Supplementary Material for Efficient Consensus Maximization for Visual Localization by Globally Optimal Rotation Search

Jiawei Cai, Yanmei Jiao, Dibin Zhou, Xiumei Li, Rong Xiong, and Yue Wang

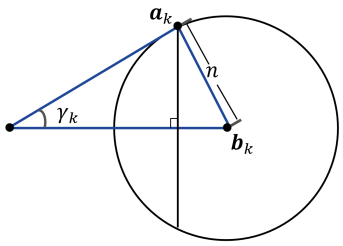


Fig. 1. Relating the Euclidean and the angular error.

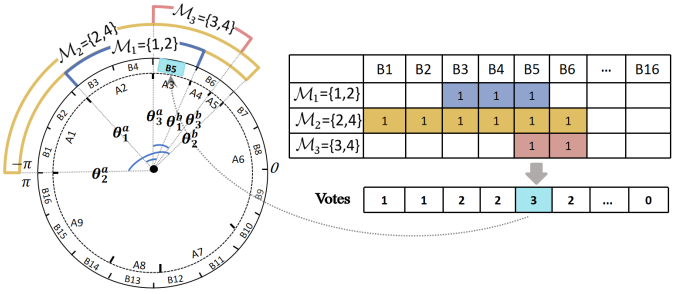


Fig. 2. The illustration of proposed voting and acceleration algorithm.

I. Method

A. Rotation search in SO(3)

As shown in Fig. 1, the inlier threshold γ_k (in radians) is data dependent and constraint by

$$\cos\gamma_k = \frac{\|\mathbf{a}_k\|^2 + \|\mathbf{b}_k\|^2 - n^2}{2\|\mathbf{a}_k\|\|\mathbf{b}_k\|}. \tag{1}$$

Therefore, for all inliers that satisfy $|s\widetilde{\mathbf{u}}_k - \mathbf{R}\mathbf{p}_k| \leq n$, we can transform the error metric from Euclidean distance to angular distance as

$$\angle(\mathbf{b}_k, \mathbf{R}\mathbf{a}_k) \leq \gamma_k, \quad k \in \mathcal{I}, \tag{2}$$

where $\mathbf{b}_k := s\widetilde{\mathbf{u}}_k$, $\mathbf{a}_k := \mathbf{p}_k$.

*1) **Dimensionality reduction of rotation**:* Inspired by GORE, we can achieve dimensionality reduction for rotation. Specifically, the process of iterating over the remaining measurements i to align with measurement k can be detailed as follows. First, the rotation that aligns measurement k according to (2) is identified as $\mathbf{R}_k$. This rotation is then decomposed into two separate rotations to explain it more clearly:

$$\mathbf{R}_k = \mathbf{A}\mathbf{B}, \tag{3}$$

where, $\mathbf{B}$ is a rotation determined by (2), $\mathbf{A}$ represents a rotation by an angle $\theta \in [-\pi, \pi]$ about the axis $\mathbf{B}\mathbf{a}_k$.

Since $\mathbf{A}$ preserves $\mathbf{B}$, the constraint on measurement k imposed by (2) is always satisfied. Then, to enable $\mathbf{R}_k$ to also transform $\mathbf{a}_i$ to $\mathbf{b}_i$, we can solve for $\mathbf{A}$. Specifically, for each measurement i, a corresponding $\mathbf{A}$ can be determined. If we denote the rotation $\mathbf{A}$ that aligns the maximum number of measurements i with measurement k as $\mathring{\mathbf{A}}$, then $\mathring{\mathbf{A}}$ must satisfy the following condition:

$$\begin{aligned} &\max_{\mathbf{R}\in SO(3), \mathcal{I}_k \subseteq \mathcal{M}\setminus\{k\}} |\mathcal{I}_k| + 1 \\ &\text{subject to} \quad \angle(\mathbf{b}_i, \mathring{\mathbf{A}}\mathbf{B}\mathbf{a}_i) \leq \gamma_i, \quad i \in \mathcal{I}_k. \end{aligned} \tag{4}$$

For the rotation $\mathbf{B}$, we define the rotation of $\hat{\mathbf{B}}$ perfectly aligned $\mathbf{a}_k$ to $\mathbf{b}_k$, denoted by

$$\mathbf{b}_k = \hat{\mathbf{B}}\mathbf{a}_k. \tag{5}$$

It is important to note that $\hat{\mathbf{B}}$ is not identical to $\mathbf{B}$. The rotation $\hat{\mathbf{B}}$ aligns $\mathbf{a}_k$ and $\mathbf{b}_k$ perfectly, whereas $\mathbf{B}$ is a rotation that satisfies (2), meaning it only aligns $\mathbf{a}_k$ with $\mathbf{b}_k$ within an angular error γ_k. This relationship is also illustrated in Fig. 1(a) in the manuscript. More formally, we can define the following region:

$$S_\gamma(\mathbf{b}) := \{\mathbf{a} \in \mathbb{R}^3 | \|\mathbf{a}\| = 1, \angle(\mathbf{a}, \mathbf{b}) \leqslant \gamma\}. \tag{6}$$

Thus, $S_{\gamma_k}(\mathbf{b}_k)$ represents the spherical region generated by rotating the measurement $\mathbf{b}_k$ within an angular distance of γ_k. Consequently, the relationship $\mathbf{B}\mathbf{a}_k \in S_{\gamma_k}(\mathbf{b}_k)$ is self-evident.

Additionally, to clearly indicate the dependency of $\mathbf{A}$ on the rotation axis $\mathbf{r}$ and angle θ, we now denote it as $\mathbf{A}_{\theta,\mathbf{r}}$:

$$\mathbf{A}_{\theta,\mathbf{r}} = \{exp(\theta[\mathbf{r}]_\times)\}. \tag{7}$$

Then we can denote the possible region obtained by applying such a rotation $\mathbf{A}$ to any unit norm point $\mathbf{p}$ as $circ(\mathbf{p}, \mathbf{r})$:

$$circ(\mathbf{p}, \mathbf{r}) := \{\mathbf{A}_{\theta,\mathbf{r}}\mathbf{p} | \theta \in [-\pi, \pi]\}. \tag{8}$$

Based on the above decomposition of $\mathbf{R}_k$, $\mathbf{B}\mathbf{a}_k$ is the rotation axis of $\mathbf{A}$, so the interior of $S_{\gamma_k}(\mathbf{b}_k)$ is also the set of possible rotation axis of $\mathbf{A}$. Then applying the same rotational procedure to $\mathbf{a}_i$ as illustrated in Fig. 1(a) in the manuscript, the possible region in which $\mathbf{R}_k\mathbf{a}_i = \mathbf{A}_{\theta,\mathbf{r}}\mathbf{B}\mathbf{a}_i$ lies can be denoted as:

$$L_k(\mathbf{a}_i) := \{circ(\mathbf{p}, \mathbf{r}) | \mathbf{p} \in S_{\gamma_k}(\hat{\mathbf{B}}\mathbf{a}_i), \mathbf{r} \in S_{\gamma_k}(\mathbf{b}_k)\}. \tag{9}$$

Therefore, when $L_k(\hat{\mathbf{B}}\mathbf{a}_i) \bigcap S_{\gamma_i}(\mathbf{b}_i) = \emptyset$, the pair $(\mathbf{a}_i, \mathbf{b}_i)$ cannot be aligned by any rotation $\mathbf{R}_k$ satisfying (2) with

$(\mathbf{a}_k, \mathbf{b}_k)$, the corresponding relationship $(\mathbf{a}_i, \mathbf{b}_i)$ can be safely removed without affecting the results. At the same time, the problem of solving for the 3D rotation $\mathbf{R}$ is thus reduced to solving for a one-dimensional angle θ.

*2) **Interval voting**:* A more illustrative voting case is shown in Fig. 2, on the left side, we visually show how $\mathcal{M}_1$, $\mathcal{M}_2$, and $\mathcal{M}_3$ sequentially cover different continuous intervals from bin 1 to bin 6. After populating the binary table on the right side of Fig. 2 with these entries, we then count the votes for all bins. It becomes evident that bin 5 corresponds to the maximum intersection of $\mathcal{M}_1$, $\mathcal{M}_2$, and $\mathcal{M}_3$, receiving votes from all three. Therefore, the point pairs corresponding to bin 5 represent the union of the point pairs contained within $\mathcal{M}_1$, $\mathcal{M}_2$, and $\mathcal{M}_3$, which corresponds to the index set of point pairs $\{1, 2, 3, 4\}$.

The acceleration is primarily achieved through the use of a large binary table. However, since this table only stores binary values, it does not require significant space complexity. Additionally, the discretization resolution provides a balance between memory usage and accuracy. We have found that extremely high resolutions often decrease accuracy, as small intervals can disperse the voting results. Noise leads to uncertainty intervals that are similar but not completely identical, allowing for the merging of intervals corresponding to different sets at lower resolutions. Since this approximation only occurs near numerical values, its impact on the final average estimate of the inlier set is limited and acceptable.

*3) **Rotation search algorithm**:* The time complexity analysis of rotation search algorithm is as follows. For a given k, by traversing all $i \neq k$, we obtain K angular intervals $\{\Theta_i\}$. In response to this, we propose an improved discretized interval voting algorithm. Since this method requires frequent operations on a discretized array, we leverage the advantages of difference arrays in such scenarios to further reduce the time complexity. Overall, the time complexity of our algorithm is $O(K + L)$, where L is the length of the discretized interval. Therefore, line 9 in Alg. 2 in the manuscript takes $O(K + L)$ time. Typically, we can evenly divide the interval $[-\pi, \pi]$ using a discretization step of 0.5° or 1°, resulting in L being 360 or 720, respectively. Additionally, Gore uses interval stabbing to handle these K angular intervals, with a complexity of $O(K \log K)$. In the worst case, for all k, the overall time complexity of Alg. 2 in the manuscript is $O(K(K + L))$. Compared to Gore's $O(K^2 \log K)$, the advantage of our algorithm becomes more pronounced as K increases, such as when using dense feature matching like LoFTR [1].

B. Search in SE(3)

*1) **Robustness of rotation search against translation perturbation**:* In this section, we provide the proofs related to (13) and (14) in the manuscript. As defined in the manuscript, $\mathbf{x} := \mathbf{a} - \mathbf{p}_{corg}$ and $\mathbf{x}' := \mathbf{a} - (\mathbf{p}_{corg} + \epsilon)$. Then, for the new pairs $(\mathbf{x}'_i, \mathbf{b}_i)$ and $(\mathbf{x}'_k, \mathbf{b}_k)$, we have:

$$\mathbf{x}'_i = \mathbf{x}_i - \epsilon, \tag{10}$$

$$\mathbf{x}'_k = \mathbf{x}_k - \epsilon. \tag{11}$$

Then, we first need to solve for $\hat{\mathbf{B}}$ that perfectly aligns $\mathbf{x}_k$ to $\mathbf{b}_k$, which allows us to obtain:

$$\mathbf{b}_k = \hat{\mathbf{B}}\mathbf{x}_k. \tag{12}$$

Similarly, after introducing the perturbation ϵ to $\mathbf{p}_{corg}$, it becomes:

$$\mathbf{b}_k = \hat{\mathbf{B}}'\mathbf{x}'_k. \tag{13}$$

Substituting (11) into (13) and combining with (12), we can obtain:

$$\hat{\mathbf{B}}'\epsilon = (\hat{\mathbf{B}}' - \hat{\mathbf{B}})\mathbf{x}_k. \tag{14}$$

Next, for the point pair i, combining (10) and (14), we have:

$$\begin{aligned}\hat{\mathbf{B}}'\mathbf{x}'_i &= \hat{\mathbf{B}}'(\mathbf{x}_i - \epsilon) \\ &= \hat{\mathbf{B}}'\mathbf{x}_i + \hat{\mathbf{B}}\mathbf{x}_k - \hat{\mathbf{B}}'\mathbf{x}_k.\end{aligned} \tag{15}$$

Thus, we have:

$$\begin{aligned}\hat{\mathbf{B}}'\mathbf{x}'_i - \hat{\mathbf{B}}\mathbf{x}_i &= \hat{\mathbf{B}}'\mathbf{x}_i + \hat{\mathbf{B}}\mathbf{x}_k - \hat{\mathbf{B}}'\mathbf{x}_k - \hat{\mathbf{B}}\mathbf{x}_i \\ &= \hat{\mathbf{B}}'\mathbf{x}_i + \hat{\mathbf{B}}\mathbf{x}_k - \hat{\mathbf{B}}'\mathbf{x}_k - \hat{\mathbf{B}}\mathbf{x}_i + (\hat{\mathbf{B}}\mathbf{x}_i - \hat{\mathbf{B}}\mathbf{x}_i) \\ &= \hat{\mathbf{B}}'(\mathbf{x}_i - \mathbf{x}_k) - \hat{\mathbf{B}}(\mathbf{x}_i - \mathbf{x}_k).\end{aligned} \tag{16}$$

Next, based on the triangle inequality, we can derive:

$$\begin{aligned}||\hat{\mathbf{B}}'\mathbf{x}'_i - \hat{\mathbf{B}}\mathbf{x}_i|| &\leqslant ||(\hat{\mathbf{B}}' - \hat{\mathbf{B}})\mathbf{x}_i|| + ||(\hat{\mathbf{B}}' - \hat{\mathbf{B}})\mathbf{x}_k|| \\ &\leqslant \frac{||\mathbf{x}_i||}{||\mathbf{x}_k||}||\epsilon|| + ||\epsilon|| \\ &\leqslant (1 + \frac{||\mathbf{x}_i||}{||\mathbf{x}_k||})||\epsilon||.\end{aligned} \tag{17}$$

To maintain the consensus relationship between measurements i and k, it is necessary to ensure that

$$G_1 \subseteq G_2. \tag{18}$$

Then based on (17) and (18), we can obtain:

$$(1 + \frac{||\mathbf{x}_i||}{||\mathbf{x}_k||})||\epsilon|| \leqslant r_1 + r, \tag{19}$$

where r_1 is the axis corresponding to angle $\delta(\theta')$, while r is the axis corresponding to angle γ_i. Fig. 4 in the manuscript also provides a more intuitive depiction of this relationship.

In practical applications, both γ_i and γ_k represent small angular errors, and their values are assumed to be consistent. Then for angle $\delta(\theta')$, we can further derive from $\delta(\theta) := 2|\theta| \sin\left(\frac{\gamma_k}{2}\right) + \gamma_k$ which is mentioned in Sec. II-B3 of the main text:

$$\delta(\theta') \approx (|\theta'| + 1)\gamma_k. \tag{20}$$

It is important to note that, both $\mathbf{x}_i$ and $\mathbf{x}_k$ are normalized three-dimensional vectors. Thus, utilizing the Euclidean and angular errors (1), we can obtain:

$$||\epsilon|| \leq \sqrt{2(1 - \cos(|\theta'| + 2)\gamma_k)}. \tag{21}$$

Obviously, the range of the translational perturbation that rotation search can tolerate depends on both θ' and γ_k. Theoretically, the translational perturbation may be small. However, in practical applications, we can choose a fairly large size to achieve higher efficiency.

TABLE I
SUCCESS RATE COMPARISON OF OPENLORIS-SCENE.

	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
deg	2 / 5 / 8	2 / 5 / 8	2 / 5 / 8	2 / 5 / 8	2 / 5 / 8	2 / 5 / 8
P3P	**39.9** / 64.1 / 67.6	32.9 / 49.7 / 55.0	**36.5** / 60.7 / 67.1	38.6 / 52.7 / 57.8	9.9 / 17.2 / 21.4	19.9 / 25.6 / 29.9
EPnP	39.3 / 63.8 / 66.2	31.3 / 47.8 / 52.6	34.0 / 57.3 / 64.8	36.9 / 50.7 / 55.2	9.0 / 15.3 / 19.3	18.2 / 23.2 / 26.6
AP3P	39.8 / 64.1 / 67.5	32.5 / 49.2 / 54.4	36.3 / 60.3 / 66.6	38.2 / 52.3 / 56.2	9.9 / 16.8 / 21.3	19.4 / 24.9 / 29.1
ours	37.6 / **65.7** / **70.5**	**33.1** / **53.2** / **59.3**	35.5 / **70.2** / **80.4**	**41.2** / **55.0** / **58.8**	**11.9** / **19.2** / **24.5**	**22.8** / **30.1** / **33.6**
	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
deg	2 / 5 / 8	2 / 5 / 8	2 / 5 / 8	2 / 5 / 8	2 / 5 / 8	2 / 5 / 8
P3P	64.6 / 79.8 / 84.9	51.4 / 55.3 / 57.2	**69.6** / 72.9 / 74.4	**70.3** / 83.0 / 89.0	57.1 / 77.0 / 88.9	82.1 / **87.6** / **87.8**
EPnP	64.4 / 78.5 / 83.7	48.9 / 54.2 / 56.1	69.5 / 72.9 / 74.1	70.1 / 82.5 / 88.1	53.9 / 74.2 / 85.4	**82.8** / 87.2 / 87.5
AP3P	64.4 / 79.7 / 84.7	50.6 / 55.2 / 57.5	69.4 / 72.8 / 73.9	69.9 / 83.1 / 88.3	57.6 / 77.0 / 88.4	81.9 / 87.4 / 87.7
ours	**67.7** / **82.1** / **87.6**	**52.5** / **56.9** / **57.8**	68.5 / **73.9** / **77.5**	69.3 / **83.4** / **90.6**	**59.9** / **80.3** / **92.2**	82.3 / 86.6 / 87.6

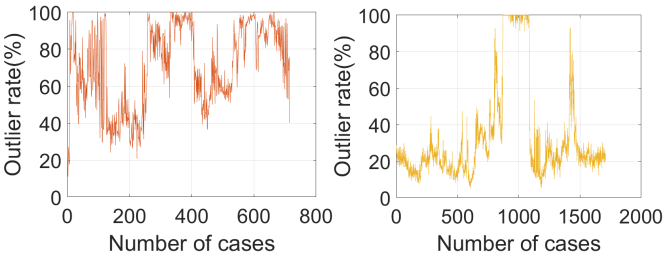
II. EXPERIMENTAL RESULTS

A. Real world experiments

*1) **Data generation of the Newer College dataset and ZJG dataset**:* We generate a usable set of 2D-3D point correspondences through a series of operations on the two datasets, using the same data generation pipeline. This includes map construction, image retrieval, and feature matching. Specifically, we utilize COLMAP [2] to construct a 3D map from the reference sequences. For reference image retrieval, we employ NetVLAD [3] to identify the most similar reference images to the query images. Feature extraction is carried out using SuperPoint [4], and feature matching is performed using the Nearest Neighbor method. Based on the aforementioned operations, we can obtain 2D-3D data.

*2) **OpenLORIS dataset**:* The OpenLORIS dataset [5] is a visual localization dataset designed for wheeled robots, which is collected from robots equipped with RealSense D435i sensors. The main features of this dataset are its rich variability and the wide range of real-world application scenarios. The OpenLORIS dataset covers a wide range of common scenarios such as homes, offices, hallways, and cafes, which illustrate the challenges that robots may encounter in their daily lives. These challenges mainly come from variations in lighting conditions, different observation viewpoints, and the presence of dynamic objects and humans in the scene. The detailed process of 2D-3D data generation is as follows. Specifically, for each query image, we use NetVLAD to retrieve the top three reference images from the map session. We then apply SuperPoint and KNN matching [6] on the retrieved map images and the query image to establish 2D-2D feature correspondences. For map construction, we triangulate scene points with the observation from multiple frames and refine the reconstruction with bundle adjustment. This process ultimately yields the 2D-3D data needed for our task.

The results in Table. I demonstrate that our method outperforms other compared methods in most scenarios. Notably, the improvement is most significant in the home5 scene. For instance, in the home5 scene, within an error range of 1 meter and 8°, our method demonstrates an improvement of over 13% compared to other comparison methods. In contrast, the enhancement in the cafe1 scene is less pronounced. We also conducte an analysis to investigate the reasons for this discrepancy. We show the ground truth outlier rates for both scenarios in Fig. 3. It can be observed that the proportion of cases with extreme outlier rates is significantly higher in the home5 scenario than in the cafe scenario. Specifically, in home5, 38% of cases have an outlier rate above 80%, and 17% of cases have an outlier rate exceeding 95%. For the cafe scene, cafe1, where the improvement is less significant, our analysis reveals that, in this scenario, cases with an outlier rate below 50% account for over 90%. Therefore, in such scenarios, our algorithm's advantage cannot be fully demonstrated. The strength of our method lies in its ability to guarantee deterministic optimality even under high outlier conditions.



(a) Home5 Outlier Rate (b) Cafe1 Outlier Rate

Fig. 3. Line charts of actual outlier rates for scenes home5 and cafe1.

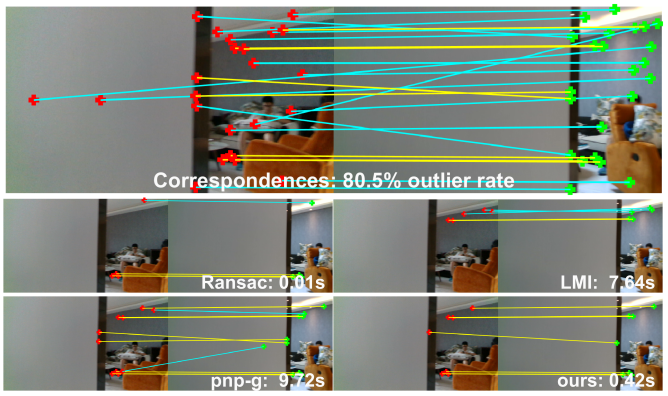


Fig. 4. The yellow lines indicate inliers and cyan lines are outliers.True inliers are defined as correspondences with reprojection errors of less than 10 pixels [7]. There are only 8 inliers among total 39 correspondences due to the significant environmental changes. The comparison of identified inliers by LMI, png-g, Ransac and our proposed method. We also show the computation time of each method.

REFERENCES

[1] J. Sun, Z. Shen, Y. Wang, H. Bao, and X. Zhou, "Loftr: Detector-free local feature matching with transformers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 8922–8931.

[2] A. Fisher, R. Cannizzaro, M. Cochrane, C. Nagahawatte, and J. L. Palmer, "Colmap: A memory-efficient occupancy grid mapping framework," *Robotics and Autonomous Systems*, vol. 142, p. 103755, 2021.

[3] 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*, 2016, pp. 5297–5307.

[4] M. Mera-Trujillo, S. Patel, Y. Gu, and G. Doretto, "Self-supervised interest point detection and description for fisheye and perspective images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 6497–6506.

[5] 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)*. IEEE, 2020, pp. 3139–3145.

[6] Q. Chen, D. Li, and C.-K. Tang, "Knn matting," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 9, pp. 2175–2188, 2013.

[7] Y. Jiao, Y. Wang, X. Ding, B. Fu, S. Huang, and R. Xiong, "2-entity random sample consensus for robust visual localization: Framework, methods, and verifications," *IEEE Transactions on Industrial Electronics*, vol. 68, no. 5, pp. 4519–4528, 2021.