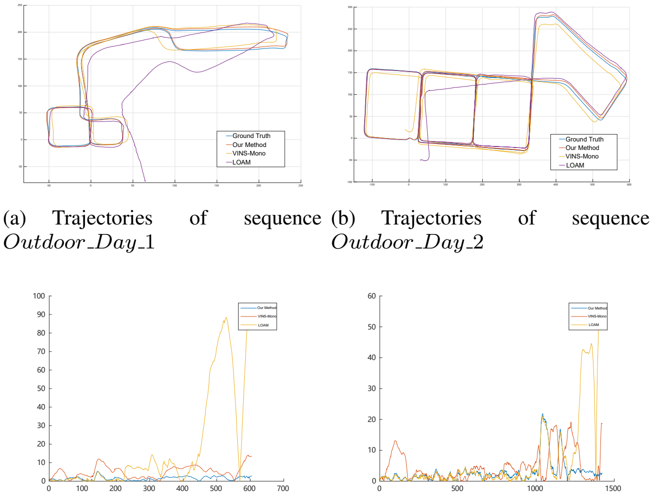
VII. EXPERIMENT

We evaluate our method using the MVSEC dataset [19]. The sensor suite contains a Velodyne Puck LITE and a VI-sensor [20]. These sensors are rigidly mounted in different frames, and the extrinsics between all sensors are calibrated. The outdoor driving sequences are collected by car at speeds up to 12 m/s, and only in these sequences, both VI-sensor and Velodyne lidar are enabled. So experiments are performed on the outdoor driving sequences. The experimental video is provided as an attachment to this paper.

A. Odometry Accuracy

In this section, we compare our odometry method with the state-of-the-art laser odometry method LOAM [1] and visual-inertial state estimator VINS-Mono [2]. Neither our method nor VINS-Mono includes loop closure or proximity detection in this experiment. The trajectories of sequence $Outdoor\_Day\_1$ and sequence $Outdoor\_Day\_2$ are shown in Fig.6(a) and Fig.6(b) respectively. Absolute Trajectory Error (ATE), which gives the root mean square measure of error between the ground truth and the pose estimation, is used for evaluating the performance of motion estimation. ATE of of sequence $Outdoor\_Day\_1$ and sequence $Outdoor\_Day\_2$ are shown in Fig.6(c) and Fig.6(d) respectively.

Combine the results on these two sequences, we can find that VINS-Mono has a larger drift than our odometry method in the whole process. This is because the scan matching refines the motion estimation effectively. The performance of LOAM is similar to our method in early stages of these two sequences, but at some special locations, especially when the device rotates violently, LOAM suffers from severe drift. From which we can infer that visual-inertial motion approximation provide a favorable constraint for scan matching. The ATE of three methods on all five sequences are shown in Table I. For each sequence, the column of the method with the smallest error is shown in bold. It is obvious that our method is much better than VINS-Mono in all sequences, and better than LOAM in most sequences.



(a) Trajectories of sequence $Outdoor\_Day\_1$ (b) Trajectories of sequence $Outdoor\_Day\_2$

(c) ATE versus time of sequences $Outdoor\_Day\_1$ (d) ATE versus time of sequences $Outdoor\_Day\_2$

Fig. 6. Odometry accuracy

TABLE I
RMSE OF ABSOLUTE TRAJECTORY ERROR

Sequence	VINS-Mono	LOAM	Our Method
Outdoor_Day_1	6.783734	53.259610	**2.406597**
Outdoor_Day_2	9.359266	20.418958	**2.326079**
Outdoor_Night_1	16.548410	**1.505892**	1.939628
Outdoor_Night_2	8.459014	3.540939	**2.672621**
Outdoor_Night_3	6.028355	1.619103	**1.585172**

B. Robustness Verification

We filter the point clouds for a period of time in the sequence $Outdoor\_Day\_1$ to simulate a structure-less scene. As shown in Fig.7, the structure-less situation starts from the red dot position and ends at the black dot position. In our approach, due to the lack of geometric features, few points have been registered to the map, however visual-inertial odometry is used to maintain accurate motion estimation, as shown in Fig. 7(b). On the contrary, severe drift occurs on LOAM as shown in Fig. 7(a), and when the environment no longer lacks structural information, point clouds are incorrectly registered to the map (the green points in Fig. 7(a)). The ground truth of this sequence is shown in Fig.6(a).

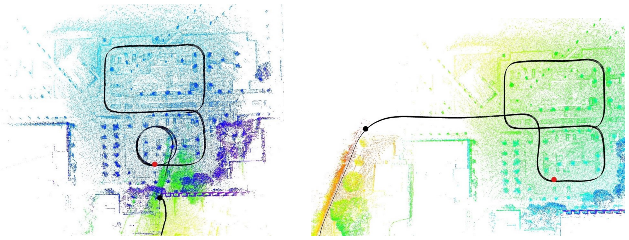
Similarly, in order to simulate the situation of camera degradation, we blurred several consecutive images in the sequence $Outdoor\_Day\_1$. As shown in Fig.8, the VIO module fails at the red dot, and VINS-Mono attempts to reinitialize. At the black dot position, VINS-Mono reinitialize successfully. However, due to the blank state in the tracking loss period, there is no way to splice the two trajectories into a complete trajectory. Our system is able to handle such situation, scan to scan matching is used to estimate the motion until the recovery of visual-inertial tracking.

C. Loop Closure and Proximity Detection

We evaluate loop closure and proximity detection on sequence $Outdoor\_Day\_2$. As shown in Fig. 9, we use VILO to denote the proposed odometry method, neither loop closure nor proximity detection is performed. VIL-Proximity represents the method with proximity detection only, and VIL-Pro-Loop represents the method with both loop closure and proximity detection. Although the VILO reaches high accurate performance, loop closure and proximity optimization can further improve the accuracy. The ATE of VILO, VIL-Proximity and VIL-Pro-Loop on sequence $Outdoor\_Day\_2$ are shown in Table II, the column of the method with the minimum error is shown in bold. Proximity detection reduces the ATE by 5.62%, and the combination of loop closure and proximity detection reduces the ATE by 8.44%.

VIII. CONCLUSION

A visual-inertial-laser fusion SLAM system is proposed in this paper. The system features an odometry method with



(a) LOAM

(b) Our method

Fig. 7. The performance of LOAM and our method when suffers from lidar degradation.

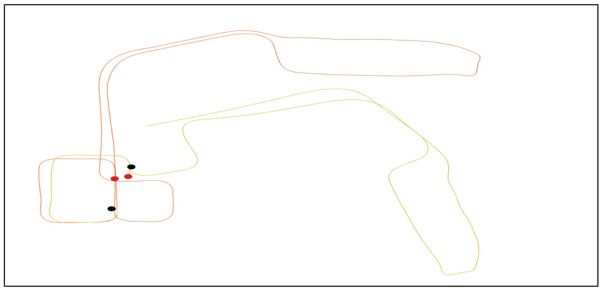


Fig. 8. Trajectories of VINS-Mono and our method when VIO fails. The trajectory estimated by our method is indicated by the orange curve, and the trajectory estimated by LOAM is indicated by the yellow curve.

high robustness and low drift, flexible system structure and global optimization. The odometry method using visual-inertial tightly-coupled method to output motion approximation, and further refines it by scan matching, which significantly improve the accuracy of motion estimation. The system structure, which can be reconfigured automatically and flexibly, makes it robust enough to handle sensor degradation. Loop closure and proximity constrains are used to remove drift accumulation. Experiments on public dataset prove that our method has a dramatic performance both in accuracy and robustness. Experimental results also show that our method outperforms the state-of-the-art visual-inertial SLAM and laser-based SLAM.

References

[1] J. Zhang and S. Singh, "Loam: Lidar odometry and mapping in real-time." in *Robotics: Science and Systems*, vol. 2, p. 9, 2014.

[2] T. Qin, P. Li, and S. Shen, "Vins-mono: A robust and versatile monocular visual-inertial state estimator," *IEEE Transactions on Robotics*, no. 99, pp. 1–17, 2018.

[3] A. I. Mourikis and S. I. Roumeliotis, "A multi-state constraint kalman filter for vision-aided inertial navigation," in *Robotics and automation, 2007 IEEE international conference on*, pp. 3565–3572. IEEE, 2007.

[4] M. Bloesch, S. Omari, M. Hutter, and R. Siegwart, "Robust visual inertial odometry using a direct ekf-based approach," in *Intelligent Robots and Systems (IROS), 2015 IEEE/RSJ International Conference on*, pp. 298–304. IEEE, 2015.

[5] 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.

[6] R. Mur-Artal and J. D. Tardós, "Visual-inertial monocular slam with map reuse," *IEEE Robotics and Automation Letters*, vol. 2, no. 2, pp. 796–803, 2017.

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

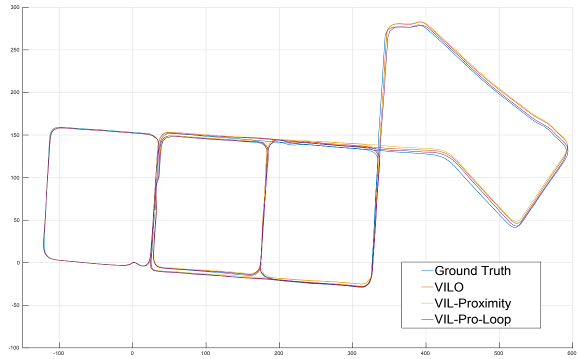


Fig. 9. The result of global pose optimization when loop closure or proximity is detected.

TABLE II
RMSE OF ABSOLUTE TRAJECTORY ERROR

Sequence	VILO	VIL-Proximity	VIL-Pro-Loop
Outdoor_Day_2	2.306079	2.176554	**2.111371**

[8] S. Lynen, M. W. Achtelik, S. Weiss, M. Chli, and R. Siegwart, "A robust and modular multi-sensor fusion approach applied to mav navigation," in *2013 IEEE/RSJ international conference on intelligent robots and systems*, pp. 3923–3929. IEEE, 2013.

[9] D. Gálvez-López and J. D. Tardos, "Bags of binary words for fast place recognition in image sequences," *IEEE Transactions on Robotics*, vol. 28, no. 5, pp. 1188–1197, 2012.

[10] M. Bosse, R. Zlot, and P. Flick, "Zebedee: Design of a spring-mounted 3-d range sensor with application to mobile mapping," *IEEE Transactions on Robotics*, vol. 28, no. 5, pp. 1104–1119, 2012.

[11] J. A. Hesch, F. M. Mirzaei, G. L. Mariottini, and S. I. Roumeliotis, "A laser-aided inertial navigation system (l-ins) for human localization in unknown indoor environments," in *2010 IEEE International Conference on Robotics and Automation*, pp. 5376–5382. IEEE, 2010.

[12] J. Zhang and S. Singh, "Visual-lidar odometry and mapping: Low-drift, robust, and fast," in *Robotics and Automation (ICRA), 2015 IEEE International Conference on*, pp. 2174–2181. IEEE, 2015.

[13] S. Scherer, J. Rehder, S. Achar, H. Cover, A. Chambers, S. Nuske, and S. Singh, "River mapping from a flying robot: state estimation, river detection, and obstacle mapping," *Autonomous Robots*, vol. 33, no. 1-2, pp. 189–214, 2012.

[14] D. Droeschel, J. Stückler, and S. Behnke, "Local multi-resolution representation for 6d motion estimation and mapping with a continuously rotating 3d laser scanner," in *2014 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 5221–5226. IEEE, 2014.

[15] D. Holz and S. Behnke, "Mapping with micro aerial vehicles by registration of sparse 3d laser scans," in *Intelligent Autonomous Systems 13*, pp. 1583–1599. Springer, 2016.

[16] J. Zhang and S. Singh, "Laser–visual–inertial odometry and mapping with high robustness and low drift," *Journal of Field Robotics*, vol. 35, no. 8, pp. 1242–1264, 2018.

[17] V. Lepetit, F. Moreno-Noguer, and P. Fua, "Epnp: An accurate o (n) solution to the pnp problem," *International journal of computer vision*, vol. 81, no. 2, p. 155, 2009.

[18] S. Agarwal, K. Mierle *et al.*, "Ceres solver," 2012.

[19] A. Z. Zhu, D. Thakur, T. Ozaslan, B. Pfrommer, V. Kumar, and K. Daniilidis, "The multi vehicle stereo event camera dataset: An event camera dataset for 3d perception," *IEEE Robotics and Automation Letters*, vol. 3, no. 3, pp. 2032–2039, 2018.

[20] J. Nikolic, J. Rehder, M. Burri, P. Gohl, S. Leutenegger, P. T. Furgale, and R. Siegwart, "A synchronized visual-inertial sensor system with fpga pre-processing for accurate real-time slam," in *Robotics and Automation (ICRA), 2014 IEEE International Conference on*, pp. 431–437. IEEE, 2014.