

Associating Landmarks from SLAM’s Visual Structure

Matthew Bradley

*Department of Systems Design Engineering
University of Waterloo
Waterloo, Canada
m7bradle at uwaterloo.ca*

John Zelek

*Department of Systems Design Engineering
University of Waterloo
Waterloo, Canada
jzelek at uwaterloo.ca*

Abstract—Place recognition is the online task of detecting revisits to previously seen locations and is a key to many navigational systems. In Simultaneous Localization and Mapping, recovering the relative camera pose between recognized visit and revisit (e.g. using bundle adjustment) allows for global map optimization, improving localization accuracy. Visual SLAM recovers structure to estimate camera movement but it is typically not used for visual place recognition. Limited past work which adapted LiDAR place recognition descriptors to SLAM-recovered physical structure found superior robustness to visual effects vs appearance-based VPR, but overall had poorer recall. It was found that LiDAR descriptors’ whole-scan matching assumes excellent 360 degree pointcloud coverage while cameras have limited FoV. We observe that SLAM-tracked points congregate on objects and distinct elements, resulting in sparsity that impacts whole-scan matching. To us this also suggests use of clustering to extract these aggregate congregations as landmarks whose configuration can be matched. Exploring this approach we found that the landmarks generated still vary in detected position, but a far more significant hurdle is that the same landmarks may not be repeatedly clustered each time a scene is visited. This is due to large-scale clustering still being sensitive to instability in the individual SLAM points. This was improved significantly but not sufficiently through visual semantic labeling of the initial 3D points, helping to provide more stable, guided clustering solutions. Still, single missing or “outlier” landmarks are detrimental to successful association between landmark sets. To address this instability in future work we recommend careful selection of salient points from those collected by SLAM, for those which can be expected to be the most stable and repeatably detected. This is expected to provide more stable landmarks than large-scale clustering of detected points which relies on a center-of-mass approach.

Keywords—Place Recognition; SLAM; VPR; LiDAR; Navigation;

I. INTRODUCTION

Place recognition refers to the task of determining when a previously visited place is observed again, accomplished by matching measurable features of the environment. It is critical to navigation systems that operate for extended periods, including SLAM (Simultaneous Localization and Mapping) which underpins virtually all uses of mobile robots. In the case of SLAM, the global map can be optimized using the relative pose between both visits to a place (obtained

through bundle adjustment), ensuring more accurate large-scale navigation.

Structural place recognition has traditionally been applied to the domain of LiDAR scans, where the physical structure of the environment is the primary information available. Here a 360-degree scan of the environment is summarized into a descriptor for later comparison with subsequent scans. Handcrafted methods like Scan Context [1], M2DP [2], and DELIGHT [3] are favored for their efficiency as LiDAR scans often contain many tens or even hundreds of thousands of points, although descriptors based on neural networks have been proposed [4].

Parallel to this, visual place recognition has formed a key component of image-based SLAM on regular cameras. Cameras are relatively inexpensive and common sensors and visual SLAM has seen a great deal of development with many systems proposed [5] [6] [7] [8]. Visual place recognition is appearance based, and typically consists of either handcrafted visual vocabulary or is based on global descriptors which increasingly employ neural networks. Visual vocabulary seeks to categorize and describe the local visual features present in SLAM keyframes, producing a compact summary descriptor. Global descriptors extract features at larger scales. They are generally far more robust to illumination changes but consume more computational resources and tend to be sensitive to the particular viewpoint that an image was captured from [9]. So far a good compromise between these methods has not been found, though some visual vocabularies have been extended to larger image regions [10].

Unlike the environment’s appearance, the underlying 3D structure of the environment does not change with illumination or other visual effects. Physical surfaces remain the same and in the same position regardless of how visual effects might change their appearance or where visual features are detected on them. This structure is partially recovered by visual SLAM systems while estimating 3D camera motion, but the use of 3D structure in visual place recognition has been rarely explored. One of the most recent and successful efforts was [11] who adapted 3D points tracked by a SLAM system [12] for use with a series of preexisting handcrafted LiDAR descriptors. They found

these descriptors significantly outperformed SoTA visual place recognition under challenging illumination conditions, however they were overall less able to find matching places in general. A significant disadvantage found by [11] was that the limited field of view of cameras provided significantly less overall coverage than a 360-degree LiDAR scanner, resulting in large sections of their imitation-scans which were incomplete. As the preexisting descriptors used assume full coverage, with all regions of the covered volume are reflected in their output description vectors, this led to false negative matches when the pattern of visual coverage differed.

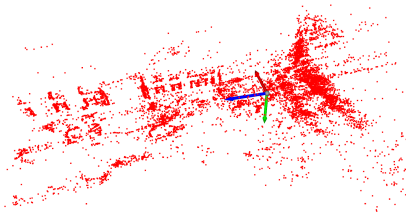


Figure 1. An overhead view of 3D points recovered by a SLAM system, representing tracked visual features. They tend to clump on distinct objects and structures rather than appear evenly distributed.

Our contribution is the exploration of a new approach to integrate SLAM-recovered structure into visual place recognition, in hopes of overcoming the difficulties of this previous approach. To overcome the problem of coverage encountered by [11], we associate individual landmarks extracted from imitation scans using SoTA graph association methods [13]. The goal is that that full spatial coverage should not be required, instead associating landmarks detected in the area of overlap. This is inspired by our observation that the points collected from SLAM are not evenly distributed and instead “clump” around particular objects and structures, including windows, doors, changes in surface texture, edges, and small objects like utility boxes (Figure 1). As the individual points are unreliable due to various visual effects, we employ clustering in an effort to get a “center of mass” of each clump and mitigate the varying detection of individual points. Points recovered by SLAM are also more sparse than those of LiDAR, compared to typical inter-object distances, and so it can be challenging to segment nearby objects from each other. It can also be difficult to constrain the growth of clusters to particular structures and objects. We mitigate this problem through the use of visually-derived semantic labels for the 3D points, providing additional separation that constrains cluster growth and renders it less sensitive to the selection of particular parameters.

Despite these strategies and improvements, we have found that the repeatability of the landmarks generated through clustering in this way is insufficient for reliable association between visits to a scene. The exact number and position

of landmarks is reliant on the way that underlying SLAM points are clustered and despite the improvements we obtain in segmentation the clustering approach is still affected by the repeatability of underlying SLAM-captured 3D points. We examine properties of the landmarks we generate such their repeatability and error in their estimated position, as well as the sensitivity of landmark association to these properties. For the future research making use of SLAM-recovered structure for VPR we recommend salient filtering of the points collected from SLAM to obtain those belonging to truly stable objects and structures in the environment and which can be expected to be the most stable between observations.

II. BACKGROUND

A. Structural Place Recognition on LiDAR Pointclouds

Navigational systems frequently need to know when they return to a previous place and those using LiDAR are inherently structural, operating on 3D scans and using structure-encoding descriptors. As in other forms of place recognition, these descriptors reduce a LiDAR scan to a description vector which can be efficiently compared with vectors from other places. Handcrafted descriptors are popular as the large number of points in a scan makes neural networks and other computationally-demanding techniques difficult to run in real-time. When neural networks are used, voxel grids are frequently used to reduce the input size as with MinkLoc3D [14] and LCDNet [4].

The work of [11] relies on three handcrafted LiDAR descriptors, Scan Context [1], M2DP [2], and DELIGHT [3]. They share the common approach of dividing space around the LiDAR scanner into volumetric cells, describing their contents, and assembling these descriptions into the full output description vector. Scan Context divides the X-Y plane into cells arranged in concentric rings, recording the maximum Z height of points in each cell. The descriptors are slid over each other during comparison in a window fashion to overcome unknown rotation, testing each possible alignment. DELIGHT [3] divides space into eight semi-spherical sections, then further into inner and outer shells. Histograms of LiDAR reflectivity values for points in each volume form the overall descriptor. M2DP is slightly different, where the points are projected into a set of 2D views. A series of 2D concentric rings then divides each view into cells where the number of points in each cell is recorded and compressed using Singular Value Decomposition. All of these methods establish a 1-to-1 correspondence between all sub-regions of the described volume and each part of the descriptor. [11] found that this means that scans of the same location with different coverage (due to visibility or traversal differences) can result in mismatching descriptors and lower overall recall of recognized places.

B. Structural Place Recognition on SLAM Pointclouds

Given illumination and viewpoint changes can disrupt visual appearance-based descriptors [9] [15], we are interested in structural information which is robust to visual effects like illumination. Regardless of how environment's appearance changes, its physical surfaces remain constant. This remains true even as visual effects might alter where on surfaces particular visual features are detected. The benefits of using SLAM-recovered structure for visual place recognition is relatively unexplored, however [11] have applied off-the-shelf LiDAR descriptors to pointclouds collected from visual SLAM. They found that these descriptors far outperformed SoTA methods of VPR under extreme illumination changes, but under normal conditions their performance was less competitive. The descriptors considered (Scan Context [1], M2DP [2], and DELIGHT [3]) operate by partitioning the volume around the LiDAR sensor (or in this case the camera center) into spatial cells, with a direct correspondence between the cells and portions of the descriptor. [11] found that the limited viewpoint of the camera frequently resulted in mismatched coverage between visits, leading to significant areas of difference and thus differences in the resulting place descriptors and lower recall. Here we explore an alternative way to leverage SLAM-recovered structure, based on landmarks to make associations on a more granular level and overcome differences in coverage.

C. Visual Effects on SLAM Pointclouds

We are not the first to encounter stability problems with the image features extracted by SLAM systems to estimate 3D position. While the underlying physical structure of the environment may remain constant, visual effects can impact the specific places where visual techniques detect feature points. A method is proposed by [16] uses local symmetry to judge the potential future stability of a feature point. This follows their finding that feature points detected by algorithms like SIFT [17] are very susceptible to noise. Feature points were often not re-detected upon returning to a previous location, and while many might be detected at any moment only a subset tend to persist across multiple frames. In our case, we seek to mitigate these effects using clustering to get an aggregate, overall view of the detected points.

An analysis conducted by [18] across four point detectors/descriptors found that changes in illumination affected the repetition rate of feature points, the rate at which these features were found in similar positions over multiple frames. They also find that matching features across frames can be impacted by spatial transformations like scale changes and rotations. In order for SLAM to determine a point's 3D position it must be identified in multiple frames. The method [18] used to assess repeatability was proposed in a previous analysis of point-feature detectors and descriptors [19] which was further extended in [20], and [21] where

homographies were used to compare distributions of points in images taken from different viewpoints.

III. METHODOLOGY

A. Collecting 3D SLAM Points

Our collection of 3D slam points follows largely the same process as that proposed by [11], differing in the following stages. Where [11] simply feed collected points to LiDAR descriptors, we generate landmarks through clustering and attempt association. Utilizing the same modified SLAM system proposed by [11], we collect the estimated 3D coordinates of detected feature points in every SLAM-selected keyframe image in KITTI [22] sequences 00, 02, 05, and 06 (those with meaningful revisits). We focus on KITTI as it is the dataset which is the least visually challenging, found by [11] to be where visual methods outperform their proposed structural method(s). Having points for each keyframe, we generate an "imitation LiDAR scan" for each keyframe by accumulating points from the last 100 frames in the same way [11] do. This pointcloud is clipped to within 45 meters of the camera center. When identifying revisits to the same place we apply the same criteria as [11]. They must be within 10 meters and occur at least 100 frames apart. The sequences with a non-trivial number of valid revisits are sequences 00, 02, 05, and 06, with 1353, 753, 791, 453 qualifying frames, respectively.

As we use semantic labelling for some approaches, we also apply You Only Segment Once [23] to every keyframe which provides pixel-wise semantic labels. Before accumulation into imitation scans, the 3D points in each keyframe are projected to 2D using camera intrinsics so the nearest pixel label can be applied to them. Most labels are either "building," "vegetation," "road," "car," and "sidewalk." We use buildings, vegetation, and cars as these three are common but also have the most distinct, self-contained boundaries.

B. Clustering Approaches

Observing that the visual SLAM points collected are not uniformly distributed like those of a LiDAR scan, but are instead "clumped" around objects (like windows, utility boxes, etc.), we explore clustering to convert them into landmarks. For each cluster the corresponding landmark point is the average position of its member points. We explored use of geometric structures like lines and planes, however the points in an imitation scan are too sparse for feasible recovery of these geometric primitives.

Two main clustering algorithms are explored, DBSCAN [24] and BIRCH [25]. DBSCAN seeds clusters and then grows them by finding nearby points. BIRCH operates through tree-based partitioning and is provided with a target total, where as DBSCAN generates a widely varying number of clusters. GrassGraph [13] requires a consistent total number, so we must randomly add (within the same

bounds) or subtract some when using it. We ensure 120 landmark points with regular clustering and 20 with semantic clustering, based on the average number typically found in the first 100 frames of sequence 00. We find that this is a significant limitation, as GrassGraph is sensitive to outliers and random landmark addition/removal exacerbates to this.

To improve separation of nearby elements (e.g. trees from the surrounding earth), we explore providing additional separation using semantic labels. To apply this to DBSCAN [24] we cluster three times, exclusively on points from each of the three classes above. The resulting clusters are then combined and the total number of landmark points corrected as before. With BIRCH [25] we wish to maintain the its ability to maintain a target total. We collect the three labeled point sets into one point cloud but physically separate them with large vertical offsets to discourage clustering points labeled with different semantic classes together.

C. Grassmannian Association

To find associations between sets of landmarks from different visits we make use of GrassGraph [13], a SoTA method for finding associations between matching sets of points when their relative transformation is unknown. In return we receive a partial set of landmark-to-landmark associations and a recovered alignment transformation matrix. GrassGraph relies on forming an equivalence between the set of all linear combinations of the points in a pointset and those of a translated and/or rotated image of the set. The result are coordinates for each point in a new space that is invariant to affine transformations. [13] overcome a lingering rotational ambiguity from the use of Singular Value Decomposition by using eigenvectors to describe each point, derived from a graph of the distances between the new point coordinates. GrassGraph is meant for small sets of keypoints (runtime grows with the cube of the number of points) and so our conversion of messy 10,000+ point pointclouds into fewer landmarks is critical to feasibly using it for association. One can also cache eigenvectors and other data for each collection of points and reuse this data for multiple rounds of association.

D. Measuring Landmark Repeatability

To gauge the repeatability of the landmark points we obtain, we repeatedly compare landmarks from pairs of a visit and closest later revisit. We measure both the percentage of initial-visit landmarks which lack a nearby corresponding landmark in the revisit set, as well as the distance between the two (visit and revisit landmarks) when one exists. We give the latter as the standard deviation of an assumed half-normal distribution which randomly displaced points follow. This process provides two quantitative measures for the quality of our landmarks.

As there is no ground truth for what landmarks should exist and where, it is challenging to set a criteria for when

a landmark has no nearby match during a revisit. Due to the greatly varying density of collected points from visual SLAM a fixed threshold for closeness is infeasible. Instead we use a mutual-nearest neighbor criterion: two landmarks correspond, one from the initial visit to a place and one from a later revisit, if they are mutually each others' nearest neighbors. Given a pair of landmarks detected in roughly the same location, one may qualify as the nearest to a more distant outlier. This relationship will not be reciprocal, however, as the outlying point is more distant to that point than the point's proper partner. In practice this method has proven reasonably robust, although it rarely pairs landmark points when they are both distant outliers.

E. Measuring GrassGraph Sensitivity

We measure the performance of GrassGraph [13] when associating landmarks to quantify what effects make the task difficult. The performance of GrassGraph itself is measured in terms of the number of associations it recovers, as well as the accuracy of the estimated aligning transformations. This is done using paired sets of visit and synthetic revisit landmarks.

For initial visit landmarks we use those from BIRCH [25] clustering on all previously qualifying KITTI [22] keyframes. We also conduct tests using the 3D object pointsets (SHREC [26]) with which GrassGraph [13] was initially proposed (randomly subsampled to 120 points, for parity). To produce the "revisit" landmarks in the synthetic sets we take the initial set of "visit" landmarks and apply a random rotation/translation. We add outliers by replacing a specified percentage of landmark points with random points. We add noise by moving the landmarks in random directions, with the magnitude sampled from a half-normal distribution with the desired standard deviation. The number of outliers is varied in first small, then larger steps between 0% and 50% while the standard deviation of the noise is sampled in increments from 0 to 6 meters with additional steps in the 1.5 meter range. This is roughly twice the standard deviation of inter-landmark distances seen during characterization of semantically-assisted clustering and four times that of unassisted clustering. These standard deviations are in meters and so are scaled when using SHREC [26] pointsets by their typical diameter, compared with the diameter of the imitation scans derived from SLAM.

IV. RESULTS AND DISCUSSION

A. Characterizing the Clustered Landmarks

Here we provide an analysis of the repeatability of the clustering-based landmarks we extract. This illustrates the properties of the methods explored as well as the benefits of using semantic labels to inform better clustering solutions. In conjunction with our following analysis of the sensitivity of the GrassGraph [13] point association method used, this also sheds light on why we were not able to obtain sufficiently

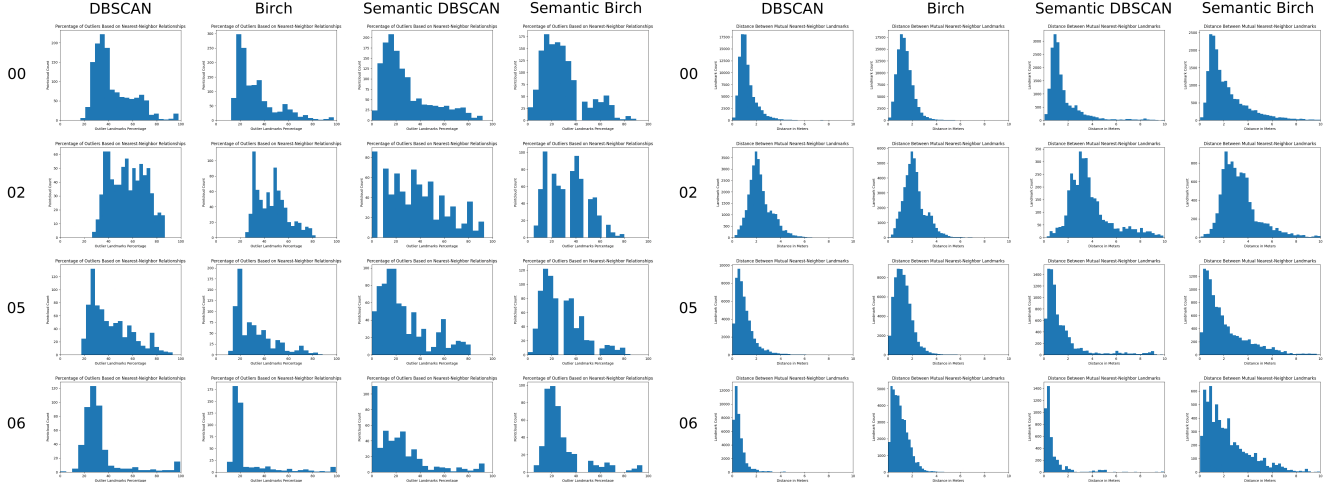


Figure 2. Histograms for each method of clustering and sequence in KITTI. On the left is the number of SLAM keyframes with a particular percentage of outlier landmarks. On the right is the distribution of distances between matching landmarks, across all keyframes. These cases are before correction of the total number of landmarks for association.

reliable landmark association for a place recognition system. In our analysis we focus on two properties of the landmark sets, the number of unpaired “outlier” landmarks which are not seen in a later closest revisit, and the positional error which causes the positions of landmarks to deviate when redetected.

In Figure 2 are distributions of how many outlier landmarks per frame were typical (left 4x4 grid), as well as distributions of the distance that landmarks have been displaced between visit and revisit (right 4x4 grid). Numerical figures can be found in Tables I and II respectively. Focusing first on the unassisted DBSCAN [24] and BIRCH [25] columns, we can see a typical peak at 20% outliers which trails off to the right, raising the average percentage of landmarks which are outliers to 25-56% depending on the method and sequence. This is a high percentage of outlier points for point association methods, as most are not tested past 50%. BIRCH achieves an about 10% lower percentage of outliers and a smaller standard deviation, though one standard deviation rises close to 50% in some cases. BIRCH targets a specific total number of landmarks, avoiding the correction penalty DBSCAN pays for generating a variable number (see the second line under DBSCAN numbers, Table I). When measuring the position error of unassisted methods, a standard deviation of around 1.5 meters is typical, as presented in Table II. The standard deviation assumes an expected half-normal distribution, reasonable when measuring distance between randomly-displaced points. With sequences 00 and 02 which have cases of worse repeatability (the mean displacement is not as zero-centered) we provide regular Gaussian distribution figures at the bottom of the table. Upon inspection sequence 02 stands out as being nearly featureless, depicting an empty road surrounded on

both sides by continuous thick vegetation. Sequence 00 is the longest and perhaps the most varied, and so may incorporate these or other difficult sections as well.

To reduce the prevalence of outliers, we employ semantic labels to constrain clustering and produce two enhanced methods. Observing the semantically-assisted clustering methods in Figure 2 as well as Tables I and II we can see an 8-10% reduction in the total number of outlier landmarks. This reduces the range from 25-56% to 22-37% and manifests as a leftward shift in the distribution of outlier percentages. The pre-total-correction results of semantic methods are also very comparable, meaning the exact method of clustering is less important with better separation. However, the noise standard deviation (the typical distance landmarks move when redetected) has doubled to about 3m. The most likely cause is an effective change in scale, as semantically-guided clustering tends to cluster entire objects and not just feature-rich substructures (e.g. houses instead of windows and doors). This also results in fewer overall clusters (in this case 20 were more typical than 120). Larger clusters (and increased distance between them) could lead to more error in estimating their cluster center.

B. Characterizing the GrassGraph Association

In addition to characterizing generated landmarks by number of outliers and the drift in redetected landmarks’ positions, we examine the performance of GrassGraph [13] SoTA point association with respect to these metrics. This allows us to better understand its frequent failure to associate generated landmarks. To do so we take starting sets of landmarks and introduce controlled quantities of outliers and positional error to simulate a range of conditions. We do so not only with landmarks we derive from the KITTI [22] dataset, but also pointsets drawn from the SHREC

		DBSCAN	BIRCH	Sem-DBSCAN	Sem-BIRCH
Seq 00	Min	16.66	12.50	0.00	0.00
		27.50		0.00	
	Avg	43.67	32.73	28.23	28.87
		46.67		33.43	
Seq 02	Std Dev	15.86	16.32	20.49	18.07
		13.74		19.81	
	Min	26.47	24.16	0.00	5.00
		29.16		0.00	
Seq 05	Avg	55.88	46.16	36.61	32.97
		55.26		42.84	
	Std Dev	14.51	12.88	24.56	16.16
		14.42		21.67	
Seq 06	Min	17.59	10.00	0.00	0.00
		20.00		0.00	
	Avg	42.28	31.97	26.52	28.05
		44.26		32.31	
Seq 02	Std Dev	17.16	16.01	20.67	16.98
		16.90		22.03	
	Min	0.00	9.16	0.00	5.00
		23.33		0.00	
Seq 05	Avg	33.22	25.39	21.73	27.64
		40.38		30.29	
	Std Dev	17.88	18.49	21.84	17.99
		16.17		21.94	

Table I

STATISTICS FOR THE AVERAGE NUMBER OF OUTLIER LANDMARKS GENERATED ACROSS THE FOUR METHODS EXPLORED AND THE FOUR SEQUENCES IN KITTI WITH SIGNIFICANT REVISITS. SECONDARY NUMBERS BELOW THOSE FOR DBSCAN ILLUSTRATE THE EFFECT OF CORRECTION OF THE TOTAL NUMBER OF LANDMARKS (WHICH TENDS TO RAISE THE NUMBER OF OUTLIERS), WHILE BIRCH STATISTICS REMAIN VIRTUALLY UNCHANGED.

		DBSCAN	BIRCH	Sem-DBSCAN	Sem-BIRCH
Seq 00		1.49	1.58	2.13	2.78
		1.64		2.35	
Seq 02		2.50	2.38	5.61	3.82
		2.73		4.53	
Seq 05		1.31	1.47	2.72	2.94
		1.55		2.47	
Seq 06		0.85	1.24	2.14	3.01
		1.13		2.47	
Average		1.57	1.67	3.15	3.14
		1.76		2.95	
Avg (excl. 02)		1.22	1.43	2.33	2.91
		1.44		2.43	

		DBSCAN	BIRCH	Semantic DBSCAN	Semantic BIRCH
Seq 00	Avg	1.26	1.43	1.61	2.23
		1.37		1.77	
	Std Dev	0.78	0.68	1.39	1.66
Seq 02		0.90		1.54	
	Avg	2.31	2.20	4.51	3.40
		2.45		4.00	
Seq 05	Std Dev	0.96	0.92	3.31	1.73
		1.19		2.10	

Table II

STANDARD DEVIATION OF THE DISTANCE BETWEEN MATCHED VISIT-REVISIT LANDMARK PAIRS, FOR VARIOUS SEQUENCES AND CLUSTERING METHODS. A HALF-NORMAL DISTRIBUTION IS ASSUMED FOR THESE ERROR DISTANCES. FOR 00 AND 02 THE PEAK IS NOT VERY WELL ZERO-CENTERED, SO SECONDARY VALUES ARE ALSO GIVEN ASSUMING A NORMAL DISTRIBUTION. FOR DBSCAN THE VALUES BELOW INCORPORATE THE EFFECT OF CORRECTING THE TOTAL NUMBER OF LANDMARKS, WHERE BIRCH IS VIRTUALLY UNAFFECTED.

[26] dataset on which GrassGraph was initially proposed. The results are presented in Figures 3 and 4. The overall health of GrassGraph’s ability to associate landmark sets is in the form of the number of associations found though we also give other metrics based on the recovered alignment transformation.

1) *On the KITTI Dataset:* Looking at the top two plots in in Figure 3 which presents results on landmarks from KITTI [22], it can be seen that introducing noise or outliers has an immediate effect on the performance of GrassGraph [13], reducing the number of associations found. The degradation due to positional noise is more gradual, while 5% outliers decreases the percentage of associations found to approximately 3%. This only occurs at the higher end of the noise scale. Adding individual outlier landmarks as in the bottom-most plot of either figure, we see that even one outlier has a similarly strong effect of reducing the percentage of associations found to around 10%. At the same time, with one outlier the Frobenius norm error of the recovered alignment matrix rises from 0.0 to the range of 15-19 across the four sequences, and at 5% outliers rises to the range of 35-37 (Figure 5). This demonstrates a very strong sensitivity to outliers, making the reliable redetection of outliers even more important than their precise positioning. Both, however, have a large enough effect to

require consideration.

2) *On the SHREC Dataset:* The initial proposal of GrassGraph [13] displayed a susceptibility to outliers, particularly in 3D space, but consistency of testing is important. For this reason we repeat our outlier and noise testing on the SHREC [26] dataset of common objects with which GrassGraph was initially demonstrated. The results are presented in Figure 4 to accompany Figure 3. Here the same impact of noise and outliers can be seen, however the impact with SHREC pointsets is slightly less. At 5% outliers, 13% of the landmarks are associated instead of just 3%. At 5% outliers the Frobenius norm error is 4.9 instead of ranging from 35-37 (Figure 6). The impact of noise is more comparable, being close to the range seen on the different sequences of KITTI [22].

C. Discussion of Results and Recommendations for Future Work

It has become clear that a strong intolerance to landmarks not found reliably is intrinsic to the GrassGraph [13] method of associating points, and has a abrupt and strong effect on the number of associations made. A moderately higher tolerance to outliers on the SHREC [26] dataset also suggests that, all things being equal, the shape of the point sets themselves has an effect on association. Visually, the common objects in SHREC are very self contained

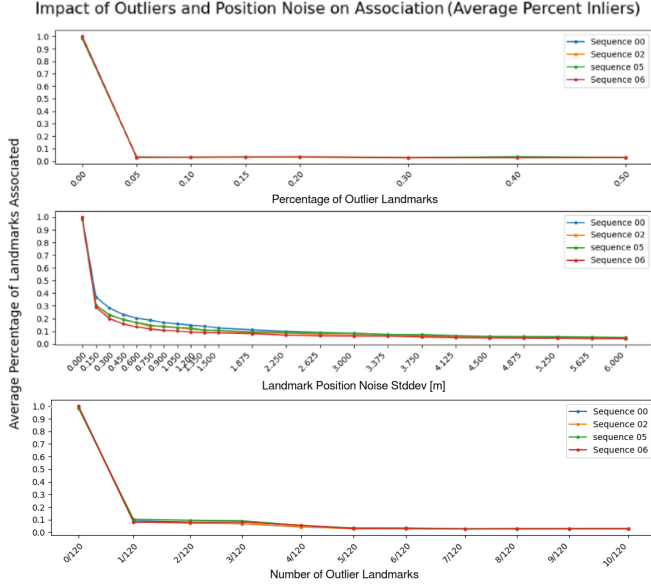


Figure 3. Percentage of point associations found by GrassGraph [13] when associating landmarks derived from SLAM 3D points (SLAM on the KITTI [22] dataset), in the presence of outlier points (top), position noise (middle), and a plot where one outlier point is added at a time (bottom).

(Figure 7), being much closer to a rounded closed shape than the long and branching city streets seen in recovered SLAM structure (Figure 1). It appears the nature of the task, for example an inward-looking object-alignment task or an outward-looking navigation task in a large environment, has a previously unaddressed effect on association through pointset shape when using GrassGraph.

It is also clear that more reliable detection of landmarks is required, beyond what has been explored with clustering. Clustering of SLAM points, even when semantically guided, does not provide the needed repeatability. This is in a large part due to the semi-random nature of the underlying points gathered from SLAM, whose positions are approximate and detection unreliable due to various visual effects. While points tend to be detected repeatedly on specific structures, the boundaries can become indistinct leading to multiple object instances becoming merged or single instances being divided during clustering. It is also possible that few points may be detected on an object, causing it to be lost during clustering. All of these effects alter not just the precise location of landmarks, but also may split them or cause them to not be found at all.

To address these issues we recommend further exploration towards selecting which points will be stable landmarks using various saliency measures. A simple extension would be to move from semantic labeling to semantic instance detection, to help disambiguate object instances. There is also promising work on the segmentation of salient objects from backgrounds, even in the case of unknown object

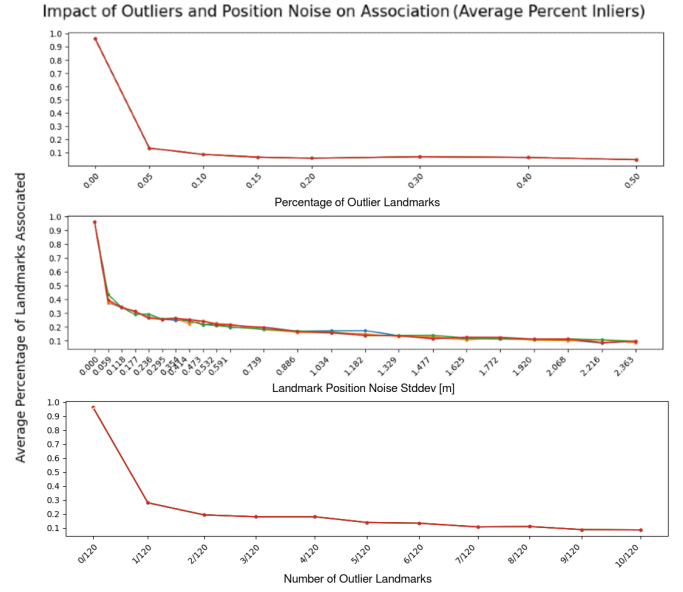


Figure 4. Percentage of point associations found by GrassGraph [13] when associating landmarks derived from the SHREC [26] 3D object dataset, in the presence of outlier points (top), position noise (middle), and a plot where one outlier point is added at a time (bottom). Noise was tested multiple times, thus the multiple colored lines.

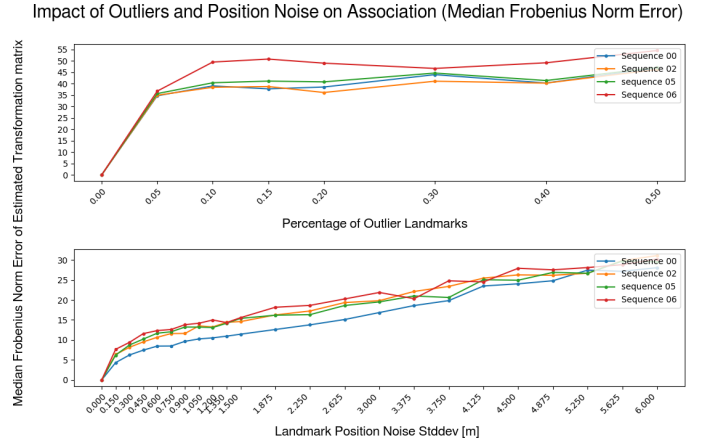


Figure 5. Median Frobenius norm error of the alignment matrix recovered by GrassGraph [13] when associating landmarks derived from SLAM 3D points (SLAM on the KITTI [22] dataset), in the presence of outlier points (top) and position noise (bottom).

classes [27] [28]. Even parts of previously-unseen objects can be automatically segmented [29]. Neural networks have also been employed to select stable and salient features [30] in appearance-based place recognition. There is existing work to learn at runtime what observed features are or are not stable, for example by [31]. In this work features are promoted from short term to long term memory through repetition. By filtering for the most stable and reliably redetected feature points in the environment, we expect the resulting landmarks will be significantly more repeatable

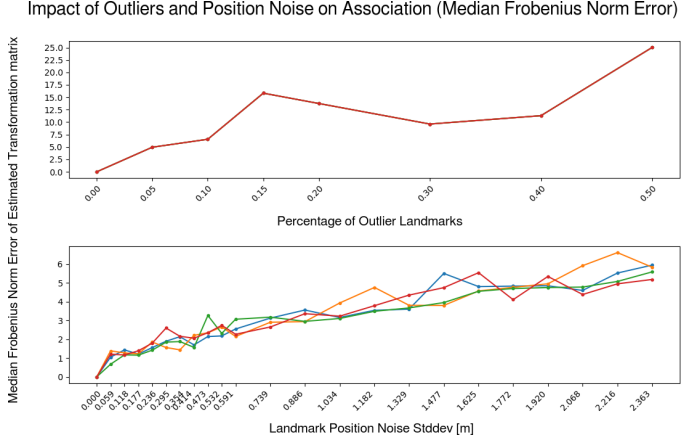


Figure 6. Median Frobenius norm error of the alignment matrix recovered by GrassGraph [13] when associating landmarks derived from the SHREC [26] 3D object dataset, in the presence of outlier points (top) and position noise (bottom). Noise was tested multiple times, thus the multiple colored lines.

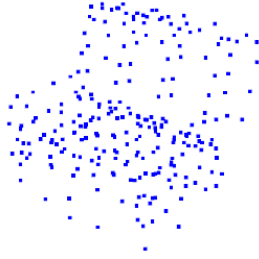


Figure 7. A typical object (a chair) from the SHREC [26] dataset.

compared with simply clustering all available features and relying on a center-of-mass style approach. We found repeatability of detection has the most impact on the ability to associate landmarks, so more stable landmarks will be much improved for association between place visits.

V. CONCLUSIONS

Applying structural place recognition to visual SLAM promises to alleviate core challenges in appearance-based visual place recognition. The underlying structure of the environment has previously been found to be robust to illumination and other visual effects. Differences in coverage have frustrated attempts to apply whole-scan structural LiDAR descriptors, while visual points naturally accumulate on objects and textures in clumps. This would suggest that clustering into landmarks for association should be explored instead, however we have found that the underlying instability of detected points from SLAM leads to unstable landmark detection. This was improved significantly but not sufficiently through semantic labeling. To address these difficulties we recommend selection instead of only the most salient and stable points, for landmarks that can be expected to be repeatedly redetected. This will allow for reliable

place recognition for visual SLAM that is rooted in physical structure, robust to challenging visual effects.

ACKNOWLEDGMENT

We would like to thank the Natural Sciences and Engineering Research Council of Canada and the Ontario Graduate Scholarship Program for supporting this research.

REFERENCES

- [1] G. Kim and A. Kim, "Scan context: Egocentric spatial descriptor for place recognition within 3d point cloud map," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 4802–4809.
- [2] L. He, X. Wang, and H. Zhang, "M2dp: A novel 3d point cloud descriptor and its application in loop closure detection," in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2016, pp. 231–237.
- [3] K. P. Cop, P. V. Borges, and R. Dubé, "Delight: An efficient descriptor for global localisation using lidar intensities," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 3653–3660.
- [4] D. Cattaneo, M. Vaghi, and A. Valada, "Lcdnet: Deep loop closure detection and point cloud registration for lidar slam," *IEEE Transactions on Robotics*, vol. 38, no. 4, pp. 2074–2093, 2022.
- [5] 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.
- [6] C. Campos, R. Elvira, J. J. G. Rodríguez, J. M. Montiel, and J. D. Tardós, "Orb-slam3: An accurate open-source library for visual, visual-inertial, and multimap slam," *IEEE Transactions on Robotics*, vol. 37, no. 6, pp. 1874–1890, 2021.
- [7] J. Engel, V. Koltun, and D. Cremers, "Direct sparse odometry," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 3, pp. 611–625, 2017.
- [8] R. Wang, M. Schworer, and D. Cremers, "Stereo dso: Large-scale direct sparse visual odometry with stereo cameras," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 3903–3911.
- [9] S. Lowry, N. Sünderhauf, P. Newman, J. J. Leonard, D. Cox, P. Corke, and M. J. Milford, "Visual place recognition: A survey," *IEEE Transactions on Robotics*, vol. 32, no. 1, pp. 1–19, 2015.
- [10] S. Hausler, S. Garg, M. Xu, M. Milford, and T. Fischer, "Patch-netvlad: Multi-scale fusion of locally-global descriptors for place recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 14 141–14 152.
- [11] J. Mo and J. Sattar, "A fast and robust place recognition approach for stereo visual odometry using lidar descriptors," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 5893–5900.

- [12] —, “Extending monocular visual odometry to stereo camera systems by scale optimization,” in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2019, pp. 6921–6927.
- [13] M. Moyou, A. Rangarajan, J. Corring, and A. M. Peter, “A grassmannian graph approach to affine invariant feature matching,” *IEEE Transactions on Image Processing*, vol. 29, pp. 3374–3387, 2019.
- [14] J. Komorowski, “Minkloc3d: Point cloud based large-scale place recognition,” in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 1790–1799.
- [15] S. Garg, T. Fischer, and M. Milford, “Where is your place, visual place recognition?” *arXiv preprint arXiv:2103.06443*, 2021.
- [16] G. Kootstra, S. De Jong, and L. R. Schomaker, “Using local symmetry for landmark selection,” in *International Conference on Computer Vision Systems*. Springer, 2009, pp. 94–103.
- [17] D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.
- [18] C. Hua, Y. Yu, and Z. Wang, “Study on feature extraction algorithm of mobile robot vision slam under dynamic illumination,” in *AOPC 2019: Optical Sensing and Imaging Technology*, vol. 11338. SPIE, 2019, pp. 446–451.
- [19] K. Mikolajczyk and C. Schmid, “An affine invariant interest point detector,” in *Computer Vision—ECCV 2002: 7th European Conference on Computer Vision Copenhagen, Denmark, May 28–31, 2002 Proceedings, Part I 7*. Springer, 2002, pp. 128–142.
- [20] —, “A performance evaluation of local descriptors,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 27, no. 10, pp. 1615–1630, 2005.
- [21] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. V. Gool, “A comparison of affine region detectors,” *International journal of computer vision*, vol. 65, pp. 43–72, 2005.
- [22] A. Geiger, P. Lenz, and R. Urtasun, “Are we ready for autonomous driving? the kitti vision benchmark suite,” in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.
- [23] J. Hu, L. Huang, T. Ren, S. Zhang, R. Ji, and L. Cao, “You only segment once: Towards real-time panoptic segmentation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 17 819–17 829.
- [24] M. Ester, H.-P. Kriegel, J. Sander, X. Xu *et al.*, “A density-based algorithm for discovering clusters in large spatial databases with noise,” in *kdd*, vol. 96, no. 34, 1996, pp. 226–231.
- [25] T. Zhang, R. Ramakrishnan, and M. Livny, “Birch: an efficient data clustering method for very large databases,” *ACM sigmod record*, vol. 25, no. 2, pp. 103–114, 1996.
- [26] B. Li, T. Schreck, A. Godil, M. Alexa, T. Boubekeur, B. Bustos, J. Chen, M. Eitz, T. Furuya, K. Hildebrand *et al.*, “Shrec’12 track: Sketch-based 3d shape retrieval,” *Eurographics 2012 Workshop on 3D Object Retrieval*, 2012.
- [27] G. Li, Y. Xie, L. Lin, and Y. Yu, “Instance-level salient object segmentation,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2386–2395.
- [28] X. Liang, L. Lin, Y. Wei, X. Shen, J. Yang, and S. Yan, “Proposal-free network for instance-level object segmentation,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 12, pp. 2978–2991, 2017.
- [29] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, “Segment anything,” *arXiv preprint arXiv:2304.02643*, 2023.
- [30] Z. Xin, Y. Cai, T. Lu, X. Xing, S. Cai, J. Zhang, Y. Yang, and Y. Wang, “Localizing discriminative visual landmarks for place recognition,” in *2019 International conference on robotics and automation (ICRA)*. IEEE, 2019, pp. 5979–5985.
- [31] B. Bacca, J. Salvi, and X. Cufí, “Appearance-based mapping and localization for mobile robots using a feature stability histogram,” *Robotics and Autonomous Systems*, vol. 59, no. 10, pp. 840–857, 2011.