Voxel-Based Representation Learning for Place Recognition Based on 3D Point Clouds

Sriram Siva†, Zachary Nahman†, and Hao Zhang

***Abstract*— Place recognition is a critical component towards addressing the key problem of Simultaneous Localization and Mapping (SLAM). Most existing methods use visual images; whereas, place recognition using 3D point clouds, especially based on the voxel representations, has not been well addressed yet. In this paper, we introduce the novel approach of voxel-based representation learning (VBRL) that uses 3D point clouds to recognize places with long-term environment variations. VBRL splits a 3D point cloud input into voxels and uses multi-modal features extracted from these voxels to perform place recognition. Additionally, VBRL uses structured sparsity-inducing norms to learn representative voxels and feature modalities that are important to match places under long-term changes. Both place recognition, and voxel and feature learning are integrated into a unified regularized optimization formulation. As the sparsity-inducing norms are non-smooth, it is hard to solve the formulated optimization problem. Thus, we design a new iterative optimization algorithm, which has a theoretical convergence guarantee. Experimental results have shown that VBRL performs place recognition well using 3D point cloud data and is capable of learning the importance of voxels and feature modalities.**

I. Introduction

For several decades, one of the core robotics challenges has been Simultaneous Localization and Mapping (SLAM). A critical component of SLAM is place recognition (also referred to as loop closure detection). Place recognition is the capability of a robot to recognize a previously visited location. It enables the robot to accurately localize itself within a global map by correcting incremental pose drifts that are accumulated during robot exploration in an environment. Place recognition, along with SLAM, has been applied in a wide variety of real-world applications, including assistive robotics [1], [2], environment exploration [3], [4] and autonomous driving [5], [6].

Most previous research utilizes images of surroundings, obtained from a visual camera installed on a robot to perform visual place recognition [7]. Long-term visual place recognition has received extensive attention during the past few years [8], [9]. It addresses the key challenge that robot environments are dynamic and change over time. For example, indoor places can experience environmental changes in human activity, lighting, and arrangement on a daily basis. Outdoor environments can look drastically different in summer versus winter and in the morning versus night.

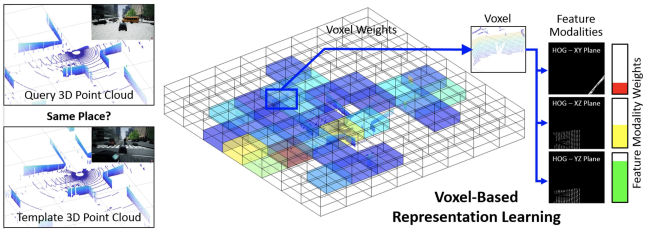


Fig. 1. Illustration of the proposed VBRL approach for place recognition on 3D point cloud data. VBRL divides each 3D point cloud into multiple voxels in the 3D space and extracts multi-modal features from each voxel. Then, VBRL performs joint learning of representative voxels and feature modalities to represent places and integrates the representation for place recognition in a unified regularized optimization formulation.

Given the recent advances in visual place recognition and SLAM, the use of cameras in some environments is difficult or inappropriate. For instance, in the low light or complete darkness (e.g., in subterranean environments), utilizing visual images for place recognition would necessitate bringing light sources to illuminate the environment, which may not be feasible. LiDAR sensors can offer a solution to perceive the environment independent of lighting conditions. By actively projecting laser light and measuring the reflected light, LiDAR sensors measure the distance to a target and provide a 3D point cloud representation of the environment, which can be used by a robot operating in the dark.

Although visual place recognition was widely studied, long-term place recognition based on 3D point clouds (e.g., obtained from LiDAR sensors) has received limited attention. Many previous LiDAR-based place recognition methods directly matched a query scan with a database of previous template scans to determine the best match of places [10]. Keypoint voting [11] and histogram-based matching (e.g., based on normal distributions transform [12]). Recently, [13] used deep learning to perform point cloud based place recognition. However, apart from [13], these methods generally rely on manually defined features without using learning and cannot address place recognition during long-term environmental changes. Also, while voxel-based 3D representations have been commonly utilized in 3D SLAM and robot navigation [14], [15], [16], 3D place recognition using voxel-based representations has not been well addressed.

In this paper, we introduce a novel *Voxel Based Representation Learning* (VBRL) approach to address the problem of place recognition using 3D point clouds acquired by LiDAR sensors. The 3D data is obtained from a 360-degree field of view by a LiDAR sensor, and our approach divides each 3D

*This work was partially supported by NSF CAREER IIS-1942056, NSF IIS-1849348, and DOE DE-FE0031650.

†Authors contributed equally to this work.

S. Siva, Z. Nahman, and H. Zhang are with the Human-Centered Robotics Lab at Colorado School of Mines, Golden, CO 80401. Email: {sivasriram, znahman, hzhang}@mines.edu.

point cloud into multiple voxels in the 3D space. Multiple feature types from each of the voxels are then extracted. Given the multiple types (modalities) of features from all 3D voxels, our proposed VBRL method automatically learns the importance of voxels and feature modalities, and integrates all features in a unified regularized optimization formulation in order to best represent places. For learning the importance of voxels, our approach is inspired by the insight that a specific subset of voxels are typically more representative to encode a place. For example, voxels closer to the LiDAR sensor can be more useful, since it can include 3D points that describe the environment with more details as objects are closer to the sensor. Learning voxel importance is achieved by VBRL through introducing structured sparsity-inducing norms as regularizations into the optimization formulation. Similarly, the VBRL approach is able to learn the importance of the different feature modalities in the same regularized optimization formulation.

There are two main contributions of this paper:

- We propose a novel formulation and the VBRL method to perform simultaneous learning of representative voxels and feature modalities to represent places for place recognition using 3D point cloud data.
- An optimization algorithm is implemented to solve the regularized optimization problem that has a theoretical guarantee to converge to the global optimal solution.

II. Related Work

In this section, we first present the state-of-the-art in long-term visual place recognition, followed by an overview of the latest place recognition techniques utilizing 3D data.

A. Long-Term Visual Place Recognition

The state of the art visual place recognition techniques can be broadly classified into methods based on local features, global features, or learning based on representations.

Local features apply a detector to detect points of interest (e.g., geometric shapes such as lines, corners, e.t.c.) in an image and thus encode the local information around each of the detected points of interest. The Scale-Invariant Feature Transform (SIFT) local features were used in combination with a bag-of-words (BoW) approach to detect previously visited places in a 2D image [17]. In [18], binary BRIEF and FAST features are applied to get a BoW representation and perform loop closure detection. ORB features have also been successfully used to perform place recognition [19]. Local visual features usually differ for the same location based on the different scenarios (e.g., daytime vs night, light flares, or different seasons). Accordingly, these features change with different scenarios and are not effective to encode all scene scenarios for image matching in long-term place recognition.

Global features portray the holistic representation of the scene. For example, Histogram of Gradients (HOG) features group pixels in a grid and stores unsigned gradient changes within each of the pixels in a histogram. GIST [20] features, constructed steerable Gabor filters at different orientations and scales to perform place recognition [21]. Local Binary Pattern is used to represent the whole image using intensity and gradient differences within the image to calculate a binary string [22]. Given the recent advancement in deep learning, Convolutional Neural Networks (CNNs) were also utilized to extract deep representative features to match image sequences to perform long-term place recognition [23], [24], [25]. Many other approaches [26], [27] have also used global features to successfully perform long-term place recognition. These global features without using dictionary-based quantization can encode the whole image information. It has been observed that, while performing long-term place recognition, global features outperform local features [7], [27], [28].

The third category is based on representation learning. Several techniques utilize convolutional neural networks in order to learn representative features from visual images [29]. More recently, [8], [9] address the challenges of the Long-term Appearance Change (LAC) problem by using a learning method to learn which visual feature extraction modalities are shared between different scene scenarios.

Although visual place recognition showed promising performance in good lighting conditions, in harsh environments with low lighting, high-quality images can be hard to obtain. Place recognition approaches based on LiDAR data becomes necessary in such conditions.

B. 3D Place Recognition

3D place recognition methods fall in four broad categories: global scene descriptors, histogram feature binning, keypoint voting, and learning approaches. Constructing a global scene descriptor and comparing it to a database of previously seen locations has been well studied. In [10] and [30], range image similarity is calculated to recognize places. [31] constructs a feature from sparse triangulated landmarks. Other techniques utilize histograms of point cloud features to perform place recognition. In [12], histograms of normal distribution transform are constructed based on surface orientation and smoothness. In [32], [33], histograms are constructed of simple global features extracted from LiDAR scans. In [11], [34], keypoints are 3D descriptors that are calculated from a subset of 3D data. The keypoint downsampling and extraction allows matching to be performed quickly. Other approaches utilize local features called segments, and implement a segment matching algorithm [35]. There has also been research into extracting learned deep features from point cloud data [36], [37], [38]. The PointNetVlad approach extracts deep features from the point cloud and also uses a separate deep network for place recognition [13]. CNNs have also been successfully used to infer 3D structures [39], [40] and generate local key-point descriptors for point clouds [38].

Previous methods for 3D place recognition are generally based on manually defined features without representation learning. Long-term variations have also not been explicitly addressed by the methods using 3D point clouds. In addition, while voxel-based 3D representations have been commonly used for 3D SLAM, 3D place recognition using voxel-based representations has not been well addressed.

III. The VBRL Approach

Notations: Given a matrix $\mathbf{M} = \{m_{ij}\} \in \Re^{u \times v}$, we refer to its i-th row and j-th column as $\mathbf{m}^i$ and $\mathbf{m}_j$. Its Frobenius norm is computed by $\|\mathbf{M}\|_F = \sqrt{\Sigma_{i=1}^{u} \Sigma_{j=1}^{v} m_{ij}^2}$. Given a vector $\mathbf{m} \in \Re^v$, its ℓ_2-norm is defined as $\|\mathbf{m}\|_2 = \sqrt{\mathbf{m}^\top \mathbf{m}}$.

A. Problem Formulation

Given a set of point cloud instances acquired during long-term LiDAR-based navigation over different scenarios, each point cloud is divided into a set of voxels. Then, multiple feature types are extracted from each of these voxels and we defined a modality as the features computed from a specific feature descriptor. The multi-modal features extracted from all voxels are denoted as $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_n] \in \Re^{d \times n}$. $\mathbf{x}_i \in \Re^d$ is the vector of features extracted from all the voxels of the i-th 3D point cloud, which is a concatenation of features from all m modalities, such that $d = \Sigma_{i=1}^{m} \Sigma_{j=1}^{v} d_{ij}$, where d_{ij} is the dimensionality of the i-th feature modality in the j-th voxel, and v is the total number of voxels. The corresponding long-term scenarios (e.g., summer and winter) are represented as $\mathbf{Y} = [\mathbf{y}_1, \dots, \mathbf{y}_n] \in \Re^{n \times c}$, where c denotes the number of scenarios and $\mathbf{y}_i$ is the scenario indicating vector, with each element $y_{ij} \in \{0, 1\}$ denoting that the i-th 3D point cloud is collected from j-th scenario.

Then, we formulate place recognition based on 3D point clouds as a regularized optimization problem:

$$\min_{\mathbf{W}} \mathcal{L}(\mathbf{X}, \mathbf{Y}; \mathbf{W}) + \lambda \mathcal{R}(\mathbf{W}) \tag{1}$$

where $\mathcal{L}(.)$ is the loss function, $\mathcal{R}(.)$ is the sparsity-inducing regularization term, and $\lambda \geq 0$ is a trade-off hyperparameter to balance the loss function and the regularization term. The model parameter $\mathbf{W}$ is a weight matrix, which represents the importance of the features in $\mathbf{X}$ to represent the scenarios $\mathbf{Y}$ in general. By learning the weight matrix $\mathbf{W}$ in Eq. (1), we learn features that are more important towards place recognition. That is, the features that are more important towards place recognition have higher weights and the less important features have weights closer to zero. The loss function is designed to encode the error of using the learned model to represent the scenarios, which can be defined as $\mathcal{L}(\mathbf{X}, \mathbf{Y}; \mathbf{W}) = \min_{\mathbf{W}} \|\mathbf{X}^\top \mathbf{W} - \mathbf{Y}\|_F^2$.

The solution to the optimization problem defined in Eq. (1) is $\mathbf{W} = [\mathbf{w}_1, \dots, \mathbf{w}_c] \in R^{d \times c}$, where $\mathbf{w}_i \in \Re^d$ denotes the weights of features from all views and modalities to represent the i-th scenario. Since $\mathbf{w}_i$ contains the weights of features from m-modalities in all voxels, it can be further denoted as $\mathbf{w}_i = [\mathbf{w}_i^1, \dots, \mathbf{w}_i^m]^\top$. In addition, since each $\mathbf{w}_i^j$ includes the weights of features (extracted from the j-th modality with respect to the i-th scenario) from all voxels, it can be further divided into v parts as $\mathbf{w}_i^j = [\mathbf{w}_i^{j^1}, \dots, \mathbf{w}_i^{j^v}] \in \Re^{d_{ij}}$, where $\mathbf{w}_i^{j^k}$ denotes the weights of features extracted from the k-th voxel and j-th modality with respect to the i-th scenario.

B. Learning Representative Voxels and Feature Modalities

When performing place recognition, we hypothesize that some voxels within the 3D point cloud are more representative than others. To identify representative voxels for place recognition, we introduce a regularization term called a voxel norm. Formally, this norm is a sparsity-inducing norm that can be mathematically defined as $\mathcal{R}_V(\mathbf{W}) = \Sigma_{i=1}^{v} \|\mathbf{W}^i\|_F$. This voxel norm $\mathcal{R}_V$ is used as a regularization term in our optimization formulation to enforce the grouping effect of the multimodal features shared among different scenarios and promote sparsity among different voxels.

Different feature modalities usually capture different characteristics of a place. Some feature modalities can be more representative to describe a place than others. Thus, it is also beneficial to identify the importance of feature modalities to improve long-term place recognition performance. Accordingly, we also propose a regularization term to identify representative feature modalities under the unified regularized optimization framework, which is named modality norm. It is mathematically defined as:

$$\mathcal{R}_M(\mathbf{W}) = \sum_{i=1}^{m} \|\mathbf{W}^i\|_F + \sum_{i=1}^{d} \|\mathbf{w}^i\|_2 \tag{2}$$

which is a combination of two structured sparsity-inducing norms. The first term applies the Frobenius norm within each modality and then applies a group ℓ_1-norm across different modalities. By enforcing sparsity among modalities, this term allows the VBRL method to identify representative modalities that have larger weights, and to make the weights of non-representative features tend towards 0. The second term in Eq. (2) denotes the $\ell_{2,1}$-norm (i.e., a ℓ_2-norm for each column and ℓ_1-norm across different columns) used to enforce the sparsity of the columns of $\mathbf{W}$ and grouping effect of the weights in each column. By enforcing sparsity of individual features, this norm helps recognize representative individual features and assign a zero value to the weights of non-representative features (e.g., from noise).

Incorporating both of the regularization terms to identify representative voxels and feature modalities, our final formulation of learning voxel-based multimodal representations for place recognition can be defined as the following regularized optimization problem:

$$\min_{\mathbf{W}} \mathcal{L}(\mathbf{X}, \mathbf{Y}; \mathbf{W}) + \lambda_1 \mathcal{R}_V(\mathbf{W}) + \lambda_2 \mathcal{R}_M(\mathbf{W}) \tag{3}$$

where λ_1 and λ_2 denote trade-off hyperparameters to govern the balance between the loss function and the structured sparsity-inducing norms.

C. Voxel-Based Multimodal Place Recognition

Once the formulated regularized optimization problem in Eq. (3) is solved based on Algorithm 1, the optimal weight matrix $\mathbf{W}^*$ is obtained. Given the feature vector $\mathbf{x} \in \Re^d$ that is extracted from all voxels and feature modalities in a query 3D point cloud, and a feature vector from a template 3D point cloud $\tilde{\mathbf{x}} \in \Re^d$, we compute a similarity score between this query and template point clouds as follows:

$$s = \sum_{i=1}^{m} \sum_{j=1}^{v} w_M(i) * w_V(j) * (|\mathbf{x}_{ij} - \tilde{\mathbf{x}}_{ij}|) \tag{4}$$

Algorithm 1: The proposed iterative algorithm to solve the formulated problem in Eq. (3)

```
Input  : $\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_n] \in \Re^{d \times n}$ and
         $\mathbf{Y} = [\mathbf{y}_1, \cdots, \mathbf{y}_n]^\top \in \Re^{n \times c}$
1 Let $t = 1$. Initialize $\mathbf{W}(t)$ by solving
    $\min_{\mathbf{W}} \mathcal{L}(\mathbf{X}, \mathbf{Y}; \mathbf{W})$.
2 while not converge do
3     Calculate the block diagonal matrix $\mathbf{D}(t+1)$,
        where the $k$-th diagonal block of $\mathbf{D}(t+1)$ is
        $\frac{\mathbf{I}_v}{2\|\mathbf{W}^k\|_F}$.
4     Calculate the block diagonal matrix $\tilde{\mathbf{D}}(t+1)$,
        where each element of the matrix $\tilde{\mathbf{D}}(t+1)$, is
        given as $\big(\frac{\mathbf{I}_m}{2\|\mathbf{W}^i\|_F} + \frac{1}{2\|\mathbf{w}^i\|_2}\big)$.
5     For each $\mathbf{w}_i (1 \le i \le c)$,
        $\mathbf{w}_i(t+1) = \big((\mathbf{X}\mathbf{X})^\top + \lambda_1 \mathbf{D}(t+1)$
        $+ \lambda_2 \tilde{\mathbf{D}}(t+1)\big)^{-1} (\mathbf{X}\mathbf{y}_i)$.
6     $t = t + 1$.
Output: $\mathbf{W} = \mathbf{W}(t) \in \Re^{d \times c}$
```

where $\mathbf{x}_{ij}$ denotes the vector features from the i-th modality and the j-th voxel, $w_M(i)$ is sum of all weights of features in the i-th feature modality, and $w_V(j)$ is sum of all weights of features in the j-th voxel. When this score is above a user-defined threshold, the query 3D point cloud is determined as a match with the template 3D point cloud.

D. Optimization Algorithm

The objective function in Eq. (3) is composed of three non-smooth regularization terms. Generally, this is challenging to solve. Accordingly, we implement an iterative algorithm to solve the formulated regularized optimization problem.

Taking the derivative of the objective function with respect to the columns of $\mathbf{W}$ and setting it to zero gives us:

$$\mathbf{X}\mathbf{X}^\top \mathbf{w}_i - \mathbf{X}\mathbf{y}_i + \lambda_1 \mathbf{D}\mathbf{w}_i + \lambda_2 \tilde{\mathbf{D}}\mathbf{w}_i = 0 \qquad (5)$$

where $\mathbf{D}$ is a diagonal matrix with the i-th diagonal element defined as $\frac{\mathbf{I}_v}{2\|\mathbf{W}^i\|_F}$, $\tilde{\mathbf{D}}$ denotes a diagonal matrix with the i-th diagonal element defined as $\frac{\mathbf{I}_m}{2\|\mathbf{W}^i\|_F} + \frac{1}{2\|\mathbf{w}^i\|_2}$, and $\mathbf{I}_v$ and $\mathbf{I}_m$ are identity matrices with size v and m respectively. Then, we obtain:

$$\mathbf{w}_i = \left(\mathbf{X}\mathbf{X}^\top + \lambda_1 \mathbf{D} + \lambda_2 \tilde{\mathbf{D}}\right)^{-1} (\mathbf{X}\mathbf{y}_i) \qquad (6)$$

Since the matrices $\mathbf{D}$ and $\tilde{\mathbf{D}}$ are dependent on the weight matrix $\mathbf{W}$, we implement an iterative algorithm to solve the formulated regularized optimization problem, as described in Algorithm 1, which holds a theoretical guarantee to converge to the global optimal solution[1].

Complexity. Since the optimization problem in Eq. (3) is convex, Algorithm 1 converges to the global optimal solution fast. In each of the iterations, computing Step 3 and Step 4 is trivial. Step 5 can be computed through solving a system of linear equations with quadratic complexity.

IV. Experimental Results

In our implementation, each 3D point cloud scan from LiDAR is divided into many voxels. From each voxel five different feature descriptors are extracted including (1) covariance of points contained within the voxel, (2) Histogram of Oriented Gradients (HOG) features of a snapshot of the point cloud in the XY plane, (3) XZ plane, (4) YZ plane, and (5) Subvoxel Occupancy.

The subvoxel occupancy feature is obtained simply by dividing a voxel into 8 equal subvoxels. If the subvoxel is occupied by any points, a 1 is written to the feature matrix. Otherwise a 0 is written. As opposed to concatenating these features together from each voxel, VBRL operates with the intuition that learning a shared representation of the overall scene from multiple data instances and weighting the feature matrix accordingly will fuse the feature modalities more effectively for loop closure detection.

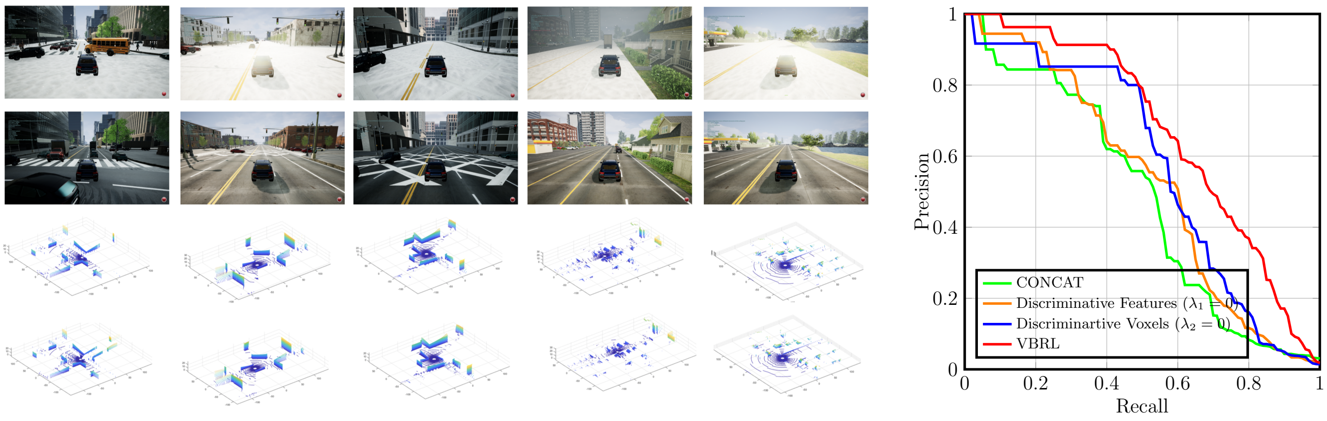
Experiments are evaluated both qualitatively and quantitatively. To showcase that VBRL learns a better representation of a LiDAR scan than feature extraction alone, we compare VBRL ($\lambda_1 = 10$ and $\lambda_2 = 0.1$) to performing loop closure detection with features concatenated together ($\lambda_1 = 0$ and $\lambda_2 = 0$), voxel learning only ($\lambda_1 = 10$ and $\lambda_2 = 0$), and modality learning only ($\lambda_1 = 0$ and $\lambda_2 = 0.1$).

A. Results on Autonomous Driving Simulation

At first, we evaluate the performance of our VBRL approach to perform 3D point cloud based long-term place recognition by extensive experimenting on data obtained from the AirSim simulator. AirSim [41] is an autonomous driving simulator developed by Microsoft to facilitate the development of self-driving vehicle methods in a virtual environment. We collect the dataset in AirSim's cityscape environment with roads, skyscrapers, parks, and dynamically moving cars and pedestrians. A virtual LiDAR sensor is installed on top of a vehicle to record the point cloud data from the virtual environment. The point cloud based LiDAR scans are collected from 210 unique locations within the environment. These scans are first collected in clear, sunny weather. This set of point cloud scans constitute one scenario of training our VBRL approach. Then, point cloud scans are collected from these locations from the self-driving vehicle during snow and fog conditions forming the second scenario. All of the 210 locations were distinctive to one another and there was no overlap. To perform the experiments on simulated data and evaluate our approach, we used 160 point cloud scans for training and a disjoint set of 50 point clouds are designated for testing. It is to be noted that the training and testing data doesn't have any overlap to make sure that our approach can learn a robust weight matrix $\mathbf{W}$, that can be used to perform place recognition in new and unseen locations.

The main challenges associated with this dataset are the dynamic cars and pedestrians. The LiDAR scans are robust to changes in lighting conditions and are not affected by the virtual snow. However, because fog, as well as snow, could reflect lasers, certain LiDAR points may be represented as

[1] Proof is available at: `http://hcr.mines.edu/publication/VBRL_Supp.pdf`



(a) Examples of matched point clouds

(b) Precision-recall curves

Fig. 2. Qualitative and quantitative experimental results over the AirSim simulations.

noise, affecting the representation of the scene adversely. This is a key challenges in present-day place recognition using point clouds because autonomous vehicles need to operate in snow, fog, rainy seasons. However, this dataset doesn't provide data with variation in vegetation. Illustrated in Figure 2 are the place recognition results based on our VBRL approach and its comparison to baseline approaches. The qualitative results on 3D point cloud scan matches are illustrated in Figure 2(a). The template point clouds from the snow scene that obtain the maximum matching score are shown in the top row, while the query scenes from the clear scene are shown in the bottom row. It is observed that our VBRL approach can match point clouds, despite changes in lighting conditions and weather, thus proving the capability to perform long-term place recognition.

The classification problem is analyzed quantitatively using the standard precision-recall curve. Figure 2(b) shows the precision-recall performance of VBRL when compared with features concatenated together, discriminative voxels alone, and discriminative features alone. We observe that using feature concatenation alone we achieve minimal performance in point cloud-based place recognition. Using discriminative features increases the performance, as the area under the curve increases. Introducing the discriminative voxel learning approach increase the performance even more. Finally, we observe the performance of our VBRL approach, where it obtains the maximum area under the curve when compared with previous methods, indicating the best performance. Therefore, by fusing multiple feature modalities together and weighting them based on importance, the VBRL approach yields the best results for loop closure detection.

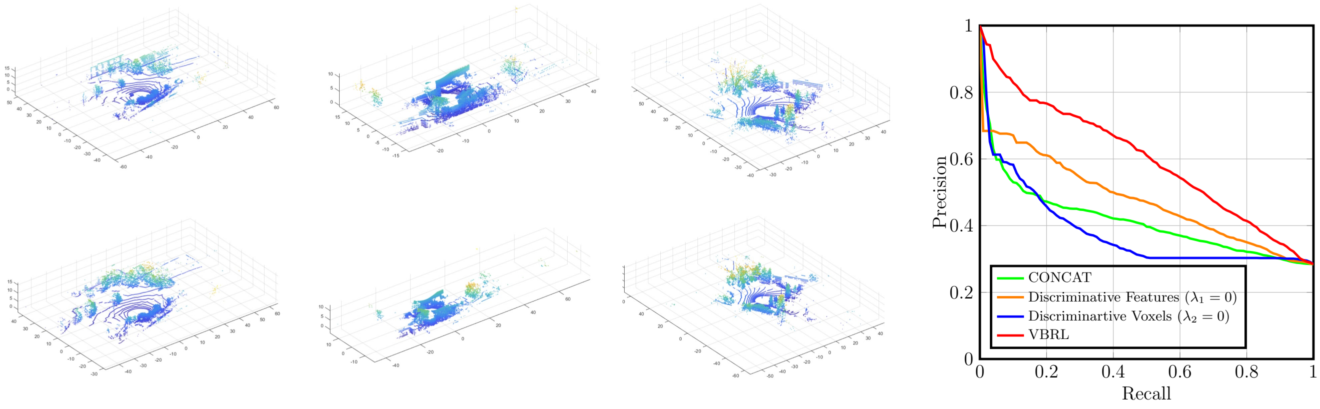
The modality weights learned automatically for the AirSim dataset are shown in Figure 4(a). The Subvoxel Occupancy feature is the most important with a weight of 30% and the covariance feature is the second most important with a weight of 29%. The three HOG feature importances range from 4% for HOG-XY to 28% for HOG-XZ.

The learned voxel weights are shown via a color map above in Figure 1. Voxels occurring more towards the center of the workspace are learned to be weighted as more important in place recognition. This makes sense as the center voxels are most likely to be occupied because they are closest to the sensor origin and in a LiDAR scan point clouds are more populated in the center. Apart from this, we also analyze the relative importance of the different layers of voxels (top, middle, and bottom) when performing place recognition using point clouds in the AirSim dataset. It was observed that the relative importance of bottom, middle and top layer was 37.08%, 42.22%, and 20.72% respectively. This indicated that the bottom and middle layer were critical towards point cloud based place recognition

B. Results on the NCLT Dataset

The North Campus Long Term (NCLT) [42] dataset is collected at the University of Michigan by a mobile robot driven around the campus. There are 27 separate sessions with varying robot routes in the dataset, which occur over the course of 15 months and span multiple times of day and seasons. The dataset contains long-term changes in lighting conditions, vegetation, and weather. Two sessions are chosen: one collected in June and the other in December. These seasons are selected as they have overlapping routes and seasonal changes. A Velodyne HDL-32E LiDAR sensor was used to collect 3D point cloud data of the environment and was mounted on the mobile robot. This dataset has dynamic pedestrians and also has vegetation changes. Change is vegetation is typically observed with seasonal changes and is important to be addressed in the LAC problem, to perform long term place recognition. The NCLT dataset includes 850 LiDAR scans from the month of June and a corresponding 850 LiDAR scans from the month of December. For this set of experiments, we choose 700 instances of point clouds for training and for testing a disjoint set of 150 point cloud scans are taken. Again, it is to be noted that the training and testing data doesn't have any overlap in order to make sure that our approach is robust.

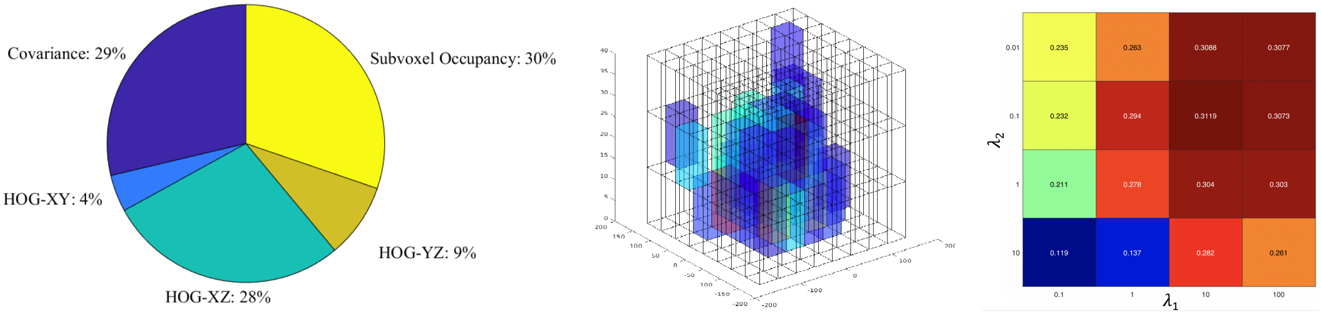
The qualitative results of the performance of our VBRL approach are provided in Figure 3(a) in which query scans



(a) Examples of matched point clouds

(b) Precision-recall curves

Fig. 3. Experimental results over the NCLT dataset for long-term 3D point cloud based place recognition in different seasons.



(a) Modality importance for AirSim

(b) Voxel weights for NCLT

(c) Hyperparameter analysis

Fig. 4. Experimental results over the NCLT dataset in different seasons. Figure 4(a) shows the importance of feature modalities for the AirSim simulations. Figure 4(b) shows the importance of voxels for long-term place recognition using the NCLT dataset, where the robot is located in the center of the point cloud at position $(0, 0)$. Figure 4(c) illustrates the performance variations of our VBRL approach given different hyperparameter values over NCLT.

from the data collected in June are shown on the bottom row and resulting matches from December are shown in the top row. Our VBRL approach is able to recognize scenes from 3D point cloud data despite vegetational, seasonal and structural changes (such as leaves falling off of trees).

Figure 3(b) shows the qualitative precision-recall analysis of VBRL on the NCLT dataset. Once again, it is observed that VBRL yields greater area under the precision-recall curve than discriminative voxels, discriminative features, or feature concatenation. Additionally, the learned modality weights obtained are shown. The learned voxel weights are also shown in Figure 4(b) and results obtained are similar to the AirSim dataset in that the center voxels are learned to be of more importance than the outer voxels.

The learned voxel weights are also shown in Figure 4(b). An analysis of weights of the different voxel layers showed that the bottom, middle and top layer have their relative importance as 33.92%, 54.62%, and 11.46% respectively. Quite contrasting to the results obtained in the AirSim dataset, we see that the top layer has very little importance. The bottom layer's importance also decreases. However, the middle layer plays a major role in place recognition. The NCLT dataset was majorly collected when the robot traverses over open areas and didn't have any tall structures throughout its route. Thus, the bottom and top layer's importance is less, whereas, the middle layer has the most importance as it has the maximum number of important features that are critical in place recognition as opposed to the AirSim point cloud data which had point cloud scans of tall buildings nearby.

C. Discussion

Importance of Voxels and Feature Modalities: Our VBRL approach can automatically estimate the importance of each of the voxels and feature modalities while training. The relative importance of voxels is illustrated in Figure 4(b). Intuitively, points closer to us are more important towards performing place recognition. It is analogous to the fact that humans also use nearby points such as street signs and buildings to recognize places rather than using mountains in the distance. Accordingly, our approach indicates that point clouds near the center are of more importance. On the other hand, voxels far away from the center are of least importance and thus their weights are close to zero. The importance of feature modalities are illustrated in Figure 4(a). The pie chart here indicates the relative importance of different feature modalities towards performing voxel-based place recognition. It is observed that Subvoxel occupancy, Covariance and HOG-XZ have an importance of 30%, 29%

and 28% respectively and are equally important in general, whereas, HOG-XY is of least importance.

Hyperparameter Selection: The hyperparameters λ_1 and λ_2 in our formulation of the final objective function, Eq. (3), are designed to control the strength of regularization norms over learning descriptive voxels and feature modalities respectively. Their optimal values can be determined using cross-validation during training. From the result in Figure 4(c), we observe that when $\lambda_1 = 10$ and $\lambda_2 = 0.1$, VBRL statistically obtains the best accuracy while performing 3D point cloud based place recognition. In general, the range $\lambda_1 \in \{1, 100\}$ and $\lambda_2 \in \{0.01, 1\}$ can result in satisfactory results, which demonstrates that both of the regularization terms are useful to improve performance.

V. Conclusion

In this paper, we study the key problem of long-term place recognition using 3D point clouds, through proposing a novel Voxel-Based Representation Learning (VBRL) method. Our approach divides each 3D point cloud scan into multiple voxels in the 3D space and extracts multiple modalities of features from each of the voxels. Then, our VBRL approach performs joint learning of representative voxels and feature modalities to represent places and integrates the representation for place recognition in a unified regularized optimization formulation. Due to the presence of two non-smooth sparsity-inducing norms, our formulated optimization problem is hard to solve. Therefore, we design an iterative solver that has a convergence guarantee. Experiment results have validated that VBRL obtains promising performance on long-term place recognition using 3D point clouds.

References

[1] Lili Meng, C. W. de Silva, and Jie Zhang, "3D visual slam for an assistive robot in indoor environments using RGB-D cameras," in *ICCSE*, 2014.

[2] C. De la Cruz, T. F. Bastos, F. A. A. Cheein, and R. Carelli, "Slam-based robotic wheelchair navigation system designed for confined spaces," in *ISIE*, 2010.

[3] S. S. Belavadi, R. Beri, and V. Malik, "Frontier exploration technique for 3D autonomous slam using k-means based divisive clustering," in *AMS*, 2017.

[4] R. Sim and N. Roy, "Global a-optimal robot exploration in slam," in *ICRA*, 2005.

[5] A. Singandhupe and H. La, "A review of slam techniques and security in autonomous driving," in *IRC*, 2019.

[6] R. W. Wolcott and R. M. Eustice, "Visual localization within lidar maps for automated urban driving," in *IROS*, 2014.

[7] S. M. Lowry, N. Sünderhauf, P. Newman, J. J. Leonard, D. D. Cox, P. I. Corke, and M. Milford, "Visual place recognition: A survey," *TRO*, vol. 32, pp. 1–19, 2016.

[8] F. Han, X. Yang, Y. Deng, M. Rentschler, D. Yang, and H. Zhang, "SRAL: Shared representative appearance learning for long-term visual place recognition," *R-AL*, 2017.

[9] S. Siva and H. Zhang, "Omnidirectional multisensory perception fusion for long-term place recognition," in *ICRA*, 2018.

[10] S. Grzonka, B. Steder, and W. Burgard, "3D place recognition and object detection using a small-sized quadrotor," in *RSS*, 2011.

[11] M. Bosse and R. Zlot, "Place recognition using keypoint voting in large 3D lidar datasets," in *ICRA*, 2013.

[12] M. Magnusson, H. Andreasson, A. Nuchter, and A. J. Lilienthal, "Appearance-based loop detection from 3D laser data using the normal distributions transform," in *ICRA*, 2009.

[13] M. A. Uy and G. H. Lee, "Pointnetvlad: Deep point cloud based retrieval for large-scale place recognition," *CoRR*, 2018.

[14] J. Ryde and J. A. Delmerico, "Extracting edge voxels from 3D volumetric maps to reduce map size and accelerate mapping alignment," in *CRV*, 2012.

[15] Inwook Shim, Yungeun Choe, and Myung Jin Chung, "3D mapping in urban environment using geometric featured voxel," in *URAI*, 2011.

[16] Z. Liu, H. Chen, H. Di, Y. Tao, J. Gong, G. Xiong, and J. Qi, "Real-time 6D lidar slam in large scale natural terrains for ugv," in *IV*, 2018.

[17] A. Angeli, D. Filliat, S. Doncieux, and J.-A. Meyer, "A fast and incremental method for loop-closure detection using bags of visual words," *TRO*, pp. 1027–1037, 2008.

[18] D. Gálvez-López and J. D. Tardos, "Bags of binary words for fast place recognition in image sequences," *TRO*, vol. 28, no. 5, pp. 1188–1197, 2012.

[19] R. Mur-Artal and J. D. Tardós, "Fast relocalisation and loop closing in keyframe-based slam," in *ICRA*, 2014.

[20] A. Oliva and A. Torralba, "Building the gist of a scene: The role of global image features in recognition," *Progress in Brain Research*, vol. 155, pp. 23–36, 2006.

[21] N. Sünderhauf and P. Protzel, "Brief-gist-closing the loop by simple means," in *IROS*, 2011.

[22] R. Arroyo, P. F. Alcantarilla, L. M. Bergasa, and E. Romera, "Towards life-long visual localization using an efficient matching of binary sequences from images," in *ICRA*, 2015.

[23] 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," *RSS*, 2015.

[24] N. Sünderhauf, S. Shirazi, F. Dayoub, B. Upcroft, and M. Milford, "On the performance of convnet features for place recognition," in *IROS*.

[25] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *ACM MM*, 2014, pp. 675–678.

[26] E. Pepperell, P. I. Corke, and M. J. Milford, "All-environment visual place recognition with smart," in *ICRA*, 2014.

[27] T. Naseer, L. Spinello, W. Burgard, and C. Stachniss, "Robust visual robot localization across seasons using network flows," in *IJCAI*, 2014.

[28] M. J. Milford and G. F. Wyeth, "Seqslam: Visual route-based navigation for sunny summer days and stormy winter nights," in *ICRA*, 2012.

[29] Z. Chen, F. Maffra, I. Sa, and M. Chli, "Only look once, mining distinctive landmarks from convnet for visual place recognition," in *IROS*, 2017.

[30] B. Steder, G. Grisetti, and W. Burgard, "Robust place recognition for 3D range data based on point features," in *ICRA*, 2010.

[31] T. Cieslewski, E. Stumm, A. Gawel, M. Bosse, S. Lynen, and R. Siegwart, "Point cloud descriptors for place recognition using sparse visual information," in *ICRA*, 2016.

[32] T. Röhling, J. Mack, and D. Schulz, "A fast histogram-based similarity measure for detecting loop closures in 3-d lidar data," in *IROS*, 2015.

[33] E. Fazl-Ersi and J. K. Tsotsos, "Histogram of oriented uniform patterns for robust place recognition and categorization," in *IJRR*, 2012.

[34] R. Zlot and M. Bosse, "Place recognition using keypoint similarities in 2d lidar maps," in *Experimental Robotics*, 2009.

[35] R. Dube, D. Dugas, E. Stumm, J. Nieto, R. Siegwart, and C. Cadena, "Segmatch: Segment based place recognition in 3D point clouds," in *ICRA*, 2017.

[36] H. Yin, X. Ding, L. Tang, Y. Wang, and R. Xiong, "Efficient 3D lidar based loop closing using deep neural network," in *ICRB*, 2017.

[37] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3D classification and segmentation," in *CVPR*, 2017.

[38] G. Elbaz, T. Avraham, and A. Fischer, "3D point cloud registration for localization using a deep neural network auto-encoder," in *CVPR*, 2017.

[39] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, and J. Xiao, "3D shapenets: A deep representation for volumetric shapes," in *CVPR*, 2015.

[40] P. Wohlhart and V. Lepetit, "Learning descriptors for object recognition and 3d pose estimation," in *CVPR*, 2015.

[41] S. Shah, D. Dey, C. Lovett, and A. Kapoor, "Airsim: High-fidelity visual and physical simulation for autonomous vehicles," in *FSR*, 2017.

[42] N. Carlevaris-Bianco, A. K. Ushani, and R. M. Eustice, "University of Michigan North Campus long-term vision and lidar dataset," *IJRR*, 2015.