2-Entity Random Sample Consensus for Robust Visual Localization: Framework, Methods and Verifications

Yanmei Jiao, Yue Wang, *Member, IEEE,* Xiaqing Ding, Bo Fu,
Shoudong Huang, *Member, IEEE,* and Rong Xiong, *Member, IEEE*

***Abstract*—Robust and efficient visual localization is essential for numerous robotic applications. However, it remains a challenging problem especially when significant environmental or perspective changes present, as there are high percentage of outliers, i.e. incorrect feature matches, between the query image and the map. In this paper, we propose a novel 2-entity RANSAC framework using 3D-2D point and line feature matches for visual localization with the aid of inertial measurements and derive minimal closed form solutions using only 1 point 1 line or 2 point matches for both monocular and multi-camera system. The proposed 2-entity RANSAC can achieve higher robustness against outliers as multiple types of features are utilized and the number of matches needed to compute a pose is reduced. Furthermore, we propose a learning-based sampling strategy selection mechanism and a feature scoring network to be adaptive to different environmental characteristics such as structured and unstructured. Finally, both simulation and real-world experiments are performed to validate the robustness and effectiveness of the proposed method in scenarios with long-term and perspective changes[1].**

***Index Terms*—Camera pose estimation, random sample consensus (RANSAC), robust localization.**

I. Introduction

LOCALIZATION is a fundamental capability for mobile robots with applications to driverless cars, unmanned aerial vehicles and so on [1] [2]. Visual localization attracts more attention as the cameras are low-cost, lightweight and versatile compared with Light Detection and Ranging (LiDAR). The general idea of visual localization is to recover the translation and rotation of the query camera based on feature matches using various descriptors (e.g. FAST [3], SIFT [4], ORB [5], and some learning-based descriptors [6] [7] [8] etc.). Outliers are unavoidable between the feature matches, which severely affect the accuracy of the estimated pose. The typical way to achieve robust estimation against outliers is random sample consensus (RANSAC) [9]. However, with the serious appearance changes in long-term localization [10] [11], RANSAC is neither reliable nor efficient as the outlier rate is high. There are two factors affecting the success probability of RANSAC. One is the number and rate of inliers, i.e. correct feature matches (higher is better), and the other is the minimal number of matches needed to compute the pose (smaller is better). Much effort has been paid to improve the success rate of RANSAC and in result advance the robustness of the localization.

For the first, number and rate of inliers, previous works utilize various types of features simultaneously (e.g. points, lines and planes) [12] [13]. There are also works exploiting the observation of multiple views in the multi-camera system to increase the number of inliers [14] [15]. However, these methods in fact utilize independent feature types or features from independent cameras. Thus the actual inlier number does not increase. For the second factor, people are searching for minimal solutions using minimal number of 3D-2D feature matches to estimate camera pose. Specifically, there are methods using 4 points [16], 3 points [17] [18] and 3 lines [19]. The minimal solutions combining point and line matches are proposed in [20]. By making use of inertial measurements, the minimal number can be further reduced. Thus the 2-point minimal solutions are studied in [21] and [22]. But [21] utilizes the relative measurement of yaw angle which is only available between two consecutive images, thus cannot be applied in localization. And [22] only considers point matches which is sensitive to appearance variations.

In this paper, we combine point and line features from multiple camera views and propose a mixed sampling strategy to utilize all features jointly for pose estimation as illustrated in Fig. 1. In addition, we propose a feature scoring network to effectively improve the inlier rate during feature sampling. Furthermore, we reduce the number of both point and line matches from 3 to 2 by aligning the direction of gravity with the aid of inertial measurements. Note that most existing methods solve 3D-2D problem in two steps, calculating the 3D coordinates of the features in camera frame and estimating the pose by 3D-3D registration [17] [16]. It is hard to introduce

Manuscript received August 21, 2019; revised December 8, 2019 and February 20, 2020; accepted March 23, 2020. This work was supported in part by the National Key R&D Program of China (2018YFB1309300), and in part by the National Nature Science Foundation of China (61903332, U1609210). (Corresponding author: Yue Wang.)

Yanmei Jiao, Yue Wang, Xiaqing Ding, Bo Fu and Rong Xiong are with the Department of Control Science and Engineering, State Key Laboratory of Industrial Control and Technology, Zhejiang University, Hangzhou, P.R. China (e-mail: ymjiao@zju.edu.cn, wangyue@iipc.zju.edu.cn, xqding@zju.edu.cn, fubofb@zju.edu.cn, rxiong@zju.edu.cn).

Shoudong Huang is with the Centre for Autonomous Systems, Faculty of Engineering and IT, University of Technology, Sydney, Australia (e-mail: Shoudong.Huang@uts.edu.au).

[1] https://youtu.be/Zqgxntz11hI

inertial measurements. Therefore it is non-trivial to present the minimal closed form solutions exploiting only 2 entities for both monocular and multi-camera system. Embedding these methods into RANSAC framework, we propose a 2-entity RANSAC for long-term robust visual localization. The main contributions of this paper are summarized as follows:

- We derive the minimal closed form solutions to 3D-2D pose estimation utilizing 1 point 1 line or 2 point matches with the aid of inertial measurements for both monocular and multiple camera system.
- We propose a robot visual localization system by embedding the solutions into RANSAC and propose two modules including a learning-based selection mechanism and feature scoring network to complete the system.
- The effectiveness and efficiency of the proposed method are verified on both synthetic and challenging real world sessions with seasonal and perspective changes.
- The source code of proposed monocular and multiple camera algorithms are available on github[2] which is a contribution to the community for comparative study.

This paper completes our previous work [23] by generalizing the minimal solutions from mono-camera system to multi-camera system. We also extend the strategy selection mechanism to an end-to-end learning-based method and present a new feature scoring network to perform a weighted 2-entity RANSAC. More thorough simulation and real world experiments for both monocular and multiple camera system are designed to demonstrate the practicability of the proposed method. The rest of this paper is organized as follows. In Section II we give a detailed statement of the visual localization problem and the framework of the proposed method. Section III gives the derivation of the closed form minimal solutions for monocular and multi-camera system. The description about the learning-based sampling strategy selection mechanism and the feature scoring method is given in Section IV. Then the synthetic and real world verifications are presented in Section V. Finally, we summarize the conclusion in Section VI.

II. Problem Statement and Framework

Visual localization considered in this paper refers to estimating the 6DoF (degrees of freedom) camera pose using 3D-2D feature matches between the query image and the pre-built map. The map consists of map images and 3D point and line features which are reconstructed from the visual features during map building. Then the visual localization with the aid of inertial measurements can be formulated as follows.

The 3D map is built by running a visual inertial simultaneous localization and mapping (VI-SLAM) and its reference frame is denoted as $\mathcal{W}_m$. A visual inertial navigation system (VINS) is performed in the current query session of which the origin is defined as $\mathcal{W}_q$. Denoting the reference frame of the query camera as $\mathcal{C}_q$, the visual localization problem is to estimate the pose of the query camera in the world reference frame, i.e. $T_{\mathcal{W}_m\mathcal{C}_q}$. With the aid of inertial measurements, the direction of gravity between $\mathcal{W}_m$ and $\mathcal{W}_q$ can be easily aligned, thus the pitch and roll angle of $T_{\mathcal{W}_m\mathcal{C}_q}$ are the same as the corresponding measurements of $T_{\mathcal{W}_q\mathcal{C}_q}$ denoted as $\tilde{\beta}$ and $\tilde{\gamma}$. Thanks to the global observability of pitch and roll, the localization problem can be expressed in 4DoF as

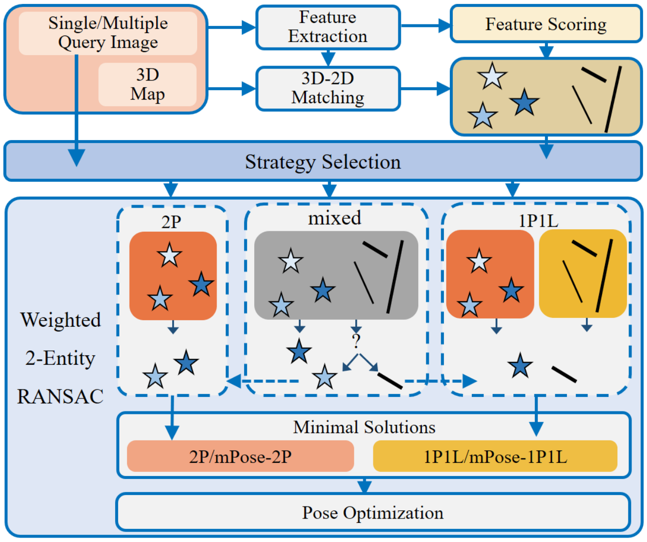


Fig. 1: The framework of 2-entity RANSAC localization.

$$T_{\mathcal{W}_m\mathcal{C}_q} = [R_z(\alpha)R_y(\tilde{\beta})R_x(\tilde{\gamma})|(T_1 \quad T_2 \quad T_3)^T] \tag{1}$$

where α and (T_1, T_2, T_3) represent the unsolved yaw angle and translation. In the rest of this paper, we intend to solve the inversion of the pose, i.e. $T_{\mathcal{C}_q\mathcal{W}_m}$ of which the unknown parameters remains the same. As one 3D-2D feature match can provide two independent constraints, the problem can be solved using two non-degenerate feature matches, which derives the 2-entity minimal solutions.

To deal with outliers, we embed the proposed minimal solutions to RANSAC [9] framework, which can give a robust result by sampling elements in the input dataset and voting for inliers. Let P be the success rate of RANSAC within k iterations, and w be the inlier rate. Assuming that the minimal number of elements to compute the model is n, we have

$$1 - P = (1 - w^n)^k \tag{2}$$

from which we could find that, with the fixed number of iterations, the higher inlier rate and the less minimal number of elements needed to solve the model, the higher success rate RANSAC can achieve. In this paper, we reduce the number n from 3 to 2 with the aid of inertial measurements and exploit a feature scoring method to improve the inlier rate w in a sample of RANSAC. Furthermore, using the multi-camera minimal solutions to combine observations from multiple views and the strategy selection mechanism to be adaptive to different environmental variations can indeed increase the inlier number. Therefore, the proposed visual-inertial localization method can get greater robustness against high percentage of outliers.

The framework of the proposed visual localization is illustrated in Fig. 1. Given the extracted 3D-2D point and line matches, we first score the feature points with the trained network and apply the score as the weight of the feature matches in sampling. Then we judge the characteristic of the

[2] https://github.com/slinkle/2-Entity-RANSAC

query image through a pre-trained convolution neural network to select the proper sampling strategy in RANSAC. With the scored feature matches and the selected sampling strategy, a weighted 2-entity RANSAC is performed. After the nonlinear pose optimization with the final point and line inliers, we get the localization result.

III. Minimal Solutions

In this section, the minimal solutions for both monocular and multi-camera system are derived. We refer to the collinearity of point matches and coplanarity of line matches to solve the 4DoF localization problem. According to the projection geometry shown in $\mathcal{C}_q$ of Fig. 2, the map point P_1^0, its projection point D_1^0 and the optical center C^0 of the camera are collinear, denoted as $\{C^0, D_1^0, R_{\mathcal{C}_q\mathcal{W}_m}P_1^0 + t_{\mathcal{C}_q\mathcal{W}_m}\}_L$. By substituting the third point into the line equation computed by the first two points, two independent constraints can be derived. In addition, the map line $L_2^0L_3^0$, its projection line segment $D_2^0D_3^0$ and the camera center C^0 are coplanar, denoted as $\{C^0, D_2^0, D_3^0, R_{\mathcal{C}_q\mathcal{W}_m}L_2^0 + t_{\mathcal{C}_q\mathcal{W}_m}\}_P$, and $\{C^0, D_2^0, D_3^0, R_{\mathcal{C}_q\mathcal{W}_m}L_3^0 + t_{\mathcal{C}_q\mathcal{W}_m}\}_P$. Similarly, by substituting the two end points of the map line into the plane computed with the first three points, another two independent constraints can be derived. Therefore, the 4DoF visual localization problem can be solved using two non-degenerate feature matches with three combinations: 1 point 1 line, 2 points and 2 lines.

A. Monocular Camera System

1) 1 point 1 line: This subsection presents the minimal solution using one point and one line match, which denoted as 1P1L. For simplification, two intermediate reference frames are introduced for camera and map, denoted as $\mathcal{C}_1$ and $\mathcal{W}_1$ respectively.

The choice of $\mathcal{C}_1$: The detailed illustration and computation of $T_{\mathcal{C}_1\mathcal{C}_q}$ is provided in the Supplementary Material [24]. As shown in Fig. 2, in $\mathcal{C}_q$, the camera center is C^0, and the image feature point is D_1^0. The end points of the image line segments are D_2^0, D_3^0. After transformation, in $\mathcal{C}_1$, the corresponding points are expressed as follows.

$$C = \begin{bmatrix} 0 \\ 0 \\ -1 \end{bmatrix}, D_1 \triangleq \begin{bmatrix} a_1 \\ a_2 \\ 0 \end{bmatrix}, D_2 = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, D_3 \triangleq \begin{bmatrix} a_3 \\ 0 \\ 0 \end{bmatrix}$$

Note that the parameters are all known which can be computed using the transformation $T_{\mathcal{C}_1\mathcal{C}_q}$.

The choice of $\mathcal{W}_1$: In $\mathcal{W}_m$, as shown in Fig. 2, the map point is P_1^0 and the end points of the map line are L_2^0, L_3^0. The transformation $T_{\mathcal{W}_1\mathcal{W}_m}$ is designed to transform the P_1^0 to the origin of $\mathcal{W}_1$. Thus in $\mathcal{W}_1$, $P_1 = \mathbf{0}_{3\times1}$, $L_{i=\{2,3\}} \triangleq \begin{bmatrix} X_i & Y_i & Z_i \end{bmatrix}^T$.

Pose estimation between $\mathcal{C}_q$ ***and*** $\mathcal{W}_m$: Let's denote the unsolved rotation and translation matrix as R and t, i.e. $R \triangleq R_{\mathcal{C}_1\mathcal{W}_1}$, $t \triangleq t_{\mathcal{C}_1\mathcal{W}_1}$. According to the collinearity of $\{C, D_1, RP_1 + t\}_L$, two equations can be derived:

$$a_1T_2 - b_1T_1 = 0 \tag{3}$$

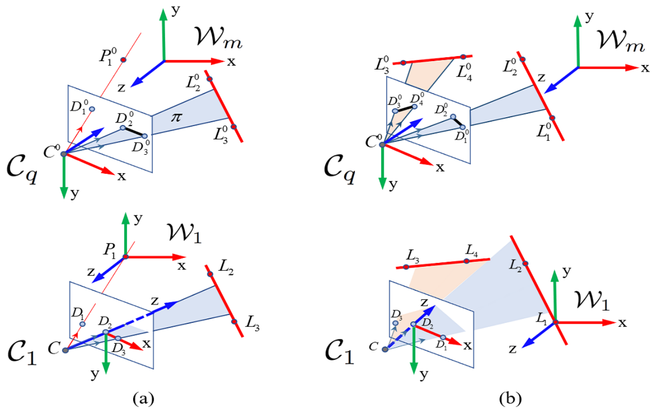


Fig. 2: The illustration of intermediate reference frame for (a) 1 point 1 line and (b) 2 lines case.

$$b_1T_3 - T_2 = -b_1 \tag{4}$$

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

$$R_{21}X_2 + R_{22}Y_2 + R_{23}Z_2 + T_2 = 0 \tag{5}$$

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

$$R_{21}X_3 + R_{22}Y_3 + R_{23}Z_3 + T_2 = 0 \tag{6}$$

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

R is only determined by the unknown yaw angle α, which can be solved by combining (5) and (6). After substituting α into (3) - (6), the translation t can also be solved. Then the localization problem can be solved as

$$T_{\mathcal{C}_q\mathcal{W}_m} = {T_{\mathcal{C}_1\mathcal{C}_q}}^{-1} \cdot T_{\mathcal{C}_1\mathcal{W}_1} \cdot T_{\mathcal{W}_1\mathcal{W}_m} \tag{7}$$

Degenerate cases: If the point lies on the line, the corresponding 1P1L case is degenerated.

2) 2 points: In this section, the minimal solution denoted as 2P using two 3D-2D point matches is derived. We will not introduce the intermediate reference frame for camera. The points in $\mathcal{C}_q$ following the notations in 1P1L are

$$C = \mathbf{0}_{3\times1},\ D_{i=\{1,2\}} \triangleq \begin{bmatrix} a_i & b_i & 1 \end{bmatrix}^T$$

where the parameters are all known which can be computed using intrinsic parameters and normalized depth.

For world reference frame, the transformation $T_{\mathcal{W}_1\mathcal{W}_m}$ is a translation which transforms P_1^0 to the origin of $\mathcal{W}_1$. Thus in $\mathcal{W}_1$, $P_1 = \mathbf{0}_{3\times1}$, $P_2 \triangleq \begin{bmatrix} X_2 & Y_2 & Z_2 \end{bmatrix}^T$.

Pose estimation between $\mathcal{C}_q$ ***and*** $\mathcal{W}_m$: Following the notations in 1P1L, two equations can be derived from $\{C, D_1, RP_1 + t\}_L$:

$$a_1T_2 - b_1T_1 = 0 \tag{8}$$

$$a_1T_3 - T_1 = 0 \tag{9}$$

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

$$\begin{aligned} &a_2(R_{21}X_2 + R_{22}Y_2 + R_{23}Z_2 + T_2) \\ &\quad - b_2(R_{11}X_2 + R_{12}Y_2 + R_{13}Z_2 + T_1) = 0 \end{aligned} \tag{10}$$

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

$T_{\mathcal{C}_q\mathcal{W}_1}$ can be solved by combining (8) - (11) and the localization result can be computed by

$$T_{\mathcal{C}_q\mathcal{W}_m} = T_{\mathcal{C}_q\mathcal{W}_1} \cdot T_{\mathcal{W}_1\mathcal{W}_m} \tag{12}$$

3) 2 lines: In this section, we are going to compute the camera pose with two 3D-2D line matches.

The choice of $\mathcal{C}_1$: The detailed illustration and computation of $T_{\mathcal{C}_1\mathcal{C}_q}$ can be found in the Supplementary Material [24]. As shown in Fig. 2, in $\mathcal{C}_q$, the camera center is C^0, the end points of the two image line segments are D_1^0, D_2^0 and D_3^0, D_4^0.

The choice of $\mathcal{W}_1$: For world reference frame, the transformation is designed to transform one end point of one map line to the origin of $\mathcal{W}_1$.

Pose estimation between $\mathcal{C}_q$ ***and*** $\mathcal{W}_m$: Following the notations in 1P1L, the following equations can be derived according to $\{C, D_1, D_2, RL_1 + t\}_P$ and $\{C, D_1, D_2, RL_2 + t\}_P$,

$$T_2 = 0 \tag{13}$$

$$R_{21}X_2 + R_{22}Y_2 + R_{23}Z_2 + T_2 = 0 \tag{14}$$

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

$$\begin{aligned} &a_2(R_{21}X_3 + R_{22}Y_3 + R_{23}Z_3 + T_2) \\ &- a_3(R_{11}X_3 + R_{12}Y_3 + R_{13}Z_3 + T_1) = 0 \end{aligned} \tag{15}$$

$$\begin{aligned} &a_2(R_{21}X_4 + R_{22}Y_4 + R_{23}Z_4 + T_2) \\ &- a_3(R_{11}X_4 + R_{12}Y_4 + R_{13}Z_4 + T_1) = 0 \end{aligned} \tag{16}$$

From (13) - (16), we can easily find that T_3 cannot be solved. In fact, one constraint provided by inertial measurement unit (IMU) is coincident with one constraint provided by line coplanarity, thus the 4DoF localization problem cannot be solved in 2 lines case.

B. Multi-Camera System

The proposed minimal solutions can be generalized to multi-camera system, which means the observation from other cameras can also be combined to estimate query camera pose. Specifically, the multi-camera can be multiple rigidly connected cameras with known extrinsic parameters by calibration or multiple temporal images of one camera with known relative transformation by VINS/odometry. In the rest of this paper, we unify the two possible conditions as “mPose” condition. According to the last section, there are also two possible minimal solutions: mPose-1P1L and mPose-2P.

1) mPose 1 point 1 line: In this section, 1 point and 1 line match observed in different cameras are utilized to solve the problem. The solution is denoted as mPose-1P1L. The camera frame which observed the point match is denoted as $\mathcal{C}_p$, and the camera frame observed the line match is denoted as $\mathcal{C}_l$. The transformation between the two camera frames $T_{\mathcal{C}_l\mathcal{C}_p}$ is known. As in 1P1L, we will introduce intermediate reference frames to simplify the equations, denoted as $\mathcal{C}_{p1}$, $\mathcal{C}_{l1}$, and $\mathcal{W}_1$.

The choice of $\mathcal{C}_{p1}$: In $\mathcal{C}_p$, the camera center C_p^0 is the origin and the image point is D_1^0. In $\mathcal{C}_{p1}$, the camera center C_p is $\begin{bmatrix}0 & 0 & -1\end{bmatrix}^T$, and the transformation $T_{\mathcal{C}_{p1}\mathcal{C}_p}$ is a translation which transforms C_p^0 to C_p. With the transformation, the image point D_1 in $\mathcal{C}_{p1}$ can be computed.

The choice of $\mathcal{C}_{l1}$: In $\mathcal{C}_l$, the camera center C_l^0 is the origin, and the end points of the detected line segment are D_2^0 and D_3^0. The choice of $\mathcal{C}_{l1}$ is the same as $\mathcal{C}_1$ in 1P1L. Then the coordinates of D_2 and D_3 in $\mathcal{C}_{l1}$ can be calculated.

The choice of $\mathcal{W}_1$: The choice is the same as in 1P1L which transforms P_1^0 to the origin. Thus in $\mathcal{W}_1$, $P_1 = \mathbf{0}_{3\times1}$, $L_{i=\{2,3\}} \triangleq \begin{bmatrix}X_i & Y_i & Z_i\end{bmatrix}^T$.

Pose estimation between $\mathcal{C}_p$ ***and*** $\mathcal{W}_m$: The transformation between the camera intermediate frame and the world intermediate frame can be computed as follows

$$T_{\mathcal{C}_{p1}\mathcal{W}_1} = T_{\mathcal{C}_{p1}\mathcal{C}_p} \cdot T_{\mathcal{C}_p\mathcal{W}_m} \cdot T_{\mathcal{W}_m\mathcal{W}_1} \tag{17}$$

$$T_{\mathcal{C}_{l1}\mathcal{W}_1} = T_{\mathcal{C}_{l1}\mathcal{C}_l} \cdot T_{\mathcal{C}_l\mathcal{C}_p} \cdot T_{\mathcal{C}_p\mathcal{W}_m} \cdot T_{\mathcal{W}_m\mathcal{W}_1} \tag{18}$$

where the elements in right hand are all known except $T_{\mathcal{C}_p\mathcal{W}_m}$.

Let us denote $R_i \triangleq R_{\mathcal{C}_{i1}\mathcal{W}_1}$, $t_i \triangleq t_{\mathcal{C}_{i1}\mathcal{W}_1}, (i = p, l)$. According to the collinearity of $\{C_p, D_1, R_pP_1 + t_p\}_L$, the coplanarity of $\{C_l, D_2, D_3, R_lL_2 + t_l\}_P$ and $\{C_l, D_2, D_3, R_lL_3 + t_l\}_P$, four equations about the four unsolved parameters in $T_{\mathcal{C}_p\mathcal{W}_m}$ can be derived. And the localization problem can be solved by combining the four equations as in 1P1L.

Degenerate cases: If the 3D point lies on the 3D line, the corresponding mPose-1P1L case is degenerated.

2) mPose 2 points: The minimal solution using 2 point matches observed in different cameras is denoted as mPose-2P. Let us denote the query camera reference frame as $\mathcal{C}_q$, and the additional camera frame $\mathcal{C}_a$. In this case, no intermediate camera reference frames are introduced, thus the point D_1 in $\mathcal{C}_q$ and the point D_2 in $\mathcal{C}_a$ are

$$C_{i=\{q,a\}} = \mathbf{0}_{3\times1},\ D_{i=\{1,2\}} \triangleq \begin{bmatrix}a_i & b_i & 1\end{bmatrix}^T$$

The transformation of the world reference is the same as in 2P. Thus in $\mathcal{W}_1$, $P_1 = \mathbf{0}_{3\times1}$, $P_2 \triangleq \begin{bmatrix}X_2 & Y_2 & Z_2\end{bmatrix}^T$.

Pose estimation between $\mathcal{C}_q$ ***and*** $\mathcal{W}_m$:

$$T_{\mathcal{C}_q\mathcal{W}_1} = T_{\mathcal{C}_q\mathcal{W}_m} \cdot T_{\mathcal{W}_m\mathcal{W}_1} \tag{19}$$

$$T_{\mathcal{C}_a\mathcal{W}_1} = T_{\mathcal{C}_a\mathcal{C}_q} \cdot T_{\mathcal{C}_q\mathcal{W}_m} \cdot T_{\mathcal{W}_m\mathcal{W}_1} \tag{20}$$

Let us denote $R_i \triangleq R_{\mathcal{C}_i\mathcal{W}_1}$, $t_i \triangleq t_{\mathcal{C}_i\mathcal{W}_1}, (i = q, a)$. According to the collinearity of $\{C_q, D_1, R_qP_1 + t_q\}_L$ and $\{C_a, D_2, R_aP_2 + t_a\}_L$, four equations about the four unsolved parameters in $T_{\mathcal{C}_q\mathcal{W}_m}$ can be derived. And the localization problem can be solved as in 2P.

IV. Strategy Selection and Scoring

In the former section, the minimal solutions for visual localization problem are derived including 1P1L, 2P, mPose-1P1L and mPose-2P. Embedding the derived solutions, we propose the 2-entity RANSAC and present three sampling strategies: 1P1L, 2P and mixed as shown in Fig. 1. The 1P1L sampling strategy refers to sampling one point match and one line match and utilizing the 1P1L or mPose-1P1L minimal solution. The 2P sampling strategy refers to sampling two point matches and utilizing 2P or mPose-2P. While the mixed refers to using the minimal solution according to the selected feature types. To be specific, one point match is selected firstly and then another feature match is selected among the remaining features. 2P or mPose-2P is performed when it is a point match, while 1P1L or mPose-1P1L for a line match. Next we will analyze the success probability of the three sampling strategies respectively.

The number of point and line matches are denoted as N_p and N_l, and the corresponding inlier numbers are denoted as IN_p and IN_l, respectively.

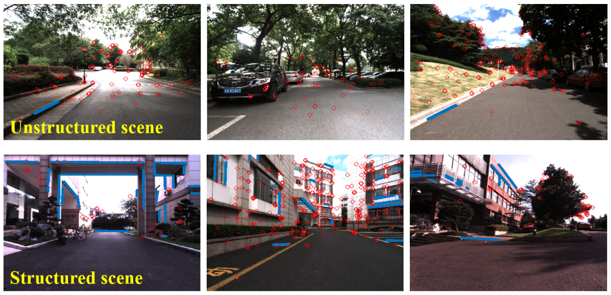


Fig. 3: Some examples of unstructured and structured scene.

$$IN_p/N_p = \lambda_p,\ IN_l/N_l = \lambda_l\,(0 \le \lambda_p, \lambda_l \le 1) \tag{21}$$

where λ_p, λ_l denote the inlier rate of point and line matches.

After comparison of the three sampling strategies' success probability, denoted as P_{1P1L}, P_{2P} and P_{mixed} respectively, of which the details are presented in Supplementary Material [24], we have the following conclusions:

$$\lambda_l \ge \lambda_p \Rightarrow P_{1P1L} > P_{mixed} > P_{2P} \tag{22}$$

$$\lambda_l < \lambda_p \Rightarrow P_{1P1L} \le P_{mixed} \le P_{2P} \tag{23}$$

A. Strategy Selection

From (22) and (23), we find that the sampling strategy selection is relevant to the relative inlier rate of point and line features, but not the absolute inlier rate. We consider that the key of strategy selection is the inlier distribution of point and line features in the query image, which is a global feature of the appearance. It thus inspires us to skip the inlier estimation and directly imitate such a feature extraction and classification process. Since we can easily get the labeled data based on (22) and (23) as well as mapping result, we model the sampling strategy selection as a convolutional neural network with full image convolution and global pooling to classify the best strategy for the input image, in a supervised learning manner.

Specifically, we select map session containing around 5000 images as training data. When labeling the image, we compute the ground truth pose of the image got by laser to count the inlier ratio of point and line matches respectively. The label is 2P if the point inlier ratio is obviously higher than lines, while 1P1L when the line inlier ratio is obviously higher. As for the circumstance the two ratios are similar, mixed label is used instead. As it is a standard image classification problem, we pick popular VGG16 [25] as the backbone followed by 3-class softmax to make the selection. In testing phase, given a query image, we run the trained network to predict the best sampling strategy, which is then utilized for feature sampling. This network brings the environment awareness to the proposed 2-entity RANSAC, of which the effectiveness is shown in the later ablative experiments.

B. Feature Scoring

Compared with sampling strategy selection, feature scoring focuses more on local appearance. Specifically, feature scoring builds a metric for feature descriptors to estimate the inlier. Several existing methods are proposed to develop position induced metric [26], or descriptor induced metric [27] to achieve this task. In the long term localization, the environmental variations significantly affect the feature detection repeatability and descriptor distance ratios, causing very low inlier rate. If we still use these methods, the metric may not reflect the inlier. To address this problem, we build the metric using multilayer perceptron (MLP), which nonlinearly embeds the descriptor into a learned metric space formed by ground truth data.

Specifically, the input of the MLP is the 32-bit descriptor of each feature point obtained during the feature extraction with LibVISO2 [28]. And the hiden layer consists of 128 nodes which is fully connected with the output layer of two nodes. The output tells the probability of the input correspondence being an inlier. To annotate the data, we refer to the mapping pose to evaluate the re-projection error for judgement. After training, we run the model for each candidate correspondence to predict the inlier probability, upon which we can non-uniformly sample the feature correspondences that prone to be inliers in 2-entity RANSAC localization, improving the chance of convergence.

V. Experimental Results

We conduct simulation and real world experiments over other state-of-the-art methods for both monocular and multiple camera system. The computing platform for all experiments is an Intel i7-7700 @3.60GHz and 8G RAM. The proposed methods are implemented in C++ using the framework of open-source library OpenGV [29]. And the code of the proposed methods is available on github[2] such that the evaluation results can be easily reproduced. We also use the implementation of the compared algorithms P3P [17], EPnP [16], UPnP [30], GP3P [31] and GPnP [31] in OpenGV. While for 2P1L [20], we implement the code on MATLAB according to the paper, so that the computational efficiency is not compared with 2P1L in Section V-A. The strategy selection and feature scoring network are implemented on Pytorch [32].

A. Simulation Experiments

The simulation experiments are conducted to illustrate the accuracy on noisy image features, the efficiency with computational time comparison, the sensitivity when the pitch and roll angles are inaccurate, and the robustness in presence of outliers. We get the 3D-2D feature matches by projecting the 3D points and lines to 2D features with varying camera poses. Then 100 iterations of RANSAC are performed for each method and the final identified inliers are sent to nonlinear optimization. The compared mono-camera algorithms include P3P [17], EPnP [16], UPnP [30] and 2P1L [20]. And multi-camera algorithms include GP3P [31] and GPnP [31]. As the UPnP and GPnP implemented in OpneGV can only deal with the situations with no outliers, we embed the solutions into RANSAC to improve robustness and denote them as UPnP(*) and GPnP(*) in the remaining of the paper. For evaluation, we compute the translation and rotation error of the estimated pose $[R|t]$ compared to the ground truth $[R_{gt}|t_{gt}]$. The translation error is expressed as $\|t - t_{gt}\|$ in meter and the rotation error is expressed as $\triangle R = RR_{gt}^T$ in degree by axis-angle representation as in [15].

TABLE I: Accuracy simulation results of mono-camera algorithms.

	Entities	10 points + 10 lines			6 points + 6 lines			3 points + 3 lines		
	Noise(pixel)	0.5	1.0	2.0	0.5	1.0	2.0	0.5	1.0	2.0
	Mean Error(mm/deg)	Trans/Rota	Trans/Rota	Trans/Rota	Trans/Rota	Trans/Rota	Trans/Rota	Trans/Rota	Trans/Rota	Trans/Rota
6DoF	P3P	2.677 / 0.020	8.161 / 0.062	22.324 / 0.165	4.563 / 0.034	13.922 / 0.098	50.448 / 0.357	- / -	- / -	- / -
	EPnP	2.417 / 0.018	7.298 / 0.055	19.052 / 0.138	3.885 / 0.028	10.671 / 0.075	25.448 / 0.184	- / -	- / -	- / -
	UPnP(*)	2.856 / 0.029	8.400 / 0.060	21.649 / 0.156	4.356 / 0.031	13.819 / 0.087	48.535 / 0.312	- / -	- / -	- / -
	2P1L	2.384 / 0.011	6.320 / 0.020	14.122 / 0.044	3.738 / 0.013	7.137 / 0.024	18.033 / 0.066	6.170 / 0.023	12.443 / 0.053	26.120 / 0.087
	1P1L	2.245 / 0.010	5.824 / 0.022	13.768 / 0.048	3.532 / 0.013	6.453 / 0.027	15.481 / 0.061	4.663 / 0.016	10.375 / 0.032	25.107 / 0.078
	2P	2.233 / 0.010	5.873 / 0.020	13.201 / 0.043	3.427 / 0.012	6.956 / 0.018	14.456 / 0.046	5.108 / 0.015	10.271 / 0.035	20.886 / 0.071
4DoF	P3P	2.509 / 0.017	6.914 / 0.049	18.279 / 0.113	4.172 / 0.029	11.246 / 0.054	33.847 / 0.118	- / -	- / -	- / -
	EPnP	2.375 / 0.015	6.631 / 0.037	16.775 / 0.096	3.581 / 0.023	7.346 / 0.046	17.373 / 0.093	- / -	- / -	- / -
	UPnP(*)	2.586 / 0.019	6.817 / 0.041	18.961 / 0.118	3.871 / 0.027	10.935 / 0.059	30.366 / 0.102	- / -	- / -	- / -
	2P1L	2.317 / 0.009	5.769 / 0.019	13.409 / 0.035	3.107 / 0.011	6.357 / 0.021	16.104 / 0.053	5.749 / 0.018	11.110 / 0.046	25.198 / 0.077
	1P1L	2.167 / 0.009	4.443 / 0.016	11.571 / 0.029	2.998 / 0.012	5.774 / 0.024	14.027 / 0.047	4.231 / 0.012	9.175 / 0.021	19.772 / 0.049
	2P	2.166 / 0.008	4.480 / 0.014	11.776 / 0.029	2.813 / 0.011	6.305 / 0.016	12.830 / 0.033	4.863 / 0.013	9.130 / 0.025	17.025 / 0.053

TABLE II: Accuracy simulation results of multi-camera algorithms.

Entities	50 points + 50 lines			30 points + 30 lines			10 points + 10 lines		
Noise(pixel)	0.5	1.0	2.0	0.5	1.0	2.0	0.5	1.0	2.0
Mean Error(mm/deg)	Trans/Rota	Trans/Rota	Trans/Rota	Trans/Rota	Trans/Rota	Trans/Rota	Trans/Rota	Trans/Rota	Trans/Rota
GP3P	2.895 / 0.000	5.498 / 0.001	13.745 / 0.002	3.225 / 0.000	6.677 / 0.001	17.365 / 0.002	3.843 / 0.001	8.592 / 0.001	23.425 / 0.003
GPnP(*)	2.786 / 0.000	5.543 / 0.001	13.545 / 0.002	3.240 / 0.000	6.615 / 0.001	16.571 / 0.002	3.791 / 0.001	7.830 / 0.001	18.780 / 0.002
mPose-1P1L	1.463 / 0.000	4.466 / 0.000	12.612 / 0.001	1.609 / 0.000	6.020 / 0.001	14.642 / 0.001	1.917 / 0.000	6.874 / 0.000	18.269 / 0.001
mPose-2P	1.422 / 0.000	4.731 / 0.001	12.734 / 0.001	1.607 / 0.000	6.078 / 0.001	14.577 / 0.001	1.859 / 0.000	6.815 / 0.001	18.314 / 0.001

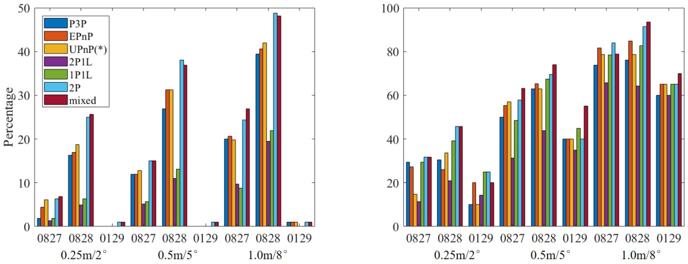
TABLE III: Computation Time Comparison (ms).

	Entities	6	106	206	306	406	506
mono-camera	P3P	0.395	1.691	2.156	2.613	3.099	3.573
	EPnP	1.649	7.169	7.687	8.073	8.516	9.025
	UPnP(*)	68.124	189.197	187.856	188.241	188.497	189.699
	2P	0.055	0.543	0.872	1.174	1.498	1.815
	1P1L	0.079	0.447	0.626	0.827	1.011	1.202
multi-camera	GP3P	0.440	1.266	1.397	1.547	1.681	1.840
	GPnP(*)	16.906	23.051	23.582	24.279	24.918	25.807
	mPose-2P	0.080	0.595	0.945	1.303	1.661	1.999
	mPose-1P1L	0.120	0.517	0.721	0.952	1.152	1.365

1) Accuracy: For accuracy quantification, we add Gaussian noise with zero mean and various standard deviations to the 2D features as in [20] and vary the number of feature matches in different levels. As the proposed method uses inertial measurements, for fair comparison we also provide pitch and roll for comparative methods. Specifically, we fix pitch and roll in refinement to perform 4DoF optimization. The results presented in Table I and Table II show that, as we can utilize both point and line inlier feature matches to do the final pose optimization, the accuracy of proposed methods is better than others, in both monocular and multiple camera cases.

2) Efficiency: The real-time performance is important for the robot localization algorithm. We vary the number of entities (including points and lines features) in the scene and count the computation time of different methods for both monocular and multiple camera pose estimation algorithms as in [30]. For the data in each test, we add the Gaussian noise with zero mean and 1.0 pixel standard deviations to the 2D features and generate 10% outliers to simulate the real data. The compared methods are all RANSAC-based algorithms and are all implemented in OpenGV which is a fair comparison. The result is shown in Table III. Note that the computation time of UPnP(*) and GPnP(*) is increased due to outliers. Results show that the proposed methods are the fastest, which can satisfy the real-time performance of the robot localization.

3) Sensitivity: It's necessary to study the impact of the quality of the pitch and roll angles on the final accuracy, as we reduce the DoF of the pose by leveraging the two angles measured by inertial sensor. We add Gaussian noise with zero mean and various standard deviations on both two angles and vary the number of feature matches in three levels: 10, 6, 3, which can be seen in Fig. 5 (the level N indicates total N point and N line matches). With enough feature matches as in level 10, our method can tolerate noise of 20 degree on both two angles, as we can perform the 6DoF optimization with the identified inliers to get higher accuracy. Empirically, the noise of pitch and roll angle is far less in real world application [33] [34]. Thus influence on the final accuracy can be ignored.



(a) Unstructured

(b) Structured

Fig. 4: Success rate on selected scene.

4) Robustness: We vary the outlier rate from 0 to 80% by adding a certain number of outliers which are generated by incorrectly projecting the features and change the inlier number in three levels: 10, 6, 3. For evaluation, all methods are performed 500 times to count the average success rate. The localization is assumed to be successful when the translation error is lower than 0.1 m and the rotation error is smaller than 0.5 degree as in [35] [15]. When the outlier rate increases as shown in Fig. 6 and Fig. 7, the performance of our method outperforms the compared methods obviously.

B. Real World Experiments

The YQ-Dataset [10] collected across weathers and seasons is utilized as real data. There are three sessions collected at summer 2017 in three days, denoted as 2017-0823, 2017-0827 and 2017-0828, and one session collected in winter 2018 after snow, denoted as 2018-0129. The 3D map is built with 2017-0823 session by running visual inertial SLAM [36]. The other three sessions are used to evaluate the localization performance. For evaluation, we compute the ground truth of the relative pose between the query camera and the map by

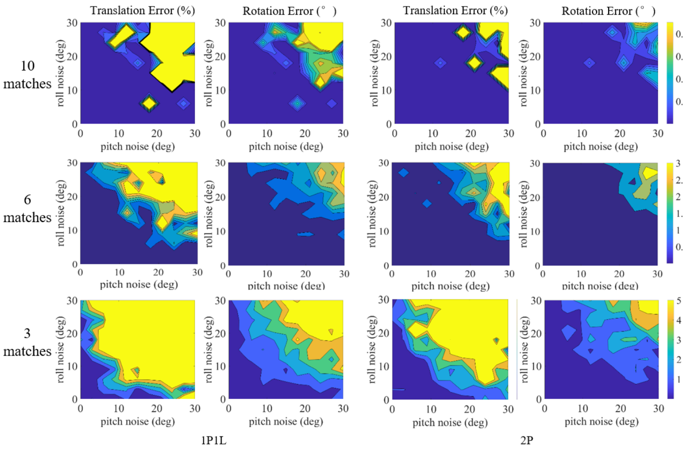
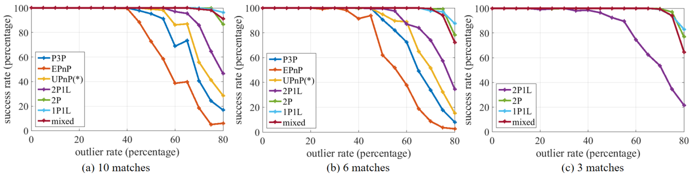


Fig. 5: Sensitivity simulation results. To present the translation and rotation error in the same color bar, we express the translation error in percentage as in [20].

aligning the synchronized LiDAR scans. For the pitch and roll angles of the query camera, we exploit the estimation of visual inertial odometry [36]. More details on dataset and platform can be found in Supplementary Material [24].

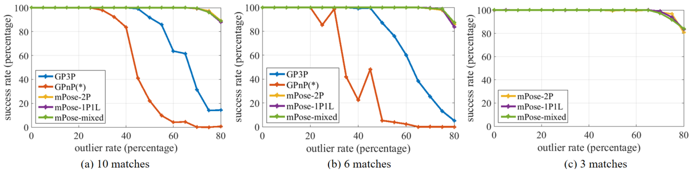
1) Evaluation on selected scene: We first evaluate the proposed methods including 1P1L, 2P and mixed against P3P [17], EPnP [16], UPnP [30] and 2P1L [20]. Due to the uneven distribution of the characteristics in the map environment, we manually select around 50 structured and 100 unstructured segments according to experience as shown in Fig. 3 to make a fair comparison between the different methods with different feature types. Then we count the successfully localized query images under different error thresholds based on the thresholds in [35] out of all selected places in each segment to compute success rate. The results are shown in Fig. 4, from which we can find that in the unstructured segments, the 2P method performs the best and in structured segments, the 1P1L is similar as 2P. In addition, the mixed method gives relatively stable performance in all segments, which confirms the analysis illustrated in Section IV. Moreover, one should notice that the better robustness of line features advances the performance of 1P1L in structured segments of 2018-0129, as the outlier rate grows due to the changing season.

2) Evaluation on whole session: Then we evaluate on whole session and count the success rate of localization over four different thresholds, of which the results are shown in Table VI and Table VIII for monocular and multi-camera algorithms. The multi-camera in this experiment refers to using multiple temporal images of which the extrinsic parameters are determined by VINS. As expected, the performance of mixed sampling strategy is relatively stable in all sessions. While 1P1L is not as good as point based methods, as there are lots of trees on each side of the road in the whole map so that point features are far more abundant than lines. Furthermore, adding another frame's observation obviously promotes the performance, as the number of potential inliers increases and the failed localization caused by few feature matches is reduced. Besides, the proposed method is obviously better than compared methods that require more features, which validates the robustness of our method in changing environment.



(a) 10 matches (b) 6 matches (c) 3 matches

Fig. 6: Robustness results of monocular camera algorithms.



(a) 10 matches (b) 6 matches (c) 3 matches

Fig. 7: Robustness results of multiple camera algorithms.

3) Ablation Study Experiment: To validate the performance of the proposed feature scoring method, we first conduct comparison experiment with the inlier score estimation methods of BLOGS [26] and BEEM [27]. All methods adopt LibVISO2 [28] for feature points extraction and descriptor computation. The ground truth criteria to judge a correspondence to be an inlier is the reprojection error less than 8 pixel, which is the same as in the implementation of *solvePnP* in OpenCV [37]. The same division for training and testing dataset are utilized for BEEM and ours (2017-0827 for training, 2017-0828 and 2018-0129 for testing). More implementation details are shown in Supplementary Material [24]. After normalization of the estimated inlier score, we unify that the correspondence with score higher than 0.8 is inlier as in [27]. Then we calculate the precision, recall and F-score of the inlier estimation on the three sessions, which is shown in Table V.

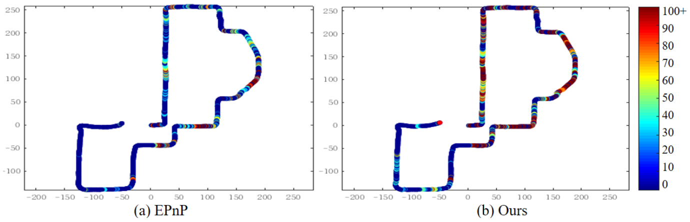
From the result, we find that the performance of BLOGS which mainly depends on position based metric is worse than descriptor based metric, BEEM and ours. And the performance of the proposed method outperforms BEEM, especially on the two test sessions (2017-0828 and 2018-0129), reflecting that the generalization of our method is better. The results validate that the large variations in changing environment seriously affect the descriptor space, and show the advantage of embedding the descriptor into data driven metric space. Therefore, the proposed feature scoring method is a better choice to guide the candidate correspondences sampling.

Then to validate the effectiveness of each component of the proposed method, we conduct an ablation study experiment shown in Table IV. The mixed method is selected to represent the 2-entity method. The 2-entity RANSAC with learning-based strategy selection is denoted as 2ESel, and if applied with the proposed feature scoring method, then denoted as 2ESelScore. The strategy selection and feature scoring network are all trained with the map data. Moreover, mPose-* refers to applying above methods to multi-camera system. From the result, we could find that with proper sampling strategy selection and feature scoring method, the performance of 2-entity RANSAC can be further promoted.

To explain the better performance of the proposed method, we analyze the distribution of the successful localization on the whole session and the number of identified inliers. We take the EPnP and our mPose-2ESelScore method for example to

TABLE IV: Ablation study experiment on whole session.

Trans Error (m) Rota Error (°)	2-Entity	Strategy Selection	Feature Scoring	Multiple Cameras	2017-0827 0.25/0.5/1.0/5.0 2.0/5.0/8.0/10.0	2017-0828 0.25/.5/1.0/5.0 2.0/5.0/8.0/10.0	2018-0129 0.25/0.5/1.0/5.0 2.0/5.0/8.0/10.0
mixed	✓				22.1 / 33.7 / 40.1 / 53.6	23.0 / 34.7 / 40.4 / 49.3	13.2 / 24.7 / 34.6 / 54.1
2ESel	✓	✓			22.3 / 34.0 / 41.9 / 54.2	24.6 / 35.7 / 41.6 / 49.1	14.1 / 24.9 / 34.9 / 54.6
2ESelScore	✓	✓	✓		22.6 / 34.3 / 41.9 / 54.6	24.8 / 36.2 / 42.3 / 50.2	14.9 / 25.4 / 35.9 / 55.0
mPose-mixed	✓			✓	23.7 / 36.8 / 45.7 / 55.1	23.1 / 36.9 / 42.8 / 52.5	16.6 / 27.6 / 37.8 / 58.6
mPose-2ESel	✓	✓		✓	24.5 / 38.2 / 46.3 / 56.7	24.2 / 37.5 / 43.8 / 53.3	17.1 / 28.3 / 38.4 / 59.5
mPose-2ESelScore	✓	✓	✓	✓	25.5 / 39.0 / 47.4 / 58.9	27.8 / 38.8 / 45.9 / 54.8	17.6 / 29.5 / 39.2 / 61.8



(a) EPnP (b) Ours

Fig. 8: The distribution of successful localization and inliers.

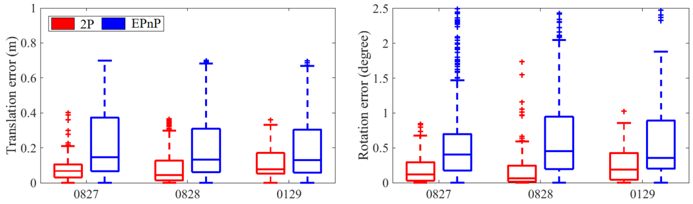


Fig. 9: Repeatability result.

compare the distribution in Fig. 8. The color bar indicates the number of inliers. Results show that, as we can utilize both point and line features and the minimal number of feature matches needed to compute the camera pose is reduced, our method can achieve more inliers and thus more successful localizations, such that the ATE (absolute trajectory error) is smaller as shown in Table VII.

4) Repeatable positioning: For the localization algorithm, the repeatable positioning in a same scenario is very important. To quantify the repeatability, we repeat the same method in the same place to present the statistic result of translation and rotation error between each two estimated results. To be specific, we test the proposed method 100 times in the same place with 100 RANSAC iterations in each test. To demonstrate the repeatability in different environmental characteristics, we utilize the selected scenes of three sessions mentioned in Section V-B for test. The result expressed in translation and rotation error is shown in Fig. 9. The repeatability is similar on three sessions. As reflected by the result of the proposed method, the mean repetition error of translation is around 0.1 m and rotation is around 0.2 degree. And the repeatability of the proposed method is better than EPnP.

C. Real-time Robot System Experiments

Finally, we conduct real-time visual localization experiments on both monocular and multiple camera robot system. The details about the robot platform and multi-camera dataset can be found in Supplementary Material [24]. For monocular camera system validation, we still use the YQ-dataset. The proposed algorithms are integrated in the existing pose estimation module of ORB-SLAM [38]. We compare the number of successful localization and inliers for robustness evaluation of EPnP and our 2P method. The criterion for successful localization is the same as in ORB-SLAM. For accuracy, we calculate the ATE as shown in Table VII. Since the relocalization module in ORB-SLAM is not robust and the localization using multiple temporal images is not supported, the performance is not as good as in Section V-B. But note that the proposed 2P method is still better than EPnP.

To further evaluate the multi-camera algorithms, we collect a new dataset with a physical five-camera robot system and there are four sessions denoted as 0325, 0329, 0331, 0333. We select the 0325 session to build the map, and the others to test the performance with real multiple images from multiple cameras. For accuracy, we compare the sequential localization estimates with the visual odometry estimates, in a similar manner to [39]. The results on this dataset are shown in Table IX. As the results show, the proposed localization method gives the best performance over all evaluations and is more robust and accurate as the mean and median error are all smaller. In addition, since there are large perspective changes in this dataset, the advantage of multi-camera algorithms is bigger than that of mono-camera algorithms, which validates the necessity of multi-camera algorithms in practice.

There are some limitations in real world application. The IMU and multiple cameras are needed which increases the costs and the calibration is also required. And the network of strategy selection needs re-training when localizing in the new map. However, if the front end of the robot system utilizes VINS, the IMU-camera configuration and calibration would also be useful for these algorithms, such that the improvement of the robustness given by our algorithm is still worth.

VI. Conclusion

In this work, the minimal closed form solutions to 3D-2D visual localization using both point and line features are derived by making use of inertial measurements. Embedding the solutions, the 2-entity RANSAC framework is proposed and verified on thorough simulation and real world experiments to show the robustness in long-term localization. In addition, the proposed mPose minimal solutions can be applied to multi-camera system, which are validated in real data with significant environmental and perspective changes. In the future, we are going to focus on the insensitive feature extraction and matching to get more reliable feature matches.

TABLE V: Inlier score estimation comparison.

	2017-0827			2017-0828			2018-0129		
	BEEM	BLOGS	Ours	BEEM	BLOGS	Ours	BEEM	BLOGS	Ours
Precision	0.723	0.701	**0.743**	0.589	0.517	**0.738**	0.460	0.417	**0.673**
Recall	0.304	0.149	**0.651**	0.433	0.182	**0.659**	0.196	0.150	**0.683**
F-score	0.375	0.229	**0.668**	0.494	0.264	**0.676**	0.238	0.197	**0.647**

TABLE VI: Success rate of mono-camera algorithms.

	2017-0827	2017-0828	2018-0129
Trans Error (m)	0.25/0.5/1.0/5.0	0.25/.5/1.0/5.0	0.25/0.5/1.0/5.0
Rota Error (°)	2.0/5.0/8.0/10.0	2.0/5.0/8.0/10.0	2.0/5.0/8.0/10.0
P3P	15.6 / 28.6 / 38.2 / 49.4	16.2 / 32.1 / 39.1 / 47.1	9.2 / 18.3 / 27.8 / 45.2
EPnP	15.9 / 29.2 / 38.6 / 50.7	17.8 / 33.0 / 39.6 / 47.6	9.8 / 20.7 / 28.8 / 45.6
UPnP(*)	19.2 / 33.1 / 39.2 / 51.8	19.6 / 34.3 / 38.2 / 47.4	10.4 / 21.3 / 28.2 / 46.8
2P1L	9.8 / 14.2 / 18.6 / 31.8	9.5 / 17.4 / 23.0 / 34.1	7.0 / 11.5 / 16.5 / 33.3
1P1L	9.8 / 14.2 / 18.6 / 31.8	9.5 / 17.4 / 23.0 / 34.1	7.0 / 11.5 / 16.5 / 33.3
2P	21.9 / 33.8 / 41.0 / 54.6	20.9 / 31.9 / 37.1 / 44.3	13.9 / 24.4 / 33.5 / 53.5
mixed	22.1 / 33.7 / 40.1 / 53.6	23.0 / 34.7 / 40.4 / 49.3	13.2 / 24.7 / 34.6 / 54.1

TABLE VII: Mono-camera robot localization.

		Success	Inlier		ATE (m)	
		Number	Number	rmse	mean	median
0827	EPnP	275	43920	1.307	1.042	0.821
	Ours	**479**	**50857**	**1.219**	**0.936**	**0.607**
0827ORB	EPnP	108	38827	2.374	1.979	1.557
	Ours	**214**	**42390**	**2.245**	**1.795**	**1.398**
0828	EPnP	363	48076	1.312	0.995	0.690
	Ours	**596**	**69652**	**1.041**	**0.744**	**0.481**
0828ORB	EPnP	187	46757	1.261	1.026	0.816
	Ours	**277**	**58020**	**1.225**	**0.957**	**0.723**
0129	EPnP	144	14479	1.515	1.228	0.948
	Ours	**268**	**31980**	**1.370**	**1.109**	**0.851**
0129ORB	EPnP	41	13884	3.712	2.598	1.963
	Ours	**62**	**24316**	**3.684**	**2.538**	**1.952**

TABLE VIII: Success rate of multi-camera algorithms.

	2017-0827	2017-0828	2018-0129
Trans Error (m)	0.25/0.5/1.0/5.0	0.25/.5/1.0/5.0	0.25/0.5/1.0/5.0
Rota Error (°)	2.0/5.0/8.0/10.0	2.0/5.0/8.0/10.0	2.0/5.0/8.0/10.0
GP3P	22.9 / 36.5 / 42.8 / 54.3	22.6 / 36.8 / 42.7 / 52.2	15.2 / 22.1 / 31.0 / 53.5
GPnP(*)	21.1 / 35.5 / 41.4 / 52.4	22.9 / 34.8 / 41.3 / 50.2	14.3 / 21.8 / 30.2 / 52.0
mPose-1P1L	12.7 / 26.8 / 32.0 / 42.6	11.7 / 23.4 / 33.3 / 40.1	13.4 / 17.7 / 25.3 / 39.8
mPose-2P	23.0 / 37.1 / 45.6 / 55.8	23.9 / 36.6 / 43.2 / 51.7	15.3 / 26.4 / 37.2 / 57.6
mPose-mixed	23.7 / 36.8 / 45.7 / 55.1	23.1 / 36.9 / 42.8 / 52.5	16.6 / 27.6 / 37.8 / 58.6

TABLE IX: Multi-camera robot localization.

		Success	Inlier	Trans Error(m)		Rota Error(°)	
		Number	Number	mean	median	mean	median
0329	EPnP	10	1033	3.88	0.22	4.30	0.89
	2P	13	1262	2.82	0.12	2.99	0.72
	GP3P	32	4704	1.37	0.18	3.11	1.07
	mPose-2P	**39**	**5850**	**0.91**	**0.06**	**1.10**	**0.28**
0331	EPnP	33	4057	0.29	0.14	3.08	0.71
	2P	36	4464	0.26	0.14	2.40	0.51
	GP3P	82	12300	0.52	0.19	3.08	0.85
	mPose-2P	**94**	**15792**	**0.16**	**0.07**	**1.25**	**0.27**
0333	EPnP	1	88	-	-	-	-
	2P	4	152	10.31	12.91	51.41	1.72
	GP3P	9	1773	3.69	0.28	20.65	1.81
	mPose-2P	**12**	**1812**	**0.27**	**0.19**	**0.94**	**0.72**

References

[1] J. Zhang, R. Liu, K. Yin, Z. Wang, M. Gui, and S. Chen, "Intelligent collaborative localization among air-ground robots for industrial environment perception," *IEEE Transactions on Industrial Electronics*, vol. 66, no. 12, pp. 9673–9681, 2018.

[2] J. Kim and W. Chung, "Localization of a mobile robot using a laser range finder in a glass-walled environment," *IEEE Transactions on Industrial Electronics*, vol. 63, no. 6, pp. 3616–3627, 2016.

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

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

[5] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "Orb: An efficient alternative to sift or surf," in *2011 International conference on computer vision*, pp. 2564–2571. Ieee, 2011.

[6] J. Addison Lee, J. Cheng, B. Hai Lee, E. Ping Ong, G. Xu, D. Wing Kee Wong, J. Liu, A. Laude, and T. Han Lim, "A low-dimensional step pattern analysis algorithm with application to multimodal retinal image registration," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1046–1053, 2015.

[7] B. D. de Vos, F. F. Berendsen, M. A. Viergever, H. Sokooti, M. Staring, and I. Išgum, "A deep learning framework for unsupervised affine and deformable image registration," *Medical image analysis*, vol. 52, pp. 128–143, 2019.

[8] J. A. Lee, P. Liu, J. Cheng, and H. Fu, "A deep step pattern representation for multimodal retinal image registration," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5077–5086, 2019.

[9] 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.

[10] X. Ding, Y. Wang, D. Li, L. Tang, H. Yin, and R. Xiong, "Laser map aided visual inertial localization in changing environment," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 4794–4801. IEEE, 2018.

[11] 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*, vol. 43, no. 1, pp. 197–211, 2019.

[12] F. Dornaika and C. Garcia, "Pose estimation using point and line correspondences," *Real-Time Imaging*, vol. 5, no. 3, pp. 215–230, 1999.

[13] C. Raposo, M. Lourenço, M. Antunes, and J. P. Barreto, "Plane-based odometry using an rgb-d camera." in *BMVC*, 2013.

[14] G. H. Lee, "A minimal solution for non-perspective pose estimation from line correspondences," in *European Conference on Computer Vision*, pp. 170–185. Springer, 2016.

[15] P. Miraldo, T. Dias, and S. Ramalingam, "A minimal closed-form solution for multi-perspective pose estimation using points and lines," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 474–490, 2018.

[16] V. Lepetit, F. Moreno-Noguer, and P. Fua, "Epnp: An accurate o (n) solution to the pnp problem," *International journal of computer vision*, vol. 81, no. 2, p. 155, 2009.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] Z. Kukelova, M. Bujnak, and T. Pajdla, "Closed-form solutions to minimal absolute pose problems with known vertical direction," in *Asian Conference on Computer Vision*, pp. 216–229. Springer, 2010.

[23] Y. Jiao, Y. Wang, B. Fu, X. Ding, Q. Tan, L. Chen, and R. Xiong, "2-entity ransac for robust visual localization in changing environment," *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2019. [Online]. Available: https://arxiv.org/abs/1903.03967

[24] Y. Jiao, Y. Wang, B. Fu, X. Ding, S. Huang, and R. Xiong, "Supplementary material for 2-entity ransac for robust visual localization: Framework, methods and verifications." [Online]. Available: https://github.com/slinkle/2-Entity-RANSAC/blob/master/Supplementary%20Material.pdf

[25] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[26] A. S. Brahmachari and S. Sarkar, "Blogs: Balanced local and global search for non-degenerate two view epipolar geometry," in *2009 IEEE 12th International Conference on Computer Vision*, pp. 1685–1692. IEEE, 2009.

[27] L. Goshen and I. Shimshoni, "Balanced exploration and exploitation model search for efficient epipolar geometry estimation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 7, pp. 1230–1242, 2008.

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

[29] L. Kneip and P. Furgale, "Opengv: A unified and generalized approach to real-time calibrated geometric vision," in *2014 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1–8. IEEE, 2014.

[30] L. Kneip, H. Li, and Y. Seo, "Upnp: An optimal o (n) solution to the absolute pose problem with universal applicability," in *European Conference on Computer Vision*, pp. 127–142. Springer, 2014.

[31] L. Kneip, P. Furgale, and R. Siegwart, “Using multi-camera systems in robotics: Efficient solutions to the npnp problem,” in *2013 IEEE International Conference on Robotics and Automation*, pp. 3770–3776. IEEE, 2013.

[32] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, “Automatic differentiation in PyTorch,” in *NeurIPS Autodiff Workshop*, 2017.

[33] S. Omari, M. Bloesch, P. Gohl, and R. Siegwart, “Dense visual-inertial navigation system for mobile robots,” in *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 2634–2640. IEEE, 2015.

[34] 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.

[35] T. Sattler, W. Maddern, C. Toft, A. Torii, L. Hammarstrand, E. Stenborg, D. Safari, M. Okutomi, M. Pollefeys, J. Sivic *et al.*, “Benchmarking 6dof outdoor visual localization in changing conditions,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8601–8610, 2018.

[36] 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.

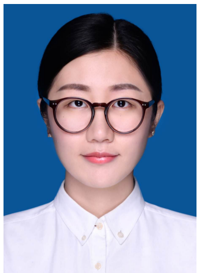
[37] G. Bradski, “The OpenCV Library,” *Dr. Dobb’s Journal of Software Tools*, 2000.

[38] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardos, “Orb-slam: a versatile and accurate monocular slam system,” *IEEE transactions on robotics*, vol. 31, no. 5, pp. 1147–1163, 2015.

[39] C. Linegar, W. Churchill, and P. Newman, “Made to measure: Bespoke landmarks for 24-hour, all-weather localisation with a camera,” in *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 787–794. IEEE, 2016.

Yanmei Jiao received her BS in Intelligence Science and Technology from the Department of Computer Science and Control Engineering, Nankai University, Tianjin, P.R. China in 2017.
She is currently a Ph.D. candidate in the Department of Control Science and Engineering, Zhejiang University, Hangzhou, P.R. China. Her research interests include computer vision and vision based localization.

Yue Wang received his PhD in Control Science and Engineering from Department of Control Science and Engineering, Zhejiang University, Hangzhou, P.R. China in 2016.
He is currently an Associate Professor in the Department of Control Science and Engineering, Zhejiang University, Hangzhou, P.R. China. His latest research interests include mobile robotics and robot perception.



Xiaqing Ding received her BS in Control Science and Engineering from the Department of Control Science and Engineering, Zhejiang University, Hangzhou, P.R. China in 2016.
She is currently a Ph.D. candidate in the Department of Control Science and Engineering, Zhejiang University, Hangzhou, P.R. China. Her latest research interests include SLAM and vision based localization.

Bo Fu received his BS in Control Science and Engineering from the Department of Control Science and Engineering, Shandong University, Jinan, P.R. China in 2017.
He is currently a Ph.D. candidate in the Department of Control Science and Engineering, Zhejiang University, Hangzhou, P.R. China. His latest research interests include multi-sensor calibration and sensor fusion.

Shoudong Huang received the Bachelor and Master degrees in Mathematics, Ph.D. in Automatic Control from Northeastern University, PR China in 1987, 1990, and 1998, respectively. He is currently an Associate Professor at Centre for Autonomous Systems, Faculty of Engineering and Information Technology, University of Technology, Sydney, Australia. His research interests include nonlinear control systems and mobile robots simultaneous localization and mapping (SLAM), exploration and navigation.

Rong Xiong received her PhD in Control Science and Engineering from the Department of Control Science and Engineering, Zhejiang University, Hangzhou, P.R. China in 2009.
She is currently a professor in the Department of Control Science and Engineering, Zhejiang University, Hangzhou, P.R. China. Her latest research interests include motion planning and SLAM.