Academia Journal of Scientific Research 7(8): 000-000, August 2019
DOI: 10.15413/ajsr.2019.0207
ISSN 2315-7712

Academia Publishing

Research Paper

High precision position solving for monocular SLAM Fused with IMU

Accepted 29th August, 2019

Wei Sun

School of Aerospace Science and Technology, Xidian University, No. 2 South Tabai Rd., Xi'an, Shaanxi, 710071, China. E-mail: wsun@xidian.edu.cn.

ABSTRACT

The scale uncertainty of monocular visual SLAM reduces pose's accuracy due to accumulation of errors during moving. This study proposed a high-precision pose solving optimization method based on direct visual odometry and inertial measurement unit. In the front-end defined as visual solution, the distortion parameters are used in pose solving by multi-parameter sliding-window optimization; the global state variables are constructed, while the multi-parameter optimization criterion is established to reduce the error accumulation of each parameter. In the process of back-end pose optimization, using the complementarity of the static zero drift of the visual sensor and the ground-truth values obtained with the IMU in a short time, the scale factor is designed to make the visual and inertia maintain the scale consistency in the pose solution; Finally, *a priori* information corresponding to the specific frame is selected to construct the overall optimization objective function based on prior information, visual information and inertial information. The global state variable is optimized by the objective function and iteratively updated to reduce the accumulation of subsequent errors and obtain the high-precision pose information. The experimental results show that compared with the monocular DSO without IMU and the current state-of-art visual-inertial method, the proposed algorithm greatly improved the accuracy of pose calculation, and reduced the error accumulation, thereby leading to the algorithm achieving improvement in construction accuracy and point cloud quality of SLAM.

Key words: Visual SLAM, Direct sparse odometry, IMU, Visual-inertial method, Pose calculation.

INTRODUCTION

The mobile robotic perception technology in the unknown environment is based on SLAM (Simultaneous Localization and Mapping, SLAM) technology, which is the key to realize the autonomous navigation (Cadena et al., 2016). Visual SLAM can be divided into direct method (Silveira et al., 2008; Engel et al., 2017; Forster et al., 2014; Engel et al., 2014) and indirect method (Mur-Artal et al., 2015; Kümmerle et al., 2011) in terms of matching principle. The mainstream indirect method is ORB-SLAM (Mur-Artal et al., 2015). Based on PTAM (Klein, 2007), the algorithm is proposed to automatically calculate the initial pose and extract high-speed feature points ORB (Oriented FAST and Rotated BRIEF, ORB) (Rublee et al., 2011) and use G2O (General Graph Optimization, G2O) (Kümmerle et al., 2011) to minimize global error, making it more suitable for large-scale scenes.

However, the tracking method relies much on the previous feature point, and update of the map too slowly and cause tracking lost during fast motion. Based on the direct method, the DSO (Direct Sparse Odometry, DSO) algorithm proposed by Engel et al. (2017) adds a photometric calibration model that directly uses pixel points to construct photometric error optimization; the computation burden is greatly simplified and can be applied

in weak texture environment with high robustness. However, in the actual application, errors accumulation will make successive pose solution have a scale proportional or wrong to the ground-truth value. Using monocular visual odometry (Zhu et al., 2018), accumulated error leads to inaccuracy pose solution, thereby drifting the scale of the point cloud and trajectory.

In visual-inertial pose solving algorithm, the main methods are tightly coupled based on filtering and optimization algorithm (Gui et al., 2015). Rovio filtering algorithm (Bloesch et al., 2015) uses the IMU and map points in the images as input and proposes a feature block to assist the EKF algorithm to solve the uncertainty of the optimization point. The prediction process of EKF estimates the block position. In the update step, the feature block is used to solve the projection point intensity difference, while the optimized state quantity is updated; but the algorithm is not perfect for the error processing and likely to cause an error cumulative phenomenon. For the optimization-based visual-inertial method, Okvis algorithm (Leutenegger et al., 2014) matches image frames by feature points to construct projection errors, and then obtains visual residual terms. It reduces the computational complexity based on key frame coupling optimization, and constructs the objective function with landmark re-projection error and the IMU bias which is used in common non-linear optimization, and the old state is marginalized to limit the complexity. However, Okvis only takes pose and deviation as solving variables. Point cloud in the constructed map is spare, while the accuracy is a little bit low and the map utilization is too low in practical applications.

The proposed method named VI-DSO (Visual-Inertial Direct Sparse Odometry) is based on IMU (Inertial Measurement Unit, IMU) and DSO which uses image frame to match pixels. The excellent characteristics which can accurately estimate the pose of moving object in a short time, and the pre-integration of IMU sample value used (Forster et al., 2015) can decrease visual's error accumulation in DSO. The proposed optimization framework is built while the IMU sample value is used for calibration, hence, the scale factor is proposed to keep the scale consistency between vision and inertia pose solution; even the bias drift problem of IMU module can be corrected by stationary camera with low drift in DSO. Using the complementary characteristics of visual and inertia method earlier mentioned, a visual-inertial SLAM high-precision pose solving method based on optimization and DSO is proposed to solve the error accumulation and scale uncertainty in monocular SLAM system. In the front-end visual-inertial pose estimation process, we construct photometric error function, and then obtain visual error terms. The distortion parameters are proposed and used to optimize the global optimization state variables. In the back-end pose optimization process, the scale factor is designed to correct the visual pose scale solution, while the global state variable is iteratively updated by the optimization solution to obtain high-precision pose solution.

These constraints are fused with a co-optimization of relative poses extracted from DSO's sliding window and IMU optimization. The experimental results show that the integrated optimization significantly reduces the accumulated rotation-, translation- and scale-drift, resulting in an overall performance comparable to state-of-the-art feature-based systems such as Okvis etc. The proposed method can utilize any image pixel with sufficient intensity gradient, which makes it robust even in featureless areas. The VI-DSO achieves high-precision pose calculation and has improved significantly in construction accuracy and point cloud quality.

VISUAL INERTIAL VARIABLE DEFINITION AND FRONT-END ESTIMATION

Visual Inertial Odometry

Figure 1 shows the pose estimation of the visual-inertial odometry integrated with IMU. Pose estimation is divided into two modules: front-end estimation and back-end optimization. The front-end estimation mainly completes the initial pose estimation of the image frame. The photometric error function between the new frame and the nearest key frame is constructed, while the initial pose is solved by non-linear optimization method. Back-end optimization integrates visual and IMU to achieve high precision pose estimation. The inter-frame error is constructed according to the key frame of the front-end initial pose selection, while the inertia error is constructed based on the IMU pre-integration value of front-end. The scale is used to keep the scale consistency between the visual and inertia. The *priori* information corresponding to a specific frame is selected to construct the back-end optimization objective function of the *priori*, inter-frame and inertia error. Gauss-Newton algorithm is used to solve the objective function, while the position information of the map points is used to build the point cloud map in SLAM.

Construction of the optimization state variables

It is known that the variables p, v, q and the inherent deviations b^g, b^a were obtained by pre-integrating the measured values of the IMU module; the frame state at time k can then be defined. The corresponding optimization variables are as shown in Equation 1:

$$x_k = [{}_W p_k, {}_W v_k, {}_W q_k, b_g, b_a, s] \tag{1}$$

The variables represent the scale factor, which makes the visual photometric and inertia error to keep the scale consistency in the pose information. The variables

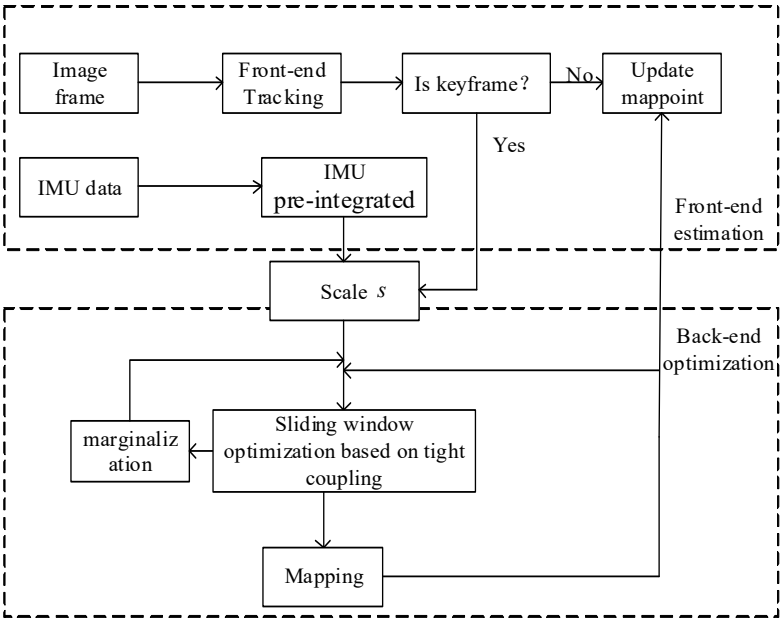

Figure 1: Diagram of the proposed Visual-Inertial Odometry.

p, v, q, b^g and b^a represent the translation, velocity, angle, gyroscope bias, and accelerometer bias of the IMU pre-integration. The rules of constructing global optimization state variables are according to the fusion characteristics of visual and inertial tight coupling (Martinelli, 2018); in back-end pose optimization process, map points are used to construct visual inter-frame photometric error, hence, the inverse depth information of the map point is component parameter of global optimization state variable. In addition, the accuracy of the camera internal parameters, distortion and other parameters included in the de-distortion inter-frame photometric error also affects the accuracy of the back-end pose optimization. Therefore, the global optimization state variable of the visual inertia odometry is defined as Equation 2:

$$x = [x_1, x_2, ..., x_m, k_1, k_2, k_c, \rho_1, ..., \rho_n] \quad (2)$$

Where k_1, and k_2 represents the distortion variable of the image, k_c represents the internal parameter variable of the monocular camera, ρ_i represents the inverse depth of the unmargined map point in the back-end sliding window, and m represents the number of active key-frames in the window.

Definition of scale factor

Figure 2 shows the motion transform from IMU to visual. B, W, and C correspond to IMU, world, and camera coordinate system, respectively. The coordinate system of dotted line indicates unreal scale coordinate system due to scale uncertainty when transforming from camera coordinate system to world coordinate system. When transforming origin of the B- system to W-system, there is a scale proportional relationship between direct conversion and C-system conversion. The T_{WB} obtained by the two schemes should be consistent under ideal conditions, but variables obtained by IMU pre-integrating the sampled values are calibration values; moreover, due to scale uncertainty of monocular vision, poses solution cannot be obtained, and the accumulated error along time cause inconsistency in the scale information of the pose, so that it will be converted to the virtual W-system by means of C-system transformation.

The black straight line in Figure 2 represents the translation process from the current coordinate system to the next coordinate. The red arc line indicates the rotation process in which each axis of the current coordinate system is rotated into three axes in a direction consistent with the coordinate system indicated by the arrow. In order to maintain the IMU pre-integration and the visual pose consistent in scale, it is assumed that there is a scale factor s, so that the IMU pre-integrated of the sample value q, p and visual poses have the following relationship as shown in Equation 3:

$$s \cdot {}_{WC}p = {}_{WB}q\,{}_{WB}p - {}_{WC}q \cdot {}_{CB}q \cdot {}_{CB}p \quad (3)$$

Where the subscript represents the conversion process (as

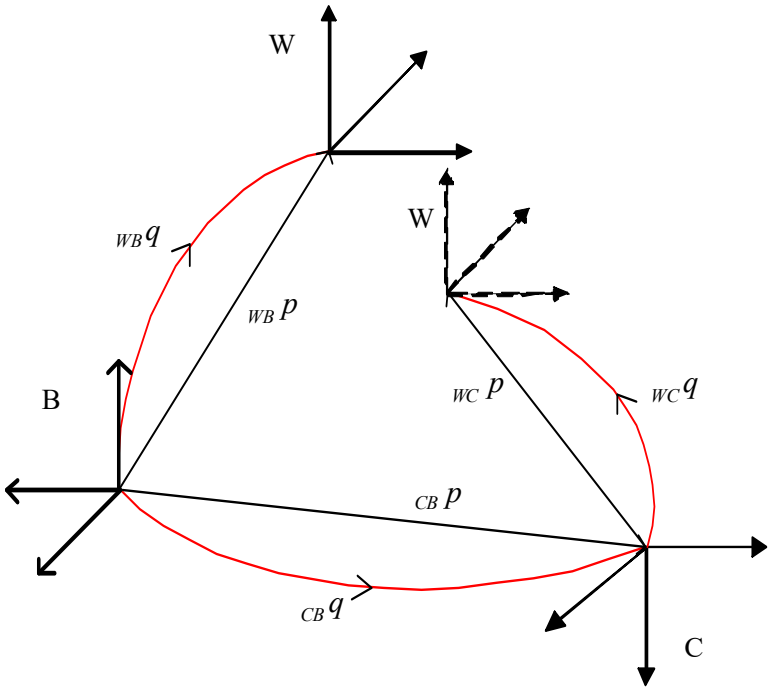

Figure 2: Coordinate system transforming.

indicated by the conversion from the C system to the W system), and the conversion process is performed by the coordinate system of the first translation and then rotated by the origin of the IMU coordinate system; finally, the difference between the coordinate values at the end of the two schemes is infinitely zero.

FRONT-END ESTIMATION, BACK-END OPTIMAL OBJECT FUNCTION CONSTRUCTION AND SOLUTION

Front-end initial pose estimation

A de-distortion inter-frame pose estimation algorithm is proposed in the process of initial pose estimation between image frames. The distortion parameters are added in the inter-frame matching process, while the photometric error function is constructed to solve the visual-inertial pose, assuming that there are two image frames: current frame and reference frame. The corresponding point p_j is obtained using projection formula for pixel point p_i in current frame as shown in Equation 4:

$$p\_j = K \cdot (1 + k_1 r^2 + k_2 r^4) \rho_j (\rho_i^{-1} \exp(\xi^{\wedge}_{k,k+1}) K^{-1} p\_i \quad (4)$$

Equation 4 uses the non-linear optimization algorithm to obtain initial pose corresponding to current frame (Strasdat, 2012). The actual measured values of IMU module are pre-integrated to obtain variables p , v , q and inherent bias b^g and b^a from module as shown in Equation 5:

$$\begin{aligned}
{}_{WB}q_{k+1}^{k} &= \prod_{i=k}^{k+1} Exp[(\omega_i - b_i^g - \eta_i^{gd})dt] \\
{}_{WB}v_{k+1}^{k} &= \sum_{i=k}^{k+1} {}_{WB}q_i^k [\tilde{a}_i - b_i^a - \eta_i^{ad} + ({}_{WB}q_i^k)^T g_w)]dt \\
{}_{WB}p_{k+1}^{k} &= \sum_{i=k}^{k+1} [{}_{WB}v_i^k dt + \frac{1}{2} {}_{WB}q_i^k [\tilde{a}_k - b_i^a - \eta_k^{ad} + ({}_{WB}q_i^k)^T g_w]dt^2]
\end{aligned} \quad (5)$$

Where the left side of the equation represents relative motion increment of inertial coordinate system relative to the world coordinate system after iterative integration of IMU measurement data independent of k at $[k, k+1]$; ω_i represents the true angular velocity measurement of IMU ith sampling. $\tilde{a}_i$ represents the true acceleration measurement value of ith IMU sampling, g_W represents gravity vector, b_i^g and b_i^a represent the gyroscope bias and acceleration bias, respectively, while η^{gd} and η^{ad} represent discretized noise, and Equation 5 is called IMU pre-integration.

The error item of visual and IMU

The main component of the optimization objective function is the error item. For back-end pose optimization algorithm,

it is necessary to construct visual and inertial error item as an overall optimization objective function to realize pose optimization process. The visual part is based on principle of gray scale invariant. Consistent with principle of front-end visual pose estimation, the photometric error is constructed for pixel point c_i on the kth frame image and point c_j on the $k+1$ th frame after projection as shown in Equation 6:

$$\begin{aligned} e_{k,k+1}^{c} &= I_k(c_i) - I_{k+1}(Kc_j) \\ &= I_k(c_i) - I_{k+1}[K\cdot(1+k_1r^2+k_2r^4)\rho_j(\rho_i^{-1}\exp(\xi_{k,k+1}^{\wedge})K^{-1}c_i)] \end{aligned} \tag{6}$$

Where $\xi_{k,k+1}$ represents Lie algebraic form of pose transformation from kth frame to $k+1$ th frame, and $(\cdot)^{\wedge}$ represents vector to matrix operation of transforming six-dimensional vector $\xi_{k,k+1}$ into a four-dimensional matrix. The maximum *a posteriori* estimate of Equation 6, and then conversion to least squares problem is as shown in Equation 7:

$$\sum_{i,j\in \mathrm{I}} \left\| r_{k,k+1}^{c} \right\|_{\Sigma_{\mathrm{I}}}^{2} = \arg\min_{x} \sum_{k} \sum_{i,j\in \mathrm{I}} \left(e_{k,k+1}^{c} \right)^{2} \tag{7}$$

The IMU module is used to pre-integrate the ideal measured value with the actual measured value to construct the error item as shown in Equation 8:

$$e_{k,k+1}^{B} = \begin{bmatrix} \Delta q_{k+1}^{k} \\ \Delta v_{k+1}^{k} \\ \Delta p_{k+1}^{k} \\ \Delta \mathrm{b}_{k,k+1}^{g} \\ \Delta \mathrm{b}_{k,k+1}^{a} \end{bmatrix} = \begin{bmatrix} \prod_{i=k}^{k+1} Exp(-\tilde{q}_{k+1}^{i\ T} J_i^l \eta_i^{gd} dt)^{T} \\ \sum_{i=k}^{k+1} [\tilde{q}_i^k \eta_i^{ad} dt - \tilde{q}_i^k (-\tilde{a}_i - \mathrm{b}_i^a + (q_i^k)^T g_w)^{\wedge} \delta q_\alpha dt] \\ \sum_{i=k}^{k+1} [\frac{1}{2}\tilde{q}_i^k \eta_k^{ad} dt^2 + \delta v_\beta dt - \frac{1}{2}\tilde{q}_i^k (\tilde{a}_k - \mathrm{b}_i^a + (q_i^k)^T g_w)^{\wedge} \delta q_\alpha dt^2] \\ \Delta \frac{\partial p_{k+1}^k}{\partial b_g} \\ \Delta \frac{\partial p_{k+1}^k}{\partial b_a} \end{bmatrix} \tag{8}$$

Where $e_{k,k+1}^{B}$ is residual item for pre-integration of IMU sample values in $[k,k+1]$ interval, including rotation difference Δq_{k+1}^{k}, velocity difference Δv_{k+1}^{k}, displacement difference Δp_{k+1}^{k}, and gyroscope bias difference $\Delta \mathrm{b}_{k,k+1}^{g}$, and accelerometer bias difference $\Delta \mathrm{b}_{k,k+1}^{a}$. $\tilde{q}_i^k$ represents actual pre-integrated value of IMU sampled value from time k to time i. In addition, b_g, b_a are also updated during pre-integration of IMU measurements. This paper assumes that bias has a small amount of δb_g, δb_a accumulate bias b_g and b_a at current sampling during iterative process of pre-integration, and updates the bias b_g, b_a in the pre-integrated ${}_{WB}q_{k+1}^{k}$, ${}_{WB}v_{k+1}^{k}$ and ${}_{WB}p_{k+1}^{k}$. The relative translation increment ${}_{WB}p_{k+1}^{k}$ is used to construct the error item for bias b_g, b_a using $\partial\, {}_{WB}p_{k+1}^{k} / \partial b_g$ and $\partial\, {}_{WB}p_{k+1}^{k} / \partial b_a$ obtained by Taylor first-order approximation at deviation. A maximum *posteriori* estimate of IMU module error item converted to a least square problem was performed as shown in Equation 9:

$$\sum_{k\in B} \left\| r_{k,k+1}^{IMU} \right\|_{\Sigma_B}^{2} = \arg\min_{x} \sum_{k} e_{k,k+1}^{B\ T} \Sigma_{k,k+1}^{-1} e_{k,k+1}^{B} \tag{9}$$

Equation 9 is the IMU inertia component that makes up the objective function, where $\Sigma_{k,k+1}$ is covariance matrix for noise item in IMU pre-integral error item. It is assumed that noise item after pre-integration is as shown in Equation 10:

$$\eta_{k,k+1}^{\mathrm{B}} = \begin{bmatrix} \delta\phi_{k,k+1}^{T} & \delta v_{k,k+1}^{T} & \delta p_{k,k+1}^{T} \end{bmatrix}^{T} \tag{10}$$

Where the noises are composed of variables such as IMU module inherent noise, η_g, η_a and pre-integration noise $\delta\phi_{k,i}$, $\delta v_{k,i}$, which are non-linearly related to the inherent noise. Since inherent noise obeys Gaussian distribution, firstly, each term is first-order approximation, while the distribution of each noise component of noise $\eta_{k,k+1}^{\mathrm{B}}$ can be linearly obtained; thereafter, the components $\delta\phi_{k,k+1}$, $\delta v_{k,k+1}$, $\delta p_{k,k+1}$ are linearly expanded, and the noise of $\eta_{k,k+1}^{\mathrm{B}}$ is composed as shown in Equation 11:

$$\eta_{k,k+1}^{B} = A_{k,i+n-1}\eta_{k,i+n-1}^{B} + B_{i+n-1}\eta_{i+n-1}^{d} \tag{11}$$

Here, we can abbreviate Equation 10 as Equation 12 given as:

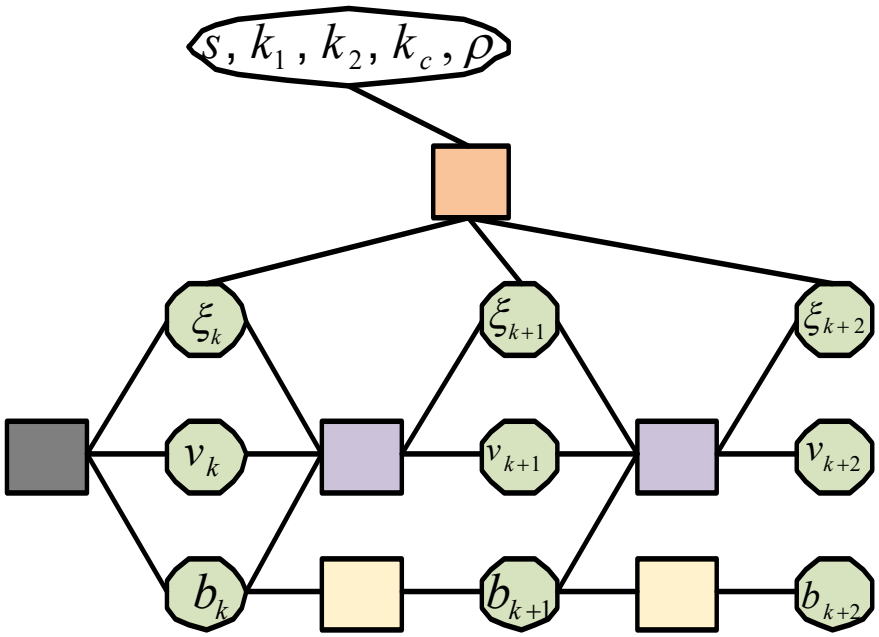

Figure 3: *Apriori* information for the factorization of the optimization process.

$$\eta^B_{k,k+1} = \begin{bmatrix} \tilde{q}^{i+n-1T}_{k+1} & 0 & 0 \\ -\tilde{q}^{k}_{i+n-1}(\tilde{a}_{i+n-1} - \mathrm{b}^a_{i+n-1} + q^{k\ T}_{i+n} g_w)^{\wedge} dt & I & 0 \\ -\frac{1}{2}\tilde{q}^{k}_{i+n-1}(\tilde{a}_{i+n-1} - \mathrm{b}^a_{i+n-1} + q^{k\ T}_{i+n} g_w)^{\wedge} dt^2 & Idt & I \end{bmatrix} \eta^B_{k,i+n-1} + \begin{bmatrix} J^{i+n-1}_l dt & 0 \\ 0 & \tilde{q}^{k}_{i+n-1} dt \\ 0 & \frac{1}{2}\tilde{q}^{k}_{i+n-1} dt^2 \end{bmatrix} \eta^d_{i+n-1} \tag{12}$$

According to the distribution of $\eta^B_{k,i+n-1}$ and η^d_{i+n-1}, the distribution of pre-integrated total noise can be obtained, while the IMU pre-integration matrix is Equation 13 given as:

$$\sum_{k,k+1} = A_{k,i+n-1} \sum_{k,i+n-1} A^T_{k,i+n-1} + B_{i+n-1} \Sigma_\eta B^T_{i+n-1} \tag{13}$$

Where $\sum_{k,i+n-1}$ is matrix pre-integrated at the previous moment, and Σ_η is the matrix of inherent noise of IMU module.

Optimization of objective function

After the analysis of error and optimization item of visual and IMU modules, objective function of back-end pose optimization can be defined as Equation 14:

$$\theta^* = \arg\min_x \frac{1}{2}\left\| f(x) \right\|^2_2 = \arg\min_x (\left\| r_0 \right\|^2_{\Sigma_0} + \sum_{i,j \in \mathrm{I}} \left\| r^c_{k,k+1} \right\|^2_{\Sigma_\mathrm{I}} + \sum_{k \in B} \left\| r^{IMU}_{k,k+1} \right\|^2_{\Sigma_B}) \tag{14}$$

The first item of optimization objective function is *a priori* information, which is related to optimization state variable obtained after back-end pose optimization at previous moment. In back-end sliding window pose optimization process, fourth frame optimization information in the window is selected as *apriori*. Figure 3 shows the optimization factor for adding *apriori* information.

The gray squares represent prior factors, which are used to provide optimization information for optimization at the kth time. It is derived from results of IMU and visual co-optimization at previous moment; the purple square represents IMU pre-integration factor, while the data between two moments is pre-integrated and the yellow square represents bias factor; it should be noted that bias is updated after pre-integration ends. The orange squares represent multi-parameters such as scale factor, internal reference, distortion and inverse depth, and visual factors for joint optimization of pose information at each moment. Figure 3 shows the association between variables in back-end optimization.

Based on the definition of optimization state variable of Equation 2 and construction of visual and IMU module error items, optimization of objective function can be

refined to solve the optimization variables of visual error and IMU module error, respectively. For each variable in visual error term of Equation (6), Jacobian matrix of visual part is obtained as shown in Equation 15:

$$J_c^k = \frac{\partial e_{k,k+1}^c}{x_{k+1}} = \begin{bmatrix} \frac{\partial e_{k,k+1}^c}{\partial q_{k+1}^k} \\ \frac{\partial e_{k,k+1}^c}{\partial v_{k+1}^k} \\ \frac{\partial e_{k,k+1}^c}{\partial p_{k+1}^k} \\ \frac{\partial e_{k,k+1}^c}{\partial b_g} \\ \frac{\partial e_{k,k+1}^c}{\partial b_a} \end{bmatrix} = \begin{bmatrix} M \cdot [\rho_j^{-1} u - p_{k+1}^k]^\wedge \\ 0 \\ M \\ 0 \\ 0 \end{bmatrix} \tag{15}$$

For each component in the global optimization state variable x, the partial derivative is given as Equation 16:

$$J_c^{k_c} = \begin{bmatrix} uc + c(q_{31}u - q_{11}) \\ vc + c(q_{32}v - q_{22}) \\ 1 + (uq_{31} - q_{11}) \cdot \frac{\rho_j}{\rho_i} c \\ 1 + (vq_{32} - q_{22}) \cdot \frac{\rho_j}{\rho_i} c \end{bmatrix} J_c^\rho = c \begin{bmatrix} f_x (p_1 - up_3) \cdot \frac{\rho_j}{\rho_i} \\ f_y (p_2 - vp_3) \cdot \frac{\rho_j}{\rho_i} \end{bmatrix} J_c^{k_{1,2}} = \begin{bmatrix} f_x ur^2 & f_x ur^4 \\ f_y vr^2 & f_y vr^4 \end{bmatrix} \tag{16}$$

Equation 16 represents Jacobian matrix of visual error for camera internal parameter, inverse depth of map point, and distortion. The residual $e_{k,k+1}^B$ of IMU module involves parameters such as components and bias of pre-integration, and there are variables at two moments, as such it is necessary to separately obtain Jacobian matrix of each optimized state variable for each moment as given in Equation 17:

$$J_B^{k,k+1} = \begin{bmatrix} -J_l^{-1}(\Delta q_{k+1}^k) \hat{q}_{k,k+1}^T q_i^T & 0 & 0 & \frac{\partial \Delta q_{k,k+1}}{\partial b_k^g} & 0 & J_l^{-1}(\Delta q_{k,k+1}) & 0 & 0 \\ [q_k^T (v_{k+1} - v_k)]^\wedge & -q_k^T & 0 & -\frac{\partial v_{k,k+1}}{\partial b_k^g} & -\frac{\partial v_{k,k+1}}{\partial b_k^a} & 0 & q_k^T & 0 \\ [q_k^T (p_{k+1} - p_k - v_k dt)]^\wedge & -q_k^T dt & -I & -\frac{\partial p_{k,k+1}}{\partial b_k^g} & -\frac{\partial p_{k,k+1}}{\partial b_k^a} & 0 & 0 & q_k^T q_{k+1} \\ 0 & 0 & 0 & I & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & I & 0 & 0 & 0 \end{bmatrix} \tag{17}$$

In Equation 17, each row represents partial derivative of $\Delta q_{k+1}^k, \Delta v_{k+1}^k, \Delta p_{k+1}^k, \Delta b_{k,k+1}^g, \Delta b_{k,k+1}^a$ for each variable in a single optimized state x_k. Since bias is not updated in interval $[k, k+1]$, only Jacobian matrix with respect to deviation term at time k is obtained. However, when IMU performs pre-integration in interval $[k, k+1]$, IMU bias item also accumulates bias b_g, b_a at current sampling in form of small quantities $\delta b_g, \delta b_a$, hence, $\Delta q_{k+1}^k, \Delta v_{k+1}^k, \Delta p_{k+1}^k$ in Equation (8) are used to obtain the partial derivative of the bias item (Timothy, 2017), which is updated after pre-integration and given as Equation 18:

$$\begin{aligned} &\frac{\partial \Delta q_{k,k+1}}{\partial b_k^g} = -J_l^{-1}(\Delta q_{k+1}^k) Exp[-\frac{\partial \bar{q}_{k,k+1}}{\partial b_k^g} \delta b_k^g] J_l (\frac{\partial \bar{q}_{k,k+1}}{\partial b_k^g} \delta b_i^g) \cdot \frac{\partial \bar{q}_{k,k+1}}{\partial b_k^g} \\ &\frac{\partial v_{k,k+1}}{\partial b_k^g} = \sum_{i=k}^{k+1} [-\bar{q}_i^k (\tilde{a}_i - b_i^a + (\bar{q}_i^k)^T g_w)^\wedge \frac{\partial \bar{q}_{k,k+1}}{\partial b_k^g} dt] \\ &\frac{\partial v_{k,k+1}}{\partial b_k^a} = \sum_{i=k}^{k+1} (-\bar{q}_i^k dt) \\ &\frac{\partial p_{k,k+1}}{\partial b_k^g} = \sum_{i=k}^{k+1} [\frac{\partial \bar{v}_{ki}}{\partial b_k^g} dt - \frac{1}{2} \bar{q}_i^k (\tilde{a}_k - b_i^a + (\bar{q}_i^k)^T g_w)^\wedge \frac{\partial \bar{q}_{k,k+1}}{\partial b_k^g} dt^2] \\ &\frac{\partial p_{k,k+1}}{\partial b_k^a} = \sum_{i=k}^{k+1} (\frac{\partial \bar{v}_{ki}}{\partial b_k^a} dt - \frac{1}{2} \bar{q}_i^k dt^2) \end{aligned} \tag{18}$$

After finding Jacobian matrix of each variable in Equation (2) for the visual and inertial error items in Equation (14), Jacobian matrix was constructed as a global Jacobian matrix J using the method described in visual inertia method (Leutenegger et al., 2014). For J, Equation (2) global state variable x and the Equation (14) objective function $f(x)$ are solved using Gauss-Newton method $J^T J \ x = -J^T f(x)$, and update value x for global state variable x is obtained. Iteratively updates each variable in Equation (2) with update value: $x_{update} \leftarrow x_{cur} + x$, updated x_{update} contains the pose solution of current time.

EXPERIMENTAL RESULTS AND ANALYSIS

Accuracy analysis definition

In accuracy analysis of the VI-DSO, the trajectory estimation values before and after fusing with IMU was compared. The high-precision solution is achieved since the output pose represents trajectory information, when error between trajectory of pose and ground-truth value is small. The error description mainly adopts two kinds of indicators: absolute trajectory error and relative pose error.

For absolute trajectory error (ATE), also known as absolute pose error (APE), the principle is to estimate the absolute distance between estimated value and ground-truth value. The two values are aligned to the same standard and the deviation is calculated as shown in Equation 19:

$$F_i = Q_i^{-1} P_i \tag{19}$$

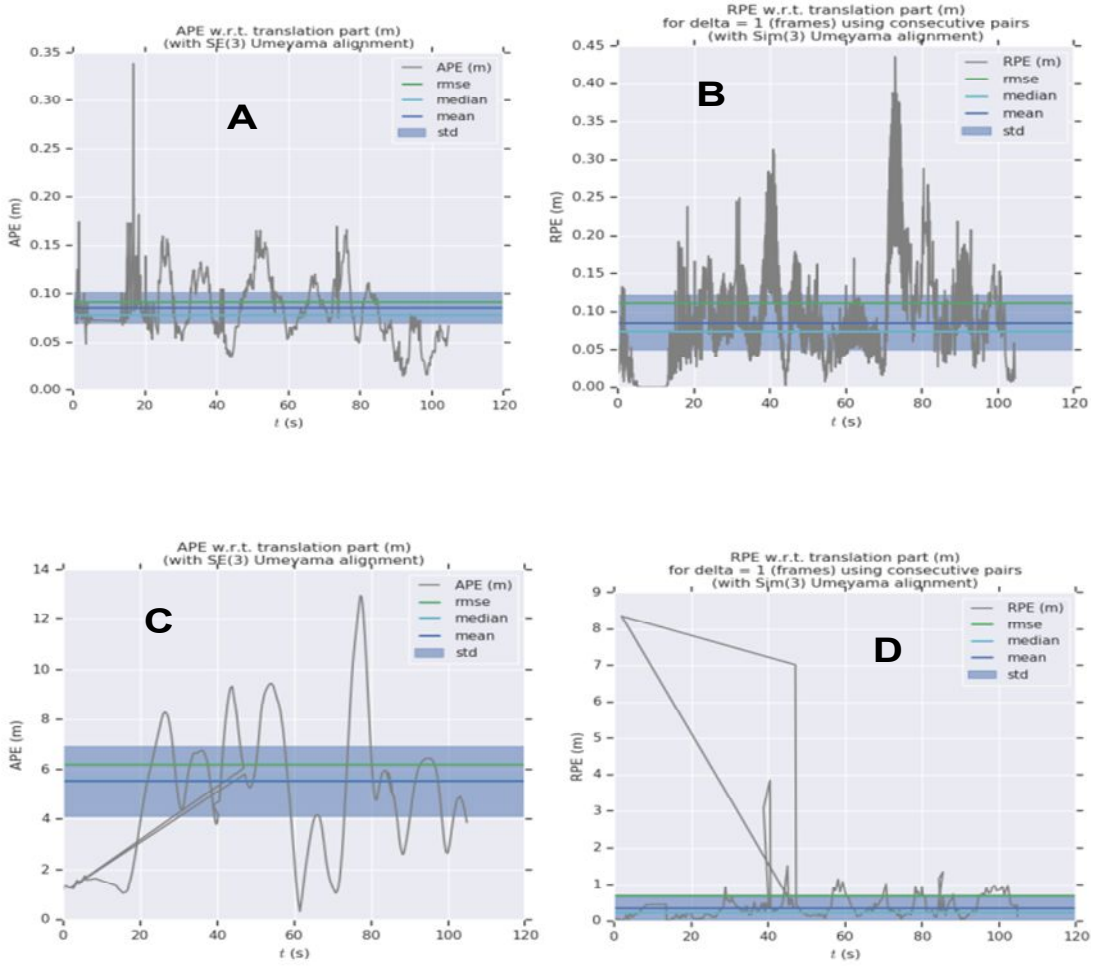

Figure 4: APE and RPE of pose estimation and true value

Where F_i denotes the overall absolute trajectory error corresponding to time i, and P_i denotes pose information optimized by the back-end of the algorithm, which consists of corresponding rotation q_i and translation p_i in Equation (1). Q_i represents ground-truth pose of the data set corresponding to P_i. At the same time, the root mean square error at all times can be calculated according to Equation 20 given as:

$$RMSE(F_{1:n}) = \left(\frac{1}{n}\sum_{i=1}^{n}\left\|trans(F_i)\right\|^2\right)^{\frac{1}{2}} \tag{20}$$

Relative pose error (RPE) does not emphasize on overall consistency of APE, but focuses on local extent difference between estimated value and ground-truth. Hence, the RPE can be defined as shown in Equation 21:

$$E_i = (Q_i^{-1}Q_{i+\Delta})^{-1}(P_i^{-1}P_{i+\Delta}) \tag{21}$$

Where Δ is size of local range alignment, and default value is 1. According to E_i we can also calculate root mean square error of RPE as shown in Equation 22:

$$RMSE(E_{1:n},\Delta) := \left(\frac{1}{n}\sum_{i=1}^{n}\left\|trans(F_i)\right\|^2\right)^{\frac{1}{2}} \tag{22}$$

Comparison of pose resolution accuracy

Taking MH_05 of EuRoc data set (Burri et al., 2016) as an example, we carry on accuracy analysis of APE and RPE based on pose estimation of the VI-DSO. As shown in Figure 4, title represents the alignment standard of the two trajectory, while the horizontal axis represents time stamp. Figure 4a shows APE between the ground-truth and estimated value of the VI-DSO. The curve in Figure 4a is connection of absolute trajectory error between estimated pose of VI-DSO and ground-truth under different time stamps. It can be observed that overall fluctuation of error value tends to be stable and error decreases as time stamp

Table 1: VIO and VO error statistics of MH_05.

MH_05	VIO_APE	VIO_RPE	VO_APE	VO_RPE
Max	0.338533	0.435904	12.960439	8.357953
Mean	0.085740	0.085652	5.536295	0.343683
Median	0.078121	0.074691	5.581579	0.224953
Min	0.14227	0.000021	0.320344	0.005898
RMSE	0.091455	0.111598	6.200442	0.704371

increases. After time stamp is 80+, error value is lower than the root mean square error value corresponding to green line; Figure 4b shows RPE of both curves.

Figure 4b shows that the curve is connection of relative trajectory error value between ground-truth and pose estimation of VI-DSO under different time-stamps. The relative trajectory error indicates error difference between adjacent frame pose, and corresponding difference is larger at horizontal axis 70 +, but overall inter-frame pose difference tends to be stable, while overall RPE root mean square error of green line is 0.11^{+}, and the error is small. Figure 4(c) is an APE error curve between ground-truth and estimated value of DSO without fusion of IMU. It can be seen that compared with curve of Figure 4(a), absolute trajectory error of ordinate corresponding to APE curve before fusing IMU is much higher than APE value of VI-DSO, and error is higher; Figure 4(d) is RPE error curve of both VI-DSO. The overall RPE root mean square error of green line is 0.6+. Compared with 0.11+ in Figure 4(b), RPE relative trajectory error before fusion IMU is much larger than the RPE value of the VI-DSO, and curve RPE has obvious hopping on horizontal axis 40+, hence, local estimation of pose is unstable.

Table 1 gives details of data of the APE and RPE accuracy comparison between ground-truth and estimated pose of fused IMU. The VIO_APE column corresponds to Figure 4(a), the VIO_RPE column corresponds to Figure 4(b), while the VO_APE column corresponds to Figure 4(c), and VO_RPE column corresponds to Figure 4(d). The RMSE item in Table 1 corresponds to green line in each figure in Figure 4 and is the main indicator for measuring accuracy of pose estimation. The conclusions drawn from Figure 4 and Table 1 are as follows:

(1) The RMSE of APE in algorithm of this paper is 0.091455, which corresponds to the APE value of green line of Figure 4(a). The RMSE of APE of monocular visual odometry DSO is 6.200442, which corresponds to the APE value of green line in Figure 4(c). The overall fluctuation of error in Figure 4(a) tends to be stable and error value decreases as time stamp increases, while absolute trajectory error of ordinate corresponding to APE curve before fusion of IMU in Figure 4(c) is much higher. Therefore, algorithm in this paper has higher global accuracy than DSO before fusion IMU.

(2) The RMSE of RPE of algorithm in this paper is 0.111598, corresponding to the RPE value of green line in Figure 4(b), and RMSE of RPE of DSO is 0.890695, corresponding to the RPE value of green line in Figure 4(d). For local trajectory comparison of RPE, we chose $\Delta = 1$, which is RPE analysis between adjacent frames. The overall fluctuation of RPE curve in Figure 4(b) indicates that inter-frame pose difference tends to be stable, while the relative trajectory error before fusion of IMU in Figure 4(d) is much larger than the RPE value of algorithm in Figure 4(b), and there is a significant jump in RPE at horizontal axis 40+, indicating a large drift, and local estimation of pose is unstable. Therefore, VI-DSO in this paper has a higher local precision and a smaller drift after the IMU is integrated.

The high precision of the algorithm after fusion of IMU is proved through accuracy analysis between estimated values of experimental results and ground-truth. The trajectory map and the three components changing on different coordinate axes were subsequently discussed. Figure 5 shows the corresponding trajectory given in Table 1 of true and error values.

The integrated trajectory diagram of Figure 5(a) consists of pose trajectories of the VI-DSO, ground-truth and monocular pose trajectory. They are aligned with the same standard; the dashed line indicates ground-truth, while the blue line represents experimental pose trajectory of the VI-DSO, and green line represents the monocular pose trajectory. It can be clearly seen that there is a scale difference of the trajectories between without fusion of the IMU and the trajectory drawn by the VI-DSO. The trace drawn by the VI-DSO can be highly consistent with the true value. In addition, Figure 5(b) and (c) represents error corresponding to the APE and RPE of the algorithm respectively, and Figure 5(d) and (e) respectively represent detailed description of error corresponding to APE and RPE of monocular trajectory. The color is used to mark the distribution of error magnitude when calculating error between estimation pose trajectory and true value. From the display of error in Figure 5(b), (c), (d) and (e), it can be concluded that error of the VI-DSO is much smaller than DSO. Moreover, there is a section of the trajectory marked by deep red in Figure 5(e), which shows that there is a significant outlier at time axis 40+ in Figure 5(d). Combining the difference between RMSE and Max in the VO_RPE term in Table 1, it can be seen that estimation error of pose in monocular odometry DSO is extremely large.

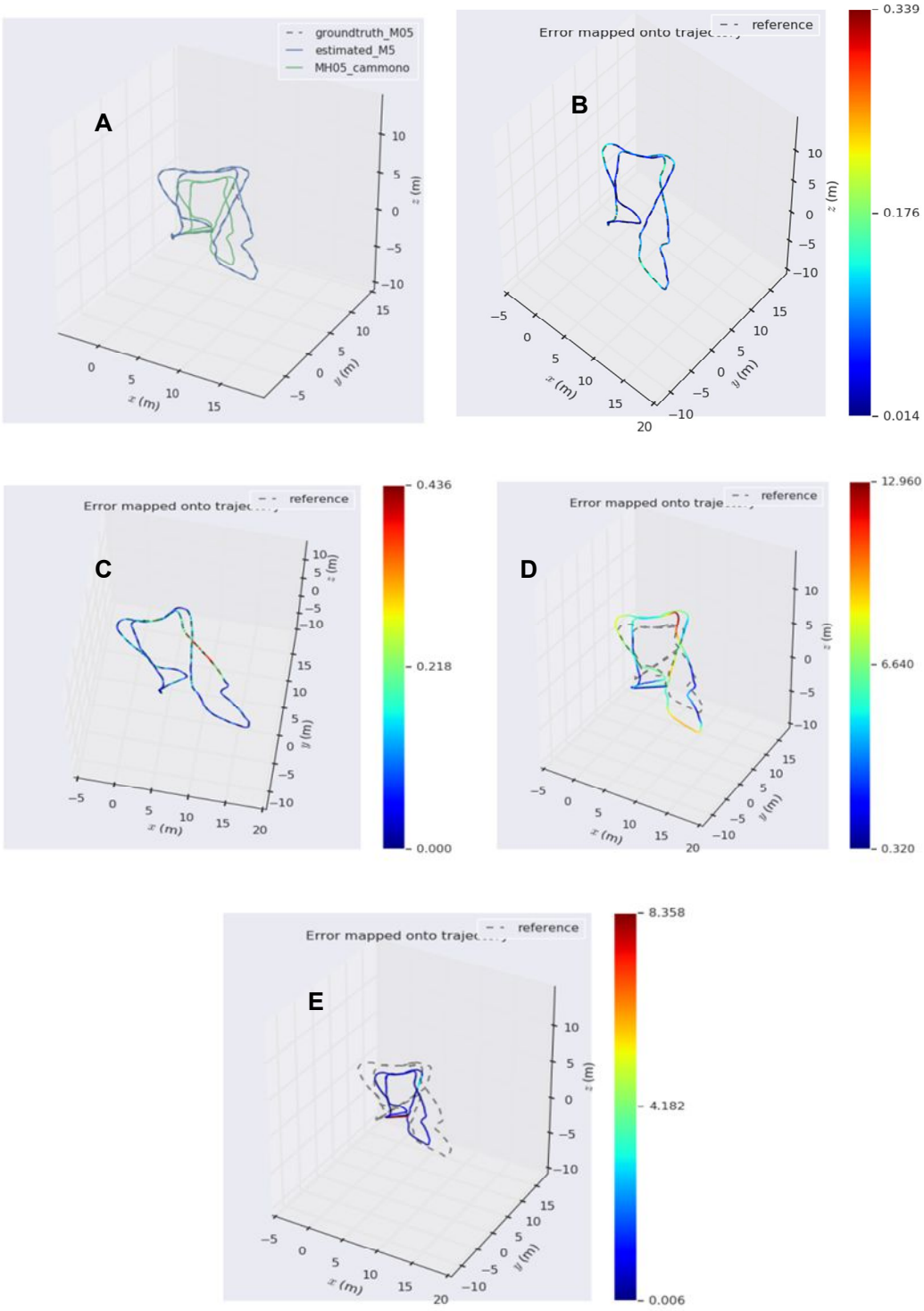

Figure 5: The trajectory and error detail of before and after fusion IMU.

Figure 6 shows the moving component diagram on space coordinate axis. Figure 6(a) shows the fitting curve of translation of MH_05 data set and pose estimation of the algorithm on the three axes x, y and z. The dotted line represents ground-truth, and solid blue line represents experimental value of the VI-DSO. As can be seen from the figure, the two are highly fitted. Figure 6(b) is a translation component drift of the algorithm and VO before the fusion of IMU on three axes of x, y and z. The VI-DSO is not only superior in magnitude to the monocular VO before fusion of IMU, but also tends to be stable in numerical variation of translation component; therefore, the high precision of

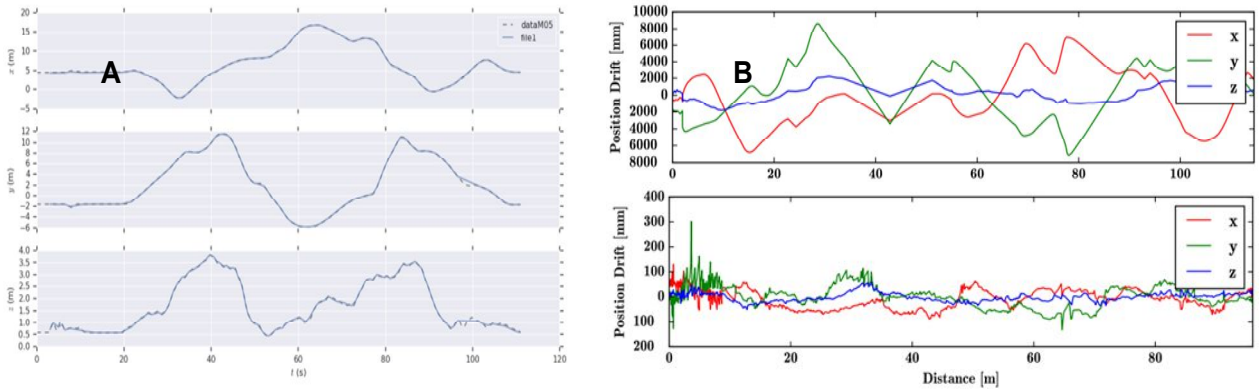

Figure 6: The translation component of the VIO\VO odometry and true value on each axis.

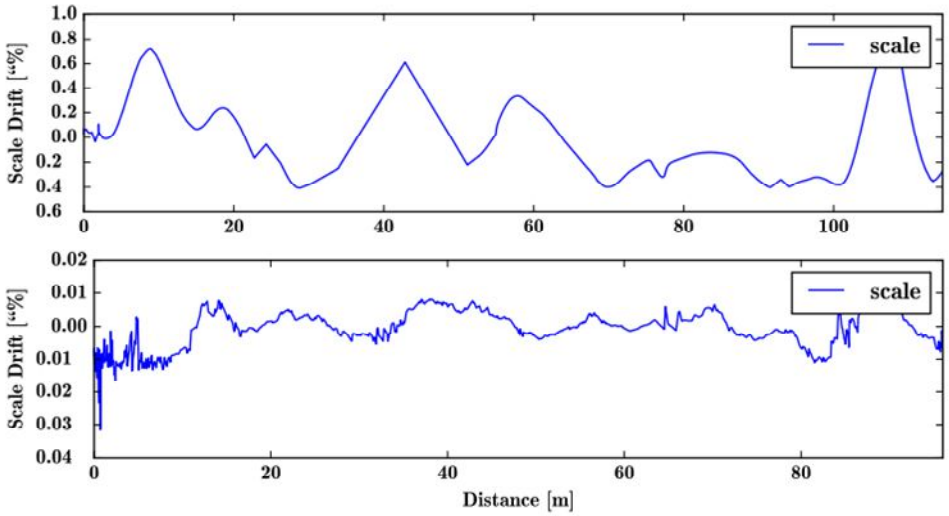

Figure 7: Drift scale of VIO\VO odometry (Top) and ground-truth (Bottom).

experimental data of the VI-DSO can be obtained from translation component.

The proposed method aims at jointly optimizing the pose of visual and inertial with scale factor. Figure 7 shows the scale factor drift between experimental data of the VI-DSO, the ground-truth and value before fusion of IMU. It can be seen from the figure that drift of the VI-DSO on scale is much smaller than that of monocular DSO before fusion of IMU, and after the VI-DSO designs scale factor and estimates it, the numerical fluctuation of scale tends to be stable. Therefore, it can be explained that the optimization of the scale factor can improve the uncertainty of the monocular scale.

Finally, the VI-DSO is compared with mainstream visual inertia method Okvis for APE. The comparison process is as a result of a data set for MH_05. Thereafter, the data sets in official EuRoC will be compared, and various APE indicators of Okvis obtained according to the report of Kasyanov et al. (2017). In the proposed method, all pixels participate in constructing error items which eliminate the computational burden of extracting feature points and matching them in Okvis. Table 2 shows the specific data. The APE data in Table 2 is a value of RMSE. As can be seen, the proposed method has a smaller RMSE on each data set than Okvis method. Based on the analysis of the experimental data, our algorithm has high precision compared with monocular DSO before IMU fusion in terms of the two major precision analysis indicators: trajectory of Figure 5, and translation variable of Figure 6.

The map constructed by our algorithm is compared with the map constructed by the DSO before fusion IMU. Figure 8 shows the constructed map. Figure 8(a) is a map constructed by a monocular without IMU, while Figure 8(b) is a map constructed by our algorithm. It can be seen from the figure that after the fusion of IMU, the constructed point cloud map is more regular. The point cloud built on walls, windows, table tops and floors are clearer and show better details.

Table 2: Accuracy comparison table on each data set.

EuRoC dataset	Degree of difficulty	VIO_APE	Okvis_APE
MH_01	Easy	0.072	0.34
MH_02		0.063	0.36
MH_03	Medium	0.095	0.30
MH_04	Difficult	0.151	0.48
MH_05		0.091	0.47
V1_01	Easy	0.067	0.12
V1_02	Medium	0.071	0.16
V1_03	Difficult	0.097	0.24
V2_01	Easy	0.056	0.12
V2_02	Medium	0.084	0.22

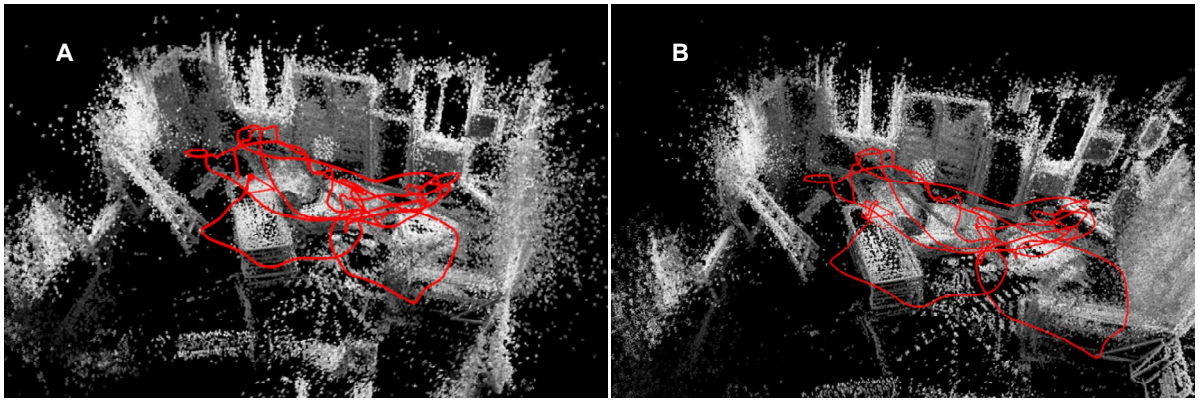

Figure 8: Map of the VIO\VO odometry.

CONCLUSION

This paper proposes a visual-inertial odometry that combines monocular DSO and IMU. The distortion parameters are added to initial pose estimation with front-end visual frame to construct visual photometric error function and solve the visual initial pose. The global optimization state variables of the algorithm are constructed to reduce the error accumulation of each variable during optimization. In the back-end pose optimization, based on the tightly coupled multi-parameter sliding window pose optimization algorithm, the scale factor is calculated to make the visual and inertia maintain the scale consistency in pose information. *A priori* information corresponding to specific frame is selected, and overall optimization objective function of prior information, visual error and inertia error function is constructed. The Jacobian matrix of global state variable is obtained for objective function, while the high-precision pose solution at the current moment is completed by updating global state variables of optimized solution iteration. Hence, the proposed method not only takes pose and deviation as solving variables, but also takes camera internal parameters, distortion and depth information of the points involved in the optimization as optimization variables. These variables are updated iteratively so that the aforementioned variables can be used in the next process of solving pose.

Finally, the accuracy comparison between the experimental data of our algorithm and the DSO before the fusion IMU and the current mainstream visual inertia method Okvis and VIO is carried out, and the improvement of the accuracy of our algorithm in constructing the point cloud map is explained. The experimental results show that the proposed method has high-precision pose calculation results after fusion of IMU.

REFERENCES

Benjamin U, Huizhong Z, Jonas U, Nikolaus M, Eddy I, Alexey D, Thomas B

(2017). "DeMoN: Depth and Motion Network for Learning Monocular Stereo", in CVPR 2017
Bloesch M, Omari S, Hutter M, et al (2015). Robust visual inertial odometry using a direct EKF-based approach. pp. 298-304.
Burri M, Nikolic J, Gohl P, et al (2016). The EuRoC micro aerial vehicle datasets. The International Journal of Robotics Research.
Cadena C, Carlone L, Carrillo H, et al (2016). Past, Present, and Future of Simultaneous Localization and Mapping: Toward the Robust-Perception Age[J]. IEEE Transactions on Robotics. 32(6):1309-1332.
Engel J, Koltun V, Cremers D (2017). Direct Sparse Odometry[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence. 40(3): 611-625.
Engel J, Sch T, Cremers D (2014). LSD-SLAM: large-scale direct monocular SLAM. Springer International Publishing.
Forster C, Carlone L, Dellaert F, et al (2015). On-Manifold Preintegration for Real-Time Visual-Inertial Odometry[J]. IEEE Transactions on Robotics. 33(1): 1-21.
Forster C, Pizzoli M, Davide S (2014). SVO: Fast Semi-Direct Monocular Visual Odometry. Hong Kong.
Gui J, Gu D, Wang S, et al (2015). A review of visual inertial odometry from filtering and optimisation perspectives. Advanced Robotics.
Kasyanov A, Engelmann F, Stückler, Jörg et al (2017). Keyframe-Based Visual-Inertial Online SLAM with Relocalization.
Keisuke T, Federico T, Iro L, Nassir N (2017). CNN-SLAM: Real-time dense monocular SLAM with learned depth prediction", in CVPR 2017
Klein G (2007). Parallel tracking and mapping for small AR workspaces.
Kümmerle R, Grisetti G, Strasdat H, et al (2011). G2o: A general framework for graph optimization.
Kümmerle R, Grisetti G, Strasdat H, et al (2011). G2o: A general framework for graph optimization[C]// IEEE International Conference on Robotics and Automation. IEEE, e43478.
Kwang MY, Eduard T, Vincent L, Pascal F (2016). "LIFT: Learned Invariant Feature Transform", in ECCV.
Leutenegger S, Lynen S, Bosse M, et al (2014). Keyframe-Based Visual-Inertial Odometry Using Nonlinear Optimization. The International Journal of Robotics Research. 34(3): 314-334.
Martinelli A (2018). Closed-form solution to cooperative visual-inertial structure from motion.
Mur-Artal R, Montiel JMM, Tardos JD (2015). ORB-SLAM: a Versatile and Accurate Monocular SLAM System[J]. IEEE Transactions on Robotics. 31(5): 1147-1163.
Mur-Artal R, Montiel JMM, Tardós JD (2015). ORB-SLAM: A Versatile and Accurate Monocular SLAM System. IEEE Transactions on Robotics. 31(5):1147-1163.
Ranftl R, Vineet V, Chen Q, Koltun V (2016). Dense monocular depth estimation in complex dynamic scenes. In International Conference on Computer Vision and Pattern Recognition (CVPR).
Rublee E , Rabaud V , Konolige K , et al (2011). ORB: An efficient alternative to SIFT or SURF[J]. 2011.
Silveira G, Malis E , Rives P (2008). An Efficient Direct Approach to Visual SLAM[J]. IEEE Transactions on Robotics. 24(5): 969-979.
Strasdat H (2012). Local Accuracy and Global Consistency for Efficient SLAM. Imperial College London.
Timothy D (2017). Barfoot. State Estimation for Robotics[M]. Cambridge University Press.
Tinghui Z, Matthew B, Noah S, David GL (2017). Unsupervised Learning of Depth and Ego-Motion from Video", in CVPR 2017.
ZHU Chaozheng, HE Ming, YANG Sheng, et al (2018). Survey of monocular Visual Odometry. Computer Engineering and Applications. 54(7): 20-28.