2019 International Conference on Robotics and Automation (ICRA)
Palais des congres de Montreal, Montreal, Canada, May 20-24, 2019

Global Localization with Object-Level Semantics and Topology

Yu Liu, Yvan Petillot, David Lane and Sen Wang

***Abstract*— Global localization lies at the heart of autonomous navigation and Simultaneous Localization and Mapping (SLAM). The appearance-based approach has been successful, but still faces many open challenges in environments where visual conditions vary significantly over time. In this paper, we propose an integrated solution to leverage object-level dense semantics and spatial understanding of the environment for global localization. Our approach models an environment with 3D dense semantics, semantic graph and their topology. This object-level representation is then used for place recognition via semantic object association, followed by 6-DoF pose estimation by the semantic-level point alignment. Extensive experiments show that our approach can achieve robust global localization under extreme appearance changes. It is also capable of coping with other challenging scenarios, such as dynamic environments and incomplete query observations.**

I. Introduction

Long-term autonomous navigation is a crucial ability for robots in many real-world scenarios. This requires a robot to handle scene changes in order to robustly localize itself over time. Although the vision-based localization problem, closely associated with place recognition, loop-closure detection and Simultaneous Localization and Mapping (SLAM), has been studied extensively, there still exist many open challenges.

Appearance-based global localization methods model an environment as a database of images captured during the mapping session. Then, query observations can be localized by retrieving image(s) from the database and computing their relative transformation. This spectrum of methods often represents images by Bag-of-Word (BoW) descriptors [1] built from local features [2]–[4]. Many state-of-the-art SLAM algorithms [5]–[9] demonstrate accurate localization when the database and query images depict the scene under similar conditions.

In spite of their remarkable results, most existing methods are affected by a number of real-world challenges. For instance, seasonal or weather changes, natural or artificial illumination variations, and the emergence of new static objects or dynamic elements drastically alter the appearance and the associated features of the scene. In the context of life-long navigation, robust localization based on visual appearance is crucial but remains a difficult problem. On the other hand, the inclusion of rich semantic information within dense maps potentially enables greater robustness for localization. This is because, similar to human perception, semantics are inherently invariant to appearance changes. For instance, a desk remains a desk regardless of being captured in the daylight or in the dark. This property makes them reliable landmarks for localization tasks. Recently, many works in place recognition and global localization have been exploring the adoption of higher level visual features that have a closer relation to the semantic description of the environment [10]–[17].

The authors are with Edinburgh Centre for Robotics, Heriot-Watt University, Edinburgh, EH14 4AS, UK. {yl81, y.r.petillot, d.m.lane, s.wang}@hw.ac.uk

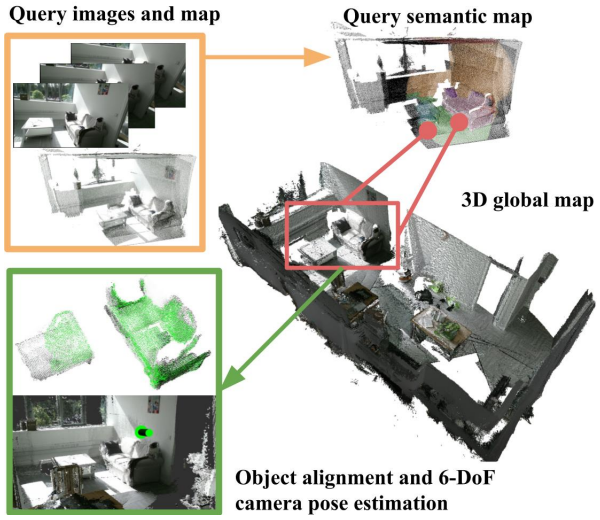


Fig. 1: An illustration of the presented global localization framework. Given a 3D global map with dense semantic features, the proposed approach estimates the 6-DoF camera pose (green arrow shown at the bottom-left) of a query observation by associating and aligning objects between the scenes at the semantic level.

In this paper, we investigate how to leverage object-level semantic information in 3D to achieve vision-based global localization. Our main contributions are three-fold:

- We creatively integrate existing techniques on dense semantics [9], [18], [19], 3D topology [14], graph matching and 3D alignment [20] into a novel object-level global localization algorithm. It achieves state-of-the-art performance on 6-DoF global localization.
- We show that object-level semantic information is beneficial for robust place recognition and global localization without the use of conventional feature descriptors. It is robust to illumination changes, scene variations, etc.
- We demonstrate that object-level alignment can handle challenging 3D point alignment and achieve accurate localization even with incomplete observations.

We evaluate our proposed method on both publicly available dataset and our collected dataset to verify its performance.

978-1-5386-6026-3/19/$31.00 ©2019 IEEE

II. Related Work

In this section we briefly review previous work on global localization and other literatures related to our work.

A. Traditional Approaches

Visual feature matching based on the BoW [1] technique has been applied in many existing localization frameworks, e.g., Fast Appearance-Based MAPping (FAB-MAP) [5]. These methods use compact descriptors built from locally invariant features (such as SIFT [2], SURF [3] and ORB [4]) to represent images, each of which is associated with a location. Global, topological localization is then approached as an image retrieval problem. These methods have established impressive results under similar perceptual condition. Descendants of these methods have further incorporated spatial information, such as building spatial dictionary and landmark covisibility graph to reduce perceptual aliasing [21]–[24].

The 6-DoF camera pose can be obtained for an exact localization by 2D-3D or 3D-3D feature matching between query and global map points and solving the Perspective-n-Point problem [25]. However, these methods relying on local (point) features could fail to localize robustly when there exist strong variations in the scene appearance. By contrast, our proposed method computes 6-DoF camera poses using high-level semantics rather than local features.

B. High-Level Features for Localization

To better cope with appearance changes, recent studies have moved toward the adoption of high-level visual cues. For instance, An et al. [26] propose a semantic-aided probablistic model to discard ambiguous local features in the urban dynamic environments. Snderhauf et al. [27] take highly salient regions from images as high-level landmarks, and use Convolution Neural Network (CNN)'s inner-layer outputs as their features. Cascianell et al. [16] extend Snderhauf's work by including the covisibility graph [21] to model an environment as a structured collection of visual landmarks. In a similar way but instead of RGB-based region proposals, Yu et al. [28] exploit clustering point cloud having similar curvature-based surfaces as high-level landmarks. Dube et al. [29] also address loop-closure detection by using clustered 3D LiDAR segments and global shape descriptors.

Localization based on object-level features have also been explored. One popular trend is to incorporate image-based object detections [15], [30]. Other attempts such as [10] detects objects by convolving pre-defined primitive kernel patches of basic shapes with the dense 3D global map. With the recent advances in learning-based semantic extraction methods, Gawel et al. [14] use CNN for semantic segmentation to obtain semantic query and global map, and then build descriptors based on the topologies of objects. McCormat et al. [19] combine state-of-the-art SLAM and semantic segmentation CNN to build consistent semantic maps by probabilistically fusing multiple semantic predictions from different viewpoints. More recently, McCormac et al. [31] apply instance segmentations to construct 6-DoF object pose graphs, further achieving SLAM by performing tracking, relocalisation and loop closure detection on objects.

C. Graph Matching

When it comes to representing a place as a constellation of visual objects, several literatures choose a graph representation to describe the objects and their topology [10], [14], [28], [32]. Therefore, the similarity between places is reduced to measuring the pair-wise similarity between their graphs. Graph matching can be approached as an assignment problem [10], [28], [32], solving for exact correspondences between nodes and edges from two graphs. Unfortunately, solving the problem in this manner is typically NP-hard [33].

Alternatively, inexact graph matching does not explicitly solve for pair-wise correspondences. For example, a graph can be summarized by its adjacency matrix to capture important topological property; graph matching is thus simplified to measuring matrix similarity [34], [35]. Others attempt to solve graph matching using graph kernels based on walks [14], [36]. In [36], the authors compute pair-wise similarity between walks' composing nodes and edges, and calculate a final matching score for the scene modeling problem. On the other hand, others such as [14] compare random walk descriptors for every node, where each descriptor encodes the local connectivity of the corresponding node. In our work we also represent scenes as semantic graphs and implement our version of random walk descriptors inspired by [14].

III. Global Localization with Object-Level Semantics and Topology

In this section, we present our global localization framework. The method builds object-based semantic graphs from query RGB images, their depth maps and semantic segmentations. From the semantic graphs, we associate object nodes by matching random walk graph descriptors between the query and global scenes. Object association and alignment are then performed to obtain 6-DoF camera pose with respect to the global map. Fig. 2 summarizes the proposed method.

A. Semantic Segmentation and Fusion

To address the global localization problem via object semantics, a global semantic map is needed. In this work, the global map is constructed from a dense SLAM algorithm during the mapping session, followed by the addition of semantic features. Similar to the Bayesian probabilistic framework as described in [19], our method simplifies the fusion step by applying a simple voting scheme. In the map, points which are perceived as a semantic class from varied viewpoints have higher confidence of being labeled correctly. Subsequent observations of the same points from different keyframes' perspectives will eventually converge and lead to labels with higher confidence. After fusing all semantics from keyframe images into the 3D map, we assign each 3D point the semantic label earning the highest votes. In the final semantic map we only keep the points which have received highly consistent semantic labeling over time.

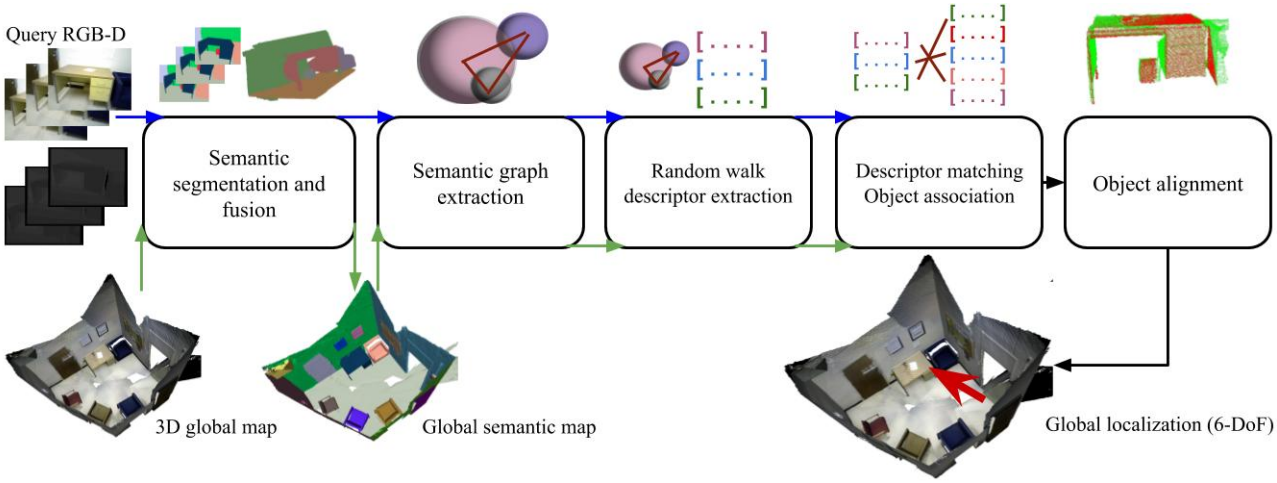


Fig. 2: Overview of the proposed approach. Our method takes a 3D global map (along with its RGB keyframes and poses), query RGB-D streams and external odometry as inputs. Semantic segmentation and fusion are then carried out to build global and query semantic maps. Next, global and query semantic graphs are extracted from the respective maps, and random walk descriptors based on [14] are obtained for all nodes in the graphs. We match each query graph's random walk descriptors with those of the global graph to find the best correspondences. Finally, associated objects are densely aligned to estimate the relative 6-DoF localization of the query within the global map.

To localize a new query, we apply the same semantic fusion technique as previously described to construct local semantic maps from query observations. Specifically, each query map is built from a small number of consecutive frames and thus is much smaller than the global map.

B. Graph Extraction

After obtaining the 3D semantic map (global or query map), we transform the map into its semantic graph representation. To add object nodes in the graph, from the semantic map we first extract nearby points having the same semantic labels with Euclidean clustering [37]. Classes of "wall", "floor", and "ceiling" are omitted as they do not introduce useful topological relations. In the graph. we choose a bounding sphere to represent each object as spheres hold the implicit property of being rotationally invariant. The size of the sphere (depicting the dimension of the object) is the distance of the furthest point away from the cluster center.

Object nodes and edges are used to describe the 3D semantic topology of a map. Undirected edges are formed between any two object nodes within a proximity distance. We also define that two nodes whose bounding spheres interact with each other always form edges regardless of the distance apart, implicitly modeling both spatial and dimensional relations. Fig. 3 shows an example of a semantic graph.

C. Random Walk Descriptors

Inspired by the simple graph descriptor in [14], we describe each node in the semantic graph with random walk descriptors. Starting from a root node, each random walker explores its connected neighboring nodes and records visited labels in sequence. The depth of each exploration is defined by the length of a walk, and the number of exploration is

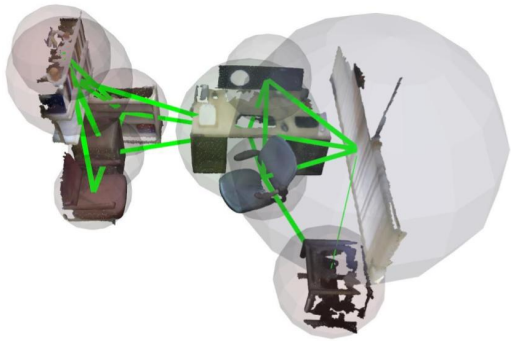


Fig. 3: Extraction of a semantic graph from its 3D semantic map (SceneNN [38] sequence 25). Each sphere represents an object. Green lines show the 3D object/semantic topology.

defined by the specified descriptor size for a node. This simple descriptor implicitly encodes the topology of objects and their neighbors. Fig. 4 illustrates the random walk descriptor with an example. In the descriptor, in addition to the visited labels, we keep track of each label's corresponding node in the graph to verify spatial consistency in the association step.

D. Object Association

Once random walk descriptors are built for the global and query graph, we perform association between their nodes based on the number of identical random walk descriptors they share. This inherently means only objects of the same semantic class will be associated. In an environment where there exist multiple objects of the same semantic class, a match between two descriptors only suggests a potential candidate. When a match is found, we trace back the corresponding nodes and perform a spatial consistency verification on the path this walk takes. If two walks share the same

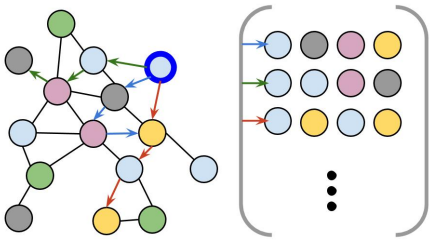


Fig. 4: 2D toy schematic of the random walk descriptor. From a root node (bold border), the random walker explores its surrounding and records the semantic labels of the visited nodes. The example demonstrates a walk of length 4.

visited labels, but the inter-node distance (between any two nodes) exceeds a threshold distance, the descriptors are not considered as a match. Finally, for any query node, the number of matched descriptors with each global node is counted. The top k global node(s) with the most shared descriptors is taken as a match, and the corresponding objects in the query and global map are associated. We allow for inexact association, to cope with scenarios where an object might be overly segmented and all its pieces need to be associated with a single object in the other map.

E. Localization based on Object-Level Alignment

After associating query objects with those in the global map, we make use of objects' geometry for pose estimation. This is achieved by densely aligning the sets of associated object point clouds from the query and global map.

To perform dense alignment, we use the Fast Point Feature Histograms (FPFH) and the Sample Consensus initial alignment method (SAC-IA) [20] to register the associated object points. The alignment result and the estimated transformation are accepted when a minimum percentage of correctly aligned points is met. The estimated transformation gives the initial alignment to localize the query objects within the map. Finally, this transformation is refined by the Iterative Closest Point algorithm (ICP) [39], resulting in the final 6-DoF query camera pose estimation.

IV. Experimental Results

In this section, we evaluate our method on a public RGB-D benchmark and our dataset collected with the Kinect v2 sensor. We demonstrate that our method using 3D semantic topology and alignment achieves robust global localization even under drastic lighting changes.

A. Dataset

The first dataset we experimented is the public SceneNN dataset [38]. It includes a number of indoor sequences of various types, scales and furnishings. Each sequence provides full-sequence RGB frames, depth maps, groundtruth pixel-wise semantic classifications, and 6-DoF camera poses. The number and types of semantic classes vary among sequences. In our experiment we use sequence 21 and 25 which represent a typical office and bedroom environment respectively.

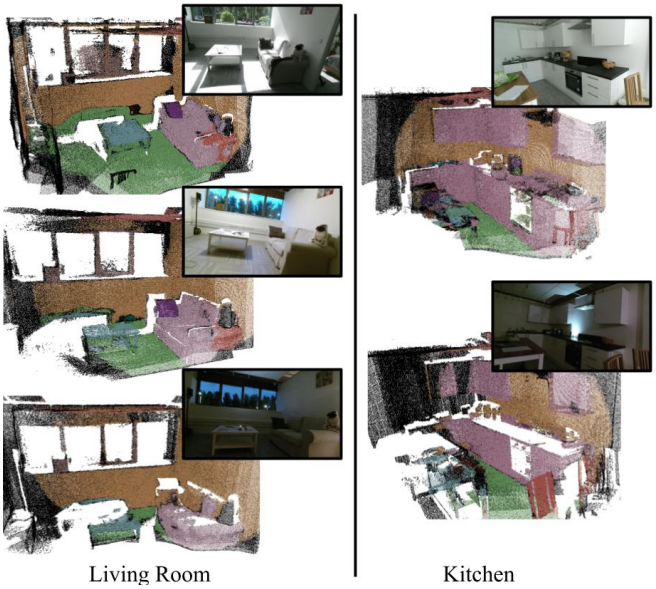


Fig. 5: Qualitative evaluation on the voting-based semantic mapping under drastic appearance variation (examples from the LKD's Day-Day and Day-Night sequence).

In order to evaluate our approach without assuming groundtruth semantic segmentations and capture scene changes which benchmarks typically do not provide, we collect our own dataset featuring a living room, kitchen and dining room (also termed "LKD"). This dataset includes several sequences taken under different lighting conditions and relocation of furnitures within the scene. The Day sequence is collected under strong daylight, while the Night sequence is collected at dawn under a significantly different lighting condition. The reference sequence is collected also in the day. For our collected dataset, we obtain semantic segmentations with the Pyramid Scene Parsing network (PSPNet) [18]. We use the off-the-shelf pre-trained model without further re-training or fine-tuning.

B. Experimental Setup

To evaluate our method in a realistic localization scenario, we first construct the global semantic map, from which each semantic query collected at a different time uses for query pose estimation. In SceneNN, each sequence has very few loops to demonstrate an incidence of re-visitation. We thus build the global semantic map by registering every 20^{th} RGB frame and its semantic segmentation projected into the 3D space with respect to the global map reference. We generate queries from the same sequence while avoiding any frame already used in the global map. Every query is generated with significantly lower number of frames and a smaller interval between any two frames. It is worth mentioning that SceneNN originally provides instance-level segmentation (i.e., unique object ID) for every object in the scene. Our approach only requires object-level classifications; thus we do not take advantage of this prior knowledge.

For the collected LKD dataset, we first construct the

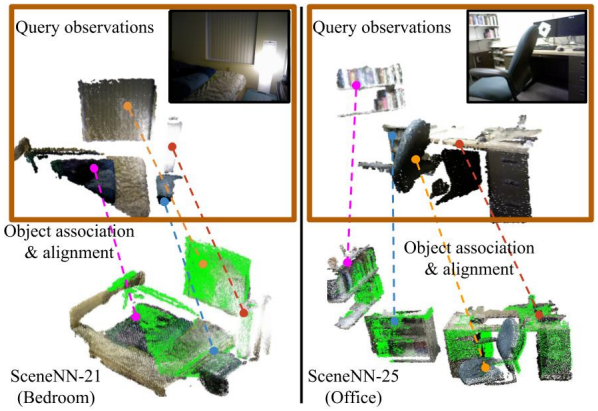
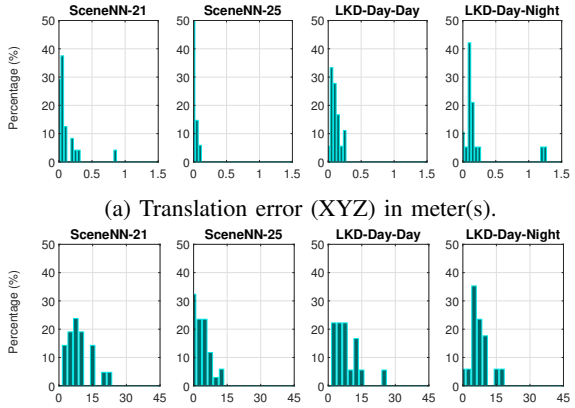


Fig. 6: Successful alignment and localization from SceneNN dataset. At the bottom of this figure, the raw RGB and green point cloud correspond to the global map objects and query objects respectively.



(a) Translation error (XYZ) in meter(s).

(b) Orientation error (RPY) in degree(s).

Fig. 7: Error distribution on estimated 6-DoF camera poses.

TABLE I: Localization results in translation and orientation.

Dataset	Translation (cm)	Orientation (○)
SceneNN-21	8.38 ± 5.90	10.67 ± 5.51
SceneNN-25	3.38 ± 3.32	4.83 ± 3.61
LKD-Day-Day	13.18 ± 6.46	9.14 ± 5.93
LKD-Day-Night	14.13 ± 6.35	8.44 ± 4.47
Total mean	8.54 ± 6.90	7.74 ± 5.26

global map using the RGB-D SLAM RTABMap [9] on the reference sequence. We apply PSPNet and the semantic fusion technique described in Section III-A to obtain the global semantic map. In the absence of groundtruth poses to evaluate localization performance on queries, we use RTABMap on query sequences to generate keyframes and groundtruth 6-DoF poses. Consequently, to form a query map, we register and perform semantic fusion from these keyframes.

C. Segmentation Performance

Our PSPNet model is pre-trained on the ADE20K dataset [40], which includes 150 indoor and outdoor classes. To prevent over-segmentation, we re-map the original 150 classes into 40 by grouping classes of similar types. The merged classes include dominant indoor furnitures (e.g., sofas, cabinets and desks, etc.).

We qualitatively assess the labeling consistency of the post-fusion semantic maps. In particular, we inspect the effect of lighting variation on semantic mapping and how the proposed voting method performs. Fig. 5 exhibits several query semantic maps from our collected sequences. As compared to the corresponding RGB images, the post-fusion semantic maps exhibit significantly higher consistency on the major objects in the scenes, such as the coffee table, sofa and the two large cabinets residing in the kitchen. This consistent labeling also enables objects to retain their dense geometry. Semantic fusion is not required on SceneNN as groundtruth segmentations are available.

D. Localization Performance

We evaluate global localization performance in terms of position and orientation by comparing the query pose estimation with the groundtruth. Every query map is built with 9 frames. The localization performance is evaluated according to the center frame of a query. We apply the relative transformation from groundtruth poses (provided by SceneNN or RTABMap), simulating data from external odometry, to register multiple frames into the query map. In addition, we employ a walk length of 4 as suggested by [14] to prevent a deeper walk from hopping among the same few nodes in a query graph while continuing to explore new ones in the much larger global graph. Finally, we use the Euclidean distance as the performance metric.

Fig. 7 describes the accuracy distribution on all tested datasets when the SAC-IA alignment's inlier threshold is fixed at 75 percent. In general, localization results are superior in the SceneNN dataset as most errors are skewed toward the lower range. Specifically, most translation and orientation errors are smaller than 0.25 meter and 10 degrees respectively. Fig. 6 displays query examples that are successfully localized in the SceneNN's global maps. The average localization accuracy for all datasets is shown in Table I. It can be seen the average translation and orientation errors of all datasets are 8.54 cm and 7.74 degrees. Especially, the proposed algorithm achieves good global localization performance on the LDK-Day-Night sequence which includes drastic illumination changes between day and night.

We further investigate the benefits of exploiting semantics for place recognition and global localization on the LKD dataset. Our approach is compared against two other methods: FAB-MAP and NetVLAD [41]. The former based on BoW is adopted in many existing SLAM frameworks and is used as the baseline algorithm. The later relies on CNN deep features trained by images depicting the same places at different viewing and lighting conditions. We run a TensorFlow implementation of NetVLAD [42] trained on the Pitts30k dataset [43]. Even though our collected sequences contain indoor scenes, they also depict the same places under

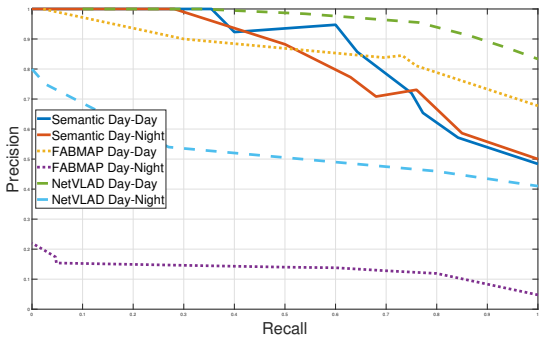


Fig. 8: Precision-recall curves of place recognition on our collected LKD datasets. Our method is termed "Semantic".

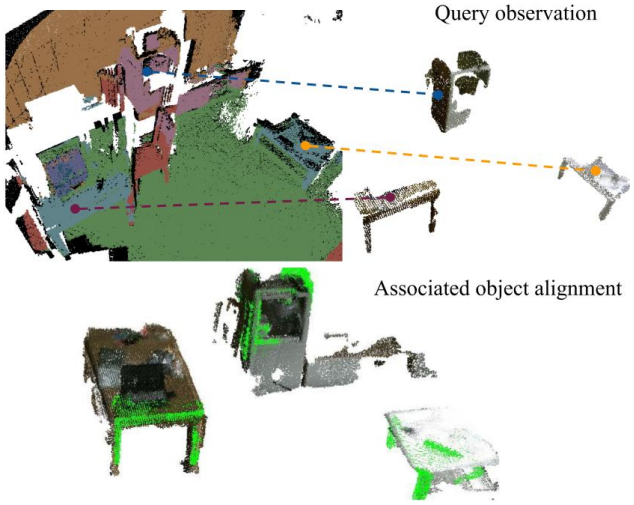


Fig. 9: Successful demonstration on challenging scenarios. Top: Query semantic map and extracted objects. Bottom: Object-level alignment between global map objects (RGB) and query objects (green). This shows reliable alignment and localization when objects are only partially observed. In addition, semantic maps offer the advantage of ignoring dynamic objects for robust localization (chairs in this case).

different lighting conditions which theoretically NetVLAD is trained to recognize.

The precision-recall curves are shown in Fig. 8. All methods achieve good performances on the LKD-Day-Day sequence where no significant appearance is present, out of which NetVLAD outperforms the rest by a perceivable margin. However, when the place recognition is conducted in the Day-Night case, i.e., matching night query to a day global map, the performance of both FAB-MAP and NetVLAD drop drastically due to significant changes in appearance and illumination (see RGB images in Fig. 5). Even though NetVLAD is trained to be robust against strong appearance changes in an outdoor environment, the model is not generic enough to make correct predictions when both illuminations and scene types vary drastically at the same time. In contrast, our proposed semantic method remains robust and achieves similar performance to the Day-Day case, which verifies high-level semantic information can be beneficial for robust place recognition.

E. Discussion on Challenging Scenarios

Appearance-based methods face difficulties coping with real-world scenarios where scenes undergo drastic changes. In contrast, our experimental results show that semantic representations, encapsulating high-level scene topology, behave as reliable landmarks even under challenging perceptual conditions. In addition, our approach also handles the following challenges well:

1) Incomplete View: Query observations only capture partial areas of the environment. Additionally, semantic segmentations also suffer from a certain degree of mis-labeling. Both of the above factors result in incomplete query segments that typically is conflicted with the complete views stored in the global map database. Nevertheless, our method captures the semantic topology using high-level graphs, which implicitly ignores the missing details during query and global object association. Furthermore, we choose to densely align multiple associated objects simultaneously, which encourages SAC-IA finding the correct transformation to properly align multiple incomplete objects with their complete correspondences.

2) Dynamic Scene: In a dynamic environment, each query can present a certain extent of variations from a global map. The inclusion or exclusion of any new or existing elements alters underlying features, creating difficulty for appearance-based methods. With high-level semantics, our method can easily ignore dynamic objects and only consider static ones for robust localization. For instance, large furniture, e.g., sofas, usually remains while others such as chairs tend to be relocated often. Both scenarios of incomplete query observation and dynamic scene co-exist in Fig. 9, which our method handles successfully.

V. CONCLUSIONS

We have presented a novel global localization approach leveraging dense semantics, 3D topology and semantic-level alignment to estimate the 6-DoF camera pose. We evaluated our approach on the public and our collected dataset with strong illumination changes. Our approach demonstrates both high accuracy and robustness under drastic appearance variations where others would be heavily affected. Furthermore, we point out our method's strengths and potentials to cope with more real-world challenging scenarios. While having only experimented in indoor scenarios, with moderate modifications our method can be expanded to outdoor applications. We believe our method furthers the possibility to address long-term localization more aligned with how humans perceive and react to the world.

ACKNOWLEDGMENT

This work was supported in part by EPSRC Robotics and Artificial Intelligence ORCA Hub (grant No. EP/R026173/1) and NVIDIA with the donation of the Titan Xp GPU.

REFERENCES

[1] J. Sivic and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos," in *ICCV*, 2003.

[2] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *IJCV*, vol. 60, no. 2, pp. 91–110, 2004.

[3] H. Bay, T. Tuytelaars, and L. Van Gool, "SURF: Speeded up robust features," in *ECCV*, 2006, pp. 404–417.

[4] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: An efficient alternative to SIFT or SURF," in *ICCV*, 2011.

[5] M. Cummins and P. Newman, "FAB-MAP: Probabilistic localization and mapping in the space of appearance," *IJRR*, vol. 27, no. 6, pp. 647–665, 2008.

[6] 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.

[7] R. Mur-Artal and J. D. Tardós, "ORB-SLAM2: an open-source SLAM system for monocular, stereo, and RGB-D cameras," *IEEE Transactions on Robotics*, vol. 33, no. 5, pp. 1255–1262, 2017.

[8] P. Henry, M. Krainin, E. Herbst, X. Ren, and D. Fox, "RGB-D mapping: Using Kinect-style depth cameras for dense 3D modeling of indoor environments," *IJRR*, vol. 31, no. 5, pp. 647–663, 2012.

[9] M. Labbé and F. Michaud, "Online global loop closure detection for large-scale multi-session graph-based SLAM," in *IROS*, 2014, pp. 2661–2666.

[10] R. Finman, L. Paull, and J. J. Leonard, "Toward object-based place recognition in dense RGB-D maps," in *ICRA Workshop on Visual Place Recognition in Changing Environments*, 2015.

[11] R. F. Salas-Moreno, R. A. Newcombe, H. Strasdat, P. H. Kelly, and A. J. Davison, "SLAM++: Simultaneous localisation and mapping at the level of objects," in *CVPR*, 2013, pp. 1352–1359.

[12] E. Stenborg, C. Toft, and L. Hammarstrand, "Long-term visual localization using semantically segmented images," in *ICRA*, 2018, pp. 6484–6490.

[13] P. Parkhiya, R. Khawad, J. K. Murthy, B. Bhowmick, and K. M. Krishna, "Constructing category-specific models for monocular object-SLAM," in *ICRA*, 2018, pp. 1–9.

[14] A. Gawel, C. Del Don, R. Siegwart, J. Nieto, and C. Cadena, "X-View: Graph-based semantic multi-view localization," *IEEE Robotics and Automation Letters*, vol. 3, no. 3, pp. 1687–1694, 2018.

[15] S. L. Bowman, N. Atanasov, K. Daniilidis, and G. J. Pappas, "Probabilistic data association for semantic SLAM," in *ICRA*, 2017, pp. 1722–1729.

[16] S. Cascianelli, G. Costante, E. Bellocchio, P. Valigi, M. L. Fravolini, and T. A. Ciarfuglia, "Robust visual semi-semantic loop closure detection by a covisibility graph and CNN features," *Robotics and Autonomous Systems*, vol. 92, pp. 53–65, 2017.

[17] R. Finman, T. Whelan, L. Paull, and J. J. Leonard, "Physical words for place recognition in dense RGB-D maps," in *ICRA Workshop on Visual Place Recognition in Changing Environments*, 2014.

[18] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *CVPR*, 2017, pp. 2881–2890.

[19] J. McCormac, A. Handa, A. Davison, and S. Leutenegger, "SemanticFusion: Dense 3D semantic mapping with convolutional neural networks," in *ICRA*, 2017, pp. 4628–4635.

[20] R. B. Rusu, N. Blodow, and M. Beetz, "Fast point feature histograms (FPFH) for 3D registration," in *ICRA*, 2009, pp. 3212–3217.

[21] C. Mei, G. Sibley, and P. Newman, "Closing loops without places," in *IROS*, 2010, pp. 3738–3744.

[22] R. Paul and P. Newman, "FAB-MAP 3D: Topological mapping with spatial and visual appearance," in *ICRA*, 2010, pp. 2649–2656.

[23] E. Johns and G.-Z. Yang, "Feature co-occurrence maps: Appearance-based localisation throughout the day," in *ICRA*, 2013, pp. 3212–3218.

[24] E. S. Stumm, C. Mei, and S. Lacroix, "Building location models for visual place recognition," *IJRR*, vol. 35, no. 4, pp. 334–356, 2016.

[25] D. Nistér and H. Stewénius, "A minimal solution to the generalised 3-point pose problem," *JMIV*, vol. 27, no. 1, pp. 67–79, 2007.

[26] L. An, X. Zhang, H. Gao, and Y. Liu, "Semantic segmentation–aided visual odometry for urban autonomous driving," *IJARS*, vol. 14, no. 5, 2017.

[27] N. Sünderhauf, S. Shirazi, A. Jacobson, F. Dayoub, E. Pepperell, B. Upcroft, and M. Milford, "Place recognition with ConvNet landmarks: Viewpoint-robust, condition-robust, training-free," in *Robotics: Science and Systems*, 2015.

[28] H. Yu, H.-W. Chae, and J.-B. Song, "Place recognition based on surface graph for a mobile robot," in *URAI*, 2017, pp. 342–346.

[29] R. Dubé, D. Dugas, E. Stumm, J. Nieto, R. Siegwart, and C. Cadena, "SegMatch: Segment based loop-closure for 3D point clouds," in *ICRA*, 2017.

[30] S. Ardeshir, A. R. Zamir, A. Torroella, and M. Shah, "GIS-assisted object detection and geospatial localization," in *ECCV*, 2014, pp. 602–617.

[31] J. McCormac, R. Clark, M. Bloesch, A. J. Davison, and S. Leutenegger, "Fusion++: Volumetric object-level SLAM," in *3DV*, 2018, pp. 32–41.

[32] J. Oh, J. Jeon, and B. Lee, "Place recognition for visual loop-closures using similarities of object graphs," *Electronics Letters*, vol. 51, no. 1, pp. 44–46, 2014.

[33] F. Zhou and F. De la Torre, "Factorized graph matching," in *CVPR*, 2012, pp. 127–134.

[34] E. Stumm, C. Mei, S. Lacroix, and M. Chli, "Location graphs for visual place recognition," in *ICRA*, pp. 5475–5480.

[35] E. Stumm, C. Mei, S. Lacroix, J. Nieto, M. Hutter, and R. Siegwart, "Robust visual place recognition with graph kernels," in *CVPR*, 2016, pp. 4535–4544.

[36] M. Fisher, M. Savva, and P. Hanrahan, "Characterizing structural relationships in scenes using graph kernels," *ACM Transactions on Graphics*, vol. 30, no. 4, p. 34, 2011.

[37] R. B. Rusu, "Semantic 3D object maps for everyday manipulation in human living environments," *KI-Künstliche Intelligenz*, vol. 24, no. 4, pp. 345–348, 2010.

[38] B.-S. Hua, Q.-H. Pham, D. T. Nguyen, M.-K. Tran, L.-F. Yu, and S.-K. Yeung, "SceneNN: A scene meshes dataset with annotations," in *3DV*, 2016, pp. 92–101.

[39] P. J. Besl and N. D. McKay, "Method for registration of 3-D shapes," *Sensor Fusion IV: Control Paradigms and Data Structures*, vol. 1611, pp. 586–607, 1992.

[40] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba, "Scene parsing through ADE20K dataset," in *CVPR*, 2017.

[41] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "NetVLAD: CNN architecture for weakly supervised place recognition," in *CVPR*, 2016, pp. 5297–5307.

[42] T. Cieslewski, S. Choudhary, and D. Scaramuzza, "Data-efficient decentralized visual SLAM," in *ICRA*, 2018, pp. 2466–2473.

[43] A. Torii, J. Sivic, T. Pajdla, and M. Okutomi, "Visual place recognition with repetitive structures," in *CVPR*, 2013, pp. 883–890.