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

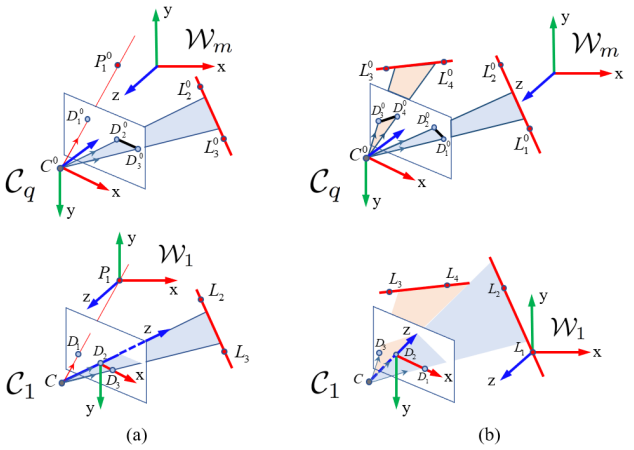


Fig. 1: The illustration of intermediate reference frame for (a) 1 point 1 line and (b) 2 lines case.

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 $\mathcal{W}_1$

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

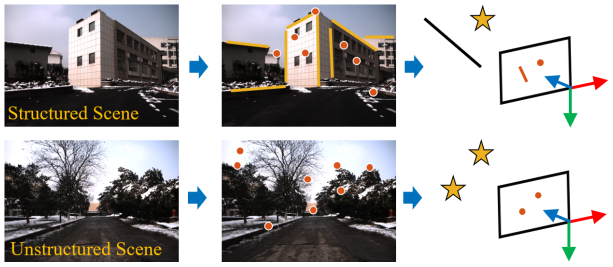


Fig. 2: Some examples of the unstructured and structured scene which are selected manually according to experience.

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.

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

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

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

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

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

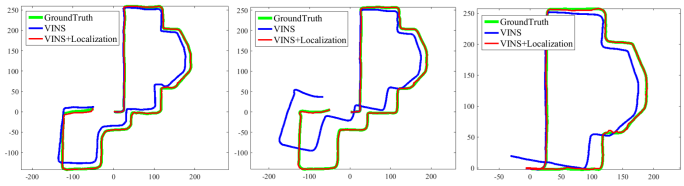
Generally, $p-1 \gg 1-\lambda$ holds for most real world applications, which means a is a small positive number close to 0. Thus, with proper scaling, the following conclusion can also hold

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

III. Results with Real Data

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

Fig. 3: Trajectory comparison of each session.

1) ATE comparison: The trajectories of each query session with the proposed 2Pose2ESelScore localization method can be seen in Fig. 3.

There are some cases to illustrate the effectiveness of the proposed model selection mechanism and the "2-pose" localization method which can be seen in Fig. 4 and Fig. 5. The sampling strategy of both two query images in Fig. 4 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 "2-pose" method, Fig. 5 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 "2-pose" localization method.

(a) 2P: 0.35m/0.52°
1P1L: 0.17m/0.42°

(b) 2P: 0.42m/0.39°
1P1L: 0.29m/0.24°

Fig. 4: 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.

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.

(a) mono: 6.23m/61.15°, 2-pose: 0.15m/0.94°

(b) mono: 2.03m/15.82°, 2-pose: 0.87m/2.41°

Fig. 5: Cases for "2-pose". The left image shows the current query and map image. And the right shows the former one.

TABLE I: Success rate on the selected scene of YQ-Dataset.

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