2-Entity RANSAC for robust visual localization in changing environment

Yanmei Jiao, Yue Wang, Bo Fu, Xiaqing Ding, Qimeng Tan, Lei Chen and Rong Xiong

Abstract— **Visual localization has attracted considerable attention due to its low-cost and stable sensor, which is desired in many applications, such as autonomous driving, inspection robots and unmanned aerial vehicles. However, current visual localization methods still struggle with environmental changes across weathers and seasons, as there is significant appearance variation between the map and the query image. The crucial challenge in this situation is that the percentage of outliers, i.e. incorrect feature matches, is high. In this paper, we derive minimal closed form solutions for 3D-2D localization with the aid of inertial measurements, using only 2 pairs of point matches or 1 pair of point match and 1 pair of line match. These solutions are further utilized in the proposed 2-entity RANSAC, which is more robust to outliers as both line and point features can be used simultaneously and the number of matches required for pose calculation is reduced. Furthermore, we introduce three feature sampling strategies with different advantages, enabling an automatic selection mechanism. With the mechanism, our 2-entity RANSAC can be adaptive to the environments with different distribution of feature types in different segments. Finally, we evaluate the method on both synthetic and real-world datasets, validating its performance and effectiveness in inter-session scenarios.**
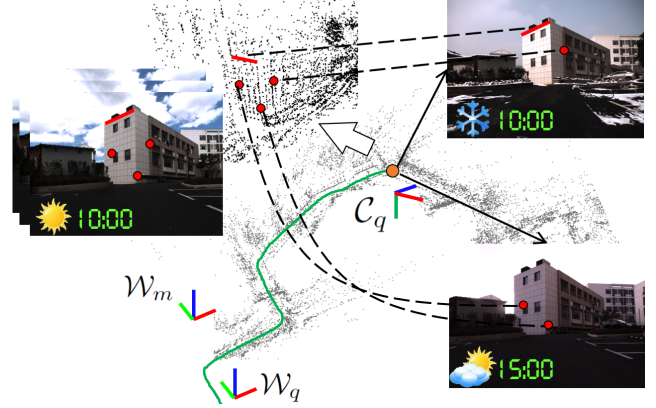

Fig. 1: The green line indicates the current session localizing in the black pre-built 3D map with 1 point 1 line or 2 points minimal solution. The map was built in summer, while the query sessions are under different weather or seasons.

I. INTRODUCTION

Localization is essential for autonomous robot navigation. Compared to the LiDAR-based localization methods, visual localization is favorable in many applications as the low-cost and stable cameras are widely available. A typical workflow of the visual localization task is to detect the feature points in the query image (FAST [1], SIFT [2]), match the image feature points to the map points (ORB [3], SIFT [2]), and estimate the pose based on a set of matches. Match outliers are hardly avoidable in this setting, leading to inaccurate pose estimation. Random sample consensus (RANSAC) [4] is a popular method to achieve robust estimation by randomly sampling the matching set and voting for the inliers. However, RANSAC is limited to serious appearance changes in the environment, in which the percentage of outliers may grow significantly [5] [6]. Therefore, reliable visual localization robust to the weather, illumination or seasonal changes remains a challenging problem.

Two factors are crucial for RANSAC to find inliers, i.e. the number of inliers in the set (higher is better), and the minimal number of matches required to estimate a pose (lower is

better). For the first factor, some previous works propose to increase the number of inliers by utilizing multiple types of features (e.g. points, circles and lines), as the feature defined on larger image regions usually leads to higher robustness against the illuminative variation. In this paper, we also consider multiple types of features following the practice of previous methods [7]. For the second factor, the well-known 3D-2D localization methods typically require 3 or 4 feature matches such as perspective-3-points (P3P) [8], efficient perspective-n-points (EPnP) [9], the minimal solutions of 3 lines [10] [11] or a mixture of 3 entities (i.e. points and lines) [7]. Our key observation of this paper is that the number of matches can be further reduced for both point and line features in robotics navigation scenario, leading to higher robustness against outliers.

Specifically, with the growing trend of visual inertial navigation in robots, the direction of gravity between the pre-built map and the current query image can be easily aligned. Therefore, we aim to reduce the number of point feature matches from 3 to 2 for closed form pose estimation, by making use of the directly observed pitch and roll angle. Furthermore, the closed form solution is also derived for the case of 1 point and 1 line, so that line feature matches can be considered. As a result, we propose a 2-entity RANSAC for persistent visual localization as illustrated in Fig. 1. In theory, as 2 observations (2D) can provide constraints for 4 degrees of freedom (DoF), the proposed solution should be minimal. Compared with the popular 2-point RANSAC [12]

Yanmei Jiao, Yue Wang, Bo Fu, Xiaqing Ding and Rong Xiong are with the State Key Laboratory of Industrial Control and Technology, Zhejiang University, Hangzhou, P.R. China. Qimeng Tan and Lei Chen are with the Beijing Key Laboratory of Intelligent Space Robotic System Technology and Applications, Beijing Institute of Spacecraft System Engineering, Beijing, P.R. China. Yue Wang is the corresponding author wangyue@iipc.zju.edu.cn. Rong Xiong is the co-corresponding author.

[13] which is useful for consecutive pose estimation in visual inertial odometry or simultaneous localization and mapping, our method focuses on 3D-2D matches without referring to measurement of relative yaw angle, making it applicable for localization. In summary, we present the contributions of the paper as follows: contribution

- We derive the closed form solutions to pose estimation using minimal 3D-2D feature matches, namely 2 pairs of point matches or 1 pair of point match and 1 pair of line match.
- We propose a 2-entity RANSAC by embedding the minimal closed form solutions into the RANSAC framework. We also proposed three sampling strategies in RANSAC that select different types of feature matches for pose estimation.
- We analyze the success probability for different sampling strategies in the 2-entity RANSAC, and propose a selection mechanism that adaptively select strategy depending on environmental characteristics, i.e. structured or unstructured.
- The proposed method is evaluated on both synthetic and multiple real world data sessions. The results validate the effectiveness and efficiency of our method on visual localization with seasonal and illuminative changes.

The remainder of this paper are organized as follows: In Section II we review some relevant visual pose estimation methods. Section III gives the overview of the proposed method and the derivation of the closed form solutions. The analysis about the criterion to select the sampling strategy in RANSAC is addressed in Section IV. We then evaluate our method in Section V on synthetic data and challenging real world data. Finally, the conclusion is drawn in Section VI, which completes this paper.

II. RELATED WORKS

The minimal solution with efficient 2D-2D point matches for camera pose estimation has been studied extensively for decades. There are many well-known methods, such as 8-point method for fundamental matrix estimation which yields a unique solution [14]. In addition, there are also 7-point [15], 6-point [16] and 5-point [17] methods which focus on further reduction of minimal number of point matches required for pose estimation. In recent years, with the development of visual inertial navigation, this number is further reduced by considering the measurement of angular velocity and the gravity. Based on this, a 3-point method was proposed in [18]. More recently, the 2-point RANSAC [12] [13] was proposed by employing the gyro to calculate the relative rotation within short duration, which was employed in several visual inertial odometry softwares. By further constraining the motion pattern of the camera, only 1 pair of point match is sufficient to derive the pose [19].

In localization, 3D-2D matches are available but there is no measurement of relative yaw angle. In this case, the minimal number of required feature matches is reduced compared with scenario of 2D-2D. Specifically, there are several 3-point methods [8] [20] [21] [22] [23] that give up to four

solutions in closed-form and other 3-line solvers [10] [11] that give up to eight solutions which require slower iterative methods. With the aid of inertial measurement, the 2-point RANSAC solution for 3D-2D matches is also investigated in [13]. But they utilized the relative measurement of yaw angle, thus cannot be applied in localization. As significant change can appear in long term, it's difficult to find enough point matches or line matches alone to guarantee the success rate of localization. Thus the methods utilizing both point and line features are studied in [7] including two possible cases: 2 points 1 line case yielding 4 solutions and 1 point 2 lines case yielding 8 solutions. However, to the best of our knowledge, the minimal solution for localization using 3D-2D point or line features with the aid of inertial measurements has not been studied yet, which is the focus of this paper.

III. MINIMAL SOLUTIONS

The definition of coordinates in this paper is shown in Fig. 1. A 3D map is built by running a visual inertial simultaneous localization and mapping (VI-SLAM), of which the world reference frame is denoted as $\mathcal{W}_m$. In the current query session, a visual inertial navigation system (VINS) is utilized with its origin defined as $\mathcal{W}_q$. Denoting the coordinates of the current query camera as $\mathcal{C}_q$, the localization problem is stated as the estimation of the query image pose with respect to the world, i.e. $T_{\mathcal{C}_q \mathcal{W}_m}$. Thanks to the observability of pitch and roll angle, $\tilde{\beta}$ and $\tilde{\gamma}$, given by the inertial measurements, the direction of gravity can be aligned between $\mathcal{W}_m$ and $\mathcal{W}_q$ [24], leading to the reduction of DoF in pose as

$$T_{\mathcal{C}_q \mathcal{W}_m} = [R_z(\alpha)R_y(\tilde{\beta})R_x(\tilde{\gamma})|(t_1 \quad t_2 \quad t_3)^T]$$

where α and (t_1, t_2, t_3) denote the yaw angle and three translations to be estimated. In sequel, for the purpose of clearance, we are going to estimate the inversion of the pose, i.e. $T_{\mathcal{W}_m \mathcal{C}_q}$. Note that the number of unknown DoFs in $T_{\mathcal{W}_m \mathcal{C}_q}$ remains the same.

To solve the 4DoF localization problem, we refer to two geometric constraints: collinearity of 3D-2D point matches and coplanarity of 3D-2D line matches. In an image, each pixel encodes a 3D projection ray. As shown in Fig. 2, according to the projection geometry, the optical center C of the camera, the map point P and its projection point D on the image lie on the same 3D line, which is denoted as $\{C, D, R_{\mathcal{W}_m \mathcal{C}_q}P + t_{\mathcal{W}_m \mathcal{C}_q}\}_L$. By solving the line equation with the first two points C and D, the third point can be substituted into the line equation to form two independent constraints. In addition, the camera center C, the map line $L_1 L_2$ and its projection line segment $D_1 D_2$ lie on the single plane π, which is denoted as $\{C, D_1, D_2, R_{\mathcal{W}_m \mathcal{C}_q}L_1 + t_{\mathcal{W}_m \mathcal{C}_q}\}_P$, and $\{C, D_1, D_2, R_{\mathcal{W}_m \mathcal{C}_q}L_2 + t_{\mathcal{W}_m \mathcal{C}_q}\}_P$. Similarly, by solving the plane equation from the first three points C, D_1 and D_2, the two end points of the map line can be substituted into it to form another two independent constraints. As a result, two pairs of non-degenerate feature matches should be sufficient to solve the 4DoF localization problem with three possible combinations: 1 point 1 line, 2 points and 2 lines in theory.
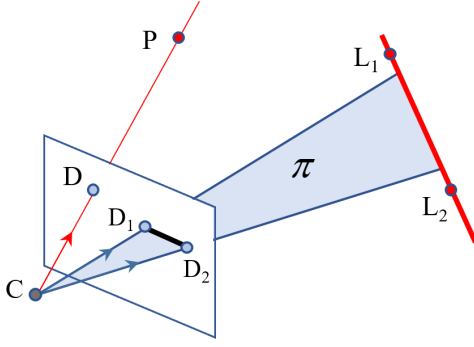
Fig. 2: The collinearity and coplanarity geometric constraints.

A. 1 point 1 line

In this section, we are going to solve the 4DoF localization problem given 1 pair of 3D-2D point match and 1 pair of 3D-2D line match. The solution is denoted as 1P1L. Given the 2D coordinates of the detected image feature point, we can compute the corresponding 3D projection ray according to the calibrated intrinsic parameters. The same process is applied to the two end points of the line segment to obtain two 3D projection rays. The 3D projection rays are expressed in the camera reference frame C_q, and the corresponding 3D coordinates of the features are expressed in the world reference frame W_m. To simplify the form of the unsolved equations, we introduce two intermediate reference frames denoted as C_1 and W_1 for camera and map respectively.

The choice of C_1: As shown in Fig. 3 (a), in the original camera reference frame C_q, the origin is the camera center C^0, the camera projection ray associated with the 2D point is given by its normalized direction vector $\vec{d_1}$, and projection rays of the two end points of the 2D line are given by normalized direction $\vec{d_2}$, $\vec{d_3}$.

In the intermediate camera frame C_1, the projection ray associated with the 2D point is denoted by CD_1, and the line, CD_2 and CD_3. Specifically, the intermediate reference frame C_1 should satisfy the following conditions:

- The new camera center C is $(0, 0, -1)$.
- One of the projection rays of the line end point CD_2 lies on the z axis such that $D_2 = (0, 0, 0)$.
- The other projection ray of the line end point CD_3 lies on the xz plane and the point D_3 is the intersection point between the x axis and the ray.
- The point D_1 in the projection ray of the point feature lies on the xy plane.

After the transformation, we have following results

$$C = \begin{bmatrix} 0 \\ 0 \\ -1 \end{bmatrix}, D_2 = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, D_3 = \begin{bmatrix} \tan(\arccos(\vec{d_2} \cdot \vec{d_3})) \\ 0 \\ 0 \end{bmatrix}$$

Then the points in the original camera frame are calculated as follows

$$C^0 = \mathbf{0}_{3 \times 1}, D_2^0 = C^0 + \vec{d_2}, D_3^0 = C^0 + \frac{\vec{d_3}}{\vec{d_2} \cdot \vec{d_3}}$$

The transformation $T_{C_q C_1}$ can be computed by transforming the three points (C^0, D_2^0, D_3^0) to (C, D_2, D_3). After that, the point $D_1 \triangleq (a_1, b_1, 0)$ can also be computed.

The choice of W_1: The transformation of the world reference is a translation which transforms the 3D point to the origin of W_1

$$T_{W_m W_1} = \begin{bmatrix} \mathbf{I}_{3 \times 3} & -P_1^0 \\ \mathbf{0}_{1 \times 3} & 1 \end{bmatrix}$$

Thus in W_1 we have:

$$P_1 = \mathbf{0}_{3 \times 1}, L_{i = \{2,3\}} \triangleq \begin{bmatrix} X_i & Y_i & Z_i \end{bmatrix}^T$$

Note that these points are all known.

Pose estimation between C_1 and W_1: Let us denote the rotation matrix and the translation matrix to be solved as R and t, that is $R \triangleq R_{W_1 C_1}$, $t \triangleq t_{W_1 C_1}$. According to the collinearity of $\{C, D_1, RP_1 + t\}_L$, the following equations are derived:

$$a_1 T_2 - b_1 T_1 = 0 \tag{1}$$

$$b_1 T_3 - T_2 = -b_1 \tag{2}$$

As for the coplanarity of $\{C, D_2, D_3, RL_2 + t\}_P$:

$$R_{21} X_2 + R_{22} Y_2 + R_{23} Z_2 + T_2 = 0 \tag{3}$$

And the coplanarity of $\{C, D_2, D_3, RL_3 + t\}_P$:

$$R_{21} X_3 + R_{22} Y_3 + R_{23} Z_3 + T_2 = 0 \tag{4}$$

where R_{mn} denotes the m-th row and n-th column entry of R, T_m denotes the m-th entry of t.

One should notice that the matrix R is determined only by α. Combining (3) and (4), we can solve α, which is then substituted in (1) - (4) for the translation t. Put all the equations together, the closed form localization problem can be solved as

$$T_{W_m C_q} = T_{C_q C_1}^T \cdot T_{W_1 C_1} \cdot T_{W_m W_1}$$

Degenerate cases: For 1 point 1 line case, if the point lies on the line, the corresponding case is degenerated.

B. 2 points

In this section, we are given 2 pairs of 3D-2D point matches, and the solution is denoted as 2P. In this case, the intermediate camera reference frame remains the same as the original one, thus the points in C_q and C_1 following the notations in 1P1L are

$$C = \mathbf{0}_{3 \times 1}, D_{i = \{1,2\}} = \begin{bmatrix} a_i & b_i & 1 \end{bmatrix}^T$$

where a_1, a_2, b_1, b_2 are known parameters computed using intrinsic parameters and normalized depth.

The transformation of the world reference is a translation which transforms one of the two 3D points to the origin of the intermediate world frame. Thus in W_1 we have:

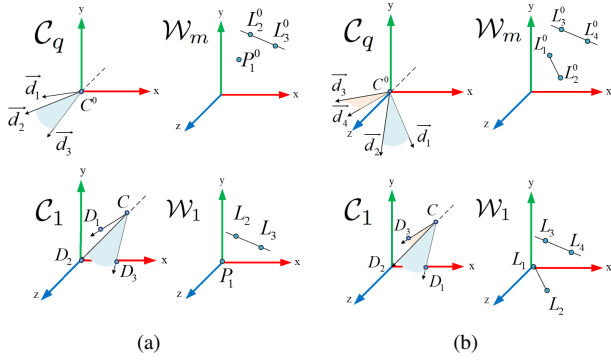$$P_1 = \mathbf{0}_{3 \times 1}, P_2 = \begin{bmatrix} X_2 & Y_2 & Z_2 \end{bmatrix}^T$$

Fig. 3: (a) The illustration of intermediate reference frame $\mathcal{C}_1$ and $\mathcal{W}_1$ for the 1 point and 1 line case. (b) The illustration of intermediate reference frame for the 2 lines case.

Pose estimation between $\mathcal{C}_1$ and $\mathcal{W}_1$: Following the notations in 1P1L, according to the collinearity of $\{C, D_1, RP_1 + t\}_L$, the following equations are derived:

$$a_1 T_2 - b_1 T_1 = 0 \tag{5}$$

$$a_1 T_3 - T_1 = 0 \tag{6}$$

As for the collinearity of $\{C, D_2, RP_2 + t\}_L$, we have:

$$a_2(R_{21}X_2 + R_{22}Y_2 + R_{23}Z_2 + T_2)$$
$$- b_2(R_{11}X_2 + R_{12}Y_2 + R_{13}Z_2 + T_1) = 0 \tag{7}$$

$$a_2(R_{31}X_2 + R_{32}Y_2 + R_{33}Z_2 + T_3)$$
$$- (R_{11}X_2 + R_{12}Y_2 + R_{13}Z_2 + T_1) = 0 \tag{8}$$

Combining (5) - (8), we can solve $T_{\mathcal{W}_1 \mathcal{C}_1}$. Thus the localization result can be obtained by

$$T_{\mathcal{W}_m \mathcal{C}_q} = T_{\mathcal{W}_1 \mathcal{C}_1} \cdot T_{\mathcal{W}_m \mathcal{W}_1}$$

C. 2 lines

In this section, we are given 2 pairs of 3D-2D line matches as shown in Fig. 3 (b).

The choice of $\mathcal{C}_1$: In the original camera reference frame $\mathcal{C}_q$, the camera center is denoted as C^0, the camera projection rays associated with the two 2D lines are given by pairs $(\overrightarrow{d_1}, \overrightarrow{d_2})$ and $(\overrightarrow{d_3}, \overrightarrow{d_4})$, respectively. In the intermediate camera frame $\mathcal{C}_1$, the 3D projection rays associated with the two 2D lines are represented by (CD_1, CD_2), and (CD_2, CD_3). Specifically, $\mathcal{C}_1$ should satisfy the following conditions:

- The new camera center C is $(0, 0, -1)$.
- The intersection line of the two interpretation planes represented by projection ray CD_2 lies on the z axis such that $D_2 = (0, 0, 0)$.
- The projection ray CD_1 lies on the xz plane and the point D_1 is the intersection point between the x axis and the ray.
- The point D_3 in the projection ray of one line lies on the xy plane.

The unit normal vectors of the two planes formed by $(C^0, \overrightarrow{d_1}, \overrightarrow{d_2})$ and $(C^0, \overrightarrow{d_3}, \overrightarrow{d_4})$ can be computed as follows

$$\overrightarrow{n_1} = \overrightarrow{d_1} \times \overrightarrow{d_2}, \overrightarrow{n_2} = \overrightarrow{d_3} \times \overrightarrow{d_4}, \overrightarrow{d_{12}} = \overrightarrow{n_1} \times \overrightarrow{n_2}$$

where $\overrightarrow{d_{12}}$ is the direction vector of the intersection line CD_2. After such a transformation, we have the following results

$$C = \begin{bmatrix} 0 \\ 0 \\ -1 \end{bmatrix}, D_1 = \begin{bmatrix} \tan(\arccos(\overrightarrow{d_1} \cdot \overrightarrow{d_{12}})) \\ 0 \\ 0 \end{bmatrix}, D_2 = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

Then the corresponding points in the original camera frame can be calculated as follows:

$$C^0 = \mathbf{0}_{3\times1}, D_1^0 = C^0 + \frac{\overrightarrow{d_1}}{\overrightarrow{d_1} \cdot \overrightarrow{d_{12}}}, D_2^0 = C^0 + \overrightarrow{d_{12}}$$

The transformation $T_{\mathcal{C}_q \mathcal{C}_1}$ can be computed by transforming the three points (C^0, D_1^0, D_2^0) to (C, D_1, D_2). After that, the point $D_3 \triangleq (a_1, b_1, 0)$ can also be computed.

The choice of $\mathcal{W}_1$: The transformation of the world reference is a translation which transforms one end point of the 3D line to the origin of the intermediate world frame. Thus in $\mathcal{W}_1$ we have

$$L_1 = \mathbf{0}_{3\times1}, L_{\{i=2,3,4\}} = \begin{bmatrix} X_i & Y_i & Z_i \end{bmatrix}^T$$

Pose estimation between $\mathcal{C}_1$ and $\mathcal{W}_1$: Following the notations in 1P1L, according to the coplanarity of $\{C, D_1, D_2, RL_1 + t\}_P$ and $\{C, D_1, D_2, RL_2 + t\}_P$, the following equations are derived

$$T_2 = 0 \tag{9}$$

$$R_{21}X_2 + R_{22}Y_2 + R_{23}Z_2 + T_2 = 0 \tag{10}$$

As for the coplanarity of $\{C, D_2, D_3, RL_3 + t\}_P$ and $\{C, D_2, D_3, RL_4 + t\}_P$:

$$a_1(R_{21}X_3 + R_{22}Y_3 + R_{23}Z_3 + T_2)$$
$$- b_1(R_{11}X_3 + R_{12}Y_3 + R_{13}Z_3 + T_1) = 0 \tag{11}$$

$$a_1(R_{21}X_4 + R_{22}Y_4 + R_{23}Z_4 + T_2)$$
$$- b_1(R_{11}X_4 + R_{12}Y_4 + R_{13}Z_4 + T_1) = 0 \tag{12}$$

From (9) - (12), we can easily find that the constraints are not sufficient to solve T_3. In fact, one of the constraints provided by IMU is coincident with one of constraints provided by coplanarity, thus the 2 lines case cannot solve the 4DoF localization problem.

IV. MODEL SELECTION

In the last section, we discussed the possible minimal closed form solutions of 4DoF localization problem when inertial measurements exist, and showed that the 1P1L and 2P are solvable. Based on the two minimal solutions, we propose the 2-entity RANSAC, in which there are three possible sampling strategies according to the corresponding solution: 1P1L, 2P and mixed. The mixed refers to using both 1P1L and 2P solutions according to the randomly selected features. Specifically, we first select a pair of point match
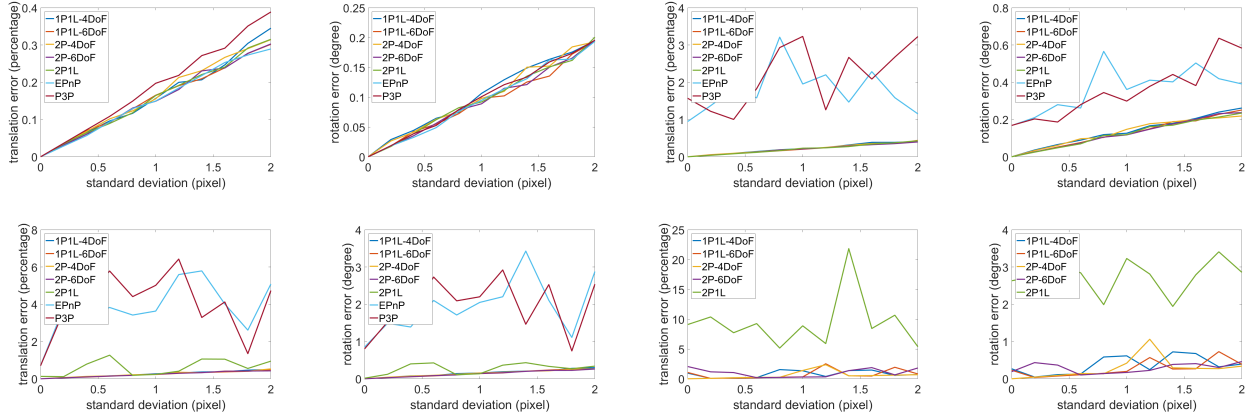
Fig. 4: Accuracy simulation results: the translation and rotation error for 10, 5, 4, 3 feature matches of different methods. Notice that EPnP and P3P are not computed in 3 matches case, for they all need at least four point matches in the scene.

and then randomly select another feature match from points and lines. If it is a point match, we perform 2P, if it is a line match, we perform 1P1L. In this section, we derive the success probability of the three sampling strategies to provide insights of sampling in 2-entity RANSAC.

Let's denote p as the number of pairs of point matches, l as the number of pairs of line matches, λ as the point inliers rate, and γ as the line inliers rate.

$$\frac{m}{p} = \lambda, \frac{n}{l} = \gamma \, (0 \le \lambda, \gamma \le 1), \frac{l}{p} = \varepsilon \, (0 \le \varepsilon \le l)$$

where m, n denote the point inliers number and line inliers number respectively. Then the success probability of different sampling strategies during one iteration in RANSAC is derived as follows:

$$P_{1p1l} = \frac{m}{p} \cdot \frac{n}{l} = \lambda \cdot \gamma$$

$$P_{2p} = \frac{m}{p} \cdot \frac{m-1}{p-1} = \lambda \cdot \frac{\lambda p - 1}{p-1}$$

$$P_{mixed} = \frac{m}{p} \cdot \frac{m+n-1}{p+l-1} = \lambda \cdot \frac{\lambda p + \gamma l - 1}{p+l-1}$$

Then we have

$$P_{1p1l} - P_{mixed} = \lambda(\gamma - \frac{\lambda p + \gamma l - 1}{p+l-1}) \propto \gamma - (\lambda - a)$$

$$P_{mixed} - P_{2p} = \lambda(\frac{\lambda p + \gamma l - 1}{p+l-1} - \frac{\lambda p - 1}{p-1}) \propto \gamma - (\lambda - a)$$

where $a = \frac{1-\lambda}{p-1}$.

Generally, $p - 1 \gg 1 - \lambda$ holds for most real world applications, which means a is a small positive number close to 0. Thus, the following conclusion can be derived:

$$\gamma \ge \lambda \Rightarrow P_{1p1l} > P_{mixed} > P_{2p}$$

Furthermore, with proper scaling, the following conclusion can also hold:

$$\gamma < \lambda \Rightarrow P_{1p1l} \le P_{mixed} \le P_{2p}$$

The percentage of line features and point features is highly related to the environment. According to the above discussion, we could find that the mixed sampling strategy is a more robust selection to the environmental variations as its performance is always the average one. Empirically, when localizing in the unstructured scene, there might be more point feature matches which indicates a higher point inlier rate than line features, leading to the choice of 2P sampling strategy. In the structured scene, although there are more line features compared with the unstructured scene, there are also many good point features. Considering the reconstruction error of the 3D lines can be higher than that of 3D points, the performance of 1P1L may not be significantly better than the 2P method in the structured scene. In order to be adaptive to the changing environment, an automatic selection mechanism is proposed by statistic estimation of the inlier rate using historical data. Specifically, the map is divided into structured or unstructured segments according to the inlier rate of point and line features. Then, when the robot is in the segment, the corresponding sampling strategy is selected, e.g. 2P in unstructured environment.

V. EXPERIMENTAL RESULTS

The proposed methods are evaluated on both synthetic and real data to demonstrate the advantages over other baseline methods in visual localization. We start by simulation experiments to illustrate the accuracy when the image feature is noisy, the sensitivity with respect to inaccurate pitch and roll angle, and the robustness against outliers. Then on real world data, we demonstrate the success rate comparison to show the effectiveness of the proposed methods and the model selection mechanism.

A. Results with synthetic data

In the simulation experiments, we generated some 3D points and lines in the cube $[-1, 1]^3$ and computed 2D projections for varying camera poses to get 3D-2D feature matches. For each method, 100 iterations of RANSAC are performed. The final identified inliers are sent to a nonlinear
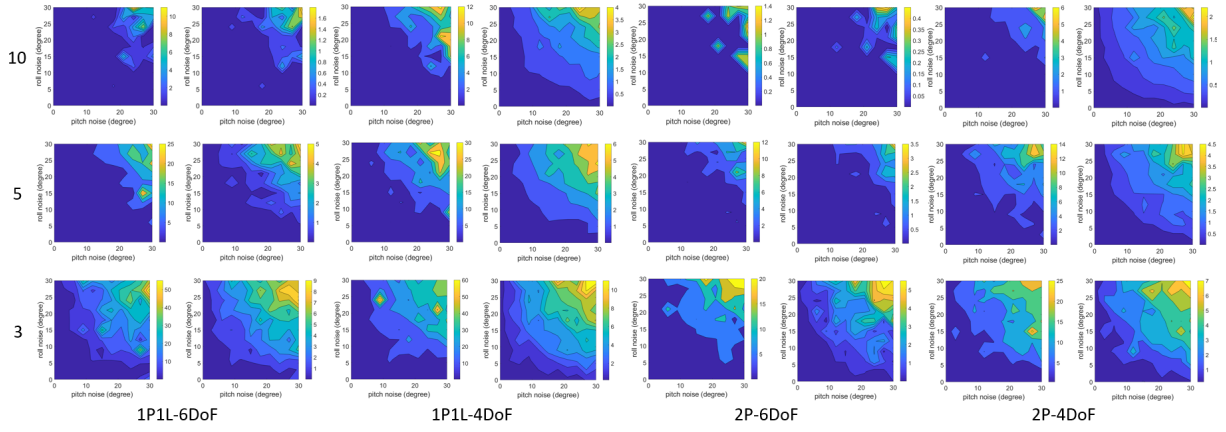
Fig. 5: Sensitivity simulation results: the translation and rotation error for 10, 5, 3 feature matches, respectively.

minimization for high accuracy. In optimization, there are two possible ways: 4DoF or 6DoF optimization. The 4DoF optimization means we fix the pitch and roll angle provided by VINS, and only optimize the other four variables in the pose. This may be useful when there are little inliers to optimize.

The minimal solutions evaluated in simulation experiments are EPnP [9], P3P [8], 2P1L [7] and our proposed solutions including 1P1L-6DoF, 1P1L-4DoF, 2P-6DoF, 2P-4DoF. The mixed-6DoF and mixed-4DoF are also evaluated in the robustness against outliers experiment. We exploited the translation and rotation error to evaluate the estimated result $[R|t]$ with respect to the ground truth $[R_{gt}|t_{gt}]$. Following [7], for translation error, we computed $\|t - t_{gt}\|/t_{gt}$ in percentage. For rotation error, we first computed the relative rotation $\triangle R = R R_{gt}^T$ and represented it in degrees.

1) Accuracy: To quantify the accuracy of different minimal solutions, we added Gaussian noise with zero mean and varying standard deviations for the 2D projections and changed the number of pairs of feature matches in four levels: 10, 5, 4, 3 (the 10 case means there are 10 pairs of point matches and 10 pairs of line matches in the scene). The results are shown in Fig. 4. We could find that when the feature matches are sufficient (see the 10 case), the proposed methods can achieve the same accuracy over other baseline methods as the standard deviation of the noise increases. While the number of pairs of feature matches decreases, the advantage of our 2-entity methods becomes bigger compared with 2P1L, EPnP and P3P. One can notice that when the number of feature matches decreases to 4, the error of our methods is slightly larger than that of the 10 matches case, while others grows severely.

2) Sensitivity: With the inertial sensor aided, we reduce the degree of the localization problem to 4 by leveraging the pitch and roll angle provided by the sensor. Therefore, it's necessary to investigate the influence of the quality of the two angles on the final accuracy. We added Gaussian noise with zero mean and varying standard deviations on the pitch and roll angle and studied the performance of the proposed methods in three feature matches levels: 10, 5, 3. The results

are shown in Fig. 5.

When there are enough feature matches (see the 10 case), the proposed methods can tolerate almost 25 degrees of noise on both pitch and roll angles. Empirically in real world application, the noise on pitch and roll angle is far less, thus the noise on pitch and roll should have limited impact on the accuracy. We also note that the 6DoF optimization outperforms the 4DoF in this case. However, as the number of pairs of feature matches decreases, the difference between the 6DoF and 4DoF methods becomes smaller. This is reasonable because when the inliers number is few, the error caused by the additional DoFs in optimization may be larger than that caused by noisy attitude estimation. Therefore, if there are very limited reliable feature matches when long term change occurs, 4DoF optimization is a good choice.

3) Robustness: To validate the robustness of the proposed methods, we designed a few experiments by varying the inlier rate from 0 to 80% with different number levels of inliers: 10, 5, 4, 3. Under each inlier number level, although the outlier rate is increasing, the number of inliers is unchanged. The outlier is generated by incorrectly associating the features in the original data. When the translation error of the estimated pose result is less than 10% and the rotation error is smaller than 5 degrees, the localization is assumed to be successful. We did 200 trails for each method to average the success rate. The results can be seen in Fig. 6. As expected, the proposed 2-entity RANSAC outperforms the other minimal solutions when the outlier rate increases, which is more obvious when the number of inliers decreases. Note that with a sufficient number of inliers, the proposed methods can achieve a success rate of more than 90% when the outlier rate is 80%.

B. Results with real data

For real world experiments, YQ-Dataset is utilized, which was collected in different weather conditions or seasons. Three of the datasets were collected during three days in summer 2017, denoted as 2017-0823, 2017-0827, 2017-0828, and the other was collected in winter 2018, denoted as 2018-0129. We select the 2017-0823 dataset to build the 3D map and utilize the other three datasets to evaluate the
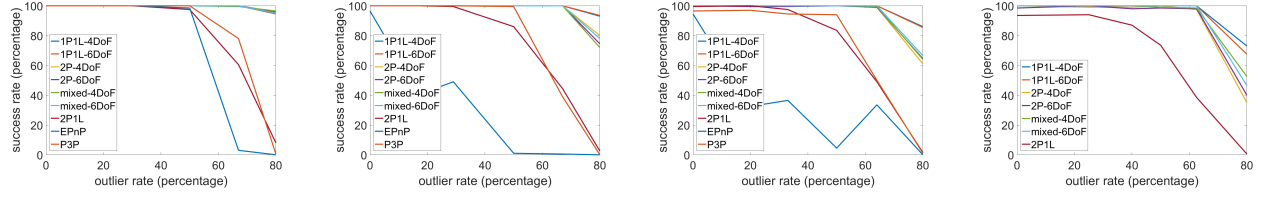
Fig. 6: Robustness simulation results: the success rate for 10, 5, 4, 3 inliers case of different methods.

TABLE I: Evaluation of the localization on the selected scene of YQ-Dataset.

	2017-0827		2017-0828		2018-0129	
	unstructured scene	structured scene	unstructured scene	structured scene	unstructured scene	structured scene
m	.25/.5/1.0/5.0	.25/.5/1.0/5.0	.25/.5/1.0/5.0	.25/.5/1.0/5.0	.25/.5/1.0/5.0	.25/.5/1.0/5.0
deg	2/5/8/10	2/5/8/10	2/5/8/10	2/5/8/10	2/5/8/10	2/5/8/10
1P1L-4DoF	0.0 / 1.9 / 3.8 / 15.0	7.9 / 23.7 / 42.1 / 81.6	1.9 / 9.4 / 15.0 / 29.4	13.0 / 32.6 / 54.3 / 87.0	0.0 / 0.0 / 0.0 / 3.8	5.0 / 15.5 / 35.5 / 85.0
1P1L-6DoF	1.9 / 5.7 / 8.8 / 18.8	29.5 / 48.6 / 78.4 / 91.9	6.3 / 13.1 / 21.9 / 35.0	39.1 / 67.4 / 82.6 / 91.3	0.0 / 0.0 / 0.0 / 4.9	25.0 / 45.0 / 65.0 / 90.0
2P-4DoF	2.5 / 11.3 / 18.8 / 36.3	20.5 / 31.8 / 50.0 / 70.5	11.3 / 29.4 / 42.5 / 56.3	21.7 / 52.2 / 67.4 / 93.5	0.0 / 0.0 / 1.0 / 6.8	10.0 / 20.0 / 30.0 / 80.0
2P-6DoF	6.3 / 15.0 / 24.4 / 40.0	31.8 / 57.9 / 83.9 / 86.8	25.0 / 38.1 / 48.8 / 58.1	45.7 / 69.6 / 91.3 / 95.7	1.0 / 1.0 / 1.0 / 9.7	25.0 / 40.0 / 65.0 / 95.0
mixed-4DoF	4.4 / 13.8 / 26.9 / 44.4	13.6 / 31.8 / 47.7 / 72.7	10.6 / 30.0 / 37.5 / 56.9	21.7 / 43.5 / 63.0 / 95.7	0.0 / 0.0 / 1.0 / 7.8	5.0 / 25.0 / 35.0 / 80.0
mixed-6DoF	6.9 / 15.0 / 23.4 / 38.8	31.8 / 63.2 / 78.9 / 89.6	25.6 / 36.9 / 48.1 / 58.1	45.7 / 73.9 / 93.5 / 97.8	1.0 / 1.0 / 1.0 / 7.8	20.0 / 55.0 / 70.0 / 95.0
P3P	1.9 / 11.9 / 20.0 / 37.5	29.5 / 50.0 / 73.7 / 86.8	16.3 / 26.9 / 39.4 / 52.5	30.4 / 63.0 / 76.1 / 91.3	0.0 / 0.0 / 1.0 / 2.9	10.0 / 40.0 / 60.0 / 90.0
EPnP	4.4 / 11.9 / 20.6 / 35.7	27.3 / 55.3 / 81.6 / 89.5	16.9 / 31.3 / 40.6 / 56.3	26.1 / 65.2 / 84.8 / 93.5	0.0 / 0.0 / 1.0 / 2.9	20.0 / 40.0 / 65.0 / 90.0

localization performance. The ground truth relative pose is provided by aligning the synchronized 3D LiDAR scans. The evaluated methods are EPnP, P3P, and our methods including 1P1L-4/6DoF, 2P-4/6DoF, mixed-4/6DoF. To get the 3D-2D feature matches between the query image and the map, we exploited the following three steps:

- Obtain the camera poses and the 3D-2D point matches in the map using visual inertial SLAM software [25].
- Run Line3D++ algorithm [26] to get the 3D-2D line matches in the map with the camera poses obtained in the first step as input.
- For the query session, we get the 3D-2D points/lines match based on the descriptors of LibVISO2 [27] and LBD [28], which are fed to the RANSAC with different solutions.
- For the pitch and roll angle of the query image, we directly use the observed corresponding values of the query camera frame relative to the query world frame.

The evaluation for real world experiments follows [29]. We first counted the correctly localized query images out of all keyframes in the query session using different methods, which can be seen in Tab. II. As expected, the performance of methods with 6DoF optimization are still better than 4DoF optimization. Besides, the 2-entity RANSAC is better than RANSAC with P3P and EPnP which calls for more number of matches. The proposed 2P and mixed are obviously better than EPnP and P3P, which indicates the robustness of the proposed methods when dealing with different weathers and seasons. Note that 1P1L is not as good as point features based methods, of which the reason is that there are many trees on both sides of the road in the map environment so that point features are far more abundant than line features.

In order to make a fair comparison between the 1P1L and the 2P methods, we manually picked some structured and unstructured segments from the map, in which some images

are shown in Fig. 7. There are around 100 places selected in unstructured segments, and 50 places in structured segments. The success rates of localization with multiple methods on the selected segments are show in Tab. I. The results show that the 2P method performs best in the unstructured scene and the performance of 1P1L is similar with 2P in the structured scene but obviously worse than that of 2P in unstructured segments, confirming the fact that the feature distribution in the environment has unavoidable effect on the method selection. Besides, from the results of the structured scene in 2018-0129, we can find that when outlier percentage grows due to the changing season, the better robustness of line feature matches promotes the performance of 1P1L method. The mixed-6DoF, as derived in theory, gives relatively stable performance compared with 2P and 1P1L in all segments.

To further verify the correctness of the model selection mechanism, we did an extra leave-one-out experiment on the 2017-0828 dataset. According to the localization results of 2017-0827 and 2018-0129 dataset on the 2017-0823 map, we automatically labeled the whole dataset with structured and unstructured segments by the inlier rate of the point and line features acquired by the estimated pose using 2P-6DoF. If the inlier rate of lines was higher than the points, 1P1L is utilized in that segment, otherwise, 2P. After that, we counted the success rate on the whole 2017-0828 dataset again and calculate the AUC value of each method, which is shown in Tab. III. Results show that with the selection mechanism, the performance is further promoted as expected.

VI. CONCLUSIONS

In this paper, the visual localization problem is reduced to 4DoF with the aid of inertial measurements. Based on this, the unified 2-entity RANSAC pipeline is developed and validated on both synthetic and real data to demonstrate the robustness and efficiency of the methods. Furthermore,

TABLE II: Evaluation on the whole YQ-Dataset.

m deg	2017-0827 .25/.5/1.0/5.0 2/5/8/10	2017-0828 .25/.5/1.0/5.0 2/5/8/10	2018-0129 .25/.5/1.0/5.0 2/5/8/10
1P1L-4DoF	2.1 / 5.2 / 9.7 / 21.5	2.6 / 7.6 / 14.3 / 26.1	2.5 / 6.0 / 10.6 / 21.8
1P1L-6DoF	6.3 / 10.2 / 15.3 / 24.3	7.7 / 14.4 / 20.4 / 28.6	6.0 / 9.3 / 12.5 / 23.7
2P-4DoF	3.6 / 10.0 / 17.2 / 31.1	8.6 / 19.3 / 28.0 / 39.3	0.4 / 5.9 / 12.6 / 25.4
2P-6DoF	9.3 / 17.1 / 24.3 / 34.4	15.7 / 25.3 / 32.5 / 39.8	5.9 / 11.5 / 16.6 / 27.3
mixed-4DoF	5.7 / 14.4 / 22.4 / 34.1	7.1 / 17.3 / 25.6 / 35.5	3.9 / 9.3 / 15.6 / 28.5
mixed-6DoF	12.1 / 18.7 / 24.4 / 33.3	13.7 / 21.9 / 28.5 / 35.1	7.3 / 12.6 / 17.0 / 27.7
P3P	6.7 / 14.7 / 21.3 / 32.2	10.1 / 19.8 / 28.2 / 37.1	4.3 / 9.7 / 12.9 / 22.8
EPnP	7.6 / 16.6 / 23.5 / 32.3	9.6 / 20.3 / 28.4 / 38.4	5.3 / 11.1 / 15.0 / 23.8

TABLE III: The automatic model selection experiment.

Method	1P1L 4DoF	1P1L 6DoF	2P 4DoF	2P 6DoF	mixed 4DoF	mixed 6DoF	selection
AUC	0.51	0.58	0.78	0.81	0.71	0.72	0.84

Fig. 7: Some examples of the unstructured scene with more point features, and the structured scene with more line features. The selected scenes are labeled manually according to experience.

we have investigated the proper mechanism to select the appropriate sampling strategy and showed the practicability in the real world application.

REFERENCES

[1] E. Rosten and T. Drummond, "Machine learning for high-speed corner detection," in *European conference on computer vision*, pp. 430–443, Springer, 2006.

[2] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[3] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "Orb: An efficient alternative to sift or surf," 2011.

[4] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[5] X. Ding, Y. Wang, D. Li, L. Tang, H. Yin, and R. Xiong, "Laser map aided visual inertial localization in changing environment," *arXiv preprint arXiv:1803.01104*, 2018.

[6] L. Tang, Y. Wang, X. Ding, H. Yin, R. Xiong, and S. Huang, "Topological local-metric framework for mobile robots navigation: a long term perspective," *Autonomous Robots*, pp. 1–15, 2018.

[7] S. Ramalingam, S. Bouaziz, and P. Sturm, "Pose estimation using both points and lines for geo-localization," in *ICRA 2011-IEEE International Conference on Robotics and Automation*, pp. 4716–4723, IEEE Computer Society, 2011.

[8] X.-S. Gao, X.-R. Hou, J. Tang, and H.-F. Cheng, "Complete solution classification for the perspective-three-point problem," *IEEE transactions on pattern analysis and machine intelligence*, vol. 25, no. 8, pp. 930–943, 2003.

[9] V. Lepetit, F. Moreno-Noguer, and P. Fua, "Epnp: Efficient perspective-n-point camera pose estimation," *International Journal of Computer Vision*, vol. 81, no. 2, pp. 155–166, 2009.

[10] M. Dhome, M. Richetin, J.-T. Lapreste, and G. Rives, "Determination of the attitude of 3d objects from a single perspective view," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, no. 12, pp. 1265–1278, 1989.

[11] H. H. Chen, "Pose determination from line-to-plane correspondences: Existence condition and closed-form solutions," in *Computer Vision, 1990. Proceedings, Third International Conference on*, pp. 374–378, IEEE, 1990.

[12] C. Troiani, A. Martinelli, C. Laugier, and D. Scaramuzza, "2-point-based outlier rejection for camera-imu systems with applications to micro aerial vehicles," in *Robotics and Automation (ICRA), 2014 IEEE International Conference on*, pp. 5530–5536, IEEE, 2014.

[13] L. Kneip, M. Chli, and R. Y. Siegwart, "Robust real-time visual odometry with a single camera and an imu," in *Proceedings of the British Machine Vision Conference 2011*, British Machine Vision Association, 2011.

[14] A. Gruen and T. S. Huang, *Calibration and orientation of cameras in computer vision*, vol. 34. Springer Science & Business Media, 2013.

[15] R. Hartley and A. Zisserman, *Multiple view geometry in computer vision*. Cambridge university press, 2003.

[16] J. Philip, "A non-iterative algorithm for determining all essential matrices corresponding to five point pairs," *The Photogrammetric Record*, vol. 15, no. 88, pp. 589–599, 1996.

[17] D. Nistér, "An efficient solution to the five-point relative pose problem," *IEEE transactions on pattern analysis and machine intelligence*, vol. 26, no. 6, pp. 756–770, 2004.

[18] F. Fraundorfer, P. Tanskanen, and M. Pollefeys, "A minimal case solution to the calibrated relative pose problem for the case of two known orientation angles," in *European Conference on Computer Vision*, pp. 269–282, Springer, 2010.

[19] D. Scaramuzza, F. Fraundorfer, and R. Siegwart, "Real-time monocular visual odometry for on-road vehicles with 1-point ransac," in *2009 IEEE International Conference on Robotics and Automation*, pp. 4293–4299, Ieee, 2009.

[20] B. M. Haralick, C.-N. Lee, K. Ottenberg, and M. Nölle, "Review and analysis of solutions of the three point perspective pose estimation problem," *International journal of computer vision*, vol. 13, no. 3, pp. 331–356, 1994.

[21] L. Kneip, D. Scaramuzza, and R. Siegwart, "A novel parametrization of the perspective-three-point problem for a direct computation of absolute camera position and orientation," 2011.

[22] P. Wang, G. Xu, Z. Wang, and Y. Cheng, "An efficient solution to the perspective-three-point pose problem," *Computer Vision and Image Understanding*, vol. 166, pp. 81–87, 2018.

[23] T. Ke and S. I. Roumeliotis, "An efficient algebraic solution to the perspective-three-point problem," *arXiv preprint arXiv:1701.08237*, 2017.

[24] M. Li and A. I. Mourikis, "High-precision, consistent ekf-based visual-inertial odometry," *The International Journal of Robotics Research*, vol. 32, no. 6, pp. 690–711, 2013.

[25] R. Mur-Artal and J. D. Tardós, "Visual-inertial monocular slam with map reuse," *IEEE Robotics and Automation Letters*, vol. 2, no. 2, pp. 796–803, 2017.

[26] M. Hofer, M. Maurer, and H. Bischof, "Efficient 3d scene abstraction using line segments," *Computer vision and image understanding*, vol. 157, pp. 167–178, 2017.

[27] A. Geiger, J. Ziegler, and C. Stiller, "Stereoscan: Dense 3d reconstruction in real-time," in *Intelligent Vehicles Symposium (IV)*, 2011.

[28] L. Zhang and R. Koch, "An efficient and robust line segment matching approach based on lbd descriptor and pairwise geometric consistency," *Journal of Visual Communication and Image Representation*, vol. 24, no. 7, pp. 794–805, 2013.

[29] P.-E. Sarlin, C. Cadena, R. Siegwart, and M. Dymczyk, "From coarse to fine: Robust hierarchical localization at large scale," *arXiv preprint arXiv:1812.03506*, 2018.