2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 25-29, 2020, Las Vegas, NV, USA (Virtual)

Collaborative Semantic Perception and Relative Localization Based on Map Matching

Yufeng Yue*, Chunyang Zhao, Mingxing Wen, Zhenyu Wu, and Danwei Wang

***Abstract*— In order to enable a team of robots to operate successfully, retrieving accurate relative transformation between robots is the fundamental requirement. So far, most research on relative localization mainly focus on geometry features such as points, lines and planes. To address this problem, collaborative semantic map matching is proposed to perform semantic perception and relative localization. This paper performs semantic perception, probabilistic data association and nonlinear optimization within an integrated framework. Since the voxel correspondence between partial maps is a hidden variable, a probabilistic semantic data association algorithm is proposed based on Expectation-Maximization. Instead of specifying hard geometry data association, semantic and geometry association are jointly updated and estimated. The experimental verification on Semantic KITTI benchmarks demonstrate the improved robustness and accuracy.**

I. Introduction

The collaborative operation of a group of robots has attracted significant attention, attempts have been made to resolve relative localization [1], collaborative 3D mapping [2], and formation control [3]. Among them, retrieving the accurate relative localization between robots is the most critical, which is obtained by matching the local maps of neighboring robots. Existing works in map matching thus far have relied on geometry features such as plane [4], lines [5], and points [6]. Recently, semantic mapping has attracted increasing attention [7]. Semantic map consists of geometry mapping and semantic information , which provides high level semantic information for map matching. This paper moves one step forward by focusing on collaborative probabilistic semantic map matching.

In general, collaborative map matching attempts to calculate the relative transformation matrix between neighboring robots by matching the 3D local maps. Current map matching approaches only unitize geometry information [1], [6], [8] and ignore high-level semantic information. Since 3D map is highly compressed from raw sensor data, the available semantic information should be modeled and combined into the registration process. Nevertheless, latest collaborative semantic mapping research has focused on combining local maps into a global representation [9], and not on how semantic information can improve the relative localization accuracy between robots.

This research project is supported by National Research Foundation (NRF) Singapore, ST Engineering-NTU Corporate Lab under its NRF Corporate Lab@ University Scheme. (*Corresponding Author: Yufeng Yue*)

All authors are with School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore (yyue001@e.ntu.edu.sg).

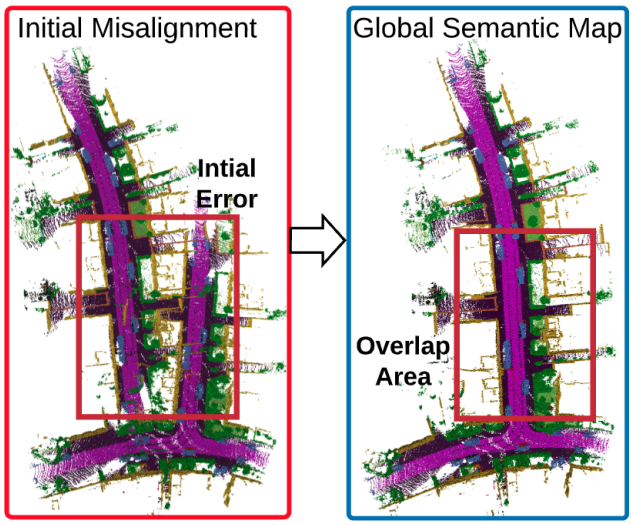


Fig. 1. Demonstration of semantic map matching with the collaborative robot system developed by ourselves, the experimental site is inside the university campus.

In the large community of data registration, dense scan registration algorithms such as ICP [10], NDT [11], GICP [12] are related to map matching. However, there are fundamental differences as they work on different data levels. Dense scan registration aims to estimate the transformation between successive raw sensor scans on single robot level, while map matching works for large 3D local maps on collaborative robots level. Therefore, scan registration is operated on dense raw data, which is essentially sample of continuous surfaces. In the mapping process, multiple scans are integrated to generate discontinuous and voxelized 3D maps. In addition, dense scan registration does not consider communication. However, it's not feasible to transmit raw sensor data between robots with limited bandwidth. Therefore, semantic map matching is considered as a promising way to estimate relative transformation. Fig. 1 illustrates an example of semantic map matching in an outdoor scene.

This paper bridges the gap between the advances in geometry map matching that rely on pure geometry features, and those in semantic map merging, focusing on semantic information integration. The key novelty of this work is the mathematical modeling of the overall collaborative semantic map matching (COSEM) problem and the derivation of its probability decomposition. More specifically, this paper proposes a semantic data association strategy based on

978-1-7281-6211-9/20/$31.00 ©2020 IEEE

Expectation-Maximization to jointly estimate semantic and geometry associations.

The rest of the paper is organized as follows. Section II reviews the related work. Section III gives an overview of the proposed framework. Section IV explains the theoretical basis for collaborative semantic map matching. Section V shows the experimental procedures and results. Section VI concludes this paper and discusses future work.

II. Related Work

In this section, the paper first gives an overview of single robot semantic mapping, and then introduces recent work on collaborative geometry map matching.

A. Single Robot Semantic Mapping

Simultaneous Localization and Mapping (SLAM) is initially proposed to address robot localization and mapping problem in unknown environment [13]. As described in Sec. I, there are essentially differences between SLAM and map matching. Therefore, SLAM will not be discussed in depth. In this paper, we only utilize SLAM to estimate odometry and generate local map for each single robot.

With the fast development of deep learning, it becomes possible to obtain the semantic information by applying CNN models. Deeplab [14], at present, is one of the best model that has excellent performance in semantic segmentation. Since semantic information is also important for environmental perception [15], [16] integrates the semantic information into mapping process. Regarding the reconstruction of moving objects, [17] incrementally fuses sensor observations into a consistent semantic map. More recently, a semantic mapping system [18] is presented that uses object-level entities to construct semantic objects and integrate them into the framework. Aforementioned approaches promote the development of single robot semantic mapping, but collaborative semantic map matching is still an open problem.

B. Collaborative Geometry Map Matching

For a group of robots in arbitrary and GPS-denied environments, the fundamental challenge is to perceive the environment and estimate relative transformation between robots [19]. The complexity of the problem comes at modeling the environment, extracting proper features and calculating transformations, while reducing communication and computational burden.

With the widespread application of 3D sensors, 3D geometry matching has received extensive attention. [1] provides a detailed evaluation of different registration algorithms for geometry map matching. For 3D occupancy mapping, [6] proposes multirobot 3-D mapping merging algorithm with octree-based occupancy grids. [20] applies submap matching approach for indoor/ outdoor robots mapping. To incorporate high-level structure information, [5] proposes a hierarchical map matching framework. However, they all use the same type of sensors. In the unstructured environment, heterogeneous sensors (Lidar, cameras) are needed to fully understand the environment. Therefore, a multiple data association strategy is proposed for heterogeneous map matching in [8]. To further address map matching in day and night environment, [2] has proposed a thermal camera-Lidar collaborative mapping framework based on multi-modal environment perception. In summary, the above algorithms have made great contributions to the problem of geometry map matching. In summary, a novel system framework and theoretical formula for semantic map matching is desired.

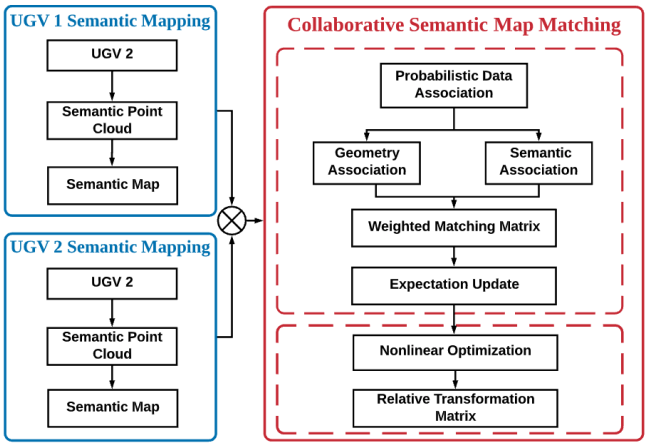


Fig. 2. The framework of collaborative semantic map matching.

III. The Framework of Collaborative Semantic Map Matching

The overview of the system architecture is depicted in Figure 2, where the framework consists of two modules: *single robot semantic mapping and collaborative semantic map matching*. Single robot semantic mapping is at the local mapping level, each robot builds its own semantic map based on multimodal semantic information fusion. Collaborative semantic map matching is at the global map level, the robots communicate with each other to transmit and fuse the local semantic maps into a global semantic map.

Single Robot Level Definition: *For each robot r, the objective is to estimate its local semantic map $\mathbf{M}_r$ given its camera observations $\mathbf{I}_r^{1:t}$, 3D laser observations $\mathbf{L}_r^{1:t}$ and robot trajectory $\mathbf{x}_r^{1:t}$.*

$$p(\mathbf{M}_r|\mathbf{I}_r^{1:t},\mathbf{L}_r^{1:t},\mathbf{x}_r^{1:t}) \tag{1}$$

In single robot level, raw sensor data $\mathbf{I}_r^{1:t}$ and $\mathbf{L}_r^{1:t}$ serves as the input. First, the multimodal information fusion algorithm is applied to generate semantic point cloud [21]. Then, the semantic point cloud is used to generate local semantic map, and the output is single robot semantic map $\mathbf{M}_r$ (see IV-A).

Collaborative Robots Level Definition: *The objective is to estimate the relative transformation matrix $\mathbf{T}_r$ given a set of semantic local maps $M_{(r,\mathbf{R}_r)}$ from neighboring robots $\mathbf{R}_r$.*

$$p(\mathbf{T}|\mathbf{M}_r,\Phi_{r_n\in\mathbf{R}_r}(\mathbf{M}_{r_n})) \tag{2}$$

In collaborative robots level, each robot r is assumed to have estimated its semantic map $\mathbf{M}_r$ based on local

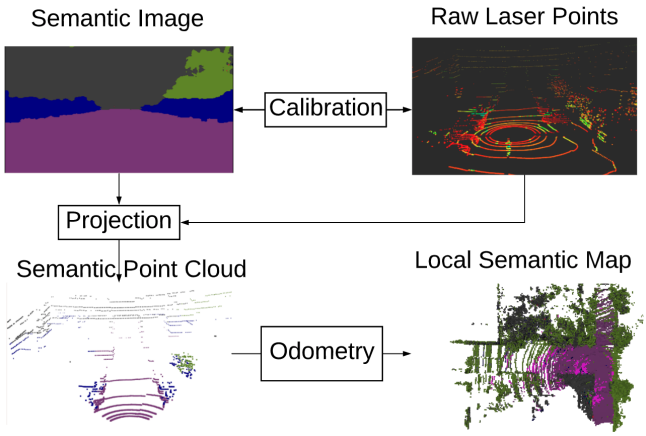


Fig. 3. Flowchart of single robot semantic mapping.

observations. Robot r receives the local maps $\mathbf{M}_{r_n}$ from all the nearby robots $r_n \in \mathbf{R}_r$. Given the received local maps, Expectation-Maximization algorithm is applied to estimate the transformation matrix by making use of semantic information (see IV-B and IV-C).

IV. Collaborative Semantic Map Matching

This section presents detailed collaborative semantic map matching framework, which is divided into three subsections, i.e., single robot semantic mapping, collaborative semantic map matching and Expectation-Maximization.

A. Single Robot Semantic Mapping

The input for collaborative semantic map matching is the local maps generated by each robot. In this paper, the semantic 3D mapping approach utilize a vision camera and a 3D Lidar as its sensing units. Firstly, the vision camera and 3D Lidar are calibrated to the same coordinate frame [21]. Then, the images are processed by Deeplab model [22] to obtain semantic images, which outputs 19 semantic classes. Following by that, the semantic image is projected to the Lidar coordinate frame to generate semantic point cloud, which serves as the direct input for semantic mapping module. Finally, we update occupancy probability and semantic label probability to generate local semantic maps [23].

B. Collaborative Semantic Map Matching

In (2), the formulation is defined as a set of local maps. Since in real applications, each robot r receives partial maps sequentially, which means the relative transformation estimation is a serial process. Therefore, we simplify the input as the current robot map $\mathbf{M}_r$ and latest incoming map $\mathbf{M}_s$ from neighboring robots. We define each map $\mathbf{M}_r = \{\mathbf{M}_r^i\}_{i=1}^{N_{\mathbf{M}_r}}$, where $N_{\mathbf{M}_r}$ is the number of voxels in map $\mathbf{M}_r$. For the matching of two maps $\mathbf{M}_r$ and $\mathbf{M}_s$, the problem can be described as estimating the most likely relative transformation matrix $\mathbf{T}$ by matching the two partial geometry maps $\mathbf{M}_r$ and $\mathbf{M}_s$. To achieve that, $\mathbf{M}_r$ is fixed as the model map, and another map $\mathbf{M}_s$ is used as the scene map, where the rigid transformation $\mathbf{T}$ transforms $\mathbf{M}_s$ to $\mathbf{M}_r$ coordinate system. $\mathbf{T} \in SE(3)$ is a rigid transformation matrix with rotation matrix $\mathbf{R} \in SO(3)$ and translation vector $\mathbf{t} \in \mathbb{R}^3$. The maximum a posterior (MAP) estimate $\hat{\mathbf{T}}$ of the transformation matrix $\mathbf{T}$ is formulated as:

$$\hat{\mathbf{T}} = \max_{\mathbf{T}} \ \log \ p(\mathbf{T}|\mathbf{M}_r, \mathbf{M}_s) \tag{3}$$

To solve this parameter maximization problem, we first model transformation matrix $\mathbf{T}$ as a 6D random variable. Then, $\mathbf{T}$ can be modeled as a multivariate Gaussian distribution, where $\mathbf{T} \sim (\mathbf{T}_{MAP}, \mathbf{T}_{\Sigma_{MAP}})$. The MAP estimate is then factorized into the product of maximum likelihood estimation (MLE) and prior estimation in (4).

$$p(\mathbf{T}|\mathbf{M}_r, \mathbf{M}_s) = p(\mathbf{M}_r|\mathbf{T}, \mathbf{M}_s) \cdot p(\mathbf{T}) \tag{4}$$

where term $p(\mathbf{M}_r|\mathbf{T}, \mathbf{M}_s)$ is a common maximum likelihood estimation (MLE) problem. $p(\mathbf{T})$ denotes the prior information, which can be given by high-level feature descriptors or GPS information. If no prior transformation matrix is given, the prior distribution is assumed to be a uniform distribution. Here, we assume that the prior transformation matrix is given and we can focus on solving the MLE problem.

To estimate $\hat{\mathbf{T}}$, voxel-wise correspondences need to be established. The main difference between geometry map matching and semantic map matching is how they establish correspondences. Previous work on geometry match only defines each voxel $\mathbf{M}_r^i$ in 3D geometry map as $\mathbf{M}_r^i = \{m_r^i\}$, where m_r^i denotes the geometry coordinate. Here, we redefine the voxel as $\mathbf{M}_r^i = \{m_r^i, l_r^i\}$. l_r^i represents the semantic label, which refers to the maximum probability of an element from a finite set of discrete class labels $\mathcal{L} = \{l_1, l_2, \cdots\}$. Here, we have geometry feature $m_r^i \in \mathbb{R}^3$ in $3D$ space and semantic feature $l_r^i \in \mathbb{R}^1$ in $1D$ space. The error metric over $p(\mathbf{M}_r|\mathbf{T}, \mathbf{M}_s)$ can be reformulated as a nonlinear optimization problem with respect to the random variable $\mathbf{T}$ (see (5)).

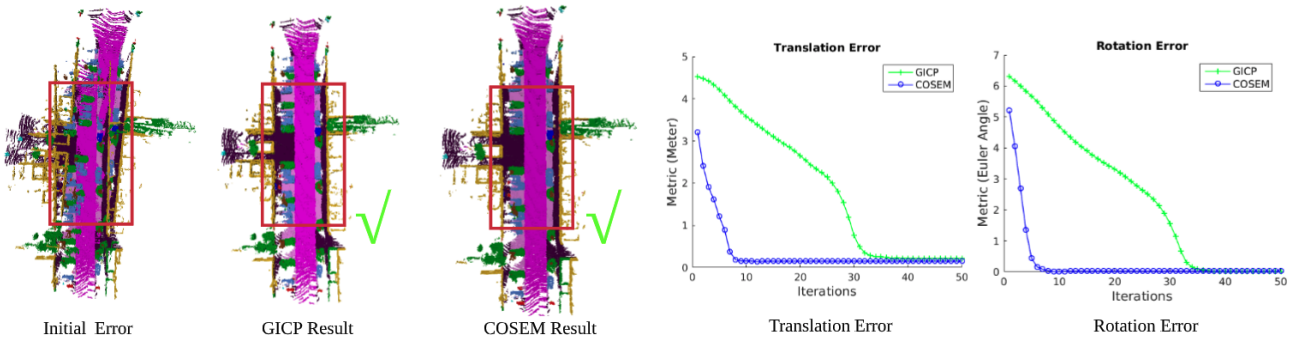
$$\hat{\mathbf{T}} = \max_{\mathbf{T}} \ \log \ p(\mathbf{M}_r|\mathbf{T}, \mathbf{M}_s) \tag{5}$$

(5) estimates the parameter $\hat{\mathbf{T}}$ that maximizes the overlap between map $\mathbf{T}$ and map $\mathbf{M}_s$. However, (5) can't be solved directly, the data association between voxels need to be estimated firstly.

C. Expectation-Maximization

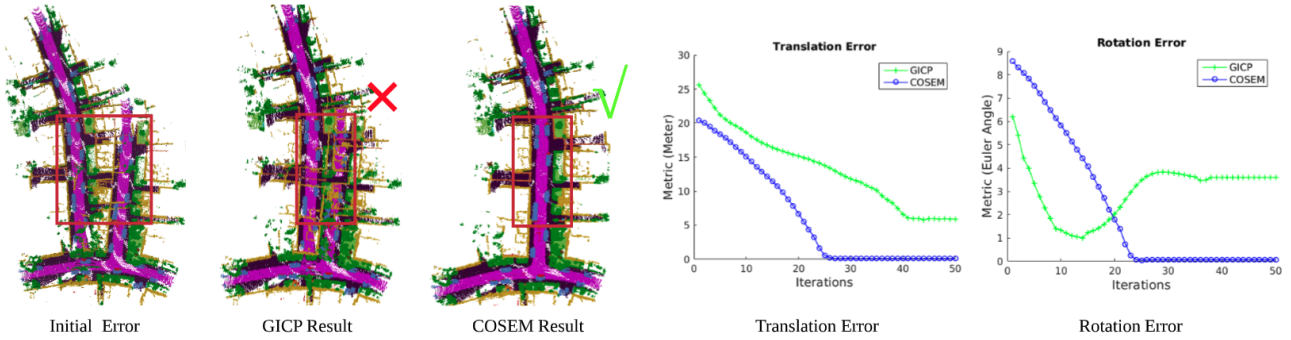
Since the voxel correspondence is not known a prior, we have to infer the hidden data association. We introduce $\mathbf{D} = \{d_{i,j}\}_{i=1}^{N_{mr}}$ to denote the set of all correspondence between map $\mathbf{M}_r$ and $\mathbf{M}_s$. Let $d_{i,j}$ denote the correspondence relationship between voxel $\mathbf{M}_r^i$ and voxel $\mathbf{M}_s^j$, such that $d_i = j$ signifies that voxel $\mathbf{M}_r^i$ corresponds to voxel $\mathbf{M}_s^j$. Then, (5) can be rewritten into the form of Expectation-Maxization problem in (6)-(6), where $\mathbf{D}$ is the set of hidden data associations.

$$\hat{\mathbf{T}}, \hat{\mathbf{D}} = \max_{\mathbf{T},\mathbf{D}} \sum_{d_i \in \mathbf{D}} p(\mathbf{D}|\mathbf{T}, \mathbf{M}_r, \mathbf{M}_s)[\log p(\mathbf{M}_r|\mathbf{T}, \mathbf{D}, \mathbf{M}_s)] \tag{6}$$



Initial Error GICP Result COSEM Result Translation Error Rotation Error

(a) The matching results on Semantic KITTI 00 dataset. Both algorithms converge to a global minimum. In contrast, COSEM converges much faster.



Initial Error GICP Result COSEM Result Translation Error Rotation Error

(b) The matching results on Semantic KITTI 02 dataset. COSEM converges to a global minimum, but GICP falls into a local minimum.

Fig. 4. The results of semantic map matching accuracy in two datasets, where COSEM is compared with GICP.

EM has the advantage of assigning probabilistic data association and update iteratively, rather than assigning hard decisions. To achieve that, there are two steps for EM: E-step efficiently estimates the hidden variables by evaluating the expectation, while the M-step solves the transformation matrix based on E-step using optimization algorithms. First, given the transformation matrix $\mathbf{T}_k$ of last step, the MLE of data association $\hat{\mathbf{D}}^{k+1}$ is computed. Then, given $\hat{\mathbf{D}}^{k+1}$, the MLE of transformation matrix $\hat{\mathbf{T}}^{k+1}$ is updated. This process will be performed iteratively until the convergence threshold is met.

1) Semantic Data Association: traditional geometry map matching approaches usually perform nearest neighbor search of Euclidean distance or match geometry hand-crafted features. Here, we introduce semantic association to increase the data association accuracy and avoid ambiguity.

$$\hat{d}^{k+1} = \max_{d_i} \prod_{i \in N_{\mathbf{M}_r}} \sum_{d_i \in \mathbf{D}} \underbrace{p(d_i | m_r^i, m_s^j, \hat{\mathbf{T}}^k)}_{\text{geometry association}} \underbrace{p(l_r^i, l_s^j | d_i, m_r^i, m_s^j, \hat{\mathbf{T}}^k)}_{\text{semantic association}} \tag{7}$$

As defined in (7), the E-step is factorized into two steps, geometry association and semantic association. In geometry association process, each voxel performs nearest neighbor search. The semantic association then updates the probability of matching voxels belong to the same semantic label.

We introduce $\omega_i = e^{-||m_s^j - f(\mathbf{T}, m_r^i)||^2_{\Sigma_g}}$ to measure the weight of the corresponding pairs, which is the negative power of the distance metric between m_s^j and m_r^i.

$$p(d_i | m_r^i, m_s^j, \hat{\mathbf{T}}^k) = \begin{cases} \omega_i, \ m_s^j \text{ is nearest voxel of } m_r^i \\ 0, \text{otherwise} \end{cases} \tag{8}$$

Then, by applying semantic association formula, those correspondences pairs with contradictory semantic labels will be rejected. For example, even m_s^j is nearest voxel of m_r^i, ω_i=0 if they do not belong to the same label.

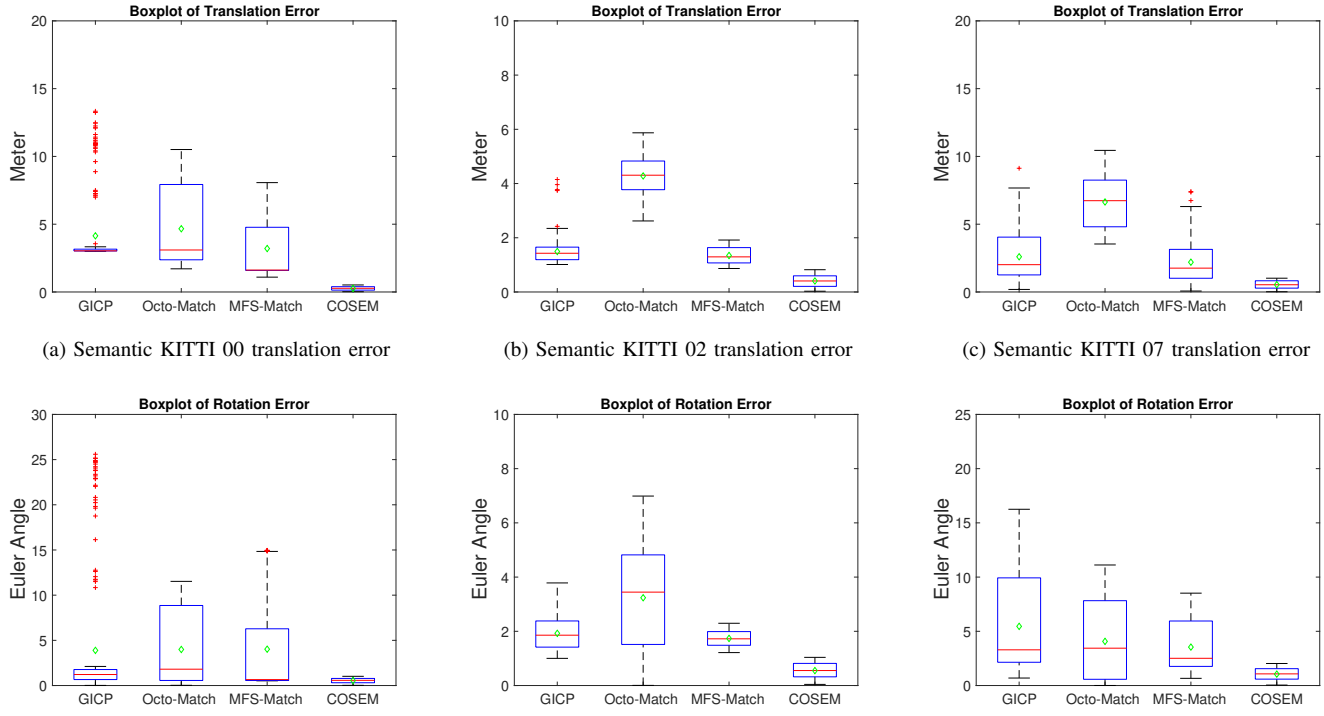
2) Nonlinear Optimization: Solving nonlinear Optimization is to minimize the log-likelihood of (9) based on the data association established in last part. Here, probabilistic weight ω_i is applied to measure the weight of each corresponding pair, where a smaller value means less weight.

$$\hat{\mathbf{T}}^{k+1} = \max_{\mathbf{T}} \prod_{i \in N_{\mathbf{M}_r}} \sum_{d_i \in \mathbf{D}} \omega_i \cdot p(\mathbf{M}_r^i | \hat{d}_i^{k+1}, \mathbf{T}, \mathbf{M}_s^j) \tag{9}$$

V. Experimental Results

Experiments are conducted on Semantic KITTI dataset [24]. The experiments on Semantic KITTI dataset is executed in a desktop PC with an Intel Core i7-6700HQ CPU @2.60GHz, 24 GB RAM in a standard C++ environment. The semantic segmentation model is trained by using the cityscapes dataset [25]. LOAM [26] is applied for single robot pose estimation, and Octomap [27] is utilized for basic 3D geometry mapping at the resolution of 0.2m. The communication between robots is established by long range Wi-Fi with limited bandwidth.

In this experiment, we select Semantic KITTI 00, Semantic KITTI 02, Semantic KITTI 07 to test the algorithm. To



(a) Semantic KITTI 00 translation error (b) Semantic KITTI 02 translation error (c) Semantic KITTI 07 translation error

(d) Semantic KITTI 00 rotation error (e) Semantic KITTI 02 rotation error (f) Semantic KITTI 07 rotation error

Fig. 5. Summary of robustness under 1331 initial perturbations. Translation error (top) and rotation error (bottom).

simulate two robots, we use a part of the dataset to generate a 3D semantic map, and split the generated map into two local maps for map matching.

Comparison Baseline: Most of the existing work either focus on single robot semantic mapping or collaborative geometry map matching. As no available work has addressed this problem, the proposed algorithm is compared against two latest geometry map matching algorithms Octo-Match [6] and MFS-Match [8]. We aslo compare it with the well-known dense registration algorithm GICP [12].

Error Metric: we first demonstrate extensive quantitative results on Semantic KITTI datasets. We first split one map into two submaps to simulate collaborative robots. The paper compares the estimated transformation T_e to the ground truth T_g. The error $\triangle T$ is calculated as $\triangle T = T_e \cdot T_g^{-1} = \{\triangle R, \triangle t\}$. The 3D translation error $e_t = ||\triangle t||$ and 3D rotation error $e_r = ||\triangle R|| = \arctan \triangle R$ are Euclidean norm of and Euler norm of the difference between ground-truth and output.

A. Matching Accuracy

According to the discussion above, there is a basic difference between map matching and dense registration. Therefore, this experiment aims to show the matching accuracy of COSEM compared with GICP. More specifically, we walked through the optimization process of each algorithm to illustrate the internal process. In the Semantic KITTI 00 dataset, the initial error is 15° rotation error on the yaw axis and $2m$ translation error in x-translation. The overlap between the two maps is approximately 60%. Both algorithms have converged to the correct global minimum (see Fig. 4a). However, the proposed algorithm requires fewer iterations than GICP. Our algorithm has converged after almost 10 iterations, while GICP requires more than 30 iterations to converge properly. All in all, our algorithm greatly improves the convergence speed. This is because our algorithm tends to find the correct data associations and rejects those misleading data associations.

In the Semantic KITTI 02 dataset, the initial misalignment are 20° rotation error in yaw axes and $10m$ translation error in x translation. The overlapping between two locals maps is just 30%, which is quite challenging for traditional dense registration algorithms. In this case, the estimation error of GICP is unstable and failed to converge to the global minimum (see Fig. 4b). The reason is due to the ambiguous data association established by geometry matching process. Compared with GICP, this algorithm makes full use of semantic information to establish the correct correspondence and gradually converges to the global minimum.

B. Robustness to Perturbations

The statistical testing results in three different dataset is presented to show the robustness to error perturbations. Since the initial transformation has random errors, we generate a total of 1331 transformations for robustness testing. To achieve that, a translation error from $-10m$ to $10m$ in steps of $2m$ is introduced along the x and y axes. In addition, the map rotates from $-25°$ to $25°$ (in steps of $5°$). For all comparisons, the number of iterations is set to be 50.

Overall, Fig. 5 shows that the proposed algorithm outperforms other algorithms in the three datasets. The boxplot

shows the median value with a red line and the mean value with a green diamond. Red crosses are marked as outliers. For the first dataset, all the algorithms achieve better results than the other two datasets. The reason is due to the high overlapping between two local maps (see Fig. 4a). However, COSEM still performs far more better than GICP, Octo-Match and MFS-Match. In the second dataset, the overlap area is much lower, as shown in Fig. 4b. Therefore, the ability to retrieve accurate transformation under this challenging environment is of great significance. As shown in Fig. 5b and 5e, COSEM still shows significant improvements compared to other algorithms. In the third dataset, the trend is aslo clear that COSEM has the best performance.

VI. Conclusion

This paper has established a collaborative probabilistic semantic map matching framework (COSEM). COSEM integrates multi-modal information, semantic data association and optimization in a general framework, which realizes collaborative localization relying on semantic maps. In the single robot level, semantic point cloud is obtained based on heterogeneous sensor fusion model and used to generate local semantic maps. In the semantic map matching process, COSEM incorporates semantic information into an Expectation-Maximization probabilistic framework. Thanks to the introduction of semantic information, the algorithm is able to establish correct data association between voxels. Extensive comparisons have been conducted on open source benchmark, the results have shown the improved accuracy and robustness. In the future, semantic information can be further integrated into the collaborative robots autonomous navigation. The robots will actively navigate at the object level, which will enable the robot to better assist human in daily life.

References

[1] A. Gawel, R. Dubé, H. Surmann, J. Nieto, R. Siegwart, and C. Cadena, "3d registration of aerial and ground robots for disaster response: An evaluation of features, descriptors, and transformation estimation," in *2017 IEEE International Symposium on Safety, Security and Rescue Robotics (SSRR)*, Oct 2017, pp. 27–34.

[2] Y. Yue, C. Yang, J. Zhang, M. Wen, Z. Wu, H. Zhang, and D. Wang, "Day and night collaborative dynamic mapping in unstructured environment based on multimodal sensors," in *Robotics and Automation (ICRA), 2020 IEEE International Conference on*, May 2020.

[3] Y. Wang, M. Shan, Y. Yue, and D. Wang, "Vision-based flexible leader-follower formation tracking of multiple nonholonomic mobile robots in unknown obstacle environments," *IEEE Transactions on Control Systems Technology*, pp. 1–9, 2019.

[4] H. Surmann, N. Berninger, and R. Worst, "3d mapping for multi hybrid robot cooperation," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Sept 2017, pp. 626–633.

[5] Y. Yue, P. G. C. N. Senarathne, C. Yang, J. Zhang, M. Wen, and D. Wang, "Hierarchical probabilistic fusion framework for matching and merging of 3-d occupancy maps," *IEEE Sensors Journal*, vol. 18, no. 21, pp. 8933–8949, Nov 2018.

[6] J. Jessup, S. N. Givigi, and A. Beaulieu, "Robust and efficient multirobot 3-d mapping merging with octree-based occupancy grids," *IEEE Systems Journal*, vol. 11, no. 3, pp. 1723–1732, 2017.

[7] J. S. Berrio, W. Zhou, J. Ward, S. Worrall, and E. Nebot, "Octree map based on sparse point cloud and heuristic probability distribution for labeled images," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 3174–3181.

[8] Y. Yue, C. Yang, Y. Wang, P. C. N. Senarathne, J. Zhang, M. Wen, and D. Wang, "A multilevel fusion system for multirobot 3-d mapping using heterogeneous sensors," *IEEE Systems Journal*, vol. 14, no. 1, pp. 1341–1352, 2019.

[9] Y. Yue, C. Zhao, R. Li, J. Zhang, M. Wen, W. Yuanzhe, and D. Wang, "A hierarchical framework for collaborative probabilistic semantic mapping," in *Robotics and Automation (ICRA), 2020 IEEE International Conference on*, May 2020.

[10] P. J. Besl and N. D. McKay, "A method for registration of 3-d shapes," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 14, no. 2, pp. 239–256, 1992.

[11] S. Ying, J. Peng, S. Du, and H. Qiao, "A scale stretch method based on icp for 3d data registration," *IEEE Transactions on Automation Science and Engineering*, vol. 6, no. 3, pp. 559–565, July 2009.

[12] A. Segal, D. Haehnel, and S. Thrun, "Generalized-icp." in *Robotics: science and systems*, vol. 2, no. 4, 2009, p. 435.

[13] L. Carlone, R. Tron, K. Daniilidis, and F. Dellaert, "Initialization techniques for 3d slam: a survey on rotation estimation and its use in pose graph optimization," in *2015 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2015, pp. 4597–4604.

[14] L.-C. Chen, G. Papandreou, I. Kokkinos, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2017.

[15] D. Kochanov, A. Ošep, J. Stückler, and B. Leibe, "Scene flow propagation for semantic mapping and object discovery in dynamic street scenes," in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2016, pp. 1785–1792.

[16] S. L. Bowman, N. Atanasov, K. Daniilidis, and G. J. Pappas, "Probabilistic data association for semantic slam," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2017, pp. 1722–1729.

[17] I. A. Bârsan, P. Liu, M. Pollefeys, and A. Geiger, "Robust dense mapping for large-scale dynamic environments," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 7510–7517.

[18] H. Yu, J. Moon, and B. Lee, "A variational observation model of 3d object for probabilistic semantic slam," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 5866–5872.

[19] Y. Wang, M. Shan, Y. Yue, and D. Wang, "Autonomous target docking of nonholonomic mobile robots using relative pose measurements," *IEEE Transactions on Industrial Electronics*, 2020.

[20] C. Brand, M. J. Schu, and H. Hirsch, "Submap matching for stereo-vision based indoor/outdoor SLAM," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, Sept 2015, pp. 5670–5677.

[21] J. Zhang, P. Siritanawan, Y. Yue, C. Yang, M. Wen, and D. Wang, "A two-step method for extrinsic calibration between a sparse 3d lidar and a thermal camera," in *2018 15th International Conference on Control, Automation, Robotics and Vision (ICARCV)*. IEEE, 2018.

[22] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[23] Y. Yue, R. Li, C. Zhao, C. Yang, J. Zhang, M. Wen, G. Peng, Z. Wu, and D. Wang, "Probabilistic 3d semantic map fusion based on bayesian rule," in *2019 IEEE International Conference on Cybernetics and Intelligent Systems (CIS) and IEEE Conference on Robotics, Automation and Mechatronics (RAM)*, Nov 2019.

[24] J. Behley, M. Garbade, A. Milioto, J. Quenzel, S. Behnke, C. Stachniss, and J. Gall, "SemanticKITTI: A Dataset for Semantic Scene Understanding of LiDAR Sequences," in *Proc. of the IEEE/CVF International Conf. on Computer Vision (ICCV)*, 2019.

[25] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016, pp. 3213–3223.

[26] J. Zhang and S. Singh, "Loam: Lidar odometry and mapping in real-time," in *Robotics: Science and Systems Conference*, July 2014.

[27] A. Hornung, K. M. Wurm, M. Bennewitz, C. Stachniss, and W. Burgard, "OctoMap: An Efficient Probabilistic 3D Mapping Framework Based on Octrees," *Autonomous Robots*, 2013.