SCIENCE CHINA
Information Sciences

• RESEARCH PAPER •

January 2021, Vol. 64 112204:1–112204:13
https://doi.org/10.1007/s11432-019-2690-0

Homography-based camera pose estimation with known gravity direction for UAV navigation

Chunhui ZHAO[1], Bin FAN[2], Jinwen HU[1*], Quan PAN[1] & Zhao XU[2]

[1]*School of Automation, Northwestern Polytechnical University, Xi'an* 710072*, China;*
[2]*School of Electronic Information, Northwestern Polytechnical University, Xi'an* 710072*, China*

Received 26 April 2019/Revised 22 July 2019/Accepted 27 September 2019/Published online 14 December 2020

Abstract Relative pose estimation has become a fundamental and important problem in visual simultaneous localization and mapping. This paper statistically optimizes the solution for the homography-based relative pose estimation problem. Assuming a known gravity direction and a dominant ground plane, the homography representation in the normalized image plane enables a least squares pose estimation between two views. Furthermore, an iterative estimation method of the camera trajectory is developed for visual odometry. The accuracy and robustness of the proposed algorithm are experimentally tested on synthetic and real data in indoor and outdoor environments. Various metrics confirm the effectiveness of the proposed method in practical applications.

Keywords unmanned aerial vehicle, UAV, pose estimation, homography, least-squares estimation

Citation Zhao C H, Fan B, Hu J W, et al. Homography-based camera pose estimation with known gravity direction for UAV navigation. Sci China Inf Sci, 2021, 64(1): 112204, https://doi.org/10.1007/s11432-019-2690-0

1 Introduction

Given a number of point correspondences between homogeneous vectors, relative pose estimation finds the pose transformation of a view with respect to a different view. Relative pose estimation is a fundamental subroutine of various applications such as simultaneous localization and mapping [1], and target localization and tracking [2, 3]. Efficient and accurate solutions to relative pose estimation have been extensively pursued in recent years, culminating in the 5-point algorithm [4, 5], 7-point algorithm [6], 8-point algorithm [7], 17-point algorithm [8], and n-point algorithm [9] among others. The noise sensitivity of these algorithms can be reduced in two ways: developing a robust framework or finding a statistical optimization solution. The random sample consensus (RANSAC) [10], which is usually employed as a robust framework, iteratively obtains the camera pose from the minimal number of random point correspondences and uses the remaining point correspondences for validation. The minimal number of point correspondences is the least number of point correspondences needed by the corresponding pose estimation algorithm; for example, the 5-point algorithm requires at least five point correspondences. Meanwhile, statistical optimization of the camera pose is usually performed nonlinearly using methods such as the Gauss-Newton (GN) or the Levenberg-Marquardt method, which obtains accurate pose estimations for any set of point correspondences.

The complexity of the RANSAC process is exponentially related to the minimal number of point correspondences necessary for the camera pose estimation; therefore, reducing the number of required point correspondences is crucial for improving the computational efficiency and robustness of RANSAC. In general, relative pose estimation within the robust framework can be tackled under some constraints [11]. When the intrinsic parameters of the camera are known, five point correspondences are sufficient to estimate the essential matrix [4]. The essential matrix is then decomposed into a relative rotation matrix and a translation vector, which describes the camera pose between two views. Using the general assumption that all three-dimensional (3D) points lie on a plane, the homography can be estimated by four point

* Corresponding author (email: hujinwen@nwpu.edu.cn)

correspondences [6]. The authors of [12] proposed an inverse projection ray-based iterative method for relative pose estimation, which requires at least four point correspondences. When a common direction is known, the camera pose can be estimated from just three point correspondences [13–16]. A closed-form solution for three points was proposed in [17]. The authors represented rotation as unit quaternions, and verified their solution using the autonomous aerial refueling system of a vision-based unmanned aerial vehicle (UAV) [18]. Further, if the rotation between two views is provided by an inertial measurement unit (IMU), the remaining translation can be recovered up to a scale using only two point correspondences [19]. In summary, the number of point correspondences required in the relative pose estimation can be reduced by enforcing some assumptions or providing additional information. In fact, a common direction can be easily obtained by the IMU or inferred from the vanishing points [20]. The plane hypothesis is reasonable because many typical plane scenarios such as indoor robots and self-driving cars are possible in nature. Some studies have combined the common direction and dominant plane assumptions to achieve relative pose estimation with a minimum of two point correspondences [21–23].

In [21–23], the minimal solution was obtained using prior knowledge of the reference plane used for the homography estimation and the gravity vector used for image alignment between two views. Guan et al. [21] decoupled the camera pose estimation by separating the point correspondences into two sets: a far set for estimating the rotation, and a near set for estimating the translation. However, the distinction between the far and near points is ambiguous, leading to a suboptimal estimation result. Although the 2-point method in [22, 23] was identical to the two aforementioned hypotheses, the 2-point method based on the three residual components of bearing vectors is redundant and even fails in some cases because the covariance matrix of the bearing vectors is singular [24]. Moreover, the 2-point method allows two possible solutions, and the subsequent verification is unclarified and difficult. As both hypotheses are beneficial in decreasing the number of point correspondences, they are further extended in our method. Our method also retains the homography scale factors, which are usually eliminated unnecessarily [21–23] with noticeable consequences on the estimation accuracy.

Therefore, we propose a least-squares solution to the relative pose estimation problem based on the homography error in the normalized image plane by assuming a known gravity direction. This paper makes three main contributions to the pose estimation literature. First, based on a known gravity direction, it introduces constant and solvable homography scale factors that improve the calculation accuracy. By contrast, traditional methods [21–23] ignore the scale factors because their values depend on the point correspondences. Second, the homography error is formulated in the normalized image plane ensuring a unique optimal solution. By contrast, traditional methods admit multiple solutions. Third, an iterative estimation method of the camera trajectory is developed for visual odometry. According to the gravity direction and the ground plane hypotheses, the proposed homography-based least squares (HLS) solution computes the camera motion using at least two point correspondences between two views. Finally, the accuracy and robustness of the proposed method is demonstrated by performing a variety of experiments. The proposed method outperforms the compared methods and is suitable and robust for practical applications.

The remainder of this paper is organized as follows: Section 2 formulates the homography-based pose estimation problem between two views; Section 3 estimates the relative camera pose via a transformed linear least squares formulation and proposes the iterative camera trajectory estimation algorithm for visual odometry; Section 4 quantitatively compares the performances of the proposed and classical methods using synthetic and real data in indoor and outdoor environments. A conclusion is drawn in Section 5.

2 Problem formulation

The images for solving the camera pose estimation problem are taken by a monocular camera at different time instants. Let $\boldsymbol{x}_i = [x_i, y_i, z_i]^{\mathrm{T}}$ be the coordinates of a given 3D point or vector $\boldsymbol{x} = [x, y, z]^{\mathrm{T}}$ in the real world camera coordinate system at time i, where i is a positive integer. The set of feature points on the ground plane [25] extracted from the image at time i is denoted as $X_i = \{\boldsymbol{x}_i^k : k = 1, \ldots, m\}$ $(m \geqslant 2)$, where $\boldsymbol{x}_i^k$ is the k-th 3D point in the camera coordinate system at time i. The ground plane is described by

$$\boldsymbol{N}_i^{\mathrm{T}} \boldsymbol{x}_i^k + d_i = 0, \tag{1}$$

where $d_i > 0$ is the distance between the origin of the camera and the ground plane at time i, and $\boldsymbol{N}_i$ is the unit normal vector of the ground plane in the camera coordinate system at time i.

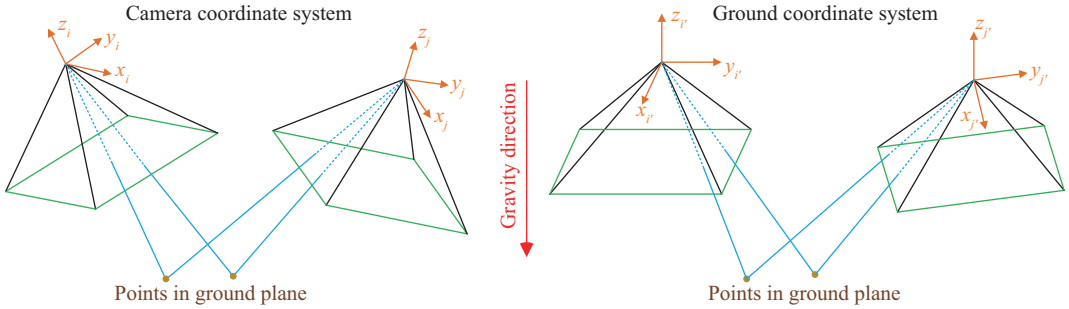

Figure 1 (Color online) Camera coordinate systems aligned with the ground plane (called the ground coordinate system). The detected 3D points lie on the ground plane.

From time i to j, we have

$$\boldsymbol{x}_j^k = \boldsymbol{R}_{ji}\boldsymbol{x}_i^k + \boldsymbol{t}_{ji}, \tag{2}$$

where $\boldsymbol{R}_{ji}$ and $\boldsymbol{t}_{ji}$ are the rotation matrix and the translation vector, respectively, owing to the motion of the camera. By (1), we get

$$\boldsymbol{x}_j^k = \boldsymbol{R}_{ji}\boldsymbol{x}_i^k - \boldsymbol{t}_{ji}\frac{\boldsymbol{N}_i^{\mathrm{T}}\boldsymbol{x}_i^k}{d_i} = \left(\boldsymbol{R}_{ji} - \frac{\boldsymbol{t}_{ji}\boldsymbol{N}_i^{\mathrm{T}}}{d_i}\right)\boldsymbol{x}_i^k. \tag{3}$$

In Figure 1, the gravity direction is assumed to be known (obtained by an IMU) [21–23]. Subsequently, the camera coordinate system can always be rotated to coincide with the ground coordinate system with the x-y plane parallel to the ground plane. The coordinates of $\boldsymbol{x}_i^k$ in the ground coordinate system can be represented as

$$\boldsymbol{x}_{i'}^k = \boldsymbol{R}_{i'i}\boldsymbol{x}_i^k, \tag{4}$$

where $\boldsymbol{R}_{i'i}$ is the corresponding rotation matrix. It holds that $\boldsymbol{R}_{i'i} = \boldsymbol{R}_{ii'}^{-1} = \boldsymbol{R}_{ii'}^{\mathrm{T}}$ and $\boldsymbol{R}_{i'i}\boldsymbol{N}_i = \boldsymbol{N} = [0,0,1]^{\mathrm{T}}$. Note that the last entry of $\boldsymbol{x}_{i'}^k$ is exactly d_i which is the height of the camera and also the depth of $\boldsymbol{x}_{i'}^k$, and this property holds for all k. Combining (3) and (4), we get

$$\boldsymbol{x}_{j'}^k = \boldsymbol{R}_{j'j}\left(\boldsymbol{R}_{ji} - \frac{\boldsymbol{t}_{ji}\boldsymbol{N}_i^{\mathrm{T}}}{d_i}\right)\boldsymbol{R}_{ii'}\boldsymbol{x}_{i'}^k = \left(\boldsymbol{R}_{j'i'} - \frac{\boldsymbol{t}_{j'i'}\boldsymbol{N}^{\mathrm{T}}}{d_i}\right)\boldsymbol{x}_{i'}^k, \tag{5}$$

where

$$\boldsymbol{R}_{j'i'} \triangleq \boldsymbol{R}_{j'j}\boldsymbol{R}_{ji}\boldsymbol{R}_{ii'}, \quad \boldsymbol{t}_{j'i'} \triangleq \boldsymbol{R}_{j'j}\boldsymbol{t}_{ji}. \tag{6}$$

In reality, the normalized vector $\bar{\boldsymbol{x}}_i^k \triangleq \boldsymbol{x}_i^k/d_i + v_i^k$ instead of the $\boldsymbol{x}_{i'}^k$ is measured, where d_i is the depth of the ground point $\boldsymbol{x}_{i'}^k$ in the ground coordinate system of the camera at time i, and v_i^k is the measurement noise. The last entry of $\bar{\boldsymbol{x}}_i^k$ should be 1.0 for all k.

From (5), we obtain

$$\bar{\boldsymbol{x}}_{j'}^k - v_j^k = \mu_{ji}\left(\boldsymbol{R}_{j'i'} - \frac{\boldsymbol{t}_{j'i}\boldsymbol{N}^{\mathrm{T}}}{d_i}\right)\left(\bar{\boldsymbol{x}}_{i'}^k - v_i^k\right), \tag{7}$$

where $\mu_{ji}^k \triangleq z_i^k/z_j^k$ is the homography scale factor which is positive. It can be simplified as

$$\bar{\boldsymbol{x}}_{j'}^k = \boldsymbol{H}_{j'i'}\bar{\boldsymbol{x}}_{i'}^k + w_{ji}^k, \tag{8}$$

where

$$\boldsymbol{H}_{j'i'} \triangleq \mu_{ji}\left(\boldsymbol{R}_{j'i'} - \frac{\boldsymbol{t}_{j'i}\boldsymbol{N}^{\mathrm{T}}}{d_i}\right), \quad w_{ji}^k \triangleq v_j^k - \boldsymbol{H}_{j'i'}v_i^k. \tag{9}$$

To minimize the squared mismatching error $\|\boldsymbol{w}_{ji}\|^2$, where $\boldsymbol{w}_{ji} \triangleq [\boldsymbol{w}_{ji}^1, \boldsymbol{w}_{ji}^2, \ldots, \boldsymbol{w}_{ji}^m]^{\mathrm{T}}$, the relative pose estimation is formulated as the following optimization problem:

$$\min_{\boldsymbol{R}_{ji},\boldsymbol{t}_{ji}} \sum_{k=1}^{m} \left\|\bar{\boldsymbol{x}}_{j'}^k - \boldsymbol{H}_{j'i'}\bar{\boldsymbol{x}}_{i'}^k\right\|^2. \tag{10}$$

Because $\boldsymbol{R}_{ji}$ is a matrix, Eq. (10) cannot be solved directly using the least squares method; however, it can be simplified and transformed into a tractable problem as described in Subsection 3.1.

3 Homography based iterative pose estimation

3.1 Least squares solution to relative pose estimation

In this subsection, we first simplify the optimization problem (10) into a standard linear least squares optimization problem, and then derive the solutions of $\boldsymbol{R}_{ji}$ and $\boldsymbol{t}_{ji}$.

The rotation $\boldsymbol{R}_{j'i'}$ has no effect on the z axis, which can be expressed as the yaw angle $\theta_{j'i'}$ from the ground coordinate system i' to j'. Accordingly, $\boldsymbol{H}_{j'i'}$ can be expanded by (9) as

$$\boldsymbol{H}_{j'i'} = \mu_{ji}\left(\begin{bmatrix}\cos\theta_{j'i'} & -\sin\theta_{j'i'} & 0\\ \sin\theta_{j'i'} & \cos\theta_{j'i'} & 0\\ 0 & 0 & 1\end{bmatrix} - \frac{\boldsymbol{t}_{j'i'}}{d_i}\begin{bmatrix}0\\0\\1\end{bmatrix}^{\mathrm{T}}\right). \tag{11}$$

Because we discuss the homography only from time i to j, we hereafter omit the indeces ji and $j'i'$ to simplify the derivations (with the exception of cases where the omission would cause confusion). From (11), we obtain

$$\boldsymbol{H} = \begin{bmatrix}\mu\cos\theta & -\mu\sin\theta & -\mu t_x\\ \mu\sin\theta & \mu\cos\theta & -\mu t_y\\ 0 & 0 & \mu(1-t_z)\end{bmatrix} \triangleq \begin{bmatrix}h_1 & -h_2 & h_3\\ h_2 & h_1 & h_4\\ 0 & 0 & h_5\end{bmatrix}, \tag{12}$$

where $\boldsymbol{t}_{j'i'}/d_i \triangleq [t_x, t_y, t_z]^{\mathrm{T}}$ and

$$\boldsymbol{h}_{j'i'} \triangleq [h_1, h_2, h_3, h_4, h_5]^{\mathrm{T}} \tag{13}$$

includes the variables to be solved. The above definitions imply that

$$h_1^2 + h_2^2 = \mu^2. \tag{14}$$

Solving $\boldsymbol{h}_{j'i'}$, we easily retrieve $\boldsymbol{R}_{ji}$ and $\boldsymbol{t}_{ji}$ in (6) as follows:

$$\begin{aligned}\boldsymbol{R}_{ji} &= \boldsymbol{R}_{jj'}\begin{bmatrix}\frac{h_{1,j'i'}}{\sqrt{h_{1,j'i'}^2+h_{2,j'i'}^2}} & -\frac{h_{2,j'i'}}{\sqrt{h_{1,j'i'}^2+h_{2,j'i'}^2}} & 0\\ \frac{h_{2,j'i'}}{\sqrt{h_{1,j'i'}^2+h_{2,j'i'}^2}} & \frac{h_{1,j'i'}}{\sqrt{h_{1,j'i'}^2+h_{2,j'i'}^2}} & 0\\ 0 & 0 & 1\end{bmatrix}\boldsymbol{R}_{i'i},\\ \boldsymbol{t}_{ji} &= d_i\boldsymbol{R}_{jj'}\begin{bmatrix}-\frac{h_{3,j'i'}}{\sqrt{h_{1,j'i'}^2+h_{2,j'i'}^2}}\\ -\frac{h_{4,j'i'}}{\sqrt{h_{1,j'i'}^2+h_{2,j'i'}^2}}\\ 1-\frac{h_{5,j'i'}}{\sqrt{h_{1,j'i'}^2+h_{2,j'i'}^2}}\end{bmatrix}.\end{aligned} \tag{15}$$

Here we must focus on the relative translation vector between two views, which is up-to-scale if d_i is unknown. From (10) and (13), we get

$$\bar{\boldsymbol{x}}_{j'}^k - \boldsymbol{H}_{j'i'}\bar{\boldsymbol{x}}_{i'}^k = \boldsymbol{A}_{i'}^k\boldsymbol{H}_{j'i'} - \boldsymbol{b}_{j'}^k, \tag{16}$$

where

$$\boldsymbol{A}_{i'}^k = \begin{bmatrix}-\bar{x}_{i'}^k & \bar{y}_{i'}^k & -1 & 0 & 0\\ -\bar{y}_{i'}^k & -\bar{x}_{i'}^k & 0 & -1 & 0\\ 0 & 0 & 0 & 0 & -1\end{bmatrix}, \quad \boldsymbol{b}_{j'}^k = \begin{bmatrix}-\bar{x}_{j'}^k\\ -\bar{y}_{j'}^k\\ -1\end{bmatrix}.$$

We further define the following augmented matrices:

$$\boldsymbol{A}_{i'} = \begin{bmatrix}\boldsymbol{A}_{i'}^1\\ \boldsymbol{A}_{i'}^2\\ \vdots\\ \boldsymbol{A}_{i'}^m\end{bmatrix}, \quad \boldsymbol{b}_{j'} = \begin{bmatrix}\boldsymbol{b}_{j'}^1\\ \boldsymbol{b}_{j'}^2\\ \vdots\\ \boldsymbol{b}_{j'}^m\end{bmatrix}. \tag{17}$$

The problem in (10) is transformed into the following linear least squares problem:

$$\min_{\boldsymbol{H}_{j'i'}} \|\boldsymbol{A}_{i'}\boldsymbol{H}_{j'i'} - \boldsymbol{b}_{j'}\|_2^2. \tag{18}$$

The unique solution of (18) is then easily obtained as

$$\boldsymbol{H}_{j'i'} = \left(\boldsymbol{A}_{i'}^{\mathrm{T}}\boldsymbol{A}_{i'}\right)^{-1}\boldsymbol{A}_{i'}^{\mathrm{T}}\boldsymbol{b}_{j'}. \tag{19}$$

The detailed computational procedure is given in Algorithm 1.

Algorithm 1 The proposed HLS method

Input: General corresponding measurements $\bar{X}_i$ and $\bar{X}_j$, and the known $\boldsymbol{R}_{i'i}$ and $\boldsymbol{R}_{j'j}$ at time i and j respectively.
Output: Rotation matrix $\boldsymbol{R}_{ji}$ and translation vector $\boldsymbol{t}_{ji}$.
1: Pre-rotate each measurement according to (4), and then normalize them to get $\bar{X}_{i'}$ and $\bar{X}_{j'}$;
2: Calculate the coefficient matrices $\boldsymbol{A}_{i'}$ and $\boldsymbol{b}_{i'}$ for all point correspondences according to (17);
3: Obtain the closed-form solution $\boldsymbol{H}_{j'i'}$ according to (19);
4: Acquire the estimation result $\boldsymbol{R}_{ji}$ and $\boldsymbol{t}_{ji}$ according to (15);
5: Return $\boldsymbol{R}_{ji}$ and $\boldsymbol{t}_{ji}$.

Remark 1. The $\boldsymbol{R}_{ji}$ directly solved by (15) may not be a valid rotation matrix because of the numerical error. Thus, we compute the single value decomposition of $\boldsymbol{R}_{ji}$ as

$$\boldsymbol{R}_{ji} = \boldsymbol{U}_{\mathrm{R}}\boldsymbol{D}_{\mathrm{R}}\boldsymbol{V}_{\mathrm{R}}^{\mathrm{T}}. \tag{20}$$

The final rotation is then estimated by minimizing the Frobenius norm [6]:

$$\boldsymbol{R}_{ji} = \boldsymbol{U}_{\mathrm{R}}\boldsymbol{V}_{\mathrm{R}}^{\mathrm{T}}. \tag{21}$$

3.2 Iterative camera trajectory estimation

According to (15), d_i should be known for solving $\boldsymbol{t}_{ji}$, which implies that monocular uncertainty can be restored. However, this measurement may require extra sensors, such as height-finding radar or a barometer equipped on the drone, which can easily estimate the initial camera height. If d_i is known, then from the definition of μ_{ji} and (14) we get

$$d_j = \frac{d_i}{\sqrt{h_{1,j'i'}^2 + h_{2,j'i'}^2}}, \tag{22}$$

implying that d_i can be iteratively estimated given the initial height d_0 of the camera. If possible, d_j should be combined with the barometer measurements to improve the robustness of the visual odometry.

The procedure for an iterative linear HLS estimation of the camera pose is described in Algorithm 2.

Algorithm 2 Iterative camera trajectory estimation

1: Initialize d_0 and set $\boldsymbol{R}_{00} = \boldsymbol{I}$, $\boldsymbol{t}_{00} = \boldsymbol{0}$;
2: $i = 0$;
3: $j = i + 1$;
4: Estimate $\boldsymbol{R}_{ji}$ and $\boldsymbol{t}_{ji}$ by Algorithm 1;
5: Updata camera pose by $\boldsymbol{R}_{0j} \leftarrow \boldsymbol{R}_{0i}\boldsymbol{R}_{ji}^{\mathrm{T}}$ and $\boldsymbol{t}_{0j} \leftarrow \boldsymbol{t}_{0i} - \boldsymbol{R}_{0i}\boldsymbol{R}_{ji}^{\mathrm{T}}\boldsymbol{t}_{ji}$;
6: Update d_j according to (22);
7: Return camera trajectory $\{[\boldsymbol{R}_{0j}|\boldsymbol{t}_{0j}]\}$;
8: Iterate time sequence $i \leftarrow i + 1$;
9: Repeat from step 3 to step 8.

4 Experimental results

In this section, we evaluate the performance of the proposed method primarily by assessing the errors in the relative pose estimation method (Subsection 3.1). We also determine the error in the homography scale factor between two views. It is noteworthy that the feasibility of our algorithm is verified by a plain

and heuristic visual odometry. First, the efficiency of the algorithm is tested in the presence of synthetic image noise and/or IMU noise. Second, to evaluate the robustness of the proposed method in the presence of real noise, the algorithm is evaluated on real data from the ETH dataset [26], and its performance is compared with that of the 2-point method [22]. Third, a UAV based navigation experiment is performed to validate the applicability of the proposed method. The proposed method is then fully validated by the synchronous application of various error metrics.

4.1 Experiment with synthetic noise

To evaluate the algorithms on synthetic noise, the following experimental procedure is conducted. The average distance between the ground plane and the center of the first camera center is set to 1 m. The baseline between two cameras is set to 0.1 m, the focal length is fixed at 1000 pixels, and the field of view is 120°. The ground plane is constructed from $m = 400$ randomly sampled 3D points, 100 of which are placed close to the periphery of the imaging plane (i.e., far from the camera). The accuracy of the proposed method is evaluated on synthetic data subjected to different image noise and increasing IMU noise, and subsequently compares with those of 2-point [22] and 5-point [5] methods. Each algorithm is evaluated either along the x- or y-axis of the first camera, which represents the forward and sideways camera motions, respectively. Meanwhile, the second camera is rotated around each axis. Each data point in the plots represents the first two intervals of the 5-partitions (quintiles) of 1000 measurements with points randomly generated from the ground plane.

To evaluate the robustness of the proposed relative pose estimation method, we compare the relative translation and rotation between times j and i, respectively. The rotational error compares the angle difference between the true and estimated rotations. When d_i is unknown, the estimated translation from time i to j is known only up to a scale, i.e., only the translational direction is obtained, so we compare the angle difference between the true and estimated translations. The rotational and translational errors are computed as follows [21].

(1) Rotational error:

$$e_{\boldsymbol{R}} = \arccos\left(\left(\mathrm{Tr}\left(\boldsymbol{R}_{ji}^{gt}\hat{\boldsymbol{R}}_{ji}^{\mathrm{T}}\right)-1\right)\Big/2\right);$$

(2) Translational error:

$$e_{\boldsymbol{t}} = \arccos\left(\left({\boldsymbol{t}_{ji}^{gt}}^{\mathrm{T}}\hat{\boldsymbol{t}}_{ji}\right)\Big/\left(\|\boldsymbol{t}_{ji}^{gt}\|\,\|\hat{\boldsymbol{t}}_{ji}\|\right)\right),$$

where $\boldsymbol{R}_{ji}^{gt}, \boldsymbol{t}_{ji}^{gt}$ denote the ground-truth transformations and $\hat{\boldsymbol{R}}_{ji}, \hat{\boldsymbol{t}}_{ji}$ are the corresponding estimated transformations.

Figure 2 shows the results for forward and sideways motion with varying image noise and precise IMU measurements. Both the HLS and 2-point methods outperform the 5-point method in terms of accuracy, and the HLS method performs slightly better than the 2-point method in terms of precision. Figure 3 compares the results dering forward and sideways motions under different IMU noise and constant image noise (0.5 pixels standard deviation). As shown in Figures 3(a), (b), (e), and (f), the precision of the estimation rotation decreases slightly for HLS with GN optimization (depicted as HLS+GN), and the HLS method and the 2-point method for rotation estimation are decreasing slightly. Figures 3(c), (d), (g), and (h) clearly demonstrate the higher translation estimation performances of the HLS and HLS+GN methods compared to the 2-point method. Additionally, the HLS method is more robust for distant 3D points compared to other methods in the presence of varying image and IMU noise. Figure 4 shows the estimation error in the homography scale factor under different noise. In these experiments, 1000 scale estimation errors are recorded at each error level. Although the uncertainty of the scale estimation error increases with the noise level, the proposed method is robust toward image and IMU noise.

The time consumptions of the HLS and the 2-point methods are of the same order of magnitude in all experiments. In both methods, the camera pose between two views is estimated in approximately 0.1 ms, which is sufficiently fast for real-time applications. In conclusion, it is easier to acquire a more accurate relative pose estimation using the HLS method.

4.2 Experiment on the ETH dataset

We compare the performance of the HLS and 2-point [22] methods using the ETH dataset [26]. The ETH dataset is an indoor dataset comprising a quadrotor with a downward-looking camera with a 130-degree field of view. The dataset contains $n = 1361$ monochrome sequences recorded by the quadrotor flying

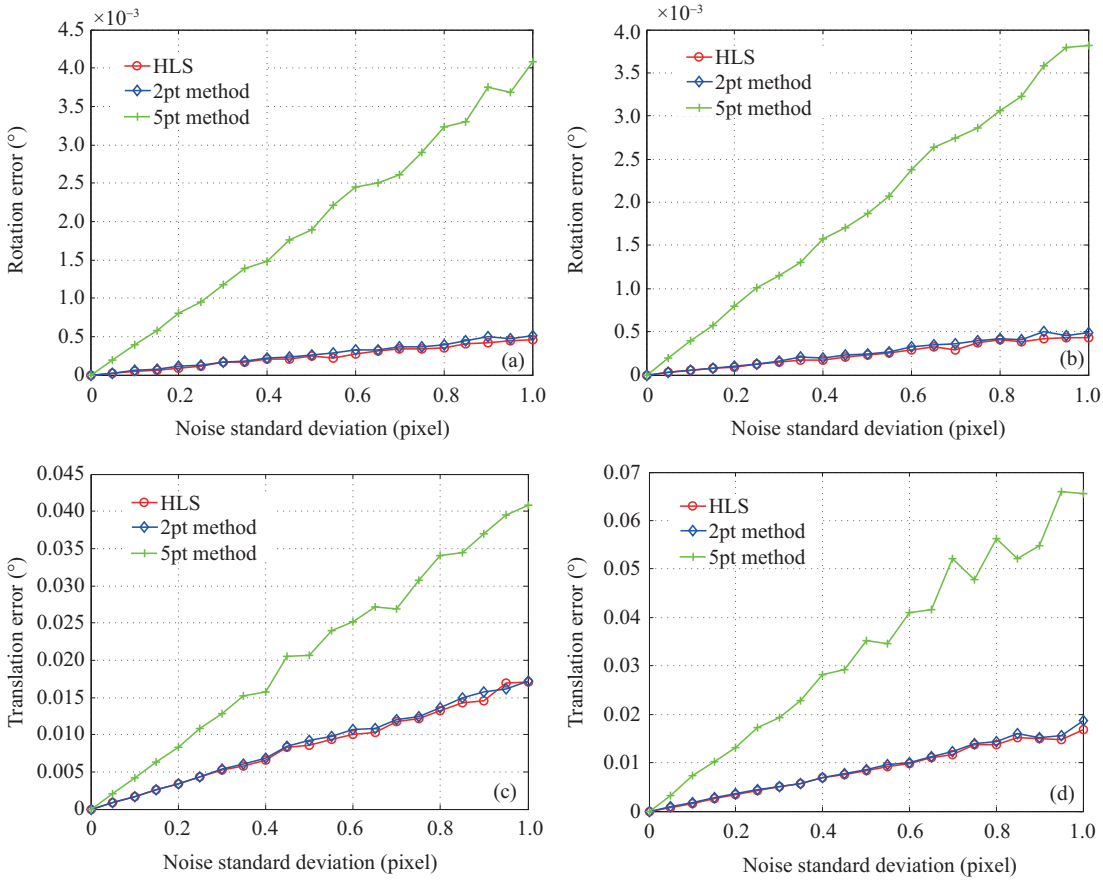

Figure 2 (Color online) Evaluation of the HLS, 2-point (2pt), and 5-point (5pt) methods during forward ((a) and (c)) and sideways ((b) and (d)) motions with varying image noise.

with a circular trajectory in an office-size environment. The single ground-plane scene and real-time IMU data present an ideal scenario for testing our method. In this experiment, we extract 1000 SIFT (scale invariant feature transform) features [27] from each image and match them based on the epipolar geometry [28] between consecutive frames. Note that these features lie directly on the ground because of the nature of the dataset. Using the IMU data of the sequences, we pre-rotate the features to align them with the ground plane. The camera trajectory is then estimated by the proposed method.

For the error evaluation, we assume that the proposed method computes a sequence of camera poses between consecutive frames:

$$\left\{\boldsymbol{T}_{1,2},\ldots,\boldsymbol{T}_{n-1,n}\in \mathrm{SE}\,(3)\mid \boldsymbol{T}_{i,i+1}=\begin{bmatrix}\hat{\boldsymbol{R}}_{i,i+1} & \hat{\boldsymbol{t}}_{i,i+1}\\ 0 & 1\end{bmatrix}, i=1,\ldots,n-1\right\},$$

and their corresponding ground-truth camera trajectories are

$$\left\{\boldsymbol{Q}_1,\ldots,\boldsymbol{Q}_n\in \mathrm{SE}(3)\mid \boldsymbol{Q}_i=\begin{bmatrix}\boldsymbol{R}_{wi}^{gt} & \boldsymbol{t}_{wi}^{gt}\\ 0 & 1\end{bmatrix}, i=1,\ldots,n-1\right\},$$

where $\hat{\boldsymbol{R}}_{i,i+1}$ and $\hat{\boldsymbol{t}}_{i,i+1}$ are the estimated rotation and translation transformations from the $(i+1)$th camera coordinate system to the ith camera coordinate system, respectively, and $\boldsymbol{R}_{wi}^{gt}$ and $\boldsymbol{t}_{wi}^{gt}$ are the true rotation and translation transformations from the ith camera coordinate system to the world coordinate system w, respectively. To estimate the trajectory drift error, the relative pose error (RPE) [29] between the ith and the $(i+1)$th frames is introduced as follows:

$$\boldsymbol{E}_{i,i+1}=\left(\boldsymbol{Q}_i^{-1}\boldsymbol{Q}_{i+1}\right)^{-1}\boldsymbol{T}_{i,i+1}.$$

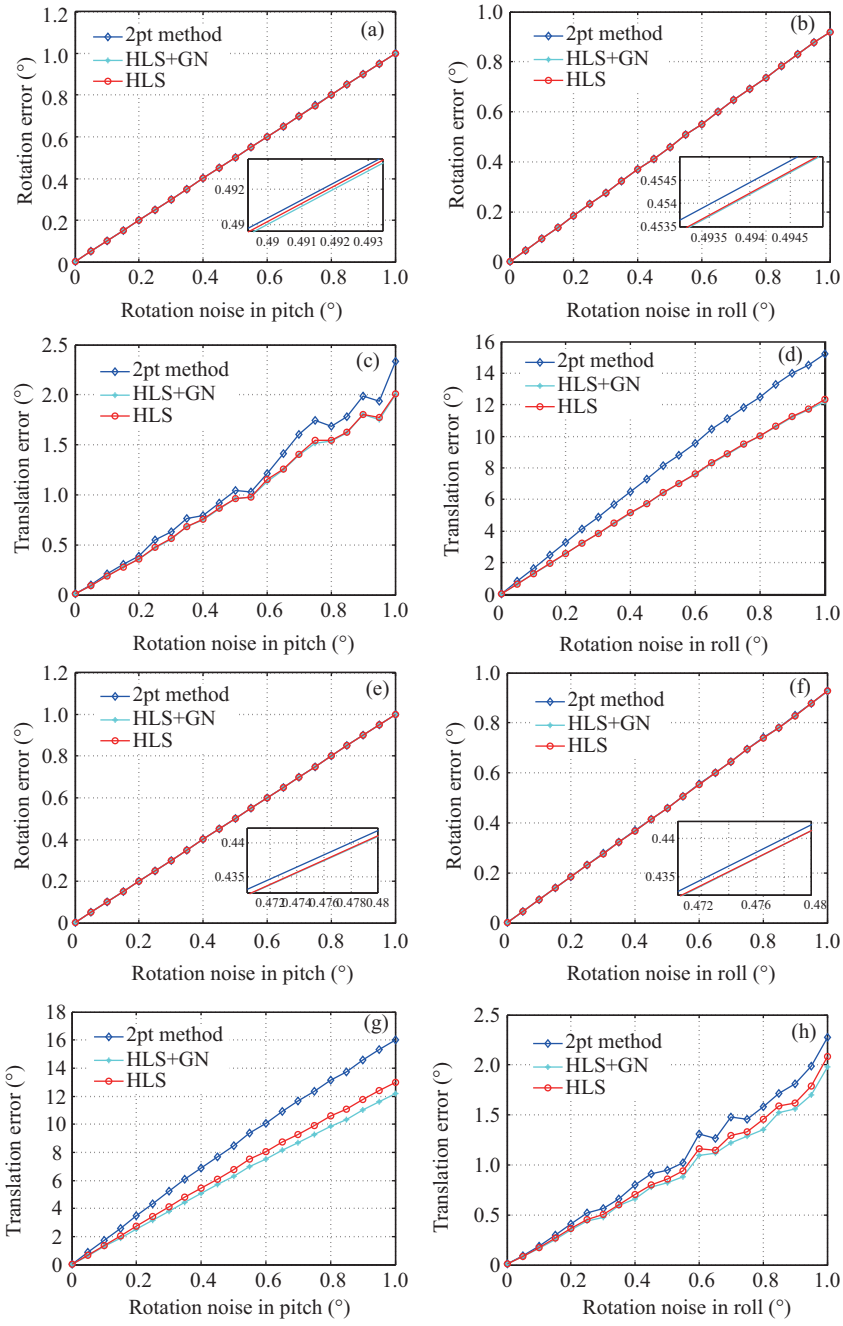

Figure 3 (Color online) Evaluation of the HLS, HLS+GN, and 2-point methods during forward ((a)–(d)) and sideways ((e)–(h)) motions with increasing IMU noise from 0° to 1°. The image noise is fixed at 0.5 pixels standard deviation.

From $n-1$ RPEs obtained between consecutive frames, the root mean squared error (RMSE) is calculated as

$$\mathrm{RMSE}\left(\boldsymbol{E}_{1:n-1}\right)=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n-1}\left\|\mathrm{trans}\left(\boldsymbol{E}_{i,i+1}\right)\right\|^{2}},$$

where $\mathrm{trans}\left(\boldsymbol{E}_{i,i+1}\right)$ represents the translational component of the RPE $\boldsymbol{E}_{i,i+1}$, i.e., the first three elements of the fourth column of $\boldsymbol{E}_{i,i+1}$.

Figure 5 shows some matched feature points in the images taken from the trajectory estimated by the HLS method. The RPEs of the HLS and 2-point methods with RMSEs of 19.3 mm and 22.7 mm, respectively, are compared in Figure 6. The RPE gradually increases after 300 frames, possibly owing to severe image distortion and/or image blurring. These results confirm the higher performance of the HLS

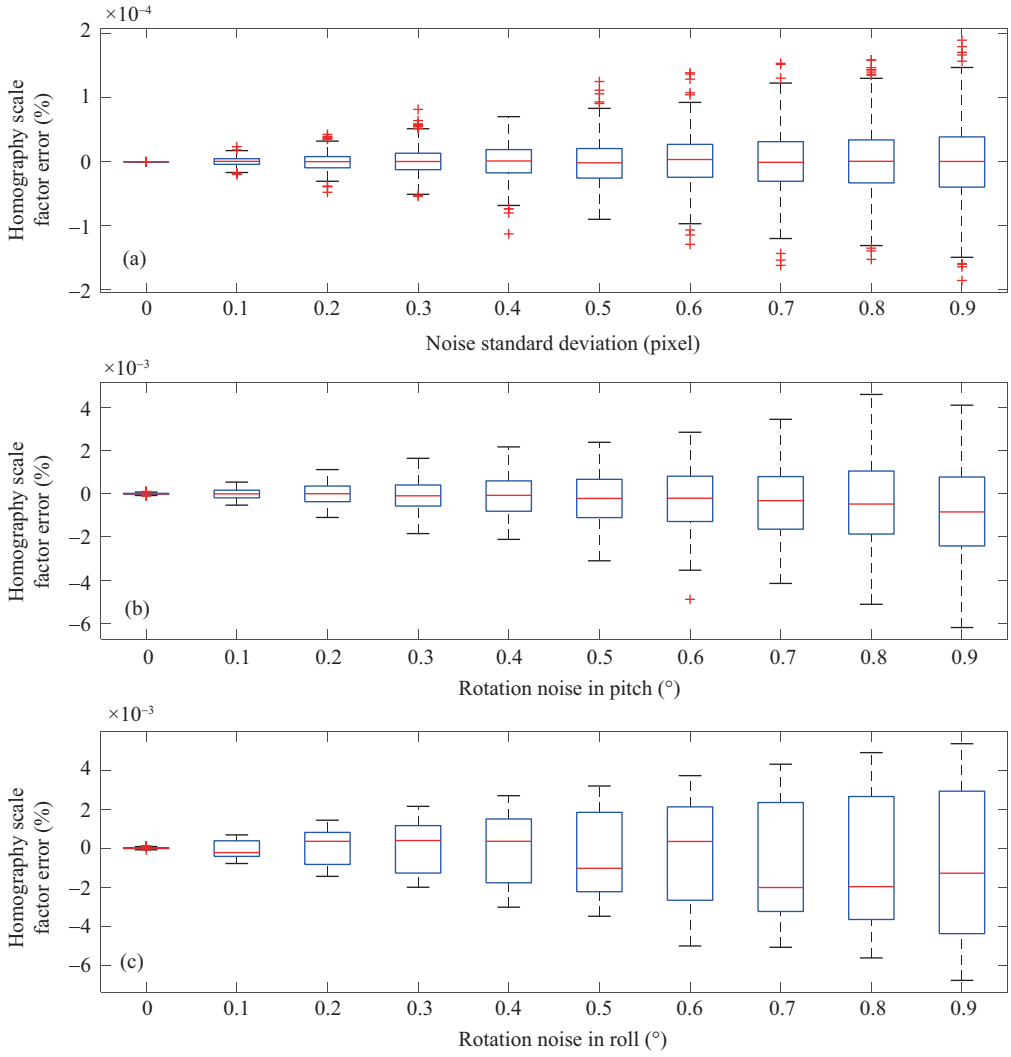

Figure 4 (Color online) Accuracy of our method in estimating the homography scale factor. We increase image noise from 0 to 1 shown in (a). Also, (b) and (c) report the effect of pitch/roll noise from 0° to 1°, and the image noise is fixed at 0.5 pixels standard deviation.

method compared to the 2-point method.

4.3 UAV navigation experiment

In the practical application experiment, a drone flies in a circular trajectory heading parallel to the tangent of the circle. This setup is used to validate the applicability of the proposed method in practice because when a drone flies horizontally at a constant slow speed, its roll, pitch and yaw angles can be obtained by airborne IMUs and magnetometers. Starting from the origin of the world coordinate system, an ordinary visual odometry [30] is performed incrementally relative to the beginning frame, and this process is relatively affected by cumulative errors. As shown in Figure 7, the trajectory $\boldsymbol{t}_{wi}$ of the ith frame with a scale is roughly obtained from the estimated yaw angle $\hat{\theta}_{wi}$ and flying radius r, given that the initial estimated translation $\hat{\boldsymbol{t}}_{wi}$ is known up to a scale:

$$\boldsymbol{t}_{wi} = \frac{r \cdot \sin(\hat{\theta}_{wi})}{\sin(\arccos(\boldsymbol{N}_y^{\mathrm{T}} \hat{\boldsymbol{t}}_{wi}) + \hat{\theta}_{wi})} \cdot \hat{\boldsymbol{t}}_{wi},$$

where $\boldsymbol{N}_y = [0, 1, 0]^{\mathrm{T}}$ is the unit y-axis coordinate.

Because rotational errors in moving camera images manifest as translational errors, they can be considered as translational errors [29]. To assess the global consistency of the proposed method, we calculate the

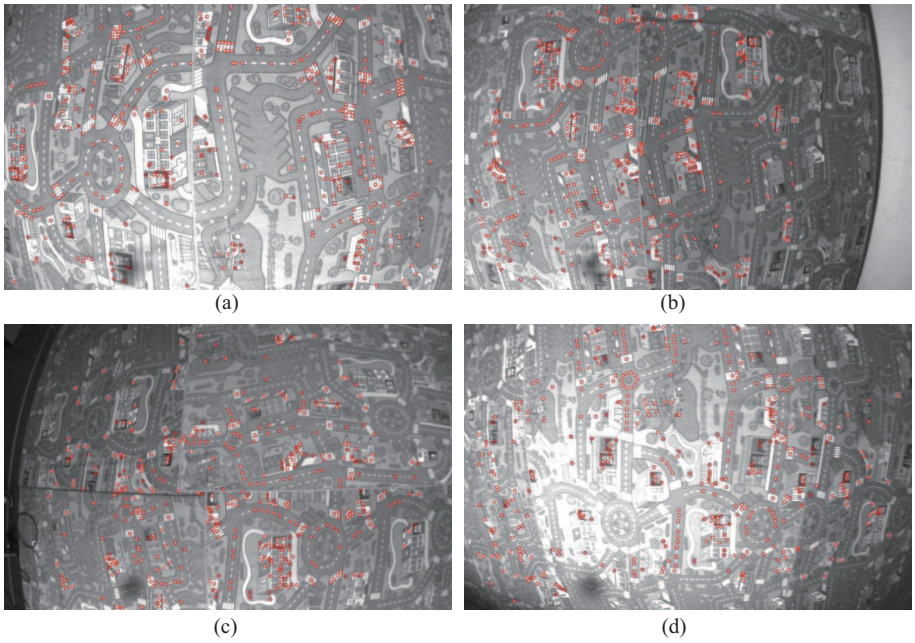

Figure 5 (Color online) A few matched feature points in the images taken from the trajectory estimated by the HLS method. (a) The 1st frame; (b) the 400th frame; (c) the 800th frame; (d) the 1200th frame.

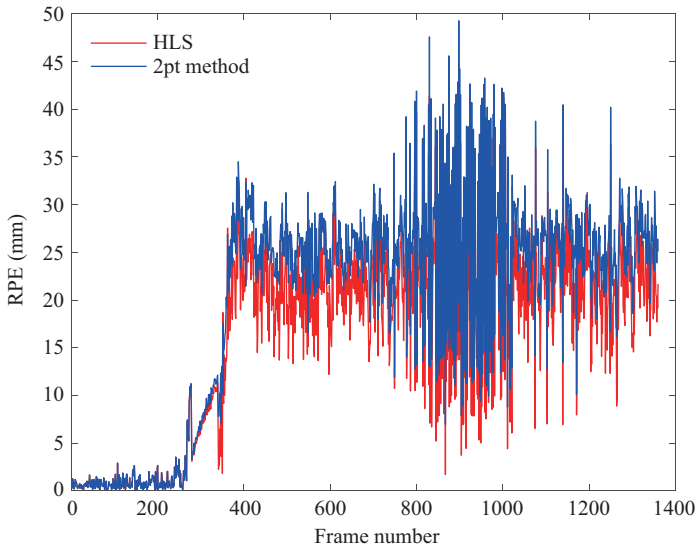

Figure 6 (Color online) Evaluating the trajectory drift errors via the RPE of the HLS method and the 2-point (2pt) method based on the ETH dataset.

absolute trajectory error (ATE) [29] as the absolute distance between the true and estimated trajectories in each frame:

$$\zeta_i = \left\| \boldsymbol{t}_{wi}^{gt} - \boldsymbol{t}_{wi} \right\|.$$

Similarly to the RPE, the RMSE over all frames is calculated as follows:

$$\mathrm{RMSE}\left(\zeta_{1:n}\right) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\zeta_i^2},$$

where n represents the number of consecutive frames.

The testing platform of our UAV experiments is an Intel Aero RTF Drone[1)] (see Figure 8(a)), flying at a height of 5 m (on average) above ground at approximately 0.6 m/s. The control frequency of the

1) Intel aero ready to fly drone. 2018. https://software.intel.com/enus/aero/drone-kit.

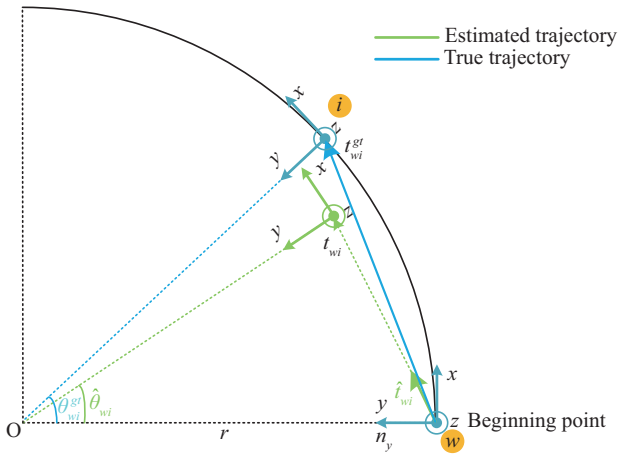

Figure 7 (Color online) Relationship between the estimated and true trajectories in outdoor experiment settings.

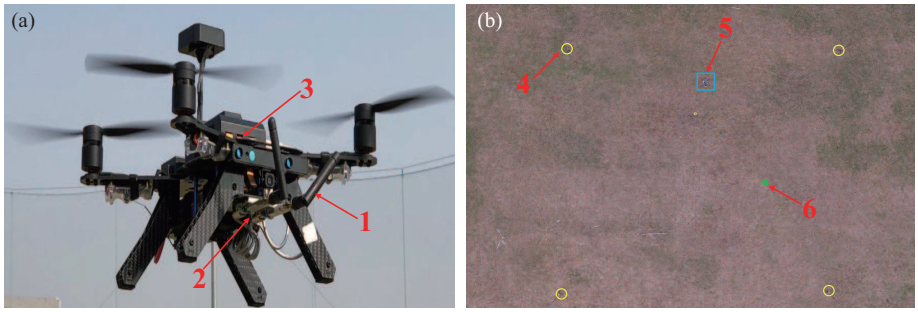

Figure 8 (Color online) Experimental platform (a) and its environment (b): (1) UWB receiver, (2) downward-looking camera, (3) flight controller with IMU, magnetometer, and other components, (4) four UWB anchors, (5) Intel drone, and (6) start point.

flight controller is 25 Hz and the yaw angle θ_{wi}^{gt} of the drone is measured by the airborne IMU and magnetometer. A downward-looking OmniVision OV7251 VGA camera[2)] with a resolution of 630×470 and an 80° field of view is attached to the bottom of the drone. The true position of the drone is revealed by an ultra-wideband (UWB) wireless localization system[3)] (see Figure 8(b)), which delivers stable and accurate positioning performance by virtue of its 50 Hz update frequency and positioning accuracy below 10 cm. The UWB anchors are fixed on tripods at the four vertices of a 15 m $\times$ 15 m square deployment region, and the UWB receiver is fixed on the drone (see Figure 8). In this setup, the $\boldsymbol{t}_{wi}^{gt}$ of the ith frame can be obtained by a calibrated rigid-body transformation and by synchronizing the timestamps between the UWB and the IMU module. Considering the size of the deployed UWB wireless localization system and the influence of the field of view and flight altitude, we set $r \approx 3.5$ m to achieve a balance between the sufficient number of overlaps and movements.

The top views of the estimated and true flight trajectories are qualitatively compared in Figure 9. The trajectory begins in the counterclockwise direction from the origin of the coordinate system. The estimated trajectory is in strong agreement with the real trajectory at the initial stage, but later drifts from the real trajectory as incremental errors accumulate in the rough visual odometry and the flight environment is interfered by breezes, textureless grass, and other disturbances. Note that the RMSE of the proposed method (0.7 m) is mainly caused by cumulative errors in the middle and late stages. Overall, the effectiveness of the proposed method is verified in practice.

5 Conclusion

We proposed a least squares solution to the relative pose estimation problem based on the homography error in the normalized image plane. Our method improves not only the accuracy of the estimated relative pose, but also the solvability of the homography scale factor. Furthermore, we developed an iterative

2) OmniVision. Ov7251. 2018. https://www.ovt.com/sensors/OV7251.
3) Ku Y. I-uwb-lps. 2017. http://www.inffuture.com/products/i-uwb-lps/.

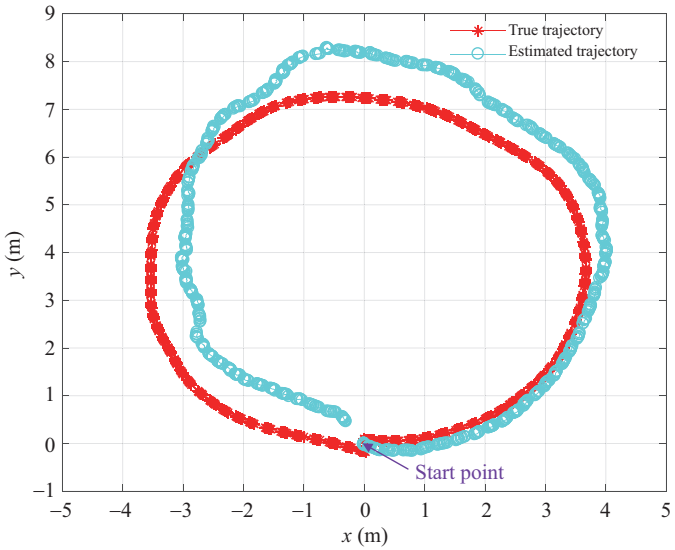

Figure 9 (Color online) Top views of the estimated and the true flight trajectories.

estimation method for the camera trajectory in visual odometry applications. Assuming a known gravity direction and a dominant ground-plane environment, the relative pose in the homography representation can be solved by linear least squares estimation, followed by a fast iterative Gauss-Newton optimization. The environmental assumptions of the proposed method are commonly met in ground-based UAVs and self-driving scenarios. Finally, synthetic as well as real data experiments in various metrics demonstrate that the proposed method not only outperforms the existing methods with respect to accuracy and robustness, but also is suitable for practical applications.

Acknowledgements This work was partially supported by National Natural Science Foundation of China (Grant Nos. 61603303, 61803309, 61703343), Natural Science Foundation of Shaanxi Province (Grant No. 2018JQ6070), China Postdoctoral Science Foundation (Grant No. 2018M633574), and Fundamental Research Funds for the Central Universities (Grant Nos. 3102019ZDHKY02, 3102018JCC003).

References

1 Mur-Artal R, Montiel J M M, Tardos J D. ORB-SLAM: a versatile and accurate monocular SLAM system. IEEE Trans Robot, 2015, 31: 1147–1163
2 Qu Y H, Zhang F, Wu X W, et al. Cooperative geometric localization for a ground target based on the relative distances by multiple UAVs. Sci China Inf Sci, 2019, 62: 010204
3 Liu S H, Wang S Q, Shi W H, et al. Vehicle tracking by detection in UAV aerial video. Sci China Inf Sci, 2019, 62: 024101
4 Nister D. An efficient solution to the five-point relative pose problem. IEEE Trans Pattern Anal Machine Intell, 2004, 26: 756–770
5 Kneip L, Siegwart R, Pollefeys M. Finding the exact rotation between two images independently of the translation. In: Proceedings of the European Conference on Computer Vision, 2012. 696–709
6 Hartley R, Zisserman A. Multiple View Geometry in Computer Vision. Cambridge: Cambridge University Press, 2003
7 Hartley R. In defense of the eight-point algorithm. IEEE Trans Pattern Anal Mach Intell, 1997, 19: 580–593
8 Li H, Hartley R, Kim J H. A linear approach to motion estimation using generalized camera models. In: Proceedings of the International Conference on Computer Vision and Pattern Recognition, 2008. 1–8
9 Kneip L, Lynen S. Direct optimization of frame-to-frame rotation. In: Proceedings of the International Conference on Computer Vision, 2013. 2352–2359
10 Fischler M A, Bolles R C. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Commun ACM, 1981, 24: 381–395
11 Kneip L, Furgale P. OpenGV: a unified and generalized approach to real-time calibrated geometric vision. In: Proceedings of the International Conference on Robotics and Automation, 2014. 1–8
12 Zhang S J, Cao X B, Zhang F, et al. Monocular vision-based iterative pose estimation algorithm from corresponding feature points. Sci China Inf Sci, 2010, 53: 1682–1696
13 Fraundorfer F, Tanskanen P, Pollefeys M. A minimal case solution to the calibrated relative pose problem for the case of two known orientation angles. In: Proceedings of the European Conference on Computer Vision, 2010. 269–282
14 Kalantari M, Hashemi A, Jung F, et al. A new solution to the relative orientation problem using only 3 points and the vertical direction. J Math Imag Vis, 2011, 39: 259–268
15 Lee G H, Pollefeys M, Fraundorfer F. Relative pose estimation for a multi-camera system with known vertical direction. In: Proceedings of the International Conference on Computer Vision and Pattern Recognition, 2014. 540–547
16 Naroditsky O, Zhou X S, Gallier J, et al. Two efficient solutions for visual odometry using directional correspondence. IEEE Trans Pattern Anal Mach Intell, 2012, 34: 818–824
17 Horn B K P, Hilden H M, Negahdaripour S. Closed-form solution of absolute orientation using orthonormal matrices. J Opt Soc Am A, 1988, 5: 1127–1135

18 Li H, Duan H B. Verification of monocular and binocular pose estimation algorithms in vision-based UAVs autonomous aerial refueling system. Sci China Technol Sci, 2016, 59: 1730–1738
19 Kneip L, Chli M, Siegwart R Y. Robust real-time visual odometry with a single camera and an IMU. In: Proceedings of the British Machine Vision Conference, 2011. 1–11
20 Bazin J, Li H D, Kweon I S, et al. A branch-and-bound approach to correspondence and grouping problems. IEEE Trans Pattern Anal Mach Intell, 2013, 35: 1565–1576
21 Guan B, Vasseur P, Demonceaux C, et al. Visual odometry using a homography formulation with decoupled rotation and translation estimation using minimal solutions. In: Proceedings of the International Conference on Robotics and Automation, 2018. 2320–2327
22 Saurer O, Vasseur P, Boutteau R, et al. Homography based egomotion estimation with a common direction. IEEE Trans Pattern Anal Mach Intell, 2017, 39: 327–341
23 Saurer O, Fraundorfer F, Pollefeys M. Homography based visual odometry with known vertical direction and weak Manhattan world assumption. In: Proceedings of IROS Workshop on Visual Control of Mobile Robots (ViCoMoR 2012), Vilamoura, 2012. 25–30
24 Urban S, Leitloff J, Hinz S. MLPNP–a real-time maximum likelihood solution to the perspective-n-point problem. ISPRS Ann Photogramm Remote Sens Spatial Inf Sci, 2016, 3: 131–138
25 Conrad D, Desouza G N. Homography-based ground plane detection for mobile robot navigation using a modified EM algorithm. In: Proceedings of the International Conference on Robotics and Automation, 2010. 910–950
26 Faessler M, Fontana F, Forster C, et al. Autonomous, vision-based flight and live dense 3D mapping with a quadrotor micro aerial vehicle. J Field Robotics, 2016, 33: 431–450
27 Lowe D G. Distinctive image features from scale-invariant keypoints. Int J Comput Vision, 2004, 60: 91–110
28 Zhao C, Fan B, Tian L, et al. Statistical optimization feature matching algorithm based on epipolar geometry (in Chinese). Acta Aeronautica et Astronaut Sin, 2018, 39: 158–166
29 Sturm J, Engelhard N, Endres F, et al. A benchmark for the evaluation of RGB-D SLAM slam systems. In: Proceedings of the International Conference on Intelligent Robots and Systems, 2012. 573–580
30 Huang A S, Bachrach A, Henry P, et al. Visual odometry and mapping for autonomous flight using an RGB-D camera. In: Proceedings of the International Conference on Robotics Research, 2017. 235–252