Scale Estimation and Refinement in Monocular Visual-Inertial SLAM System

Xufu Mu, Jing Chen(✉), Zhen Leng, Songnan Lin, and Ningsheng Huang

School of Optoelectronics, Beijing Institute of Technology, Beijing, China
muxufu@163.com, chen74jing29@bit.edu.cn

Abstract. The fusion of monocular visual and inertial cues has become popular in robotics, unmanned vehicle and augmented reality fields. Recent results have shown that optimization-based fusion strategies outperform filtering ones. The visual-inertial ORB-SLAM is optimization-based and has achieved great success. However, it takes all measurements into IMU initialization, which contains outliers, and it lacks of termination criterion. In this paper, we aim to resolve these issues. First, we present an approach to estimate scale, gravity and accelerometer bias together, and regard the estimated gravity as an indication for estimation convergence. Second, we propose a methodology that is able to use weight *w* derived from the robust norm for outliers handling, so that the estimated scale can be refined. We test our approaches with the public EuRoC datasets. Experimental results show that the proposed methods can achieve good scale estimation and refinement.

Keywords: Visual-inertial fusion · Monocular SLAM · Scale estimation

1 Introduction

The combination of vision and inertial sensors has long been a popular research field for three-dimensional structure, ego-motion estimation and visual odometry. Both monocular camera and Inertial Measurement Unit (IMU) are cheap, low-cost, low-weight and complementary. A moving camera can provide us accurate state estimation and sufficient environment 3D structure up to an unknown metric scale. While inertial sensors with high frame-rate can help us handle fast camera motion, scale ambiguity and short-term motion estimation.

Many Visual-inertial fusion strategies have been proposed, which can be divided into the loosely coupled modality and the tightly coupled one. Loosely coupled strategy is to estimate 6D pose and position separately. On the contrary, tightly coupled fusion strategy is to jointly optimize all sensor states. Most recent works concentrate on tightly-coupled visual-inertial odometry, using keyframe-based non-linear optimization [1–4] or filtering [5–8]. Non-linear optimization and tightly coupled methods have attracted much interest of researchers in recent years due to its good trade-off between accuracy and computational efficiency. This article follows this trend and focuses on the monocular unknown scale problem.

Y. Zhao et al. (Eds.): ICIG 2017, Part I, LNCS 10666, pp. 533–544, 2017.
https://doi.org/10.1007/978-3-319-71607-7_47

Visual scale estimation is a research hotspot in the monocular SLAM. The early MonoSLAM [11] initializes from a target of known size, which help to assign a precise scale to the estimated map. Filter-based methods include ROVIO [12], MSCKF [5] and [13, 14], where the scale information is added to the extended Kalman filter as an additional state variable. The paper [15] proposed a maximum-likelihood estimator for the scale of the monocular SLAM system. In [16] and visual-inertial ORB-SLAM [9], the scale is estimated within the process of optimization using methods such as Gauss-Newton. While promising, taking all visual and inertial measurements for scale estimation may contain outliers, which lead to declined accuracy of scale estimation. Besides, the method introduced in [9] is lack of robust termination criterion for IMU initialization, which results in increased computation and reducing the effect of IMU information.

In this paper, we devote to solve above problems existed in [9]. The main contribution of our research work is two-fold. Firstly, we present an approach to estimate scale, gravity and accelerometer bias together, and regard the estimated gravity as an indication for identifying convergence and termination for scale estimation procedure. Secondly, we propose a keyframe-based method that uses a weighted term to reduce the influence of large residuals, which lead to scale estimation refinement.

The remainder of this article is organized as follows. In the main Sect. 2 we explain the camera model, the IMU noise models, and the kinematics models of IMU, we also give a brief introduction about IMU pre-integration technique. In Sect. 3, we describe our approach as a whole, in particular we introduce the method for scale estimation and refinement. We also propose an automatic termination criterion. Section 4 is dedicated to show the performance of our approaches and we compare them with the ground truth. We conclude the paper in Sect. 5.

2 Preliminaries

In this section, we first introduce some notation throughout this paper: the matrix $T_{EF} = [R_{EF} \quad {}_E P_F]$ represents the transformation from reference F to reference E.

Then we will introduce some preliminary knowledge about the coordinate system, the camera model, inertial sensor model, and IMU pre-integration. Figure 1 shows the situation of the camera-IMU setup with its corresponding coordinate frames. Multiple camera-IMU units represent the consecutive states at continuous time, which is convenient for understanding the following Equations in Sect. 3.1. The camera provides the pose and the unscaled position in the camera frame C. We denote the world reference frame with W and the IMU body frame B. The transformation $T_{CB} = [R_{CB} \quad {}_C P_B]$ between camera and IMU reference systems can be calibrated using Kalibr [17].

2.1 Camera Model

Here we consider a conventional pinhole-camera model [22], which any 3D point $X_C \in \mathbb{R}^3$ in the camera reference maps to the image coordinates $x \in \mathbb{R}^2$, through the camera projection function $\pi : \mathbb{R}^3 \mapsto \mathbb{R}^2$:

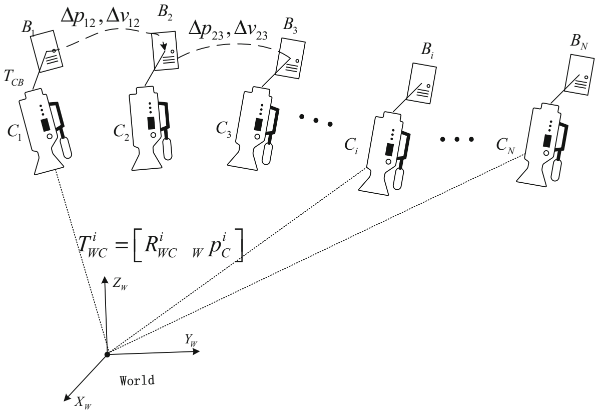

B_1
$\Delta p_{12}, \Delta v_{12}$
B_2
$\Delta p_{23}, \Delta v_{23}$
B_3
B_i
B_N
T_{CB}
C_1
C_2
C_3
C_i
C_N
$T^i_{WC} = \left[R^i_{WC} \quad {}_W p^i_C \right]$
Z_W
Y_W
World
X_W

Fig. 1. The relationship between different coordinate frames and multiple states of camera-IMU

$$\pi(X_C) = \begin{bmatrix} f_u \frac{x_C}{z_C} + c_u \\ f_v \frac{y_C}{z_C} + c_v \end{bmatrix}, \qquad X_C = \begin{bmatrix} x_C & y_C & z_C \end{bmatrix}^T \tag{1}$$

where $\begin{bmatrix} f_u & f_v \end{bmatrix}^T$ is the focal length and $\begin{bmatrix} c_u & c_v \end{bmatrix}^T$ is the principal point.

2.2 Inertial Sensor Model and IMU Kinematics Model

An IMU generally integrates a 3-axis gyroscope sensor and a 3-axis accelerometer sensor, and correspondingly, the measurements provide us the angular velocity and the acceleration of the inertial sensor at high frame-rate with respect to the body frame B. The IMU measurement model contains two kinds of noise: one is white noise n_{t}, the other is random walk noise that is a slowly varying sensor bias b_{t}, so we have:

$${}_B\widetilde{\omega}(t) = {}_B\omega(t) + b_g(t) + n_g(t) \tag{2}$$

$${}_B\widetilde{a}(t) = R^T_{WB}(t)({}_W a(t) - {}_W g) + b_a(t) + n_a(t) \tag{3}$$

where the ${}_B\widetilde{w}(t)$ and ${}_B\widetilde{a}(t)$ are the measured values expressed in the body frame, the real angular velocities ${}_B w(t)$ and the real acceleration ${}_W a(t)$ are what we need. The left subscript W denotes in the world frame. And the R_{WB} is the rotational part from the transformation $\{R_{WB} \; {}_W P\}$, which maps a point from sensor frame B to W. The dynamics of non-static bias b_t are modeled as a random process:

$$\dot{b}_g = n_{b_g} \quad \dot{b}_a = n_{b_a} \tag{4}$$

where the n_{b_g} and n_{b_a} are the zero-mean Gaussian White noises. Our goal is to deduce the motion of system from the output of IMU. For this purpose, we show the following IMU kinematics model [11]:

$$ {}_W\dot{R}_{WB} = R_{WB}\ {}_B\omega^{\wedge},\ {}_W\dot{v} = {}_W a,\ {}_W\dot{p} = {}_W v \tag{5}$$

2.3 IMU Pre-integration

The IMU pre-integration technique incorporated with SLAM framework are proposed correctly in [18]. Here we give an overview of its theory and usage within monocular visual-inertial SLAM system. The pose and velocity of IMU at time $t + \Delta t$ is obtained by integrating Eq. (5):

$$R_{WB}(t+\Delta t) = R_{WB}(t)Exp({}_B\omega(t)\Delta t) \tag{6}$$

$${}_W v(t+\Delta t) = {}_W v(t) + {}_W a(t)\Delta t \tag{7}$$

$${}_W p(t+\Delta t) = {}_W p(t) + {}_W v(t)\Delta t + \frac{1}{2}{}_W a(t)\Delta t^2 \tag{8}$$

which assumes that ${}_W a$ and ${}_B\omega$ maintain a constant in the time interval $[t, t+\Delta t]$. Equations (6)–(8) become function of the IMU measurements using Eqs. (2)–(3):

$$R(t+\Delta t) = R(t)Exp((\tilde{\omega}(t) - b_g(t) - n_g(t))\Delta t) \tag{9}$$

$$v(t+\Delta t) = v(t) + g\Delta t + R(t)(\tilde{a}(t) - b_a(t) - n_a(t))\Delta t \tag{10}$$

$$p(t+\Delta t) = p(t) + v(t)\Delta t + \frac{1}{2}g\Delta t^2 + \frac{1}{2}R(t)(\tilde{a}(t) - b_a(t) - n_a(t))\Delta t^2 \tag{11}$$

Here the coordinate frame subscripts is dropped for readability. In Eqs. (6)–(11) Δt is the sampling interval of the IMU. Assuming that the IMU is synchronized with the camera, and provides measurements at discrete times k. Integrating all IMU measurements between two consecutive keyframes at times $k = i$ and $k = j$, then the IMU pre-integration ΔR_{ij}, Δv_{ij} and Δp_{ij} are expressed as:

$$\Delta R_{ij} \doteq R_i^T R_j \ = \prod_{k=i}^{j-1} Exp((\tilde{\omega}_k - b_{gk} - n_{gk})\Delta t) \tag{12}$$

$$\Delta v_{ij} \doteq R_i^T(v_j - v_i - g\Delta t_{ij}) \ = \sum_{k=i}^{j-1} \Delta R_{ik}(\tilde{a}_k - b_{ak} - n_{ak})\Delta t \tag{13}$$

$$\begin{aligned}\Delta p_{ij} &\doteq R_i^T(p_j - p_i - v_i\Delta t_{ij} - \frac{1}{2}g\Delta t_{ij}^2) \\ &= \sum_{k=i}^{j-1}\left[\Delta v_{ik}\Delta t + \frac{1}{2}\Delta R_{ik}(\tilde{a}_k - b_{ak} - n_{ak})\Delta t^2\right]\end{aligned} \tag{14}$$

3 Scale Estimation and Refinement with a Weighted Item

In this section, we firstly introduce the process of scale estimation based on visual-inertial ORB-SLAM [9]. Since some visual-inertial measurements between two kerframes may not be exact, we propose a weighting method for outliers handling and scale estimation refinement, inspired by [10]. Next, we present a robust termination criterion for scale estimation procedure. At last, we describe the scale benchmark, which can be used to verify the accuracy of our estimated results.

3.1 Scale Estimation

In this section, we introduce the scale estimation method in details, which is able to estimate scale s, gravity ${}_Wg$, accelerometer bias b_a together. The full state vector X is defined as:

$$X = [s, {}_Wg, b_a]^T \in \mathbb{R}^{7\times 1} \tag{15}$$

In the monocular SLAM system, the camera position and 3D points are all up-to-scale. It can be solved by integrating IMU data. First we consider the following equation, which represents that it includes a visual scale s when transforming the position in the camera frame C to the IMU frame B

$$ {}_Wp_B = s{}_Wp_C + R_{WC}\ {}_Cp_B \tag{16}$$

For two consecutive keyframe i and keyframe $i+1$, the corresponding IMU position and velocity are obtained using pre-integration Eqs. (13) and (14):

$$ {}_Wp_B^{i+1} = {}_Wp_B^i + {}_Wv_B^i\Delta t_{i,i+1} + 0.5{}_Wg\Delta t_{i,i+1}^2 + R_{WB}^i(\Delta p_{i,i+1} + J_{\Delta p}^a b_a) \tag{17}$$

$$ {}_Wv_B^{i+1} = {}_Wv_B^i + {}_Wg\Delta t_{i,i+1}^2 + R_{WB}^i(\Delta v_{i,i+1} + J_{\Delta v}^a b_a) \tag{18}$$

where Jacobian $J_{(\cdot)}^a$ denotes a first-order approximation of the effect of changing accelerometer bias. Then taking Eq. (16) into Eq. (17), it becomes:

$$\begin{aligned} s{}_Wp_C^{i+1} = {} & s{}_Wp_C^i + {}_Wv_B^i\Delta t_{i,i+1} + 0.5{}_Wg\Delta t_{i,i+1}^2 + R_{WB}^i(\Delta p_{i,i+1} + J_{\Delta p}^a b_a) + (R_{WC}^i \\ & - R_{WC}^{i+1}){}_Cp_B \end{aligned} \tag{19}$$

To solve this linear system, we consider two relations (19) between three consecutive keyframes (Fig. 1 shows an example) and exploit the velocity relation in (18), we can get the following equations:

$$[\,\alpha(i) \quad \beta(i) \quad \gamma(i)\,]X = \psi(i) \tag{20}$$

where the visual scale s, gravity ${}_Wg$ and acceleration bias b_a are unknown variables. Writing keyframes i, $i+1$, $i+2$ as 1, 2, 3 for readability, we have:

$$\alpha(i) = ({}_{w}p_{c}^{2} - {}_{w}p_{c}^{1})\Delta t_{23} - ({}_{w}p_{c}^{3} - {}_{w}p_{c}^{2})\Delta t_{12} \tag{21}$$

$$\beta(i) = 0.5 I_{3\times 3}(\Delta t_{12}^{2}\Delta t_{23} + \Delta t_{23}^{2}\Delta t_{12}) \tag{22}$$

$$\gamma(i) = R_{WB}^{2} J_{\Delta p_{23}}^{a} \Delta t_{12} + R_{WB}^{1} J_{\Delta v_{12}}^{a} \Delta t_{12}\Delta t_{23} - R_{WB}^{1} J_{\Delta p_{12}}^{a} \Delta t_{23} \tag{23}$$

$$\begin{aligned}\psi(i) = {} & (R_{WC}^{1} - R_{WC}^{2})_{C}p_{B}\Delta t_{23} - (R_{WC}^{2} - R_{WC}^{3})_{C}p_{B}\Delta t_{12} - R_{WB}^{2}\Delta p_{23}\Delta t_{12} \\ & - R_{WB}^{1}\Delta v_{12}\Delta t_{12}\Delta t_{23} + R_{WB}^{1}\Delta p_{12}\Delta t_{23}\end{aligned} \tag{24}$$

Stacking all relations between every three consecutive keyframes using Eq. (20), we can get a linear overdetermined equation groups. Finally, we can solve it via Singular Value Decomposition (SVD) to get the results of the scale s, gravity ${}_{W}g$, accelerometer bias b_a. Note that we can construct $3(N-2)$ equations with 7 unknowns, where N is the number of keyframes, thus we need at least 5 keyframes.

Every time a new keyframe is inserted by ORB-SLAM, the procedure runs to get new estimated values of scale, gravity and accelerometer bias. When the termination criterion is established, the estimation procedure ends up.

3.2 Weighting Method for Scale Estimation Refinement

In the Sect. 3.1, it takes all visual-inertial measurements into the scale estimation procedure, which may contain outliers, so we utilize the weight w_i to handle outliers for estimation refinement. Simply, we exploit the initial values to weight the residual in a similar way to the Huber norm [20], and define the residual as the first moment norm:

$$r_i = |C_i X_{est} - D_i| \tag{25}$$

where X_{est} is the estimated results from Sect. 3.1, C_i and D_i are from Eq. (20) for the i-th consecutive three keyframes, and defined as:

$$C_i = [\,\alpha(i) \quad \beta(i) \quad \gamma(i)\,] \tag{26}$$

$$D_i = [\psi(i)] \tag{27}$$

The weight is associated with the residual.

$$w_i = \begin{cases} 1 & r_i < threshold \\ \frac{threshold}{r_i} & otherise \end{cases} \tag{28}$$

If the measurement is obviously wrong for our scale estimate, its w_i is set to zero. And in our experiments, we set the threshold to 0.002. With the N keyframes in the process of scale estimation, we are able to build an overconstrained linear system as:

$$\begin{bmatrix} w_1 \cdot C_1 \\ w_2 \cdot C_2 \\ \vdots \\ w_{N-2} \cdot C_{N-2} \end{bmatrix} \cdot X = \begin{bmatrix} w_1 \cdot D_1 \\ w_2 \cdot D_2 \\ \vdots \\ w_{N-2} \cdot D_{N-2} \end{bmatrix} \tag{29}$$

where C_i and D_i are from Eqs. (26) and (27) for the i-th consecutive three keyframes. Once we get the Eq. (29), the procedure runs to estimate an updated vector $\hat{X}$ by solving Eq. (29) via SVD.

3.3 Termination Criterion

In this section we propose an automatic criterion to determine when we consider the scale estimate successful. Because the norm of the nominal gravity is a constant $\sim$9.8 m/s^2, we regard it as one convergence indicator. The other is that the difference of consecutive solutions X in Sect. 3.1 is under a certain threshold for several times. The visual scale estimation terminates when both conditions above are established.

3.4 Scale Benchmark

In monocular SLAM system, the translation decomposed from essential matrix is ambiguous up to an unknown scale. To obtain a globally consistent scale factor, visual-inertial ORB-SLAM system initializes mean depth of all the feature points to one. In other words, the real visual scale is determined at the start of the system initialization. Because the first two keyframes selection and the map points generation is random in the ORB-SLAM system initialization, the initial scale is not fixed. For this reason, we need to calculate the actual scale according to the ground truth data, which is extracted by Leica MS50 and motion capture system and provide us the accurate 6D pose in the IMU body reference frame B.

Once the initialization of ORB-SLAM system completes, it outcomes an initial translation t between the first two keyframes. Meanwhile, we can calculate the actual translation ${}_{C_1}p_{C_2}$ according to their corresponding ground truth states. Then the actual scale s is computed by the following formula:

$${}_{B_1}p_{B_2} = R_{B_1C_2}\ {}_{C_2}p_{B_2} + R_{B_1C_1}\ {}_{C_1}p_{C_2} + {}_{B_1}p_{C_1} \tag{30}$$

$$s = {}_{C_1}p_{C_2}/t \tag{31}$$

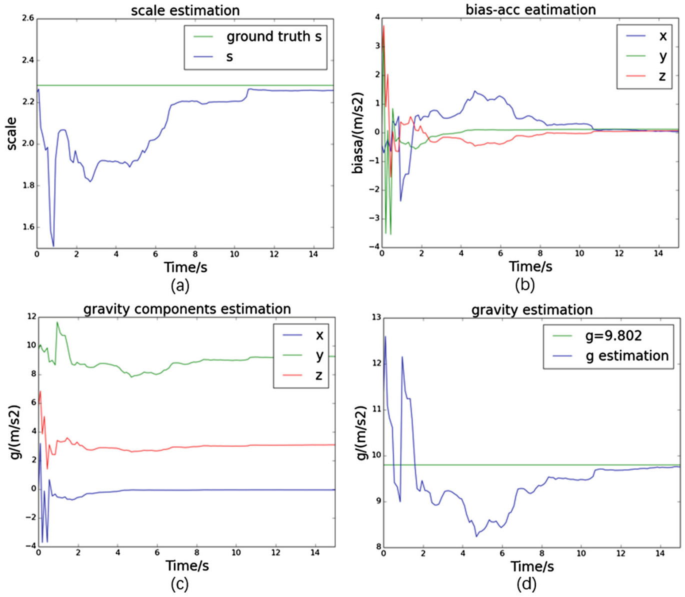
where ${}_{B_1}p_{B_2}$ is the position of B_2 in the body frame B_1, B_1 and B_2 are the IMU frames corresponding to the camera frame C_1 and C_2 at the same timestamp (see Fig. 1). $R_{B_1C_2} = R_{B_1B_2}R_{BC}$ is computed from the orientation $R_{B_1B_2}$ computed by the ground truth data and calibration R_{BC}.

4 Experimental Results

We conducted several experiments using the sequence $V1\_01\_easy$ and $V2\_01\_easy$ in the EuRoC dataset [21] to analyze the performance of our approach. It provides synchronized global shutter stereo images at 20 Hz with IMU measurements at 200 Hz and trajectory ground truth. We conduct the experiments in a virtual machine with 2 GB RAM.

4.1 Scale Estimation Results

The scale estimation procedure runs every time a new keyframe is inserted by ORB-SLAM [19]. Figure 2 shows the estimated scale, gravity and accelerometer bias. All variables are converged to stable values after 11 s. Figure 2(a) shows that the converged scale ($\sim$2.25972) is quite close to the ground truth scale (2.28132) which is the scale benchmark computed by the method that we have introduced in Sect. 3.4. Figure 2(b) indicates that the 3-axis accelerometer biases converge to almost 0. Figure 2(c) indicates that the components around x and z axes of gravity is converged quickly, and its y-axis component is converged to 9.256973 m/s^2 (near nominal gravity value). Hence the gravity direction is closed to y-axis. Figure 2(d) also shows the process of gravity estimation (depicted in blue), the green one is the nominal gravity value 9.802 m/s^2, they also come near after 11 s.



(a)

(b)

(c)

(d)

Fig. 2. The converged procedure of scale, accelerometer biases and gravity in the sequence $V1\_01\_easy$. (Color figure online)

Once we have estimated a stable and accurate scale. All 3D points in the map and the position of keyframes are updated according to the estimated scale. Figure 3(b) shows the final reconstructed sparse map, we also show a processed image in $V1\_01\_easy$.

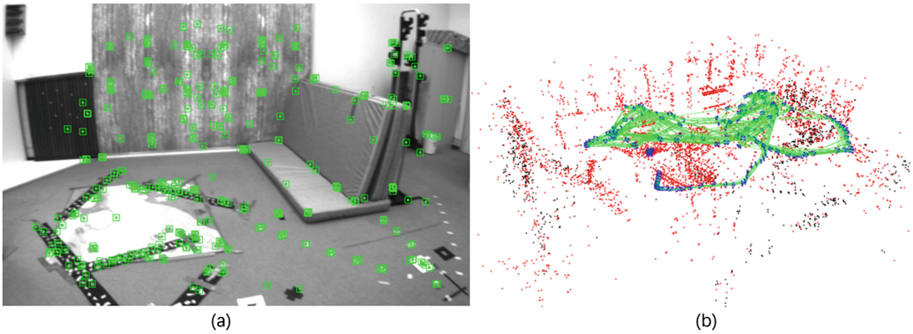


Fig. 3. A processed image and the reconstruction from sequence $V1\_01\_easy$

4.2 The Performance of Weighted Method for Scale Estimation Refinement

We evaluated the accuracy of proposed scale estimation and refinement by comparing it with the scale benchmark computed by the method in Sect. 3.4. As can be indicated in the Tables 1 and 2: for the sequence $V1\_01\_easy$ and $V2\_01\_easy$, we list the results of five tests. The second column is the scale estimation values $scale$ which is almost the same as the estimated scale s of [9], and the w_scale is the results of estimation refinement introduced in Sect. 3.3. We show the scale benchmark in the last one. The results indicate that our scale estimation refinement method can improve the accuracy of the estimated scale.

Table 1. The results of scale estimation and refinement, compared with scale benchmark for $V1\_01\_easy$

Test number	s	Scale	w_scale	Benchmark
1	2.25409	2.25972	**2.26318**	2.28132
2	2.10896	2.12254	**2.13477**	2.35991
3	2.22838	2.24186	**2.27997**	2.27073
4	2.28314	2.34132	**2.31572**	2.26904
5	2.15126	2.16206	**2.21084**	2.26057

Table 2. The results of scale estimation and refinement, compared with scale benchmark for *V*2_01_*easy*

Test number	s	Scale	w_scale	Benchmark
1	2.79302	2.80926	**2.83144**	2.92792
2	2.52695	2.52974	**2.55774**	2.59554
3	3.02662	3.01114	**3.06962**	3.11365
4	3.16472	3.17584	**3.20926**	3.35001
5	3.41774	3.41055	**3.43063**	3.46943

4.3 The Effect of Termination Criterion

Here we test our automatic criterion to determine when we consider the scale estimation successful. In the sequence *V*1_01_*easy*, the norm of recovered gravity $_{W}g$ is gradually close to the nominal gravity value ~9.8 m/s^2, after 11 s the difference is under the threshold ($0.1\,\mathrm{m}/\mathrm{s}^2$). And the other condition is established after the estimated scales come near for $n = 5$ times. Both conditions are established after the procedure runs about 11 s as depicted in the Fig. 3(a) and (d), and the scale estimation achieves convergence at that moment. And the converged speed in the paper [16] is 30 s, but its termination criterion is not mentioned.

5 Conclusions

In this paper, we showed our approaches for visual scale estimation and refinement. Firstly, we have presented an approach for the estimation of scale, gravity and accelerometer bias. Secondly, we proposed a weighting method for monocular visual scale estimation refinement, which utilizes weight w derived from the robust norm for outliers handling. Thirdly, we proposed an automatic way to identify convergence and termination for scale estimation procedure. We experimentally showed that the scale estimation is accurate, and the deduced weighting method further promotes the scale accuracy for the monocular visual map, and the termination criterion performs well, tested in the EuRoC dataset [21].

Acknowledgment. This research is supported by the National High Technology Research and Development Program of China (2015AA015902). We also would like to thank the authors of [19] for releasing the source code, and show my appreciation to Jing Wang for the code reproduction of the paper [9].

References

1. Leutenegger, S., Lynen, S., Bosse, M., Siegwart, R., Furgale, P.: Keyframe-based visual–inertial odometry using nonlinear optimization. Int. J. Robot. Res. **34**(3), 314–334 (2015)
2. Usenko, V., Engel, J., Stueckler, J., Cremers, D.: Direct visual-inertial odometry with stereo cameras. In: IEEE International Conference on Robotics and Automation (ICRA) (2016)
3. Forster, C., Carlone, L., Dellaert, F., Scaramuzza, D.: IMU preintegration on manifold for efficient visual-inertial maximum-a-posteriori estimation. In: Robotics: Science and Systems (RSS) (2015)
4. Concha, A., Loianno, G., Kumar, V., Civera, J.: Visual-inertial direct SLAM. In: IEEE International Conference on Robotics and Automation (ICRA) (2016)
5. Mourikis, A.I., Roumeliotis, S.I.: A multi-state constraint kalman filter for vision-aided inertial navigation. In: IEEE International Conference on Robotics and Automation (ICRA), pp. 3565–3572 (2007)
6. Jones, E.S., Soatto, S.: Visual-inertial navigation, mapping and localization: a scalable real-time causal approach. Int. J. Robot. Res. **30**(4), 407–430 (2011)
7. Wu, K., Ahmed, A., Georgiou, G., Roumeliotis, S.: A square root inverse filter for efficient vision-aided inertial navigation on mobile devices. In: Robotics: Science and Systems (RSS) (2015)
8. Lupton, T., Sukkarieh, S.: Visual-inertial-aided navigation for high-dynamic motion in built environments without initial conditions. IEEE Trans. Rob. 28(1), 61–76 (2012)
9. Mur-Artal, R., Tardós, J.D.: Visual-inertial monocular SLAM with map reuse. IEEE Robot. Autom. Lett. **2**(2), 796–803 (2017)
10. Yang, Z., Shen, S.: Monocular visual-inertial state estimation with online initialization and camera–IMU extrinsic calibration. IEEE Trans. Autom. Sci. Eng. **14**(1), 39–51 (2017)
11. Davison, A.J., Reid, I.D., Molton, N.D., et al.: MonoSLAM: real-time single camera SLAM. IEEE Trans. Pattern Anal. Mach. Intell. **29**(6), 1052 (2007)
12. Bloesch, M., Omari, S., Hutter, M., et al.: Robust visual inertial odometry using a direct EKF-based approach. In: 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 298–304. IEEE (2015)
13. Weiss, S., Achtelik, M., Chli, M., Siegwart, R.: Versatile distributed pose estimation and sensor self-calibration for an autonomous MAV. In: Proceeding of IEEE International Conference on Robotics and Automation (ICRA) (2012)
14. Weiss, S., Siegwart, R.: Real-time metric state estimation for modular vision-inertial systems. In: 2011 IEEE International Conference on Robotics and Automation (ICRA), pp. 4531–4537. IEEE (2011)
15. Engel, J., Sturm, J., Cremers, D.: Scale-aware navigation of a low-cost quadrocopter with a monocular camera. Robot. Auton. Syst. **62**(11), 1646–1656 (2014)
16. Tanskanen, P., Kolev, K., Meier, L., et al.: Live metric 3D reconstruction on mobile phones. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 65–72 (2013)
17. Furgale, P., Rehder, J., Siegwart, R.: Unified temporal and spatial calibration for multi-sensor systems. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 1280–1286 (2013)
18. Forster, C., Carlone, L., Dellaert, F., et al.: On-manifold preintegration for real-time visual–inertial odometry. **PP**(99), 1–21 (2016)

19. Mur-Artal, R., Montiel, J.M.M., Tardos, J.D.: ORB-SLAM: a versatile and accurate monocular SLAM system. IEEE Trans. Robot. **31**(5), 1147–1163 (2015)
20. Huber, P.J.: Robust estimation of a location parameter. Ann. Math. Statist. **35**(1), 73–101 (1964)
21. Burri, M., Nikolic, J., Gohl, P., Schneider, T., Rehder, J., Omari, S., Achtelik, M.W., Siegwart, R.: The EuRoC micro aerial vehicle datasets. Int. J. Robot. Res. **35**(10), 1157–1163 (2016)
22. Hartley, R., Zisserman, A.: Multiple View Geometry in Computer Vision, 2nd edn. Cambridge University Press, Cambridge (2004)