Head Motion Detection in Gaze Based Aiming

Minghe Cao and Jianzhong Wang✉

(School of Mechatronical Engineering, Beijing Institute of Technology, Beijing 100081, China)

Abstract: Unmanned weapons have great potential to be widely used in future wars. The gaze-based aiming technology can be applied to control pan-tilt weapon systems remotely with high precision and efficiency. Gaze direction is related to head motion, which is a combination of head and eye movements. In this paper, a head motion detection method is proposed, which is based on the fusion of inertial and vision information. The inertial sensors can measure rotation in high-frequency with good performance, while vision sensors are able to eliminate drifts. By combining the characteristics of both sensors, the proposed approach achieves the effect of high-frequency, real-time, and drift-free head motion detection. The experiments show that our method can smooth the outputs, constrain drifts of inertial measurements, and achieve high detection accuracy.

Key words: gaze aiming; head motion detection; visual-inertial information fusion

CLC number: TP 23 **Document code**: A **Article ID**: 1004-0579(2020)01-0009-07

Unmanned weapons will play an essential role in future wars. Remote control of unmanned weapon systems by Human-Machine-Interacting will be in the mainstream for a long period of time in the future. Now the teleoperation of unmanned weapon systems heavily relies on control terminals, which has several drawbacks, including the inconvenience of operation and imprecise target aiming. The teleoperation based on head-gaze motion can easily and efficiently achieve precise aiming on pan-tilt weapon systems[1].

Head and gaze motions are highly related to human inspecting behavior. It is a natural procedure that head motion can provide an extra field of view for gaze. When people are interested in an object, the head will naturally turn to the target area. When gazing at a static target, the head remains still. While tracking a dynamic target, it has different levels of movements. In the detection of eye movements, the gaze axis is based on the head. Consequently, the detection of head pose plays an essential role in gaze tracking[2-3].

The detection of head motion can be seen as a state estimation problem in six degrees of freedom. Several methods can be leveraged[4]. Sensors, including inertial sensors and visual sensors, can be attached to head or helmet directly to measure the motion. The measurements based on these sensors are easily setup, thus have broader applications[5-6].

The Inertial Measurement Unit (IMU) is a device to detect angular velocity and acceleration of a rigid body using a gyroscope and an accelerometer. The main benefit of employing IMU is that no external transmitter or receiver is needed to attach to the head when performing the detection, releasing the constrains of head motion. However, a flaw of IMU is that its outputs contain measurement noise and drift, leading to a diverge to the ground truth over time with integration. The vision sensors exploit features of environments such as edges and surfaces to acquire pose information. However, due to the high computation complexity, the algorithms based on vison can hardly achieve a relatively high real-time per-

Received 2019-11-27
Supported by Defense Industrial Technology Development Program (JCKY2017602C016)
✉ Author for correspondence, professor, Ph. D.
E-mail: cwjzwang@bit.edu.cn
DOI: 10.15918/j.jbit1004-0579.19106

formance. Besides, the vision-only methods cannot resolve series of issues like blur, occlusion, lacking features, pure rotation, and scale ambiguity, causing a failure of tracking and positioning. The fusion of visual information and inertial information can efficiently make up for corresponding shortcomings.

Visual Inertial Odometry (VIO) is a method aiming to detect the poses of a rigid body by combing the information of inertial and visual[7]. It can be classified into two categories: filter-based[8-10] and optimization-based[11-14]. The filter-based approaches estimate the current state utilizing only one previous state, while the optimization-based methods consider the impact of all states, finding solutions by minimizing errors of optimization functions. Generally, the accuracy of optimization-based approaches is superior to filter-based approaches. However, due to high volume of data, the optimization-based approaches have poor timing performance. Alternatively, based on how the visual information and inertial information are associated, the VIO can also be divided into two groups as loosely-coupled[10-11] and tightly-coupled[8-9,14]. The loosely-coupled approaches deal with visual information and inertial information separately and fuse the two in the back-end. Oppositely, the tightly-coupled techniques consider visual features and inertial measurements as one state vector and find the solution based on the state vector. In tightly-coupled optimization-based approaches, various visual information is used as the front-end. OKVIS[12] uses the Harris corner to extract keypoints and describes the keypoints using the BRISK descriptor. VINS-mono[14] employed KLT sparse optical flow. VIORB[13] exploited ORB features. Other than odometry, the VIO approaches can also be applied to applications like Visual Reality and Augment Reality[15].

In this paper, a head motion detection method based on the fusion of visual information and inertial information is proposed. The inertial information is used as a high-frequency base output to ensure a real-time performance. The rectification is performed by constructing the optimization function that minimizes the errors of reprojection and inertial bound. The rectified data is then smoothly filtered. The proposed technique achieves a high-frequency output, which has the benefits of real-time, drift-free, and smooth head motion detection.

1 Inertial Integration Model

The IMU is composed of accelerometers and gyroscopes that are usually used to measure acceleration and angular rates. We denote subscript b as the coordinate frame of IMU, and subscript w as the world coordinate frame. The IMU raw output of gyroscopes and accelerometers is $\widetilde{\boldsymbol{\omega}}_b \in \mathbb{R}^3$, $\widetilde{\boldsymbol{a}}_b \in \mathbb{R}^3$ where their ground truth is $\boldsymbol{\omega}$ and $\boldsymbol{a}$. Then in coordinate b, we have

$$\begin{aligned}\widetilde{\boldsymbol{\omega}}_b &= \boldsymbol{\omega}_b + \boldsymbol{b}_g + \boldsymbol{n}_g \\ \widetilde{\boldsymbol{a}}_b &= \boldsymbol{q}_{bw}(\boldsymbol{a}_w + \boldsymbol{g}_w) + \boldsymbol{b}_a + \boldsymbol{n}_a\end{aligned} \tag{1}$$

where $\boldsymbol{q}_{bw} = [w, x, y, z]$ is the rotation from the world frame to IMU frame, $\boldsymbol{g}_w = [0, 0, -g]^{\mathrm{T}}$ is the gravity vector in the world frame, $\boldsymbol{n}_g \sim \mathcal{N}(0, \sigma_g^2)$, $\boldsymbol{n}_a \sim \mathcal{N}(0, \sigma_a^2)$ are Gaussian noises, and $\boldsymbol{b}_g$, $\boldsymbol{b}_a$ are bias where $\dot{\boldsymbol{b}}_g = \boldsymbol{n}_g$, $\dot{\boldsymbol{b}}_a = \boldsymbol{n}_a$. If we denote the state of IMU in the frame b at time t as $\boldsymbol{x}_I^t = [\boldsymbol{p}, \boldsymbol{v}, \boldsymbol{q}, \boldsymbol{b}_a, \boldsymbol{b}_g]^{\mathrm{T}}$, where $\boldsymbol{p}$, $\boldsymbol{v}$ and $\boldsymbol{q}$ are position, velocity and rotation respectively, the time derivative of the state is

$$\dot{\boldsymbol{p}}_{wb_t} = \boldsymbol{v}_{wb_t}, \dot{\boldsymbol{v}}_{wb_t} = \boldsymbol{a}_w^t, \dot{\boldsymbol{q}}_{wb_t} = \boldsymbol{q}_{wb_t} \otimes \begin{bmatrix} 0 \\ \frac{1}{2}\boldsymbol{\omega}_{b_t} \end{bmatrix} \tag{2}$$

where $\otimes$ is quaternion multiplication. In the continuous-time model, the integration of state from current state i to next time step j can be expressed as

$$\boldsymbol{p}_{wb_j} = \boldsymbol{p}_{wb_i} + \boldsymbol{v}_{wb_i}\Delta t + \iint_{t\in[i,j]} (\boldsymbol{q}_{wb_t}\boldsymbol{a}_{b_t} - \boldsymbol{g}_w)\,\mathrm{d}t^2$$

$$\boldsymbol{v}_{wb_j} = \boldsymbol{v}_{wb_i} + \int_{t\in[i,j]} (\boldsymbol{q}_{wb_t}\boldsymbol{a}_{b_t} - \boldsymbol{g}_w)\,\mathrm{d}t$$

$$\boldsymbol{q}_{wb_j} = \int_{t\in[i,j]} \boldsymbol{q}_{wb_i} \otimes \begin{bmatrix} 0 \\ \frac{1}{2}\boldsymbol{\omega}_{b_t} \end{bmatrix} \mathrm{d}t \tag{3}$$

Typically, the output frequency of inertial sensors is higher than visual sensors', causing multiple inertial measurements between images. In optimization problems, after the image poses are adjusted at every time, the inertial measurements need to be re-integrated, resulting in computationally time-consuming. To avoid there-integration, we adopt a pre-integration model. We denote Δq_{ij} as the rotation between time i and j in the frame b, let $q_{wb_j} = q_{wb_i} \otimes \Delta q_{ij}$, from Eq. (3) we have

$$\begin{aligned} \boldsymbol{p}_{wb_j} &= \boldsymbol{p}_{wb_i} + \boldsymbol{v}_{wb_i}\Delta t - \frac{1}{2}\boldsymbol{g}_w \Delta t^2 + \boldsymbol{q}_{wb_i}\Delta\boldsymbol{\alpha}_{ij} \\ \boldsymbol{v}_{wb_j} &= \boldsymbol{v}_{wb_i} - \boldsymbol{g}_w \Delta t + \boldsymbol{q}_{wb_i}\Delta\boldsymbol{\beta}_{ij} \\ \boldsymbol{q}_{wb_j} &= \boldsymbol{q}_{wb_i}\Delta\boldsymbol{q}_{ij} \end{aligned} \tag{4}$$

where $\Delta\boldsymbol{\alpha}_{ij} = \iint_{t\in[i,j]} (\Delta\boldsymbol{q}_{it}\ \boldsymbol{a}_b^t)\ \delta t^2$, $\Delta\boldsymbol{\beta}_{ij} = \int_{t\in[i,j]} (\Delta\boldsymbol{q}_{it}\boldsymbol{a}_b^t)\delta t$, $\Delta\boldsymbol{q}_{ij} = \int_{t\in[i,j]} \Delta\boldsymbol{q}_{ij} \otimes \begin{bmatrix} 0 \\ \frac{1}{2}\boldsymbol{\omega}_{b_t} \end{bmatrix} \mathrm{d}t$ are pre-integration terms.

2 Visual Based Pose Estimation

2.1 Feature extraction

In visual information, several methods such as HOG, SIFT, SURF can be leveraged to extract environmental features[16]. Oriented FAST and Rotated BRIEF (ORB) is a fast and robust local feature detector composed of keypoint and descriptor. The Harris corner detection is first adopted to find FAST keypoints to extract ORB features. However, FAST features do not have an orientation component and multiscale features, so ORB uses a multi-scale image pyramid, which downsampling image by a scale factor in different levels and detects keypoints at each level and assigns an orientation to each keypoint. The keypoints are then scored and selected based on quality and finally described using the BRIEF descriptor.

After extracting ORB features, feature matching can be employed to find the same environment point from different image views. The similarity between two descriptors can be obtained by calculating the distance. For BRIEF, we use the Hamming Distance. Some features are inevitably mismatched when conducting feature matching. Thus RANSAC is used to cull outliers.

2.2 Mulity-view bundle adjustment

After extracting and matching features from various time, we apply the bundle adjustment (BA) to minimize the overall reprojection error by adjusting camera poses and feature point positions and therefore calculate optimized camera poses. Assume the state vector of image c_i at time i is $\boldsymbol{x}_c^i = [\boldsymbol{p}_{wc_i}, \boldsymbol{q}_{wc_i}]^{\mathrm{T}}$. f_j is the position of feature point in the world frame. Then the optimization function can be constructed based on the reprojection error

$$\arg\min_x \sum_{i=1}^{m}\sum_{j=1}^{n} \| \pi(\boldsymbol{q}_{wc_i}, \boldsymbol{p}_{wc_i}, f_j) - \boldsymbol{z}_{c_i f_j} \|^2 \tag{5}$$

where $\pi(\cdot)$ is the reprojection of features from the 3D world frame to 2D camera plane, $\boldsymbol{z}_{c_i f_j}$ is the observation of c_i to feature f_j, and x is the state vector to be optimized.

3 Proposed Approach

The head motion detection system mainly includes inertial information and visual information acquisition, visual-inertial fusion based rectification, feature point cloud initialization and update, and output filtering. The inertial measurements collect data from the accelerometer and gyroscope at a high frequency and provide short-time motion states by the integration model. The visual information acquires images and extracts features of the surrounding environment at a low frequency. The visual-inertial fusion based rectification takes the inertial bound and visual matching to construct an optimization function and then computes local drift-free pose by solving the function. After the fusion, images are selected based on certain conditions as keyframes to update the feature pointcloud. The output filter is responsible

for imposing the low-frequency rectifying data to the high-frequency inertial data, which is then smoothly filtered as the final output. The flow chart of head motion detection is illustrated in Fig. 1.

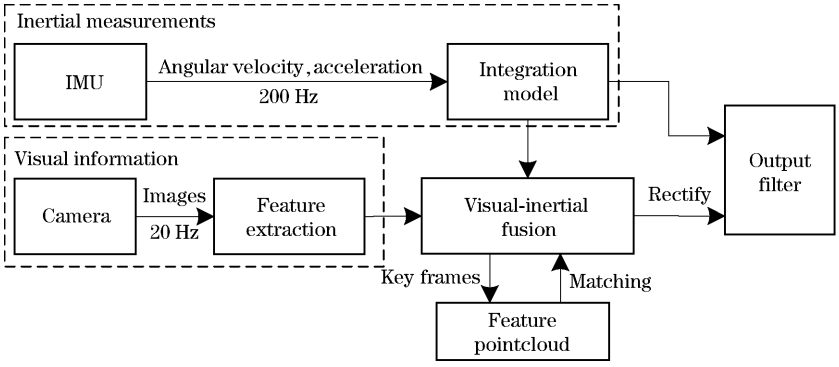

Fig. 1 Flow chart of head motion detection

3.1 Feature pointcloud

To ensure that visual information can provide a global reference in which the error of visual-inertial fusion will not accumulate with time, we construct a feature pointcloud in the world coordinate based on visual features and their position. Before the detect system running, we initialize a prior ORB feature pointcloud to ensure the accuracy and stability of the reference. The tester collects images of environments by performing head rotate and translate. After the collection, we perform the full BA optimization of collected images to find the optimized camera poses and feature positions. Finally, we stitch features together as the initialized feature pointcloud. With the pointcloud, images can take a reference to build reprojection errors.

The visual-inertial fusion will provide rectified information to inertial data. Alternatively, some images are selected from the outcome of the visual-inertial fusion as keyframes to update the pointcloud. Specifically, the select condition includes: the image has sufficient ORB features, the number of features decreased 50% compared to the last keyframe, and relative pose greater than a certain threshold.

After selecting a set of keyframes, we set the prior pointcloud fixed and employ BA to optimize keyframes pose and feature and finally update the pointcloud by inserting the optimized features.

3.2 Visual-inertial based rectification

When computing the visual-inertial rectification, the optimization-based approach will be used again. Distinct from the feature pointcloud construction, for real-time rectification, we optimize the current frame j by minimizing the feature reprojection error of all matched features and an IMU error term link to one previous frame.

Specifically, given the current image c_j, we first compute reprojection errors by extracting its ORB features and matching the features within the pointcloud. Let $\boldsymbol{x}_i, \boldsymbol{x}_j$ are the IMU states corresponding to two successive images. We can establish an optimization function

$$\arg\min_{x}\left\{\sum_{k=1}^{K}\rho\left(\left\|e_M\left(c_j, f_k\right)\right\|_{\Lambda_f}^2\right)+\rho\left(\left\|e_I\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right)\right\|_{\Lambda_I}^2\right)\right\} \tag{6}$$

where $e_M(\cdot)$ is the reprojection error term of all matched points, $e_I(\cdot)$ is an inertial error term of pre-integration, $\boldsymbol{\Lambda}_f$ and $\boldsymbol{\Lambda}_I$ are information matrices of visual information and inertial information, respectively. $\rho(\cdot)$ denote the Huber cost function. By linearizing Eq. (6), we can solve the problem using the Gauss-Newton algorithm.

3.3 Output filter

The inertial information provides high-frequency base measurements while the visual-inertial information imposes rectifications to them. However, due to the difference in frequency, a

simple fusion may result in a discontinuity, leading to an unstable control signal. Therefore, the rectified head motion is applied to a fixed window filter that weighted the poses within the window and finally smoothed the output. For a unit quaternion $\boldsymbol{q} = [w, x, y, z]^{\mathrm{T}}$, the conversion to the axis-angle representation can be expressed as

$$\begin{aligned} \theta &= 2\cos^{-1} w \\ \boldsymbol{u} &= \frac{[x, y, z]}{\sin(\theta/2)} \end{aligned} \tag{7}$$

where the equation satisfies, in which θ and $\boldsymbol{u}$ are angle and axis. Let the rotation at time i be q_i and its axis-angle representation is ρ_i. Set the window size to N. The output pose $\hat{q}$ is

$$\hat{q} = \sum_{n=1}^{N} \frac{1}{N} \rho_{i-n} \tag{8}$$

4 Experiment

Experiments were carried out to measure the head motion when remotely control a pan-tilt system to track a target using our gaze tracking system. The experiment scene is presented in Fig. 2. A tester worn the experimental system to track the target moving in front of apan-tilt system. The experiment system is composed of a head detection module, a gaze tracking module, and a ground truth IMU module. The head detection module consists of an RGB-D camera and a BMI088 IMU. The ground truth module uses a 3DM-GX5-25 IMU. In the actual operation, the head and gaze motion are combined to control the pan-tilt system. In this experiment, only head motion is considered.

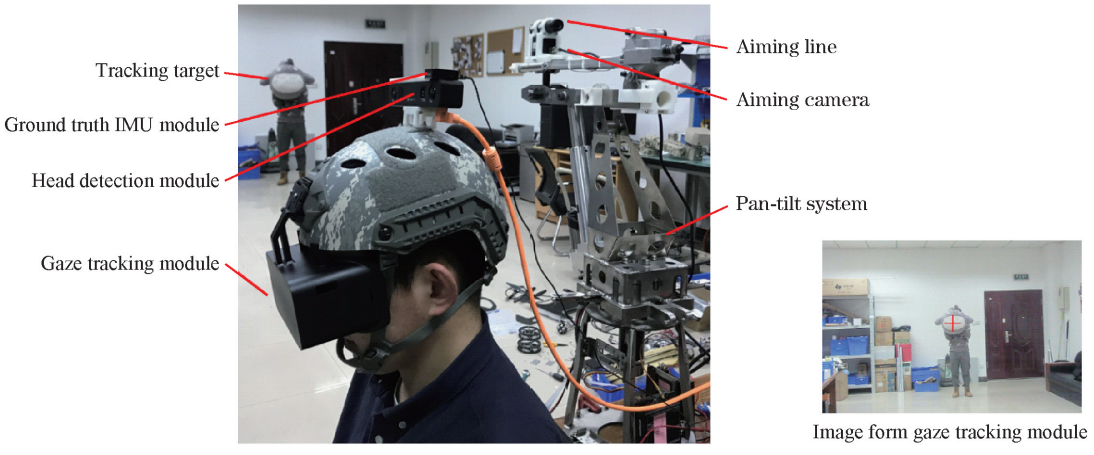

Fig. 2 Experiment scene

4.1 Output filter

Due to the fact that inertial measurements and visual-inertial fusion have different output frequencies, imposing rectification directly could result in a segmented output. Fig. 3 presents the raw output of head motion when tracking the target. The x-axis is time, and the y-axis is the pitch rotation in degrees. Fig. 3a shows the result of directly imposing rectification. Fig. 3b gives the result of filtering with a window of size $N = 10$. From the figures, it can be observed that the filtered curve made up the fracture caused by the rectification in the raw output and smoothed the raw output.

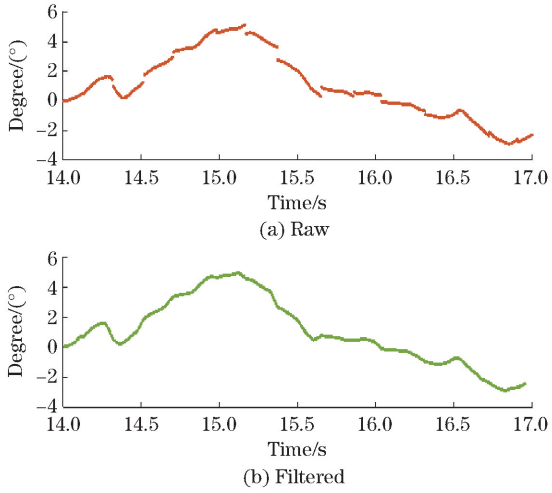

Fig. 3 Smoothed output

4.2 Detection error

The results of our method is compared to the outcome of raw IMU integration, and the errors with respect to the ground truth are shown in Fig. 4. The x-axis is experimental time and y-axis is the rotation error of pan (Fig. 4a) and tilt (Fig. 4b) directions. Fig. 4 shows that the visual-inertial approach achieved a good detection accuracy that the mean errors from both directions are less than one degree. Also, due to the measuring noises, the error for the inertial measures increases with time, while our visual-inertial method does not accumulate for the global feature pointcloud reference. The results show that our approach can efficiently rectify the inertial drift, achieved a good detection performance in both control directions.

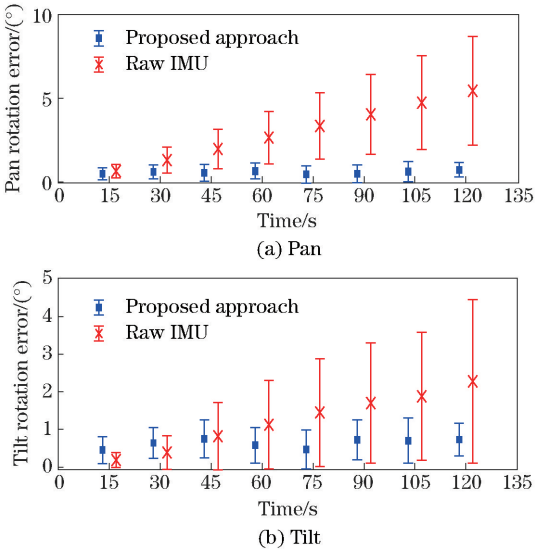

Fig. 4 Detection error

5 Conclusion

In this paper, a head motion detection method based on the fusion of visual and inertial information is proposed. The high-frequency inertial information is used as a base output in short-time for its good rotation character. The visual information is employed to construct a feature point cloud to provide a global drift-free reference. The visual information exploits inertial bounds between two images and reprojection error of the feature point cloud, and imposes rectifications to inertial measurements. Finally, the rectified data is smoothed by filtering. The experiment results show that the proposed head motion detection method with the basis on visual-inertial fusion is able to smoothly detect head motion with good accuracy.

References:

[1] Wang Jianzhong, Zhang Guangyue, Wang Hong. Mixed weighted feature method for human eye detection [J]. Transactions of Beijing Institute of Technology, 2019, 39(8): 819 – 824. (in Chinese)

[2] Nie Xiangrong. The gaze tracking system based on head rotation information and gaze information[D]. Qinhuangdao, Hebei: Yanshan University, 2017. (in Chinese)

[3] Jiang Guangyi. Research on data fusion of head movement and gaze tracking[D]. Xi'an: Xi'an Technological University, 2017. (in Chinese)

[4] Luo Bin, Wang Yongtian, Liu Yue. Multi-sensor data fusion for optical tracking of head pose[J]. Acta Automatica Sinica, 2010, 36(9): 1239 – 1249.

[5] Wang Shaomei. A method of measurement of human head motion attitude [J]. Journal of Xi'an Technological University, 2011, 31(5): 429 – 433. (in Chinese)

[6] Miezal M, Bleser G, Stricker D, et al. Towards practical inside-out head tracking for mobile seating bucks[C]// Proceedings of International Symposium on Mixed and Augmented Reality (ISMAR), Atlanta, USA, 2012.

[7] Gui J, Gu D, Wang S, et al. A review of visual inertial odometry from filtering and optimisation perspectives[J]. Advanced Robotics, 2015, 29: 1289 – 1301.

[8] Mourikis A I, Roumeliotis S I. A multi-state constraint kalman filter for vision-aided inertial navigation[C]// IEEE International Conference on Robotics & Automation, IEEE, 2007.

[9] Bloesch M, Omari S, Hutter M, et al. Robust visual inertial odometry using a direct EKF-based approach [C]//IEEE/RSJ International Conference on Intelligent Robots & Systems, IEEE, 2015: 298 – 304.

[10] Lynen S, Achtelik M W, Weiss S, et al. A robust and modular multi-sensor fusion approach applied

to MAV navigation[C]// Intelligent Robots and Systems (IROS), 2013 IEEE/RSJ International Conference on, IEEE, 2013.

[11] Falquez J M, Kasper M, Sibley G. Inertial aided dense & semi-dense methods for robust direct visual odometry[C]// 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, 2016.

[12] Leutenegger S, Lynen S, Bosse M, et al. Keyframe-based visual-inertial odometry using nonlinear optimization[J]. The International Journal of Robotics Research, 2015, 34(3):314 – 334.

[13] Mur-Artal R, Tardos J D. Visual-inertial monocular SLAM with map reuse[J]. IEEE Robotics and Automation Letters, 2017, 2(2):796 – 803.

[14] Tong Q, Peiliang L, Shaojie S. VINS-Mono: a robust and versatile monocular visual-inertial state estimator[J]. IEEE Transactions on Robotics, 2018,34(4):1004 – 1020.

[15] Wei F, Lianyu Z, Huanjun D, et al. Real-time motion tracking for mobile augmented/virtual reality using adaptive visual-inertial fusion[J]. Sensors, 2017, 17(5):1037.

[16] Zhou Pengbo, Zhao Fuqun, Wu Zhongke. Fracture surface matching method of terracotta based on geometric features[J]. Transactions of Beijing Institute of Technology, 2019, 39(5): 532 – 538. (in Chinese)

(**Edited by** Yuxia Wang)