Supplementary Material
2-Entity RANSAC for Robust Visual Localization: Framework, Methods and Verifications

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

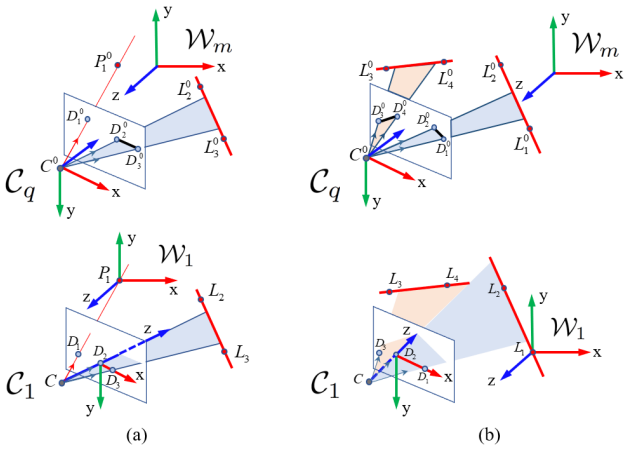


Fig. 1: The illustration of intermediate reference frame for (a) 1 point 1 line and (b) 2 lines case.

I. Minimal Solutions

A. Monocular camera system

1) 1 point 1 line: ***The choice of*** $\mathcal{C}_1$: As shown in Fig. 1 (a), in $\mathcal{C}_q$, the origin is the camera center C^0, the camera projection ray associated with the 2D point is given by its normalized direction vector $\overrightarrow{d_1}$, and projection rays of the two end points of the 2D line are given by normalized direction $\overrightarrow{d_2}$, $\overrightarrow{d_3}$.

In $\mathcal{C}_1$, the projection ray associated with the 2D point is denoted by CD_1, and the line, CD_2 and CD_3. Specifically, $\mathcal{C}_1$ should satisfy the following conditions:

- The new camera center C is $\begin{bmatrix}0 & 0 & -1\end{bmatrix}^T$.
- CD_2 lies on the z axis such that $D_2 = \mathbf{0_{3\times 1}}$.
- CD_3 lies on the xz plane and the point D_3 is the intersection point between the x axis and the ray.
- The point D_1 lies on the xy plane.

After the transformation, D_3 can be computed as follows

$$D_3 = \begin{bmatrix}\tan(\arccos(\overrightarrow{d_2}\cdot\overrightarrow{d_3})) & 0 & 0\end{bmatrix}^T$$

Then the corresponding points in $\mathcal{C}_q$ are calculated as follows

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

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

The choice of $\mathcal{W}_1$: The transformation of the world reference is a translation which transforms the 3D point to the origin of $\mathcal{W}_1$

$$T_{\mathcal{W}_1\mathcal{W}_m} = \begin{bmatrix}\mathbf{I}_{3\times 3} & -P_1^0\\ \mathbf{0}_{1\times 3} & 1\end{bmatrix}$$

Thus in $\mathcal{W}_1$

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

Note that these points are all known.

2) 2 lines: ***The choice of*** $\mathcal{C}_1$: As shown in Fig. 1 (b), in $\mathcal{C}_q$, the camera center is denoted as C^0, the camera projection rays associated with the two 2D lines are given by pairs $(\overrightarrow{d_1}, \overrightarrow{d_2})$ and $(\overrightarrow{d_3}, \overrightarrow{d_4})$, respectively. In $\mathcal{C}_1$, the 3D projection rays associated with the two 2D lines are represented by (CD_1, CD_2), and (CD_2, CD_3). Specifically, $\mathcal{C}_1$ should satisfy the following conditions:

- The new camera center C is $\begin{bmatrix}0 & 0 & -1\end{bmatrix}^T$.
- The intersection line of the two interpretation planes represented by projection ray CD_2 lies on the z axis such that $D_2 = \mathbf{0_{3\times 1}}$.
- CD_1 lies on the xz plane and the point D_1 is the intersection point between the x axis and the ray.
- The point D_3 lies on the xy plane.

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

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

where $\overrightarrow{d_{12}}$ is the direction vector of the intersection line CD_2. After such a transformation, D_1 can be computed as follows

$$D_1 = \begin{bmatrix}\tan(\arccos(\overrightarrow{d_1}\cdot\overrightarrow{d_{12}})) & 0 & 0\end{bmatrix}^T$$

Then the corresponding points in $\mathcal{C}_q$ are as follows

$$C^0 = \mathbf{0_{3\times 1}}\,,\; D_1^0 = C^0 + \frac{\overrightarrow{d_1}}{\overrightarrow{d_1}\cdot\overrightarrow{d_{12}}}\,,\; D_2^0 = C^0 + \overrightarrow{d_{12}}$$

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

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

$$L_1 = \mathbf{0}_{3\times 1},\; L_{\{i=2,3,4\}} \triangleq \begin{bmatrix}X_i & Y_i & Z_i\end{bmatrix}^T$$

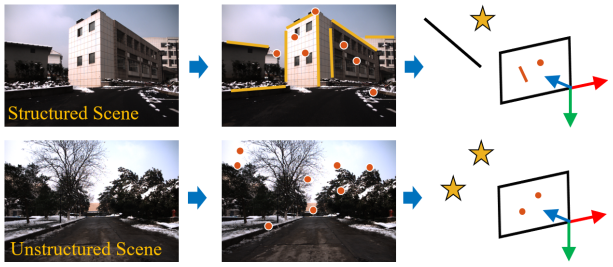


Fig. 2: Some examples of the unstructured and structured scene which are selected manually according to experience.

II. Analysis of Success Probability

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

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

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

$$P_{1p1l} = \frac{m}{p} \cdot \frac{n}{l} = \lambda \cdot \gamma$$
$$P_{2p} = \frac{m}{p} \cdot \frac{m-1}{p-1} = \lambda \cdot \frac{\lambda p - 1}{p-1}$$
$$P_{mixed} = \frac{m}{p} \cdot \frac{m+n-1}{p+l-1} = \lambda \cdot \frac{\lambda p + \gamma l - 1}{p+l-1}$$

Then we have

$$P_{1p1l} - P_{mixed} = \lambda(\gamma - \frac{\lambda p + \gamma l - 1}{p+l-1}) \propto \gamma - (\lambda - a)$$
$$P_{mixed} - P_{2p} = \lambda(\frac{\lambda p + \gamma l - 1}{p+l-1} - \frac{\lambda p - 1}{p-1}) \propto \gamma - (\lambda - a)$$

where $a = \frac{1-\lambda}{p-1} > 0$. From which we can easily derive that

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

Generally, $p-1 \gg 1-\lambda$ holds for most real world applications, which means a is a small positive number close to 0. Thus, with proper scaling, the following conclusion can also hold

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

III. Results with Real Data

For real world experiments, YQ-Dataset is utilized, which was collected in different weather conditions and seasons. The experimental platform to collect data is a four-wheel mobile robot equipped with a VLP-16 Velodyne LiDAR, a MTi 100 IMU and a pointgrey stereo camera. The robot is under remote control and each session of data is collected almost along the same way with the length of around 1,300 meters. As a campus environment, the high-dynamics include pedestrians, cyclists and moving cars. While the low-dynamics include parking cars, the shape of the trees and weather changes. Three of the four datasets used were collected during three days in summer 2017 at the different time, denoted as 2017-0823, 2017-0827, 2017-0828, and the other was collected in winter 2018 after snow, denoted as 2018-0129. The 2017-0823 dataset is selected to build the 3D map and the other three datasets are utilized to evaluate the localization performance. The ground truth relative pose is provided by aligning the synchronized 3D LiDAR scans. To get the 3D-2D feature matches between the query image and the map, we utilized the following steps:

- Obtain the camera poses and the 3D-2D point matches in the map using visual inertial SLAM software [1].
- Run Line3D++ algorithm [2] to get the 3D-2D line matches in the map.
- Get the 3D-2D points/lines matches for the query session with the descriptors of LibVISO2 [3] and LBD [4].

A. Inlier score estimation Details

To verify the advantages of the proposed feature scoring method, we conduct the inlier score estimation experiment with the real world YQ-dataset. For fair comparison, we adopt LibVISO2 [3] as the feature points extraction and descriptor computation for all methods. 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 [5]. For the inlier probability estimation in [6], we randomly sample 1000 query images from 2018-0827 session to generate the distance ratios of inlier and outlier samples (denoted as S_{in} and S_{out} in [6]). And the number of linear equations (N_R in [6]) is set to 100 to estimate the inlier ratio (α in [6]) in a least-squares way. For the inlier score estimation in [7], the similarity between each two features given the descriptors is computed with cosine distance. In the proposed method, we also randomly sample 1000 query images from 2018-0827 session to generate inlier and outlier samples as training data. After normalization of the estimated inlier score, we unify that the correspondence with score higher than 0.8 is inlier for all three methods as in [6].

B. ATE comparison

The trajectories of each query session with the proposed mPose2ESelScore localization method can be seen in Fig. 4.

There are some cases to illustrate the effectiveness of the proposed model selection mechanism and the mPose localization method which can be seen in Fig. 5 and Fig. 3. The sampling strategy of both two query images in Fig. 5 predicated by the trained CNN network mentioned in Section IV-A is 1P1L. The comparison of localization error between 1P1L and 2P of the two cases shows the correctness of the model selection mechanism. As for the mPose method, Fig. 3 shows two typical cases that the inlier rate of feature matches in the former query image is better than the current one. Therefore utilizing the two views to estimate the pose of the current query image, the performance would be obviously better than the monocular camera method, which indicates the practicability of the proposed mPose localization method.

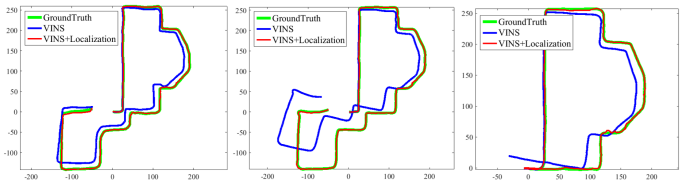
C. Multi-camera dataset

To further evaluate the multi-camera algorithms, we collect a new dataset with a physical five-camera robot system and

(a) mono: 6.23m/61.15°, mPose: 0.15m/0.94°

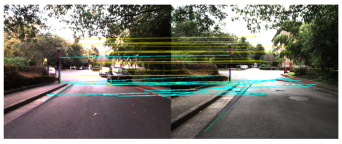
(b) mono: 2.03m/15.82°, mPose: 0.87m/2.41°

Fig. 3: Cases for "mPose". The left image shows the current query and map image. And the right shows the former one.

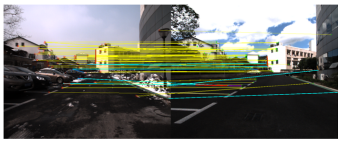


(a) 20170827 session (b) 20170828 session (c) 20180129 session

Fig. 4: Trajectory comparison of each session.



(a) 2P: 0.35m/0.52°
1P1L: 0.17m/0.42°



(b) 2P: 0.42m/0.39°
1P1L: 0.29m/0.24°

Fig. 5: Cases for model selection mechanism. The yellow line indicates the point matches between the query image (left) and the map image (right), and the cyan line indicates the line matches. The caption following the sampling strategy shows the translation error and rotation error of the query image.

there are four sessions denoted as 0325, 0329, 0331, 0333. The experimental platform is equipped with two VLP-16 Velodyne LiDAR, a Pandar40P LiDAR, a MTi-G-710 IMU and five pointgrey cameras: a stereo camera in front of the car, and the other three on the left, right and rear respectively. The experimental platform is shown in Fig. 6. There are large perspective changes in this dataset, which is challenging for visual localization.

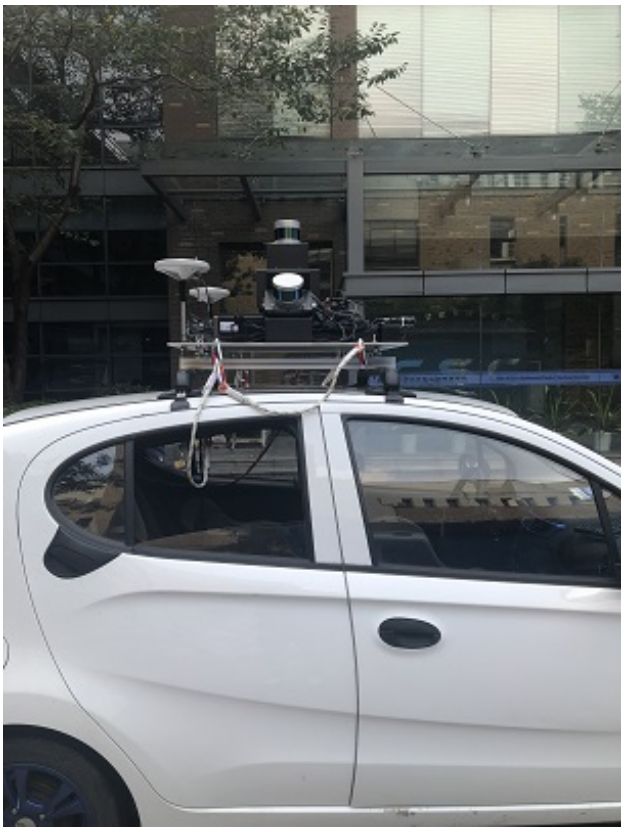


Fig. 6: The multi-camera experimental platform.

REFERENCES

[1] 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.

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

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

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

[5] https://opencv.org/.

[6] 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.

[7] 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.