Camera Ego-Positioning Using Sensor Fusion and Complementary Method

Peng-Yuan Kao[1⋆], Kuan-Wei Tseng[1], Tian-Yi Shen[1], Yan-Bin Song[1], Kuan-Wen Chen[2], Shih-Wei Hu[1], Sheng-Wen Shih[3], and Yi-Ping Hung[1]

[1] National Taiwan University, Taipei, Taiwan
[2] National Chiao Tung University, Hsinchu, Taiwan
[3] National ChiNan University, Nantou, Taiwan
zbabqr@gmail.com

Abstract. Visual simultaneous localization and mapping (SLAM) is a common solution for camera ego-positioning. However, SLAM sometimes loses tracking, for instance due to fast camera motion or featureless or repetitive environments. To account for the limitations of visual SLAM, we use sensor fusion method to fuse the visual positioning results with inertial measurement unit (IMU) data based on filter-based, loosely-coupled sensor fusion methods, and further combines feature-based SLAM with direct SLAM via proposed complementary fusion to retain the advantages of both methods; i.e., we not only keep the accurate positioning of feature-based SLAM but also account for its difficulty with featureless scenes by direct SLAM. Experimental results show that the proposed complementary method improves the positioning accuracy of conventional vision-only SLAM and leads to more robust positioning results.

Keywords: Camera ego-positioning · Sensor fusion.

1 Introduction

Ego-positioning is indispensable in many applications to enable the drone to navigate autonomously in unseen or indoor environments. With the advance of computer vision technology, visual ego-positioning is now one of the most appropriate solutions for drone navigation. This topic has been thoroughly studied and a variety of solutions are now available: two well-known visual ego-positioning categories are visual simultaneous localization and mapping (SLAM) and visual odometry (VO).

In visual SLAM and visual odometry, methods which use monocular cameras are called monocular visual SLAM and monocular visual odometry respectively. Researchers have proposed a variety methods for monocular visual SLAM and VO [10, 18, 2, 5, 3, 15, 6, 16, 1]. These can be further classified into feature-based

⋆ corresponding author: Peng-Yuan Kao

methods and direct methods. The fundamental principle of feature-based methods is detecting feature points for each frame and matching them between consecutive frames. Direct methods in turn use the entire image information, and find a suitable camera pose by minimizing the photometric error.

However, visual SLAM and VO both have drawbacks. First, when the camera moves fast, these methods can lose tracking easily due to motion blur or excessive parallax between consecutive frames. Second, for monocular methods, their map scale differs from that in the real world. Considering these drawbacks, a well-known solution is to fuse the visual ego-positioning result from the camera with inertial data from the inertial measurement unit (IMU) via sensor fusion methods [14, 22, 12, 17, 20]. In this paper, we fuse visual ego-positioning results from the camera with inertial data from the IMU via a filter-based, loosely-coupled sensor fusion method [22].

Furthermore, feature-based methods and direct methods of visual SLAM and VO have complementary advantages and disadvantages. The positioning accuracy of feature-based methods is higher than that of direct methods, but they perform poorly in featureless scenarios, usually losing tracking in this case. In contrast, although the positioning accuracy of direct methods is lower than that of feature-based methods, they still work in such featureless scenarios. Therefore, in this paper, we propose a complementary method to combine the ego-positioning results of feature-based methods and direct methods to handle featureless scenarios. Experimental results show that the proposed complementary method improves the positioning accuracy of conventional vision-only SLAM and also leads to more robust positioning results.

The rest of this paper is organized as follows. We review related work in Section 2 and present the sensor fusion method in Section 3. We present the proposed complementary ego-positioning method in Section 4, and in Section 5 we evaluate the sensor fusion method as well as the proposed complementary ego-positioning method. We conclude in Section 6.

2 Related Work

2.1 Visual Ego-Positioning

Visual ego-positioning has being well studied, two of mostly known categories of visual ego-positioning methods are visual simultaneous localization and mapping (SLAM), and visual odometry (VO). Researchers have proposed a variety methods of visual SLAM and VO these years [10, 18, 2, 5, 3, 15, 6, 16, 1]. ORB-SLAM [15] is a representative and classic feature-based method which provides positioning results in real time. ORB-SLAM uses ORB features, which can be computed and matched extremely quickly; the method is also invariant to scale, rotation, and limited affine changes. Furthermore, ORB-SLAM has a good loop-closing algorithm with which it optimizes the global map when closed loops are detected, which can effectively reduce the cumulative error. Using the loop-closing algorithm, ORB-SLAM is also able to quickly relocalize when it loses tracking.

LSD-SLAM [2] is a classic direct SLAM method which directly operates on image intensities for both tracking and mapping, instead of using feature points to find correspondences. Semi-Direct Visual Odometry (SVO) [5] is a sparse-direct method, which detects features on the consecutive frames and uses the neighbor pixels of the detected feature to do patch matching for estimating the camera pose. Direct Sparse Odometry (DSO) [1] is also a sparse-direct method. It combines a fully direct probabilistic model (minimizing a photometric error) with consistent, joint optimization of all model parameters, including geometry represented as inverse depth in a reference frame and camera motion.

2.2 Sensor Fusion of Camera and IMU

Visual SLAM and VO have two drawbacks. First, when the camera moves rapidly, the methods lose tracking easily due to motion blur or large parallax between consecutive frames. Second, for monocular methods, the map scale is different from that in the real world. Given these drawbacks, some researchers have proposed sensor fusion methods, which estimate ego-position by fusing visual sensors and inertial sensors. In general, these methods are forms of visual-inertial odometry (VIO). VIO approaches can be divided into two categories: filter-based VIO and optimization-based VIO.

Filter-based VIO [14, 22] uses a filter to fuse visual and inertial measurements. Depending on the filter state vector, filter-based methods can be classified into tightly-coupled methods and loosely-coupled methods. Tightly-coupled methods directly consider the camera pose or information on the image as part of the filter state vector input, leading to high precision but with added computational cost [14]. Mourikis et al. [14] propose extended Kalman filter (EKF)-based real-time fusion using monocular vision and IMU; this is termed multi-state constraint Kalman filter (MSCKF). Unlike conventional EKF-based methods which put features on the image frames into the state vector, MSCKF puts camera poses into the state vector to avoid the curse of dimensionality. Experimental results show that MSCKF achieves high-precision pose estimation in real time. Loosely-coupled methods, in contrast, process the visual and inertial measurements separately to reduce computational cost. Because of this feature, loosely coupled methods are typically suited for systems with very limited resources, such as drones. Weiss et al. [22] propose a framework to enable autonomous flights of micro aerial vehicles by treating visual ego-positioning as a black box. This method is suitable for implementation on drones as it is computationally efficient and can be easily used with different visual ego-positioning algorithms.

Optimization-based VIO [12, 17, 20], in turn, estimates camera pose using an objective function to minimize the reprojection residuals and IMU residuals. Leutenegger et al. [12] propose an optimization-based VIO framework called open keyframe-based visual-inertial SLAM (OKVIS). This work applies the keyframe concept to nonlinear optimization by marginalization. Qin et al. [20] propose a nonlinear optimization-based state estimator with a loop-closing algorithm, which reduces the cumulative error and thus increases positioning accuracy. Nisar et al. [19] proposed a method called visual inertial model-based odom-

etry (VIMO), which extends VINS-Mono by adding thrust measurements and dynamic residuals to the cost function to perform force estimation. They apply the concept of motion constraint which combines quadrotors dynamics and external forces. Their experiments on simulation shows 29% improvement on accuracy without increasing the computational time. Furthermore, in optimization-based VIO, pre-integration [4] is a significant method, especially for drones, which is applied to process IMU data and can reduce computational costs.

Because filter-based, loosely-coupled sensor fusion methods are computationally efficient and can be easily used with different visual ego-positioning algorithms, in this paper, we use filter-based, loosely-coupled methods as our sensor fusion methods.

2.3 Complementary Ego-Positioning

Another way to overcome the drawbacks of visual SLAM and VO is to use complementary methods, i.e., combine feature-based and direct methods. Feature-based methods and direct methods each have their pros and cons. Feature-based methods enable fast tracking but do poorly in featureless scenarios. Direct methods estimate robust pose over time but come with heavy CPU demands. Complementary methods take advantage of both methods continually for robust results. Krombach et al. [11] propose a complementary ego-positioning method called the hybrid approach. Their approach uses direct method LSD-SLAM as a keyframe register for better depth estimation values. Moreover, LIBVISO2 [9], a feature-based method, keeps tracking pose due to its fast-tracking capabilities. Their experimental results show that the approach accumulates less drift than the direct method or the feature-based method.

3 Sensor Fusion of Camera and IMU

For fast computation and easily adoption of different visual ego-positioning algorithms, which is described in Section 2.2, we use the filtered-based, loosely-coupled VIO method. Since it requires visual measurements and IMU measurements to be processed separately, we divide the sensor fusion framework into a visual positioning module and a sensor fusion module. First, the visual positioning module estimates the initial camera pose, after which the sensor fusion module fuses the initial camera pose and IMU measurements to yield a refined camera pose and the scale of the real world. Fig. 1 shows the sensor fusion framework used in this work.

3.1 Method

We use the filtered-based, loosely-coupled VIO method proposed by Weiss et al. [22], which uses an EKF framework, and estimates not only the camera pose and velocity but also the scale of the real world. The filter consists of a prediction and an update step. Below we describe in greater detail the structure of the EKF framework.

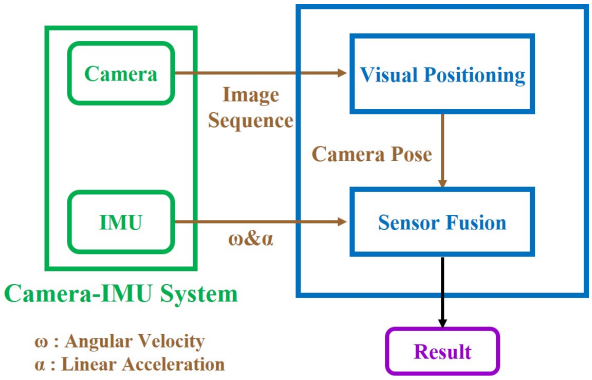

Fig. 1. Loosely-coupled sensor fusion framework

State Vector The state vector $\mathcal{X}$ of method in [22] contains eight components. The first to the third components are the IMU position $\mathbf{p}_w^i$, the velocity $\mathbf{v}_w^i$, and the rotation $\mathbf{q}_w^i$ from the inertial world coordinate system to the IMU coordinate system. The IMU rotation is expressed in quaternion form. The fourth to the sixth components are the gyroscope bias b_ω, the accelerometer bias b_α, and a measurement scale factor λ. The seventh and the eighth components are the calibration components, which are the distance $\mathbf{p}_i^c$ between IMU and camera and the rotation $\mathbf{q}_i^c$ from the IMU coordinate system to the camera coordinate system:

$$\mathcal{X} = \left[\mathbf{p}_w^i,\ \mathbf{v}_w^i,\ \mathbf{q}_w^i,\ \mathbf{b}_\omega,\ \mathbf{b}_\alpha,\ \lambda,\ \mathbf{p}_i^c,\ \mathbf{q}_i^c\right]. \tag{1}$$

Note that this method assumes the scale factor λ and the calibration components $\mathbf{p}_i^c$ and $\mathbf{q}_i^c$ remain constant over time, that is,

$$\dot{\lambda} = 0, \quad \dot{\mathbf{p}}_i^c = 0, \quad \dot{\mathbf{q}}_i^c = 0. \tag{2}$$

Prediction Step The IMU data is used in the prediction step for state propagation as the motion model in a basic Kalman filter. The measured angular velocity $\widehat{\boldsymbol{\omega}}$ and acceleration $\widehat{\mathbf{a}}$ data from the IMU are used to predict the system state by integration and double integration. The following differential equations govern the state:

$$\dot{\mathbf{p}}_w^i = \mathbf{v}_w^i, \tag{3}$$

$$\dot{\mathbf{v}}_w^i = \mathbf{R}_i^w\left(\widehat{\mathbf{a}} - \mathbf{b}_\alpha - \mathbf{n}_\alpha\right) - \mathbf{g}, \tag{4}$$

$$\dot{\mathbf{q}}_w^i = \frac{1}{2}\Omega\left(\widehat{\boldsymbol{\omega}} - \mathbf{b}_\omega - \mathbf{n}_\omega\right)\mathbf{q}_w^i, \tag{5}$$

where $\mathbf{b}_\alpha$ and $\mathbf{b}_\omega$ are the sensor bias of the accelerometer and gyroscope respectively; $\mathbf{n}_\alpha$ and $\mathbf{n}_\omega$ are the sensor noise of accelerometer and gyroscope respectively, which are modeled as white noise; $\mathbf{g}$ is the gravity vector in the world coordinate system; $\Omega()$ is the quaternion multiplication matrix of angular velocity; and $\mathbf{R}_i^w$ is the rotation matrix from the IMU coordinate system to the

world coordinate system. Note that this method models bias as a random walk, whose derivative is white noise:

$$\dot{\mathbf{b}}_\alpha = \mathbf{n}_{\mathbf{b}\alpha}, \quad \dot{\mathbf{b}}_\omega = \mathbf{n}_{\mathbf{b}\omega}. \tag{6}$$

Update Step The visual positioning result is used in the update step as the measurement in a basic Kalman Filter. For the position measurement $\mathbf{p}_w^c$ obtained from the visual positioning algorithm, we have the following measurement model:

$$\mathbf{p}_w^c = \lambda(\mathbf{R}_i^w \mathbf{p}_i^c + \mathbf{p}_w^i) + \mathbf{n}_p, \tag{7}$$

where $\mathbf{n}_p$ is the measurement noise modeled as white noise. The rotation measurement $\mathbf{q}_w^c$ obtained from the visual positioning algorithm is modeled as

$$\mathbf{q}_w^c = \mathbf{q}_i^c \otimes \mathbf{q}_w^i, \tag{8}$$

where $\otimes$ is the quaternion multiplication operator. Given the measurement model, the state estimation can be updated according to the Kalman filter procedure.

3.2 Camera-IMU System Calibration

The camera-IMU sensor fusion requires accurate calibration parameters of the camera-IMU system to maintain high performance. These parameters can be divided into camera-intrinsic parameters, camera-IMU-extrinsic parameters, and IMU noise parameters. Camera-intrinsic parameters are critical for the visual positioning module to achieve highly accurate camera-pose estimation. In this work, we use a well-known camera calibration method from Zhang [24]. Camera-IMU-extrinsic parameters are used to update the sensor fusion input state vector. To estimate these, we use the Kalibr toolbox [7, 8, 13], which is a widely used camera-IMU calibration toolbox that provides highly accurate calibration. IMU noise parameters can inform the sensor fusion system about the uncertainty of the IMU, and are important when the sensor fusion system is updating. Allan standard deviation [21] is a common method to estimate the noise parameters of the sensor.

4 Complementary Ego-Positioning

The framework of our complementary ego-positioning system is shown in Fig. 2. The framework has four main modules: a feature-based visual positioning module, a direct visual positioning module, a sensor fusion module, and a complementary fusion module. Because the feature-based method is faster and more accurate than the direct method, we use the feature-based visual positioning module as the main module and direct visual positioning module as the complementary module. First, the image sequence is fed into the main module and the complementary module to yield the camera pose. The results from the main

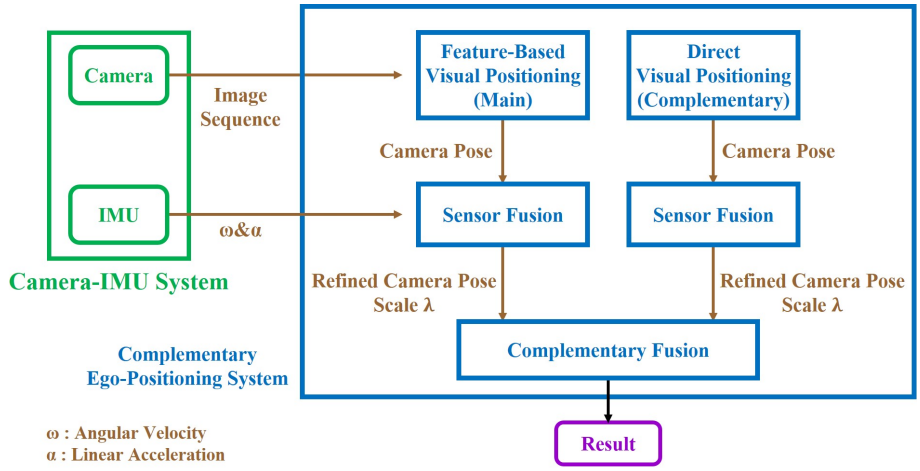

Fig. 2. Framework of proposed complementary ego-positioning system

module and complementary module are then fed into the sensor fusion module respectively, and the sensor fusion module fuses the results from these two modules with the IMU measurements to yield the main VIO result and complementary VIO result respectively. The method used in the sensor fusion module is described in detail in Section 3. Finally, as the feature-based method usually performs poorly or loses tracking in featureless scenarios, the complementary fusion module fuses the main and complementary VIO results to handle featureless scenarios, yielding a robust result as the final output.

4.1 3D Point Registration

To fuse the main and the complementary VIO results, they must be aligned. Although the estimated camera poses from the main module and the complementary module are unscaled, the scale can be acquired from the sensor fusion module because it can fuse the estimated camera poses from the main module and the complementary module with the IMU measurements to get the refined camera poses and the scale of the real world. Because the scale is estimated by the sensor fusion module, we can align the main and complementary VIO results as following. We have $\mathbf{p}_m$, which is one of the points in the main-VIO-result point set, $\mathbf{P}_m$, and $\mathbf{p}_c$, which is one of the points in the complementary-VIO-result point set, $\mathbf{P}_c$. Both $\mathbf{P}_m$ and $\mathbf{P}_c$ are finite-size point sets in a three-dimensional real vector space. The transformation from $\mathbf{p}_c$ to $\mathbf{p}_m$ can be formulated as

$$\mathbf{p}_m = \mathbf{R}\mathbf{p}_c + \mathbf{t}, \tag{9}$$

where $\mathbf{R}$ is a 3×3 rotation matrix and $\mathbf{t}$ is a 3×1 translation vector. We can solve the $\mathbf{R}$ and $\mathbf{t}$ by minimizing the error of the objective function:

$$\underset{\mathbf{R},\,\mathbf{t}}{\operatorname{argmin}} \sum \left\| \mathbf{p}_m - \mathbf{R}\mathbf{p}_c - \mathbf{t} \right\|^2. \tag{10}$$

After the transformation parameters $\mathbf{R}$ and $\mathbf{t}$ are estimated, we align the complementary VIO result to the main VIO result.

4.2 Complementary Fusion

In our framework, the main VIO and the complementary VIO execute in parallel. They continually output the scaled and the refined camera poses. At the same time, the complementary fusion module uses the underperformance detection algorithm to detect when the main module is underperforming (drift or lose tracking) and then fuses the main VIO result and the complementary VIO result. When the main module is working as expected, the complementary fusion module uses the main VIO result. When it underperforms, the complementary fusion module replaces the main VIO result with the complementary VIO result after aligning the complementary VIO result to the main VIO result. The alignment algorithm is presented in Section 4.1.

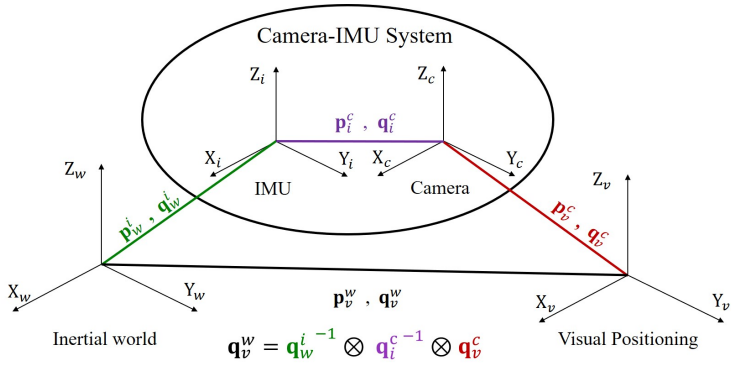

Fig. 3. Transformation relationships of coordinate systems used in sensor fusion. The red part represents the transformation relationship of the visual positioning coordinate system and camera coordinate system. The purple part represents the transformation relationship of the camera coordinate system and IMU coordinate system. The green part represents the transformation relationship of the IMU coordinate system and inertial world coordinate system.

The underperformance detection algorithm previously mentioned is from [23]. It detects behavior such as pose estimation drift or tracking loss, which often result in strange camera rotation estimations. In our framework, the camera rotation is estimated by two different modules: the visual positioning module and the sensor fusion module. Fig. 3 shows the transformation relationships of the coordinate systems used in the sensor fusion. Based on the relationships shown in Fig. 3, we estimate the rotation from the visual positioning coordinate system to the inertial world coordinate system, $\mathbf{q}_v^w$, for each EKF update step k

by the following equation:

$$\mathbf{q}_v^w(k) = \widehat{\mathbf{q}}_w^{i^{-1}}(k) \otimes \widehat{\mathbf{q}}_i^{c^{-1}}(k) \otimes \widehat{\mathbf{q}}_v^c(k), \tag{11}$$

where $\widehat{\mathbf{q}}_v^c(k)$ is the rotation from the visual positioning coordinate system to the camera coordinate system, $\widehat{\mathbf{q}}_i^c(k)$ is the rotation from the IMU coordinate system to the camera coordinate system, and $\widehat{\mathbf{q}}_w^i(k)$ is the the rotation from the inertial world coordinate system to the IMU coordinate system.

Ideally, the camera rotation estimated by the visual positioning, $\widehat{\mathbf{q}}_v^c(k)$, and the sensor fusion, $\widehat{\mathbf{q}}_w^i(k)$, is equal. However, if the visual positioning method underperforms, there is a considerable difference between $\widehat{\mathbf{q}}_v^c(k)$ and $\widehat{\mathbf{q}}_w^i(k)$, causing the variation of $\mathbf{q}_v^w(k)$ to be high because $\widehat{\mathbf{q}}_i^c(k)$ is from the calibration and is constant in the sensor fusion framework. As the variation of $\mathbf{q}_v^w$ is slow compared to the EKF update frequency, underperformance in the main module can be detected when there is an abrupt jump in the smooth $\mathbf{q}_v^w$ estimation, $\widehat{\mathbf{q}}_v^w$:

$$\widehat{\mathbf{q}}_v^w(k) = \text{Median}\left[\mathbf{q}_v^w(i)\right],\ i = k - N \rightarrow k, \tag{12}$$

where Median[] is the median filter and N is the window size. When the EKF update step comes to $k+1$, we compare the $\mathbf{q}_v^w(k+1)$ with the past M estimates $\widehat{\mathbf{q}}_v^w$. If $\mathbf{q}_v^w(k+1)$ lies outside the 3σ error bounds of the $\widehat{\mathbf{q}}_v^w$, underperformance has occurred.

5 Experiments

In this section, we evaluate the proposed complementary ego-positioning system in two scenarios: a normal scenario and a featureless scenario. The feature-based visual positioning method we used is ORB-SLAM [15], and the direct visual positioning method we used is LSD-SLAM [2]. Both methods are the classic and representative visual SLAM methods. The experimental setup is shown in Fig. 4. We built a camera-IMU system to record data for the experiment consisting of an NGIMU and a GoPro Hero4 camera, which are mounted on a box. The camera-IMU system is shown in Fig. 4(a). The sampling rate of the NGIMU was 100 Hz, and the camera was run at 30 fps. The ground truth trajectory of the camera-IMU system was taken by Vicon.

Normal Scenario In the normal scenario, we recorded data in a regular scene full of features, as shown in Fig. 4(b). The results of this scenario on the X-Y trajectory is shown in Fig. 5. The statistics of the positioning errors and scale errors are shown in Table 1. Note that all of the trajectories are aligned to the ground truth trajectories. The alignment includes scaling, rotation, and translation. The mean and standard deviation of the positioning errors in Table 1 are estimated on the aligned trajectories. The scale errors in Table 1 are the scale ratio of the positioning algorithm trajectory by the ground truth trajectory. The results show that both ORB-SLAM and LSD-SLAM have good positioning

accuracy. The positioning errors of ORB-SLAM and LSD-SLAM can be reduced by fusing with the IMU. Sensor fusion also recovers the scale of the real world. After fusing with IMU, the trajectories of ORB-SLAM and LSD SLAM become smooth and close to the ground truth trajectory. The positioning accuracy of ORB-SLAM and ORB-SLAM + IMU is higher than the positioning accuracy of LSD-SLAM and LSD-SLAM + IMU. With the complementary fusion method of the proposed complementary ego-positioning system, because the main VIO does not underperform, the final result is equal to the main VIO (ORB-SLAM + IMU) result.

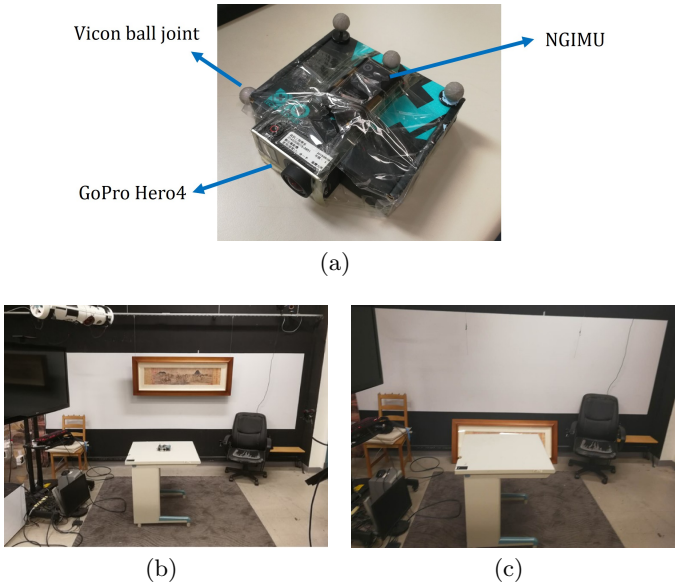

Fig. 4. Experimental setup: (a) Camera-IMU system, (b) normal scenario, and (c) featureless scenario.

Table 1. Positioning errors and scale errors in normal scenario

	Error mean (mm)	Error stdev. (mm)	Scale error
ORB-SLAM	53.5	19.2	3.13
ORB-SLAM + IMU	49.9	17.4	0.95
LSD-SLAM	58.9	30.9	2.21
LSD-SLAM + IMU	50.9	35.3	0.95
Complementary Method	49.9	17.4	0.95

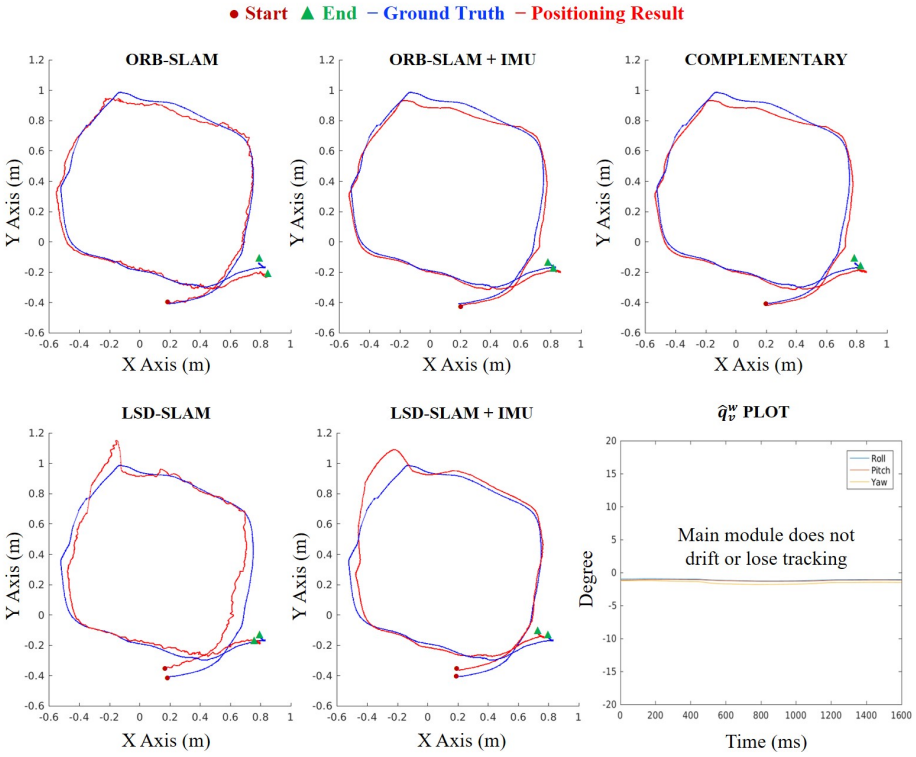

Fig. 5. X-Y trajectories of different positioning methods and plot of $\mathbf{q}_v^w$ in normal scenario

Featureless Scenario In the featureless scenario, shown in Fig. 4(c), we removed the painting in the experiment scene of the normal scenario to simulate a featureless environment with a pure white wall. The results of this scenario on the X-Y trajectory are shown in Fig. 6. ORB-SLAM loses tracking when it comes to the white wall. With the help of fusing with IMU, ORB-SLAM does not lose tracking when it comes to the white wall. However, the trajectory of the white wall part is far from the ground truth trajectory. In contrast, LSD-SLAM and LSD + IMU is robust in the whole trajectory even though their trajectories are not very close to the ground truth trajectory. The proposed complementary ego-positioning system detects the tracking loss when it comes to the white wall and uses the complementary VIO (LSD + IMU) result to replace the main VIO (ORB-SLAM + IMU) result. The statistics of positioning errors and scale errors are shown in Table 2, according to which the positioning error of the proposed complementary ego-positioning system is the lowest.

Table 2. Positioning errors and scale errors in featureless scenario

	Error mean (mm)	Error stdev. (mm)	Scale error
ORB-SLAM	N/A	N/A	N/A
ORB-SLAM + IMU	344.4	478.7	1.01
LSD-SLAM	131.9	57.2	2.37
LSD-SLAM + IMU	125.0	52.8	1.02
Complementary Method	103.0	51.3	0.95

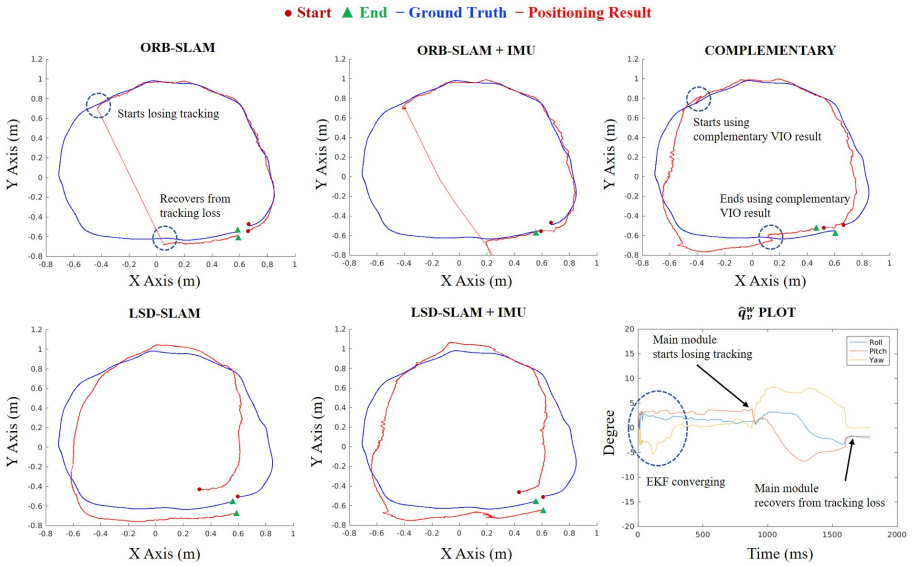

Fig. 6. X-Y trajectories of different positioning methods and plot of $\mathbf{q}_v^w$ in featureless scenario

6 Conclusion

In this paper, we propose a novel camera localization method by using sensor fusion and complementary ego-positioning. First, we use the filter-based, loosely-coupled sensor fusion to fuse the visual-positioning result with IMU measurements to yield a refined pose as well as the scale of the real world. Furthermore, we combine the ego-positioning results of feature-based SLAM and direct SLAM. Two classic and representative visual SLAM methods—ORB-SLAM and LSD-SLAM—are used in this work. ORB-SLAM is a feature-based method which is robust and accurate in normal scenes with a sufficient number of features but does poorly in featureless scenarios. LSD-SLAM is a direct method, which is less sensitive to featureless scenarios, but is less accurate than ORB-SLAM. As the two methods are complementary, the combination of ORB-SLAM and

LSD-SLAM produces more robust and accurate results. The experimental results show that in normal scenarios, sensor fusion improves the visual positioning result and estimates the scale of the real world precisely. In featureless scenarios, the direct method takes over and maintains a robust result.

Acknowledgments

This work was partially supported by MediaTek and the Ministry of Science and Technology, Taiwan.

References

1. Engel, J., Koltun, V., Cremers, D.: Direct sparse odometry. IEEE Transactions on Pattern Analysis and Machine Intelligence **40**(3), 611–625 (2017)
2. Engel, J., Schöps, T., Cremers, D.: LSD-SLAM: Large-scale direct monocular SLAM. In: European Conference on Computer Vision. pp. 834–849. Springer (2014)
3. Engel, J., Stückler, J., Cremers, D.: Large-scale direct slam with stereo cameras. In: 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 1935–1942. IEEE (2015)
4. Forster, C., Carlone, L., Dellaert, F., Scaramuzza, D.: On-manifold preintegration for real-time visual–inertial odometry. IEEE Transactions on Robotics **33**(1), 1–21 (2016)
5. Forster, C., Pizzoli, M., Scaramuzza, D.: SVO: Fast semi-direct monocular visual odometry. In: 2014 IEEE International Conference on Robotics and Automation (ICRA). pp. 15–22. IEEE (2014)
6. Forster, C., Zhang, Z., Gassner, M., Werlberger, M., Scaramuzza, D.: Svo: Semidirect visual odometry for monocular and multicamera systems. IEEE Transactions on Robotics **33**(2), 249–265 (2016)
7. Furgale, P., Barfoot, T.D., Sibley, G.: Continuous-time batch estimation using temporal basis functions. In: 2012 IEEE International Conference on Robotics and Automation. pp. 2088–2095. IEEE (2012)
8. Furgale, P., Rehder, J., Siegwart, R.: Unified temporal and spatial calibration for multi-sensor systems. In: 2013 IEEE/RSJ International Conference on Intelligent Robots and Systems. pp. 1280–1286. IEEE (2013)
9. Geiger, A., Ziegler, J., Stiller, C.: StereoScan: Dense 3D reconstruction in real-time. In: Intelligent Vehicles Symposium (IV) (2011)
10. Klein, G., Murray, D.: Parallel tracking and mapping for small AR workspaces. In: 2007 6th IEEE and ACM International Symposium on Mixed and Augmented Reality. pp. 225–234. IEEE (2007)
11. Krombach, N., Droeschel, D., Behnke, S.: Combining feature-based and direct methods for semi-dense real-time stereo visual odometry. In: International Conference on Intelligent Autonomous Systems. pp. 855–868. Springer (2016)
12. Leutenegger, S., Lynen, S., Bosse, M., Siegwart, R., Furgale, P.: Keyframe-based visual–inertial odometry using nonlinear optimization. The International Journal of Robotics Research **34**(3), 314–334 (2015)
13. Maye, J., Furgale, P., Siegwart, R.: Self-supervised calibration for robotic systems. In: 2013 IEEE Intelligent Vehicles Symposium (IV). pp. 473–480. IEEE (2013)

14. Mourikis, A.I., Roumeliotis, S.I.: A multi-state constraint Kalman filter for vision-aided inertial navigation. In: 2007 IEEE International Conference on Robotics and Automation. pp. 3565–3572. IEEE (2007)
15. Mur-Artal, R., Montiel, J.M.M., Tardos, J.D.: ORB-SLAM: a versatile and accurate monocular SLAM system. IEEE Transactions on Robotics **31**(5), 1147–1163 (2015)
16. Mur-Artal, R., Tardós, J.D.: Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras. IEEE Transactions on Robotics **33**(5), 1255–1262 (2017)
17. Mur-Artal, R., Tardós, J.D.: Visual-inertial monocular SLAM with map reuse. IEEE Robotics and Automation Letters **2**(2), 796–803 (2017)
18. Newcombe, R.A., Lovegrove, S.J., Davison, A.J.: DTAM: Dense tracking and mapping in real-time. In: 2011 International Conference on Computer Vision. pp. 2320–2327 (Nov 2011). https://doi.org/10.1109/ICCV.2011.6126513
19. Nisar, B., Foehn, P., Falanga, D., Scaramuzza, D.: Vimo: Simultaneous visual inertial model-based odometry and force estimation. IEEE Robotics and Automation Letters **4**(3), 2785–2792 (2019)
20. Qin, T., Li, P., Shen, S.: VINS-Mono: A robust and versatile monocular visual-inertial state estimator. IEEE Transactions on Robotics **34**(4), 1004–1020 (2018)
21. Vukmirica, V., Trajkovski, I., Asanovic, N.: Two methods for the determination of inertial sensor parameters. methods **3**(1) (2018)
22. Weiss, S., Achtelik, M.W., Chli, M., Siegwart, R.: Versatile distributed pose estimation and sensor self-calibration for an autonomous MAV. In: 2012 IEEE International Conference on Robotics and Automation. pp. 31–38. IEEE (2012)
23. Weiss, S., Siegwart, R.: Real-time metric state estimation for modular vision-inertial systems. In: 2011 IEEE International Conference on Robotics and Automation. pp. 4531–4537. IEEE (2011)
24. Zhang, Z.: A flexible new technique for camera calibration. IEEE Transactions on Pattern Analysis and Machine Intelligence **22**(11), 1330–1334 (2000)