2021 IEEE International Conference on Robotics and Automation (ICRA 2021)
May 31 - June 4, 2021, Xi'an, China

Robust localization for planar moving robot in changing environment: A perspective on density of correspondence and depth

Yanmei Jiao[1], Lilu Liu[1], Bo Fu[1], Xiaqing Ding[1], Minhang Wang[2], Yue Wang[1] and Rong Xiong[1]

***Abstract*— Visual localization for planar moving robot is important to various indoor service robotic applications. To handle the textureless areas and frequent human activities in indoor environments, a novel robust visual localization algorithm which leverages dense correspondence and sparse depth for planar moving robot is proposed. The key component is a minimal solution which computes the absolute camera pose with one 3D-2D correspondence and one 2D-2D correspondence. The advantages are obvious in two aspects. First, the robustness is enhanced as the sample set for pose estimation is maximal by utilizing all correspondences with or without depth. Second, no extra effort for dense map construction is required to exploit dense correspondences for handling textureless and repetitive texture scenes. That is meaningful as building a dense map is computational expensive especially in large scale. Moreover, a probabilistic analysis among different solutions is presented and an automatic solution selection mechanism is designed to maximize the success rate by selecting appropriate solutions in different environmental characteristics. Finally, a complete visual localization pipeline considering situations from the perspective of correspondence and depth density is summarized and validated on both simulation and public real-world indoor localization dataset.**

I. Introduction

Indoor service robots have been applied in various scenarios during the last decade, such as home, office, restaurant and so on [1] [2]. One of the fundamental techniques for this success is the 2D LiDAR based localization of planar moving robots [3]. To further reduce the cost, cameras are expected to provide visual localization for service robots. However, due to the sensitivity to frequent and complex environmental changes e.g. illumination, texture, objects presence, reliable visual indoor localization remains a challenge [4].

The common pipeline for visual localization is to establish the feature correspondences from the query image to the map and recover the 6 degree of freedom (DoF) camera pose through geometric estimation. In this pipeline, feature matching is vulnerable to environmental changes, causing presence of outliers. To relieve this problem, minimal solutions exploiting minimal number of correspondences for pose estimation call for research, which can be embedded into RANSAC [5] [6] to improve the robustness. For 6DoF solution, 3 correspondences are minimally required for both mono-camera [7] [8] and multi-camera scenarios [9] [10].

[1]Yanmei Jiao, Lilu Liu, Bo Fu, Xiaqing Ding, Yue Wang and Rong Xiong are with the State Key Laboratory of Industrial Control and Technology, Zhejiang University, Hangzhou, P.R. China. [2]Minhang Wang is with the Application Innovate Lab, Huawei Incorporated Company, P.R. China. Yue Wang is the corresponding author wangyue@iipc.zju.edu.cn.
This work was supported in part by the National Nature Science Foundation of China (61903332), and in part by the Science and Technology Project of Zhejiang Province (2019C01043).

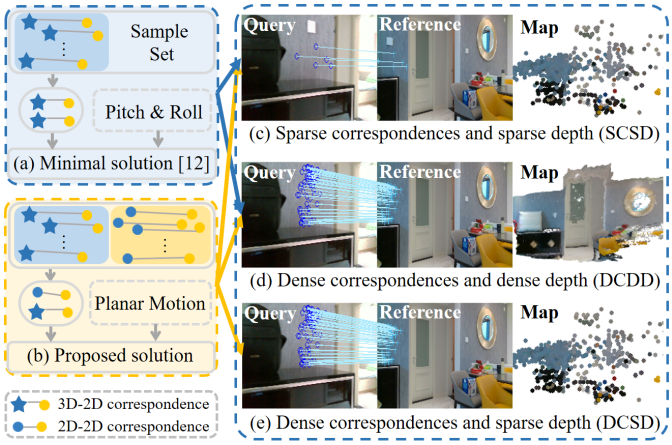


Fig. 1: Left: The sample set comparison of (a) 3D-2D feature based minimal solution and (b) the proposed minimal solution. Right: The illustration of different settings (c) SCSD, (d) DCDD and (e) DCSD. The proposed solution is the first to deal with DCSD as all correspondences with or without depth can be utilized.

Moreover, when the inertial measurement is available, the direction of gravity is known, further reducing the minimal number of correspondences to 2 as shown in [11] [12]. Note that solutions above are mainly applied in the situations where sparse correspondences and sparse depth (SCSD) are available. Whereas in indoor localization, only sparse features [13] [14] [15] may not perform well due to the textureless and highly repetitive areas (*e.g.*, walls, windows and floors) [4].

In recent years, the development of deep neural network brings the possibility for efficiently matching all pixels to generate dense correspondences [16]. In [4], with lots of correspondences, the conventional minimal solution based RANSAC is applied to eliminate the large percentage of outliers, achieving superior accuracy. We argue that this pipeline actually solves the problem of dense correspondences and dense depth (DCDD), which requires a dense 3D model in prior e.g. 3D scanning in [4]. However, such model is difficult to build as it requires large computational capacity and can be noisy. A more applicable setting is localization with dense correspondences but sparse depth (DCSD), which only requires a conventional sparse environmental model built by sparse bundle adjustment [17] [18], as shown in Fig. 1. To solve DCSD problem, still utilizing the conventional pipeline wastes those inliers among dense correspondences without depth, which should have made contributions to

outliers elimination. An ideal pipeline for DCSD is to use both 2D-3D feature matches (correspondences with depth) and 2D-2D feature matches (correspondences without depth) in RANSAC to improve robustness and accuracy.

In this paper, we propose a minimal solution utilizing one 2D-3D correspondence combined one 2D-2D correspondence by taking planar motion constraint into consideration. Specifically, according to the geometry, one point correspondence with depth (1DP) provides two independent constraints about the pose, and one point correspondence without depth (1P) provides one constraint, leading to the 1P1DP based minimal solution for the 3DoF planar motion. To the best of our knowledge, this is the first minimal solution designed for planar motion given DCSD. In addition, together with our previous work on minimal solution with two correspondences with depth (2DP) [12], we present a probabilistic analysis on the two solutions given different outlier ratio and reliable depth ratio, which guides the design of a solution selection mechanism. By composing these modules together, we further discuss the utilization of different minimal solutions in settings with SCSD, DCSD and DCDD, and propose a completed pipeline for robust visual localization. In summary, the contributions of this paper are presented as follows:

- A minimal closed-form solution namely 1P1DP in mono-camera system and MC1P1DP in multi-camera system is proposed for absolute pose estimation of planar moving robot.
- A probabilistic analysis on 1P1DP and 2DP given different environmental characteristics, SCSD, DCSD and DCDD, is presented to guide the design of a solution selection mechanism.
- A completed pipeline for robust visual localization is proposed to automatically pick the appropriate solution according to the current environment, which is then validated on lifelong indoor localization datasets.

II. Related Works

A. Density of feature correspondences

Lots of efforts have been paid to design the robust feature detectors and descriptors in geometric computer vision community. The typical pipeline for handcrafted features is to detect a keypoint and then describe it [13] [19]. Besides robustness, the real time property is also required for features [20] [14]. As convolutional neural networks show superior performance on representation, many learning-based feature detectors [21] [22] [23] and descriptors [24] [25] [26] are proposed to replace the handcrafted procedures. With the improvement of the computing capability, dense feature descriptors can be learned along with detectors to encode more information about the detected keypoints [15] [27]. Then dense feature detectors which perform detection densely on whole image and generate matches from pixel to pixel also developed [16]. This approach often capture global information for feature matching and has shown to provide better correspondences especially in indoor localization [4]. However, the construction of dense 3D map is not mature as sparse map [28] [17] [18] as it requires huge computation and storage capacity, which significantly limits the application of dense features in localization. Therefore, considering the difficulty of the dense map construction and the advantage of dense features in indoor localization, the combination of dense correspondences and sparse depth is valuable.

B. Robust pose estimation solutions

Estimating the camera pose with the feature matches and 3D map as input is typically performed by PnP solvers [7] [8] [12] embedded in robust estimator such as RANSAC [5] [6]. However, these solutions are designed only for features with depth. Therefore, the existing 2D-3D pose estimation algorithms can only cooperate with SCSD or DCDD. The other option to use dense correspondences without dense depth is to compute the pose only with 2D-2D correspondences via epipolar geometry [29]. There are also many minimal solvers for the pose estimation with pure 2D-2D correspondences, such as classical 8-point [30] and 5-point [31] and [32] [33] for reduced DoF situation. Although experiments show that it's practical for localization without depth information, it's inaccurate as it's difficult to constraint scale in the pose computation or optimization. In summary, the solutions for pose estimation are designed only for 2D-3D or 2D-2D correspondences. Therefore, a new solution is needed for utilizing both 2D-3D correspondences obtained from features matched with sparse map and 2D-2D correspondences obtained from the other dense features in DCSD, which is the focus of this paper.

III. Minimal Solution

In this paper, the planar motion property is employed to formulate the indoor localization problem. We assume the image plane is vertical to the ground such that the unknown variables of the pose between reference view r and query view q are the translation along axis x and z and the rotation around y, which are denoted as t_x, t_z and θ as illustrated in Fig. 2. Note that the assumption is easy to satisfy by rotating the appropriate pitch and roll angles obtained from calibration parameters or inertial measurements. Then the rotation and translation matrix from camera coordinate system of view r to view q can be written as:

$$R = \begin{bmatrix} cos(\theta) & 0 & sin(\theta) \\ 0 & 1 & 0 \\ -sin(\theta) & 0 & cos(\theta) \end{bmatrix}, t = \begin{bmatrix} t_x \\ 0 \\ t_z \end{bmatrix} \tag{1}$$

Constraints from 2D-3D correspondence: Given a 2D feature expressed as $p_1 = [\tilde{u}_1, \tilde{v}_1, 1]^T = K^{-1}[u_1, v_1, 1]^T$ in view q and the corresponding 3D point $P_1 = [x_1, y_1, z_1]^T$ in camera coordinate system of view r, according to the projection geometry, we have:

$$\frac{R_1 P_1 + t_x}{\tilde{u}_1} = \frac{R_2 P_1}{\tilde{v}_1} = R_3 P_1 + t_z \tag{2}$$

where $R \triangleq [R_1^T, R_2^T, R_3^T]^T$ and K is the camera intrinsic parameters. Two constraints can be derived from (2) as:

$$\tilde{v}_1 x_1 cos(\theta) + \tilde{v}_1 z_1 sin(\theta) + \tilde{v}_1 t_x - \tilde{u}_1 y_1 = 0 \tag{3}$$

$$\tilde{v}_1 z_1 cos(\theta) - \tilde{v}_1 x_1 sin(\theta) + \tilde{v}_1 t_z - y_1 = 0 \tag{4}$$

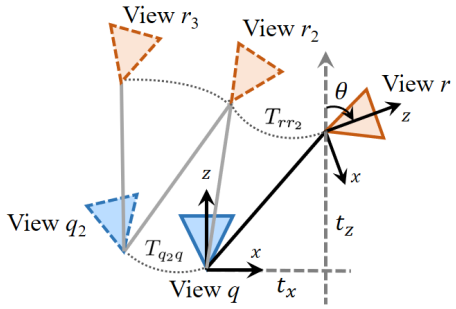


Fig. 2: The illustration of planar motion between multiple query and reference views.

It's clear that given another 2D-3D correspondence, two constraints similar to (3) - (4) can be obtained. The pose can be solved by least square with the four constraints. Since the number of constraints provided by this solution is larger than the unknowns, it can achieve higher accuracy.

Constraints from 2D-2D correspondence: Denoting the 2D-2D correspondence as p_2 from view q and p_3 from view r, the epipolar constraint can be expressed as:

$$p_2{}^T[t]_\times R p_3 = 0 \tag{5}$$

where $[\,\cdot\,]_\times$ denotes the skew symmetry matrix.

Representing unknowns t_x and t_z by θ from (3) - (4) and substituting them into (5), the pose can be solved and the solution is denoted as 1P1DP. The solution is the first to combine correspondence with and without depth for pose estimation in planar motion localization problem.

Extending to multi-reference case: As one query image may correspond to multiple reference images in map [34], integrating these correspondences from other reference images will promote the localization performance. Suppose to sample the 2D-2D correspondence between view r_2 and view q, which is different from that provides the 2D-3D correspondence. As the camera poses of reference views can be obtained from the map, the relative transformation between the two reference views T_{rr_2} is known. Then we have:

$$T_{qr_2} = T_{qr}T_{rr_2} = \begin{bmatrix} R & t \\ \mathbf{0} & 1 \end{bmatrix} T_{rr_2} \tag{6}$$

Note that the unknowns in T_{qr_2} are the same as in T_{qr}, so according to the epipolar constraint:

$$p_2{}^T[t_{qr_2}]_\times R_{qr_2} p_3 = 0 \tag{7}$$

the 1P1DP for multi-reference case can also be solved.

Extending to multi-camera system: The proposed solution can be easily extended to physical multi-camera system or temporal multi-query from continuous frames, such that more observations from other query views can assist the pose estimation of current query view. As shown in Fig. 2, the relative transformation between the two query views T_{q_2q} can be obtained by extrinsic parameters in multi-camera system or visual odometry in multi-query case.

Considering one 2D-2D correspondence provided by view q_2 and r_2, the (6) becomes

$$T_{q_2r_2} = T_{q_2q}T_{qr}T_{rr_2} \tag{8}$$

And the constraint can be obtained by substituting $R_{q_2r_2}$ and $t_{q_2r_2}$ into (7). The rest is similar to the previous 1P1DP solution. The solution for multi-camera system is denoted as MC1P1DP.

IV. Solution Selection

To deal with outliers, the proposed 1P1DP minimal solution is embedded into RANSAC [5] to achieve robust visual localization. In our previous work, we also propose another 2 points based minimal solution, 2DP [12]. To pick the appropriate one in real application, we theoretically analyze the performance of 1P1DP and 2DP with respect to environmental characteristics for guidance.

A. Success rate comparison

Recalling the procedure of obtaining the 2D-3D correspondences, we first retrieve the 2D-2D correspondences between the query image and the reference image, then extract the corresponding 3D information of the reference image in map. In the following, we first consider SCSD, under which the number of 2D-2D and 2D-3D correspondences are the same.

SCSD: Denote the inlier rate of the 2D-2D correspondences as λ. As the depth measurement can be noisy due to distance and material, we also denote the reliable depth rate of reference features as γ, $(0 \leq \lambda, \gamma \leq 1)$. Then the success rate of one trial in RANSAC with 1P1DP solution is:

$$P_{SCSD-1P1DP} = \lambda \cdot (\lambda\gamma) \tag{9}$$

and the success rate of one 2DP sample is:

$$P_{2DP} = (\lambda\gamma) \cdot (\lambda\gamma) \tag{10}$$

We have $P_{2DP} \leq P_{SCSD-1P1DP}$ and the equality exists when $\gamma = 1$, which indicates the difference of the success rate between two solutions depends on the reliable depth rate of the features in the reference image. If the reliable depth rate over the whole reference image is high, the success rate of 2DP is similar to that of 1P1DP. Note that the constraints provided by 2DP is more than that of 1P1DP and the accuracy of 2DP is also better than 1P1DP. However, when reliable depth rate is low, 1P1DP will show advantage, say most features are detected in a distant picture on the wall.

DCSD: In this situation, the inlier rate of 2D-2D correspondences is different. Denote the inlier rate of dense 2D-2D correspondences as λ_d, we have

$$P_{DCSD-1P1DP} = \lambda_d \cdot (\lambda\gamma) \tag{11}$$

Comparing with (11) and (9), the difference depends on the inlier rate of dense and sparse correspondences. When $\lambda_d > \lambda$, the advantage of 1P1DP in DCSD becomes more obvious. Note that dense correspondences are usually generated by networks that considers neighborhood consistency, higher λ_d in DCSD is thus practical.

DCDD: In this situation, γ significantly increases. Therefore, as in SCSD with high γ, 2DP can be the best choice given its better accuracy.

We find that 1P1DP and 2DP have advantages respectively given different environmental characteristics, inspiring that a

better performance can be achieved if the solution is picked correctly for complement. In summary, based on the theoretic analysis as guidance, we present a robust visual localization pipeline from the perspective on density of correspondence and depth as shown in Fig. 3.

B. Learning-based solution selection

Following the result above, we implement a learning based solution selector based on the reliability of depth in map. The solution is denoted as Mix in Fig. 3. The network takes a three-channel image as the input including depth, depth uncertainty and observation number for the features in the reference image, and predict the utilization of 1P1DP or 2DP given that reference image, as shown in Fig. 6. The uncertainty of the depth of each point is obtained from sparse bundle adjustment. ResNet-101 [35] is selected as the backbone followed by a fully connected layer with two outputs. For data construction, we run 1P1DP and 2DP for each training image, and use the solution giving better accuracy as the label. Some examples of the training data are shown in Fig. 6.

V. Experimental Results

To validate over the state-of-the-art algorithms, we perform simulation experiments with generated synthetic data and real-world experiments with public indoor localization dataset to evaluate: i) the accuracy of the proposed 1P1DP and MC1P1DP in the simulated mono-camera and multi-camera system, ii) the robustness with increasing outlier rate and decreasing reliable depth rate, iii) the success rate of the different minimal solutions and learning-based solution selection mechanism in real-world indoor localization.

A. Simulation experiments

By fixing one reference camera coordinate system as the world, we create the synthetic world by randomly generating points in $[-8, 8]^3$ cube. Then random planar motion is generated by creating the pose of query camera with random translation in range of $[-2, 2]$ along x-axis and z-axis, and random rotation in range of $[-\pi, \pi]$ around y-axis. The focal length of the virtual camera is fixed as 800 with the resolution of 1280×960 and the principal point as $(640, 480)$. For multi-camera system, three cameras with a viewing angle difference of 60 degree and a distance difference of 0.25m are bounded together. In each experimental test, 50 visible points are sampled for each camera and projected in both reference and query view, which is similar as in [9]. For evaluation, given the ground truth of the pose of query camera $[R_{gt}|t_{gt}]$, the rotation error of the estimated pose $[R|t]$ is computed as $\arccos(0.5Tr(RR_{gt}^T) - 0.5)$ in degree, and the translation error is $|t - t_{gt}|$ in meter [36].

Accuracy: To evaluate the accuracy of different algorithms, we generate Gaussian noise with zero mean and increasing standard deviation from 0 to 5 pixels to the 2D features. The Gaussian noise on the depth of the 3D points is also added and the standard deviation is fixed as 0.05m. For test on mono-camera system, besides the 2DP [12], the generally utilized 6DoF algorithms in real pose estimation tasks EPnP [8] as well as the latest improved version AP3P [37] are considered. For multi-camera pose estimation algorithms, GP3P [9] and MC2DP [12] are compared.

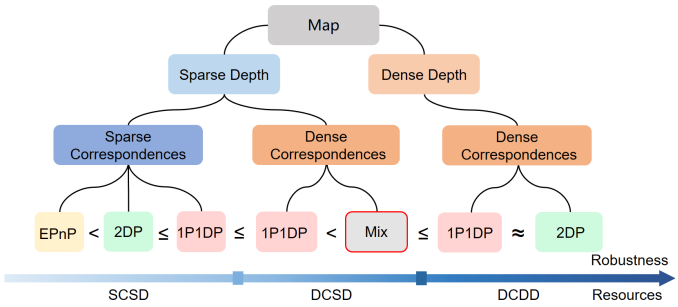
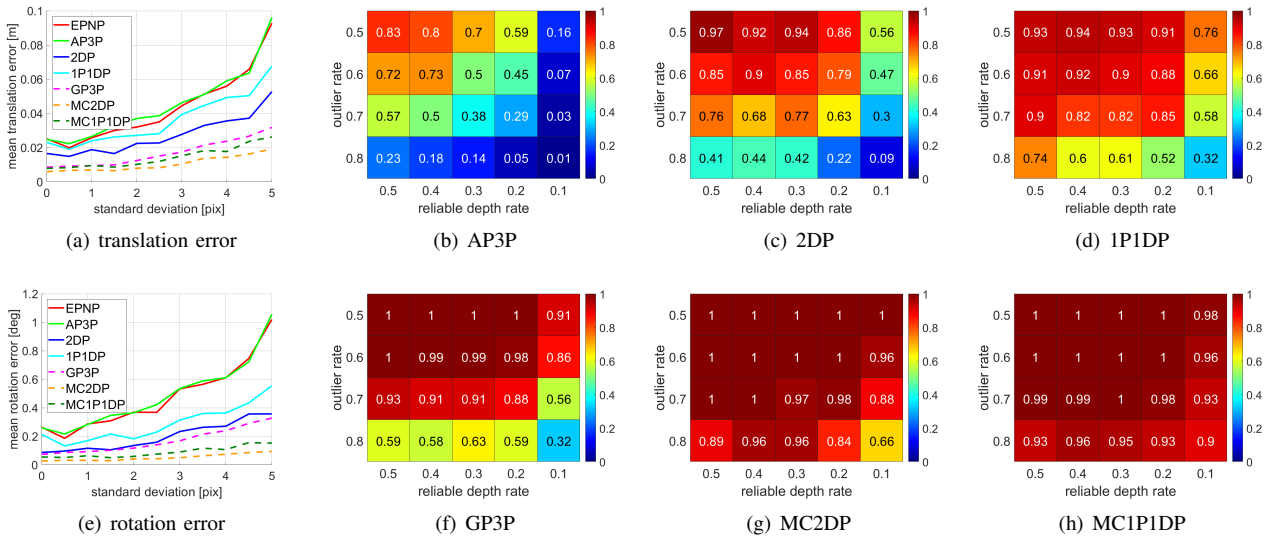


Fig. 3: Robust localization pipeline under different settings. The robustness increases from left to right as analysed in theory. The computation and storage resources also increase due to the density of the correspondences and depth.

We generate 100 experimental tests for each noise level. In each test, all mono-camera and multi-camera algorithms are embedded into RANSAC and 5000 iterations are performed, then the result with the most inliers is selected as the estimated pose. The mean error of translation and rotation over all 100 tests are drawn for mono-camera and multi-camera algorithms comparison in Fig. 4 (a) and (e). Results show that the solutions designed for reduced DoF problem achieve better accuracy, both in mono-camera and multi-camera system, as the number of correspondences required for pose estimation is reduced. And the 2DP solution which utilizes more depth information performs slightly better than proposed 1P1DP solution, which is reasonable as the constraints are stronger.

Robustness: Two experimental settings are designed to test the robustness of mono-camera and multi-camera algorithms: i) increasing outlier rate between all feature correspondences, ii) decreasing reliable depth rate to vary the ratio of correspondences with and without depth (the correspondence with unreliable depth is assumed to be no depth). The total number of feature correspondences for each camera is fixed as 50 and the outliers are generated by randomly associating a certain number of points with random incorrect camera poses. The outlier rate varies from 0.5 to 0.8. Meanwhile, the reliable depth rate is decreasing from 0.5 to 0.1, which means the number of correspondences with depth decreases. The default Gaussian noise with 2 pixel standard deviation on 2D measurements and 0.05m on 3D measurements are added for all cases. The localization result is assumed to be successful if the translation error is lower than 0.1m and rotation error is lower than 1 degree. Total 100 experimental tests for each level are generated and 500 iterations are performed for each algorithm in each test, and the success rate of different algorithms is counted and averaged for robustness comparison.

The results of mono-camera algorithms are shown in Fig. 4 (b)-(d) and multi-camera algorithms in (f)-(h). Results show that the proposed 1P1DP performs better than 2DP and other



(a) translation error (b) AP3P (c) 2DP (d) 1P1DP

(e) rotation error (f) GP3P (g) MC2DP (h) MC1P1DP

Fig. 4: (a) and (e): The accuracy comparison of mono-camera and multi-camera algorithms with increasing 2d noise. (b)-(d) and (f)-(h): The success rate comparison with increasing outlier rate and decreasing reliable depth rate of mono-camera and multi-camera algorithms.

depth correspondence based solutions, as more outliers and less depth measurements present. In addition, the success rate of multi-camera algorithms is higher than that of mono-camera algorithms, as more correspondences are available which provides more positive candidates.

B. Real-world experiments

OpenLORIS dataset [1] is employed to evaluate the feasibility of the proposed algorithm in real world planar moving robot visual localization. The dataset is collected by a wheeled robot equipped with a RealSense D435i and a RealSense T265. The dataset contains challenging lifelong variations including viewpoints, illuminations, dynamic objects changes and human occlusions. The RGBD data from RealSense D435i is utilized for evaluation in this paper. We test the success rate of different algorithms in all situations and show the result of the proposed solution selector, as the validation of the proposed pipeline.

Experimental Settings and Implementation Details: For each query image, top 3 reference images in the map session are retrieved using NetVLAD [34]. Then R2D2 [27] is utilized to get sparse correspondences between query and reference images and NCNet [16] is utilized for dense correspondences. For map building, we triangulate scene points with the observation from multiple frames and refine the reconstruction with bundle adjustment. In evaluation of multi-camera algorithms, 5 temporal images including the query image in query session are retrieved with rotation difference larger than 20 degree and translation difference larger than 0.25 m. The aligned depth image from RGBD sensor is utilized as dense depth map in DCDD test. 500 iterations are performed for each solution in RANSAC and the estimated pose is refined with nonlinear optimization for better accuracy.

For localization performance evaluation, different error thresholds are utilized to count the success rate as shown in Fig. 5 for mono-camera and Table. I for multi-camera algorithms. For better illustration, the success rate shown in Fig. 5 is normalized by the upper bound performance in each session. The upper bound is counted by manually selecting the best result with smallest error among all algorithms in each query case. For multi-camera algorithms, the success rate is the absolute value. In 1P1DP sampling, only map points with depth smaller than 5m is considered to be reliable depth points (DP). All other map points with larger depth as well as those without depth are considered to be points without depth (P). The threshold is chosen according to the inherent attributes of the RGBD sensor.

Some implementation details of solution selection network are as follows. Each channel of input data is generated as a heatmap processed with Gaussian blur and the resolution is the same as RGB image as shown in Fig. 6. Training set is composed of the reference images with difference in accuracy between 1P1DP and 2DP exceeding 0.25 m in translation. We train the network by SGD [38] solver with 12 batch size for 200 epochs. Then the trained selector is performed in each query to automatically select the best solution for pose estimation. The obtained success rate is denoted as Mix in mono-camera and MCMix in multi-camera system.

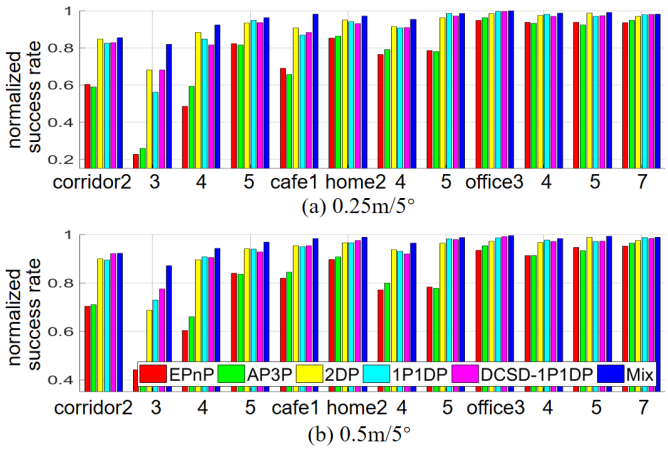
Performance on Sparse Depth: Results in mono-camera and multi-camera system of SCSD show that, 1P1DP and 2DP outperform the common 6DoF algorithms under all thresholds which confirms the importance of utilizing motion property into pose estimation. Comparing result of 1P1DP and 2DP in *cafe1* and *corridor5*, we can find that 1P1DP shows advantage in *corridor5* where the reliable depth rate is low due to the large French windows. Whereas 2DP performs better in *cafe1* where reliable depth rate is relatively higher as the scene is full of texture. On the whole, 1P1DP achieves

TABLE I: Success rate comparison of multi-camera algorithms.

Task	session	home2	home4	home5	corridor2	corridor3	corridor4
	m	0.25 / 0.5 / 1.0	0.25 / 0.5 / 1.0	0.25 / 0.5 / 1.0	0.25 / 0.5 / 1.0	0.25 / 0.5 / 1.0	0.25 / 0.5 / 1.0
	degree	5 / 5 / 5	5 / 5 / 5	5 / 5 / 5	5 / 5 / 5	5 / 5 /5	5 / 5 / 5
SCSD	GP3P	68.81 / 80.69 / 82.79	51.06 / 65.81 / 70.36	77.02 / 79.20 / 79.20	11.42 / 33.91 / 57.66	2.29 / 6.67 / 14.39	18.16 / 30.92 / 35.63
	MC2DP	77.25 / 84.39 / 85.69	71.30 / 83.33 / 87.41	98.33 / 98.46 / 98.46	18.15 / 43.14 / 62.64	3.53 / 9.34 / 19.53	22.18 / 36.43 / 43.10
	MC1P1DP (ours)	77.52 / 84.56 / 85.69	72.85 / 84.64 / 87.74	**99.87 / 99.87 / 99.87**	18.84 / 42.97 / 62.15	3.62 / 9.62 / 21.63	20.28 / 36.37 / 43.52
DCSD	DCSD-MC1P1DP (ours)	77.65 / 84.82 / 85.89	73.07 / 86.31 / 89.53	**99.87 / 99.87 / 99.87**	18.61 / 43.54 / 62.96	3.48 / 9.62 / 21.15	20.17 / 36.74 / 43.67
	MCMix (ours)	**78.89 / 85.22 / 86.69**	**74.78 / 86.75 / 90.23**	**99.87 / 99.87 / 99.87**	**18.98 / 43.60 / 63.30**	**4.48 / 11.05 / 22.49**	**23.35 / 39.33 / 45.53**
	GP3P	67.24 / 82.35 / 84.16	57.16 / 76.51 / 81.54	88.06 / 89.73 / 89.73	15.39 / 38.42 / 58.12	3.33 / 11.62 / 27.35	29.75 / 45.53 / 53.57
DCDD	MC2DP	77.79 / 85.92 / 87.06	74.26 / 88.82 / 94.03	99.23 / 99.36 / 99.36	25.77 / 46.33 / 62.67	**9.72 / 20.44 / 32.78**	34.46 / 47.49 / 54.37
	MC1P1DP (ours)	**77.95 / 86.22 / 88.06**	**74.50 / 89.24 / 94.22**	**99.87 / 100 / 100**	**27.78 / 48.09 / 62.90**	8.86 / 18.20 / 30.35	**36.05 / 48.28 / 54.53**

Task	session	corridor5	office3	office4	office5	office7	cafe1
	m	0.25 / 0.5 / 1.0	0.25 / 0.5 / 1.0	0.25 / 0.5 / 1.0	0.25 / 0.5 / 1.0	0.25 / 0.5 / 1.0	0.25 / 0.5 / 1.0
	degree	5 / 5 / 5	5 / 5 / 5	5 / 5 / 5	5 / 5 / 5	5 / 5 /5	5 / 5 / 5
SCSD	GP3P	37.00 / 59.57 / 77.78	88.33 / 88.33 / 88.33	78.97 / 79.77 / 79.77	95.17 / 95.29 / 95.29	98.86 / 98.95 / 98.95	36.87 / 71.84 / 86.30
	MC2DP	37.99 / 62.84 / 80.29	90.28 / 90.28 / 90.56	82.99 / 85.09 / 85.21	98.05 / 98.43 / 98.43	99.47 / 99.47 / 99.47	42.21 / 74.88 / 86.71
	MC1P1DP (ours)	40.70 / 63.29 / 80.36	90.83 / **91.11** / 91.11	82.64 / 85.40 / 85.52	96.60 / 96.79 / 96.79	99.56 / 99.56 / 99.56	38.17 / 73.65 / **86.89**
DCSD	DCSD-MC1P1DP (ours)	40.68 / 63.16 / 80.63	90.00 / 90.56 / 90.56	82.64 / 85.63 / 85.63	96.66 / 96.97 / 96.97	99.39 / 99.39 / 99.39	37.76 / 73.83 / 86.49
	MCMix (ours)	**41.18 / 64.64 / 81.50**	**91.11 / 91.11 / 91.39**	**83.90 / 86.78 / 86.90**	**98.24 / 98.55 / 98.55**	**99.63 / 99.63 / 99.63**	**43.33 / 77.05 / 86.89**
	GP3P	40.98 / 63.32 / 81.27	88.61 / 89.44 / 89.44	88.62 / 89.31 / 89.31	95.41 / 95.97 / 95.97	93.51 / 95.53 / 95.53	40.57 / 73.48 / 88.66
DCDD	MC2DP	47.46 / 68.98 / **85.15**	**90.00 / 91.11 / 91.11**	**92.18 / 93.45 / 93.56**	**97.29 / 97.48 / 97.48**	94.83 / **95.79 / 95.79**	**52.99 / 78.63 / 89.23**
	MC1P1DP (ours)	**49.36 / 69.51** / 85.06	88.27 / 90.00 / 90.00	87.47 / 90.80 / 90.80	96.22 / 96.79 / 96.79	**94.92 / 95.79 / 95.79**	51.23 / 78.57 / 88.70

[1] Considering the map coverage of the environment, the map sequences for each session are as follows: home1 and home3, corridor1, office1 and office2, cafe2.
[2] For testing on office6, all methods achieve 100 success rate over three thresholds.

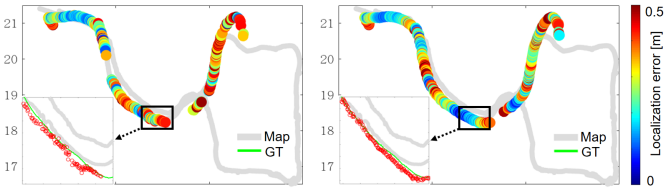


(a) 0.25m/5°
(b) 0.5m/5°

Fig. 5: The normalized success rate comparison of mono-camera algorithms in all sessions.

(a) color (b) depth (c) 1-uncertainty (d) observation num

Fig. 6: Examples of training data. Only sparse depth of the detected features are combined with the uncertainty and observation number to fed into the network.



(a) EPnP (b) Mix

Fig. 7: Localization performance comparison on *cafe1*. The proposed Mix solution achieves more success localization result with higher accuracy.

slightly better performance than 2DP which is more obvious in DCSD-1P1DP. The results reflect the fact that 2DP and 1P1DP can complement each other according to different environmental characteristics, which also explains the reason that the performance of the proposed solution selector is the best in all sessions. A comparison of localization performance on whole trajectory of *cafe1* is also drawn in Fig. 7 for clear illustration. Therefore, utilizing an appropriate strategy to combine the advantages of 2DP and 1P1DP can achieve the best performance in sparse depth condition, which is consistent with the analysis in Section IV.

Discussion about Dense Depth: In the environment with larger changes (significant illumination variations in *corridor*, frequent human activities in *home* and *cafe*), the performance of all methods with dense depth are obviously better than that with sparse depth. However, in the relatively stable *office* environment, the performance with dense depth slightly drops. That is reasonable in two aspects: i) the dense depth from the RGBD sensor is not accurate as the optimized depth in sparse map ii) sparse features perform well in this environment, then dense features will not provide more significant correspondences. And MC2DP performs better in most cases which indicates the MC2DP can be the best choice with better accuracy and robustness when dense depth map is reliable.

VI. Conclusions

In this paper, a novel minimal solution with both minimal number of feature correspondences and minimal depth information namely 1P1DP is proposed. Embedded with the solution, all the matched feature correspondences between query and reference image (with or without depth) can be integrated for pose estimation, which maximizes the sample set in RANSAC framework. Furthermore, the solution enables the combination of dense correspondences and sparse depth, taking advantages of dense features about changing environment while maintaining limited computational requirement for sparse map building. In the future, the strategy for selecting reliable depth features for 1P1DP will be optimized.

REFERENCES

[1] X. Shi, D. Li, P. Zhao, Q. Tian, Y. Tian, Q. Long, C. Zhu, J. Song, F. Qiao, L. Song, *et al.*, "Are we ready for service robots? the openloris-scene datasets for lifelong slam," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3139–3145, IEEE, 2020.

[2] J. R. Ruiz-Sarmiento, C. Galindo, and J. González-Jiménez, "Robot@ home, a robotic dataset for semantic mapping of home environments," *The International Journal of Robotics Research*, vol. 36, no. 2, pp. 131–141, 2017.

[3] S. Thrun, "Probabilistic robotics," *Communications of the ACM*, vol. 45, no. 3, pp. 52–57, 2002.

[4] H. Taira, M. Okutomi, T. Sattler, M. Cimpoi, M. Pollefeys, J. Sivic, T. Pajdla, and A. Torii, "Inloc: Indoor visual localization with dense matching and view synthesis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7199–7209, 2018.

[5] 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.

[6] S. Choi, T. Kim, and W. Yu, "Performance evaluation of ransac family," *Journal of Computer Vision*, vol. 24, no. 3, pp. 271–300, 1997.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] Y. Jiao, Y. Wang, X. Ding, B. Fu, S. Huang, and R. Xiong, "2-entity ransac for robust visual localization: Framework, methods and verifications," *IEEE Transactions on Industrial Electronics*, 2020.

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

[14] 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.

[15] D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superpoint: Self-supervised interest point detection and description," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 224–236, 2018.

[16] I. Rocco, M. Cimpoi, R. Arandjelovic, A. Torii, T. Pajdla, and J. Sivic, "Ncnet: Neighbourhood consensus networks for estimating image correspondences," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[17] 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.

[18] T. Schneider, M. Dymczyk, M. Fehr, K. Egger, S. Lynen, I. Gilitschenski, and R. Siegwart, "maplab: An open framework for research in visual-inertial mapping and localization," *IEEE Robotics and Automation Letters*, vol. 3, no. 3, pp. 1418–1425, 2018.

[19] S. Leutenegger, M. Chli, and R. Y. Siegwart, "Brisk: Binary robust invariant scalable keypoints," in *2011 International conference on computer vision*, pp. 2548–2555, Ieee, 2011.

[20] D. G. Viswanathan, "Features from accelerated segment test (fast)," in *Proceedings of the 10th workshop on Image Analysis for Multimedia Interactive Services, London, UK*, pp. 6–8, 2009.

[21] N. Savinov, A. Seki, L. Ladicky, T. Sattler, and M. Pollefeys, "Quad-networks: unsupervised learning to rank for interest point detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1822–1830, 2017.

[22] L. Zhang and S. Rusinkiewicz, "Learning to detect features in texture images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6325–6333, 2018.

[23] A. Barroso-Laguna, E. Riba, D. Ponsa, and K. Mikolajczyk, "Key. net: Keypoint detection by handcrafted and learned cnn filters," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5836–5844, 2019.

[24] V. Balntas, E. Riba, D. Ponsa, and K. Mikolajczyk, "Learning local feature descriptors with triplets and shallow convolutional neural networks.," in *Bmvc*, vol. 1, p. 3, 2016.

[25] E. Simo-Serra, E. Trulls, L. Ferraz, I. Kokkinos, P. Fua, and F. Moreno-Noguer, "Discriminative learning of deep convolutional feature point descriptors," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 118–126, 2015.

[26] K. Simonyan, A. Vedaldi, and A. Zisserman, "Learning local feature descriptors using convex optimisation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 8, pp. 1573–1585, 2014.

[27] J. Revaud, C. De Souza, M. Humenberger, and P. Weinzaepfel, "R2d2: Reliable and repeatable detector and descriptor," in *Advances in Neural Information Processing Systems*, pp. 12405–12415, 2019.

[28] J. L. Schonberger and J.-M. Frahm, "Structure-from-motion revisited," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4104–4113, 2016.

[29] Q. Zhou, T. Sattler, M. Pollefeys, and L. Leal-Taixe, "To learn or not to learn: Visual localization from essential matrices," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3319–3326, IEEE, 2020.

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

[31] 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.

[32] 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.

[33] C. C. Chou and C.-C. Wang, "2-point ransac for scene image matching under large viewpoint changes," in *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3646–3651, IEEE, 2015.

[34] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "Netvlad: Cnn architecture for weakly supervised place recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5297–5307, 2016.

[35] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[36] 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.

[37] T. Ke and S. I. Roumeliotis, "An efficient algebraic solution to the perspective-three-point problem," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7225–7233, 2017.

[38] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural computation*, vol. 1, no. 4, pp. 541–551, 1989.