Long-term Place Recognition through Worst-case Graph Matching to Integrate Landmark Appearances and Spatial Relationships

Peng Gao and Hao Zhang

***Abstract*— Place recognition is an important component for simultaneously localization and mapping in a variety of robotics applications. Recently, several approaches using landmark information to represent a place showed promising performance to address long-term environment changes. However, previous approaches do not explicitly consider changes of the landmarks, i,e., old landmarks may disappear and new ones often appear over time. In addition, representations used in these approaches to represent landmarks are limited, based upon visual or spatial cues only. In this paper, we introduce a novel worst-case graph matching approach that integrates spatial relationships of landmarks with their appearances for long-term place recognition. Our method designs a graph representation to encode distance and angular spatial relationships as well as visual appearances of landmarks in order to represent a place. Then, we formulate place recognition as a graph matching problem under the worst-case scenario. Our approach matches places by computing the similarities of distance and angular spatial relationships of the landmarks that have the least similar appearances (i.e., worst-case). If the worst appearance similarity of landmarks is small, two places are identified to be not the same, even though their graph representations have high spatial relationship similarities. We evaluate our approach over two public benchmark datasets for long-term place recognition, including St. Lucia and CMU-VL. The experimental results have validated that our approach obtains the state-of-the-art place recognition performance, with a changing number of landmarks.**

I. Introduction

Place recognition (also known as loop closure detection) is a fundamental component in visual simultaneous localization and mapping (SLAM) [1]–[3], which has been a very active research area over the past decades [4]–[6]. The purpose of place recognition is to determine whether the current visiting place has been visited before by a robot. The matched places can be employed to reduce ambiguity and accumulated errors [7]–[9] during SLAM to significantly improve the accuracy of mapping and localization, which has be widely used in a variety of robotics applications [10]–[12].

More recently, motivated by long-term autonomy applications [13]–[15], long-term place recognition has become a rapidly growing research area to perform visual SLAM over long periods of time. The goal of long-term place recognition is to identify previously visited places during long-term robot operations at different times of the day, months, and seasons. For instance, when an autonomous vehicle visits a same place over different seasons, the same place can look significantly different caused by the variations of illumination (e.g., noon versus midnight), weather (rain versus snow), and vegetation (with leaves versus without leaves).

This work was partially supported by IIS-1942056, IIS-1849348, IIS-1849359, DOT PHMSA 693JK31850005CAAP, and DOE DE-FE0031650.

Peng Gao and Hao Zhang are with the Human-Centered Robotics Lab in the Department of Computer Science, Colorado School of Mines, Golden, CO 80401. {gaopeng, hzhang}@mines.edu

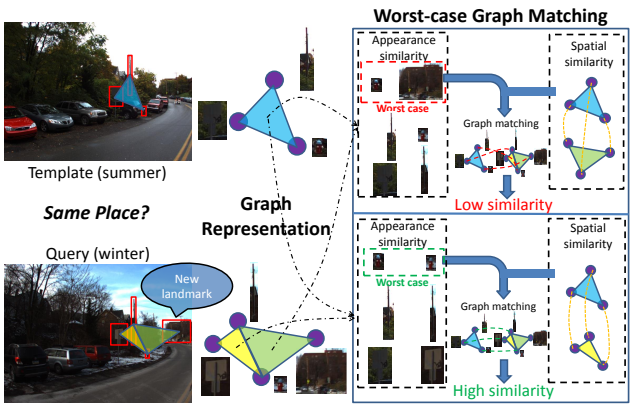


Fig. 1: Illustration of the proposed worst-case graph matching approach for long-term place recognition with newly appearing landmarks. Given an image with detected landmarks, our approach constructs a graph representation that encodes visual appearances, distance relationships, and angular relationships of the landmarks in order to represent the place. Then, our approach formulates place recognition as a graph matching problem under the worst-case scenario. It matches places by computing the similarities of distance and angular spatial relationships of the landmarks with the least similar appearances (i.e., worst-case).

Due to the importance of long-term place recognition, it has been intensively investigated [16]–[18]. Conventionally, many approaches use visual appearances of the environment to represent and match places, e.g., based on key-point features [19], region-based features [20], [21] or representative features [22]–[24]. Recently, several approaches are proposed to use landmark-based representations for place recognition, which show performance improvements and are more robust to long-term variations [25], [26]. However, the challenge of integrating both spatial relationships and appearance cues of landmarks, and the challenge caused by newly appearing or disappearing landmarks have not been well addressed yet.

In this paper, we propose a novel worst-case graph matching approach for place recognition in long-term autonomy, as demonstrated in Figure 1. Given a template or query image with detected landmarks, we generate a graph representation that simultaneously encodes the visual appearances, distance relationships, and angular relationships of the landmarks. The distance relationship is calculated as the distance between a pair of landmarks, and the angular relationship is represented

by the angles in a triangle constructed using three landmarks, which is robust to landmark scale changes. Given graph representations of places built from the query and template images, we formulate place recognition as a worst-case graph matching problem, with the goal of addressing appearing and disappearing landmarks. During long-term robot operations, landmarks within the environment can be added or removed, which means there always exists landmarks only existing in either the query or template image. For example, a building is newly constructed and a stop sign is newly removed. Our approach matches two places by computing the similarities of distance and angular spatial relationships of the landmarks that exhibit the least similar appearances (i.e., the worst-case scenario). If the worst appearance similarity of the landmarks is small, these two places are determined not as a match, even though the graph representations of the two places have high spatial relationship similarities.

The main novelty of this paper focuses on the proposal of the worst-case graph matching approach that integrates both landmark appearances and spatial relationships. Specifically, we design a unified graph representation that simultaneously encodes landmarks' appearance cues as well as distance and angular relationships, which improves the expressiveness of the representation to encode places. Second, we introduce a novel formulation of long-term place recognition as a worst-case graph matching problem, which addresses the challenge caused by newly appearing and disappearing landmarks, and is able to compute the matching score directly from the graph representations of the query and template images, instead of requiring a separate matching procedure as in most existing methods using vector-based place representations.

II. Related Work

In this section, we briefly review existing methods on landmark representations based on visual and spatial information, as well as matching paradigms for place recognition.

A. Representations for Place Recognition

For long-term place recognition, it is essential to construct a robust representation for places with challenges caused by environment variances during long periods of time [17]. We divide the existing methods into two major categories, based on visual feature of holistic environments, or based on visual or spatial information of semantic landmarks.

For representations of holistic environments, using local features were shown less effective to represent long-term place changes [20]. Thus, region-based methods using global features, such as GIST [21], HOG [20], and CNN [18], are proposed to encode the holistic environment that is observed by a robot. Based on the region-based representation, several approaches integrated multiple types of features to represent places [22], [27].

The other category of approaches using semantic landmarks have promising performance for place recognition with long-term appearance variance. [28] combined multiple local feature to generate a CNN description. Similarly, other CNN-based features [18] were used to encode visual features of landmarks detected from proposal generation [26], Edge box [29] and bounding box obtained from YOLO v2 [30]. Several representation learning-based methods were implemented to encode cues of visual landmarks into the holistic environment [31], [32].

The spatial relationship between landmarks can also be utilized for place recognition. [26] used CNN technique to generate landmark distribution descriptor to address environment and view changes. [33] introduced landmark geometry information obtained from the laser scan to visual SLAM. [25] stacked landmark features into a horizontal position in order to construct a feature descriptor to encode the spatial information of landmarks.

Most existing methods only used visual feature or simple spatial relationships of the landmarks and did not considered high order spatial relationships between the landmarks. In this paper, the proposed approach can explicitly encode visual feature and various spatial relationships of the landmarks.

B. Matching Paradigms for Place Recognition

Given the representation of places, a matching procedure is required to compute the matching score between a query observation and templates to identify the previously visited places for place recognition.

Existing matching methods can be broadly classified into two major categories; one is image-based matching and the other is sequence-based matching. For image-based matching methods, generally a similarity score between query and template image need to be calculated, and the similarity function can use the Euclidian or cosine distance [34], [35]. Another strategy utilized in image matching is based on the nearest neighbor search, including KD trees [36] and Chow Liu trees [37]. For sequence-based matching methods, the procedure of matching places is typically based on a sequence of consecutive images, instead of individual images [38]. Given the obtained vector-based representation of places, consecutive pairwise matching [39], minimizing cost flow [40], Hidden Markov Models [41], and Conditional Random Fields [42] can be used for sequence-based matching.

Our proposed worst-case graph matching method for long-term place recognition can integrate place representation and matching as a unified problem, which is different from the previous vector-based matching methods. And also we can address the challenge caused by newly appearing landmarks and spatial deformation through integrating spatial relationships and the worst appearance of landmarks.

III. Approach

In this section, we discuss the proposed principled method for worst-case graph matching that fuses the spatial relationships of landmarks with appearance cues. We also introduce the solver to address the formulated non-convex optimization problem for worst-case graph matching.

Notation. We represent matrix and tensor (i.e., 3D matrix) by bold capital letters, e.g., $\mathbf{M} = \{m_{ij}\} \in \mathbb{R}^{n \times n'}$ and $\mathbf{T} = \{t_{ijk}\} \in \mathbb{R}^{n \times n' \times n''}$, respectively. Vectors are represented

by bold lowercase letters. Furthermore, we represent the vectorized form of the matrix $\mathbf{M} \in \mathbb{R}^{n \times n'}$ using $\mathbf{m} \in \mathbb{R}^{nn'}$ that is a concatenation of the columns of $\mathbf{M}$ into a vector.

A. Problem Formulation

Given an input image, we extract landmarks to generate a graph representation, which encodes the spatial relationships of the landmarks to represent a place. Assume n landmarks are detected from the input image. Then, the positions of the landmarks in the image space are represented by the node set $\mathcal{P} = \{\mathbf{p}_1, \mathbf{p}_2, \cdots, \mathbf{p}_n\}$, where $\mathbf{p}_i = [x, y]$ denotes the central position of the i-th landmark in the image at coordinate $[x, y]$. Given the position information, we construct the spatial relationships of landmarks, which are divided into two categories, including distance spatial relationship and angular spatial relationship. The distance spatial relationships are represented by a distance set $\mathcal{E} = \{e_{i,j}\}$, where $e_{i,j}$ denotes the distance of an edge constructed by nodes $\mathbf{p}_i$ and $\mathbf{p}_j$. The angular spatial relationships are denoted by a angular set $\mathcal{T} = \{t_{i,j,k}\}$, where $t_{i,j,k} = [\theta_i, \theta_j, \theta_k], i, j, k = 1, 2, \ldots, n, i \neq j \neq k$ represents the three angles of a triangle constructed by nodes $\mathbf{p}_i$, $\mathbf{p}_j$ and $\mathbf{p}_k$. The angular relationship is robust to scale change, since angles of a triangle is invariant to scale change. Given the node set, distance set, and angular set, we can represent an input image as a graph $\mathcal{G} = (\mathcal{P}, \mathcal{T}, \mathcal{E})$.

For place recognition, given one query image and one template image, from which two graph representations $\mathcal{G} = (\mathcal{P}, \mathcal{T}, \mathcal{E})$ and $\mathcal{G}' = (\mathcal{P}', \mathcal{T}', \mathcal{E}')$ can be generated. The affinity between these two graphs can be computed by the sum up of the affinity of the distance sets ($\mathcal{E}$, $\mathcal{E}'$) and angular sets ($\mathcal{T}$, $\mathcal{T}'$).

The distance affinity $d_{ii',jj'}$ between two distances $e_{i,j}$ and $e_{i',j'}$ is generally calculated by $d_{ii',jj'} = \exp(-|e_{i,j} - e_{i',j'}|)$. The angular affinity $a_{ii',jj',kk'}$ between angles of two triangles $t_{i,j,k}$ and $t_{i',j',k'}$ can be calculated by $a_{ii',jj',kk'} = \exp(-\sum_{u,v} |\cos(\theta_u) - \cos(\theta_v)|)$, where $u \in \{i, j, k\}$ and $v \in \{i', j', k'\}$. By taking advantages of nonlinear projection function exp, the calculated affinities can be normalized to $[0, 1]$. Then, we generate the distance affinity matrix $\mathbf{D} = \{d_{ii',jj'}\} \in \mathbb{R}^{nn' \times nn'}$ and angular affinity tensor $\mathbf{A} = \{a_{ii',jj',kk'}\} \in \mathbb{R}^{nn' \times nn' \times nn'}$ given two graph representations $\mathcal{G}$ and $\mathcal{G}'$ generated from one query image and one template image.

Given the affinity matrix $\mathbf{D}$ and $\mathbf{A}$, generally, we can formulate a graph matching for place recognition as following:

$$\begin{aligned} \arg\max_{\mathbf{W}} \lambda_1 \sum_{ii'=1}^{nn'} \sum_{jj'=1}^{nn'} \sum_{kk'=1}^{nn'} a_{ii',jj',kk'} w_{ii'} w_{jj'} w_{kk'} \\ + \lambda_2 \sum_{ii'}^{nn'} \sum_{jj'}^{nn'} d_{ii',jj'} w_{ii'} w_{jj'} \qquad (1) \\ \text{s.t.} \quad \mathbf{W}\mathbf{1}_{n' \times 1} \leq \mathbf{1}_{n \times 1}, \mathbf{W}^\top \mathbf{1}_{n \times 1} \leq \mathbf{1}_{n' \times 1} \qquad (2) \end{aligned}$$

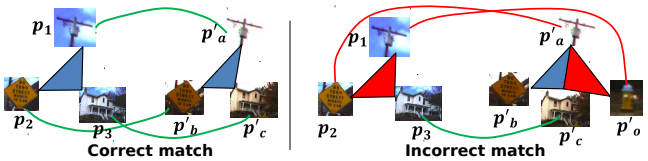


Fig. 2: Illustration of incorrect matches caused by the newly appearing landmark $\mathbf{p}'_o$ and deformation. *Left Figure*: The correct match should be $(\mathbf{p}_1 \leftrightarrow \mathbf{p}'_a)$, $(\mathbf{p}_2 \leftrightarrow \mathbf{p}'_b)$ and $(\mathbf{p}_3 \leftrightarrow \mathbf{p}'_c)$, even when the triangles t_{123} and t'_{abc} do not look the same because of landmark deformation. *Right Figure*: When a newly appearing landmark exists, represented by node $\mathbf{p}'_o$, the distance and angular affinities of subgraph constructed by $\{\mathbf{p}_1, \mathbf{p}_2, \mathbf{p}_3\}$ and $\{\mathbf{p}'_a, \mathbf{p}'_b, \mathbf{p}'_c\}$ can be smaller than the spatial affinity between subgraphs constructed by $\{\mathbf{p}_1, \mathbf{p}_2, \mathbf{p}_3\}$ and $\{\mathbf{p}'_a, \mathbf{p}'_c, \mathbf{p}'_o\}$, which results in an incorrect match $(\mathbf{p}_2 \leftrightarrow \mathbf{p}'_a)$ and $(\mathbf{p}_1 \leftrightarrow \mathbf{p}'_o)$, denoted by red lines.

We can rewrite Eq. (1) into a matrix form as following:

$$\begin{aligned} \arg\max_{\mathbf{W}} \lambda_1 \mathbf{A} \otimes_1 \mathbf{w} \otimes_2 \mathbf{w} \otimes_3 \mathbf{w} + \lambda_2 \mathbf{w}^\top \mathbf{D} \mathbf{w} \\ \text{s.t.} \quad \mathbf{W}\mathbf{1}_{n' \times 1} \leq \mathbf{1}_{n \times 1}, \mathbf{W}^\top \mathbf{1}_{n \times 1} \leq \mathbf{1}_{n' \times 1} \qquad (3) \end{aligned}$$

where $\mathbf{w} \in \mathbb{R}^{nn'}$ is the vectorized form of correspondence matrix $\mathbf{W} = \{w_{ii'}\} \in \{0, 1\}^{n \times n'}$, with $w_{ii'} = 1$ denoting the i-th node in $\mathcal{P}$ and the i'-th node in $\mathcal{P}'$ are matched, $\otimes$ is a tensor product, $\otimes_l, l = 1, 2, 3$ means multiplication between $\mathbf{w}$ and the l-th order matricization of $\mathbf{A}$ [43] and $\mathbf{1}$ is an all-ones vector. In Eq. (3), the first term denotes the accumulation of the angular similarities given the correspondence matrix $\mathbf{W}$, which is controlled by hyperparamter λ_1. Similarly, the second term represents the accumulation of distance similarities, which is controlled by λ_2. The constraint is used to enforce the one-to-one correspondence for $\mathbf{W}$, e.g. one landmark within $\mathcal{G}$ can at most have one corresponding landmark in $\mathcal{G}'$.

B. Worst-case Graph Matching

Due to long-term environmental changes, the spatial relationship of landmarks often has deformation caused by view changes (the field of view of a robot has deviation when it observes the same place) which will hurt the matching accuracy for long-term place recognition. Besides the challenge caused by spatial deformation, some landmarks will newly appear or disappear in the query and template images, like a building is occluded by trees in summer but can be seen in winter when trees have no leaves. The newly appearing landmarks will introduce useless spatial relationships which can be harmful to the matching accuracy, especially when there exists spatial deformation, as illustrated in Figure. 2.

Given these challenges, we introduce appearance cues into spatial relationships to improve the expressiveness and also propose a principled *worst-case graph matching* approach which can maximize distance and angular spatial similarities of landmarks that have least similar appearance (i.e., the worst-case scenario), in order to address the challenges

introduced by newly appearing landmarks and spatial deformation. For example, in the left figure in Figure 2, due to the challenges (the triangle constructed by nodes $\mathbf{p}'_a, \mathbf{p}'_c, \mathbf{p}'_o$), introduced by the newly appearing landmark $\mathbf{p}'_o$ and spatial deformation, the matching result will be incorrect. In this situation, the worst case can be represented by the worst similarity between a pair of landmark appearances given correspondences, e.g. the appearance similarity between the corresponding nodes $\mathbf{p}_1$ and $\mathbf{p}'_o$, which is a small value. By multiplying this worst appearance similarity, the final similarity score calculated between subgraph $\{\mathbf{p}_1, \mathbf{p}_2, \mathbf{p}_3\}$ and $\{\mathbf{p}'_b, \mathbf{p}'_c, \mathbf{p}'_o\}$ will be weakened. In other word, our proposed worst-case graph matching method maximizes the overall similarity under the worst case.

Formally, for each node $\mathbf{p}_i$ in graph $\mathcal{G}$, the appearance of its associated landmark is described as a feature vector $\mathbf{f} \in \mathbb{R}^d$, where d is the length of the feature vector. The feature can describe the shape, texture or color of the landmark. Thus, the appearance set can be denoted as $\mathcal{F} = \{(\mathbf{f}_1)^\top, (\mathbf{f}_2)^\top, \cdots, (\mathbf{f}_n)^\top\}^{n\times d}$. Thus, the appearance affinity matrix $\mathbf{Z} = \{z_{ii'}\} \in \mathbb{R}^{n\times n'}$ can be computed through $z_{ii'} = ||\mathbf{f}_i - \mathbf{f}_j||_2$. And place recognition can be formulated as the following worst-case graph matching problem:

$$\begin{aligned}\arg\max_{\mathbf{W}}\ & \lambda_1 \sum_{ii'=1}^{nn'}\sum_{jj'=1}^{nn'}\sum_{kk'=1}^{nn'} a_{ii',jj',kk'} \min\{z_{ii'}, z_{jj'}, z_{kk'}\} \\ & w_{ii'}w_{jj'}w_{kk'} \\ & +\lambda_2 \sum_{ii'}^{nn'}\sum_{jj'}^{nn'} d_{ii',jj'} \min\{z_{ii'}, z_{jj'}\} w_{ii'}w_{jj'} \\ \text{s.t.}\ & \mathbf{W}\mathbf{1}_{n'\times 1} \le \mathbf{1}_{n\times 1}, \mathbf{W}^\top \mathbf{1}_{n\times 1} \le \mathbf{1}_{n'\times 1} \end{aligned} \tag{4}$$

where $\min$ function offers the worst case during matching.

To solve this formulated optimization problem for worst-case graph matching, we implement an iterative optimization algorithm as presented in Algorithm 1. The complexity of our algorithm is $O(n^4)$. Details of the algorithm are provided in the supplementary material[1].

After solving the optimization problem utilizing Algorithm 1, we are able to obtain the optimal correspondence matrix $\mathbf{W}^* = \{w^*_{ij}\} \in \mathbb{R}^{n\times n'}$, which describes the correspondences of the landmarks in the query and template images.

C. Place Recognition

Given the correspondence matrix $\mathbf{W}^*$, we can directly compute the matching score between the query and template image as following

$$\begin{aligned} S = & \lambda_1 \sum_{ii'=1}^{nn'}\sum_{jj'=1}^{nn'}\sum_{kk'=1}^{nn'} a_{ii',jj',kk'} \min\{z_{ii'}, z_{jj'}, z_{kk'}\} \\ & w^*_{ii'}w^*_{jj'}w^*_{kk'} \\ & +\lambda_2 \sum_{ii'}^{nn'}\sum_{jj'}^{nn'} d_{ii',jj'} \min\{z_{ii'}, z_{jj'}\} w^*_{ii'}w^*_{jj'} \end{aligned} \tag{5}$$

[1] http://hcr.mines.edu/publication/ICRA20_GraphPR_Supp.pdf

Algorithm 1: The algorithm to solve the formulated non-convex optimization problem in Eq. (4).

Input : $\mathbf{A} \in \mathbb{R}^{nn'\times nn'\times nn'}$, $\mathbf{D} \in \mathbb{R}^{nn'\times nn'}$ and $\mathbf{Z} \in \mathbb{R}^{nn'}$
Output : $\mathbf{W}^* = \in \{0,1\}^{n\times n'}$

Initialize the vectorized matrix $\mathbf{w} \in \{0,1\}^{nn'}$;
Compute stochastic matrix $\mathbf{K} = \{k_{ii',jj',kk'}\}$ and $\mathbf{L} = \{l_{ii',jj'}\}$:
$k_{ii',jj',kk'} = a_{ii',jj',kk'} \min\{z_{ii'}, z_{jj'}, z_{kk'}\} / \max(\mathbf{A})$;
$l_{ii',jj'} = d_{ii',jj} \min\{z_{ii'}, z_{jj'}\} / \max(\mathbf{D})$;
while *not converge* **do**
  Compute the jump vector $\mathbf{j} = \exp(\mathbf{w}^r / \max(\mathbf{w}^r))$;
  Normalize $\mathbf{j}$ using the bistochastic normalization;
  Update $\mathbf{w}^{r+1} = \gamma(\mathbf{K} \otimes_2 \mathbf{w}^r \otimes_3 \mathbf{w}^r + \mathbf{w}^{r\top}\mathbf{L}) + (1-\gamma)\mathbf{j}$;
end
Recover $\mathbf{w}$ to $\mathbf{W}$;
Use greedy search to discretize $\mathbf{W}$ [44];
return $\mathbf{W}$

However, the number of nodes of input graph representation is always different and the matching score calculated in Eq. (5) is proportional to the number of nodes. In order to generalize our proposed method to the input graphs with different nodes, we calculate the final matching score as following:

$$\begin{aligned} S = & \frac{\lambda_1}{m^3} \sum_{ii'=1}^{nn'}\sum_{jj'=1}^{nn'}\sum_{kk'=1}^{nn'} a_{ii',jj',kk'} \min\{z_{ii'}, z_{jj'}, z_{kk'}\} \\ & w^*_{ii'}w^*_{jj'}w^*_{kk'} \\ & +\frac{\lambda_2}{m^2} \sum_{ii'}^{nn'}\sum_{jj'}^{nn'} d_{ii',jj'} \min\{z_{ii'}, z_{jj'}\} w^*_{ii'}w^*_{jj'} \end{aligned} \tag{6}$$

where $m = \min\{n, n'\}$. Due to the existence of newly appearing/disappearing landmarks, n and n' can be different. The number of matched landmarks between query and template image is dominated by the smallest number of landmarks in either query or template image. Given the optimal solution $\mathbf{W}^*$, there are m^3 angular similarities and m^2 distance similarities accumulated to the final score. Since the spatial similarity $a_{ii',jj',kk'}$ and $d_{ii',jj'}$ are between $[0,1]$, the final matching score calculated in Eq. (6) is divided by its upper bound of each term to normalize the score always between $[0,1]$. Then, if two places are matched is determined by comparing the normalized matching score with a manually set threshold. By applying Eq. (6) to obtain the normalized similarity score, our worst-case graph matching approach can compute a matching score directly from graph representations of the query and template images.

IV. Experimental Results

A. Experiment Setup

We utilize two large-scale benchmarks to evaluate our proposed method for long-term place recognition, including CMU-VL dataset and St. Lucia dataset. And the evaluation metric is the precision-recall curve that demonstrates the

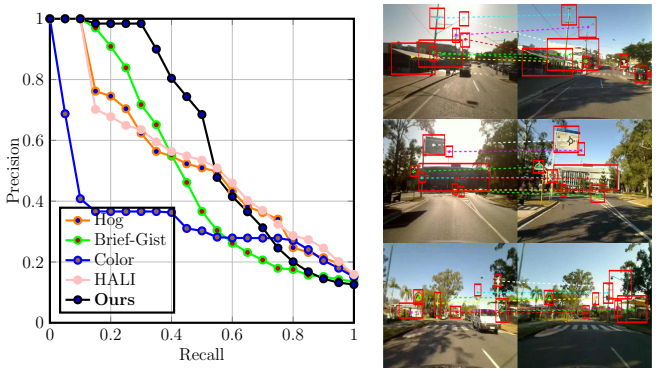
TABLE I: Description of the two public benchmark datasets for long-term place recognition.

Dataset	St. Lucia [45]	CMU-VL [46]
Scenario	Different times of a day	Different months of a year
Statistic	$10 \sim 22,000$ frames 640×480 at 15 FPS	$5 \sim 13,000$ frames 1024×768 at 15 FPS
Description	Illumination, shadow, dynamic variations	Vegetation, weather, view, dynamic variations

trade-off impact between precision and recall with variant matching threshold. Precision means the fraction of retrieved locations that are relevant and recall means the fraction of retrieved locations to all relevant locations. The performance with high recall and high precision is represented by a curve with a large area under it.

In the experiment, following recently landmark-based method [31], [32] ,only static and stable landmarks are used to construct our proposed graph representation of a place, e.g., houses, traffic signs, antenna, etc. For the appearance feature of each landmark, we use histogram of oriented gradient (HOG) feature to describe the appearance of landmarks.

We compare our proposed worst-case graph matching method with four long-term place recognition methods, which includes three appearance-based methods: **Color** that uses color feature of downsampled images [47], **HOG** that uses histogram of oriented gradient feature of downsampled images to describe the shape of landmarks [20] and **Brief-Gist** that uses brief-gist feature of downsampled images [48], and one landmark-based method: **HALI** that learns the projection from semantic landmarks to vector-based features for long-term place recognition [32].



(a) Precision-recall curves (b) Matched places

Fig. 3: Experimental results on the St. Lucia dataset. Figure 3(a) demonstrates the quantitative results based on the precision-recall curve. Figure 3(b) presents qualitative results of matched places between the query and template images, which are recorded at 3:00 PM (left) and 8:00 AM (right), respectively. The figures are best viewed in color.

B. Results on the St. Lucia Dataset

The St Lucia dataset [45] is gathered in a 9.5km circuit in Australia at different times of several days. The visual data are collected through a calibrated stereo camera mounted on a car and each video instance has 20-25 minutes. The ground truth information is offered by a GPS for experimental evaluation of place recognition. There are various scenarios included in the dataset, which contains different challenges for long-term place recognition, including the variation of illumination at various times over a day, perspective change caused by road bumps, highly dynamic objects in the street. The detail of the dataset is shown in Table I.

TABLE II: The experimental results over the St. Lucia and CMU-VL datasets. The value in $[0, 1]$ describes the area ratio under the precision and recall curve. A larger value indicates a better performance.

Approach	St Lucia Dataset	CMU-VL Dataset
Color [47]	0.3186	0.3947
Hog [20]	0.5517	0.5396
Gist-Brief [48]	0.5206	0.5530
HALI [32]	0.5569	0.5430
Ours	**0.6249**	**0.6274**

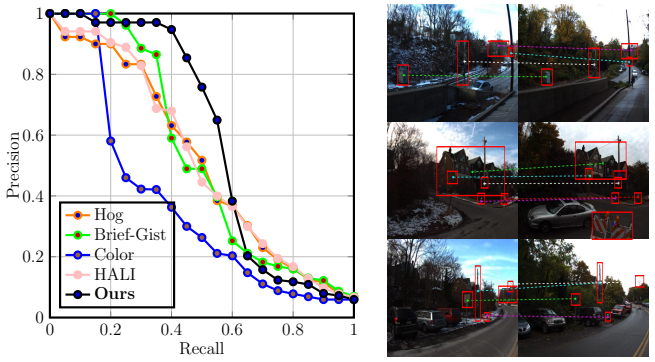
The quantitative results obtained from our method and the other baselines are demonstrated in Figure3 (a) based on the precision-recall curve. From the results, we can observe that our proposed method outperforms the visual appearance-based and landmark-based methods. To further evaluate the precision-recall curves, we calculate the ratio between the area under each curve and the whole precision-recall area space. Thus, the range of the area ratio is between $[0, 1]$ and a higher value represents better performance. The results are listed in Table II, which demonstrate that our score is 0.6249, and our method outperforms the other methods. The improvement obtained from our method is caused by our representation of spatial information that is robust to spatial deformation and also by our worst-case graph matching which is robust to objects only appeared in one image.

The qualitative results obtained from our method are demonstrated in Figure 3(b), which include three pairs of matched places between query and template images in St. Lucia dataset. Based on the results, we can see that the illumination of the same place changes a lot and the number of detected landmarks are different between query and template images. Thus, we can observe that our methods can well address the long-term variations and identify the correct matches based on the visual and spatial information of landmarks.

C. Results on the CMU-VL Dataset

The CMU Visual Localization (CMU-VL) dataset [46] is recorded on an 8.8km route in urban areas over different months of a whole year. The visual images are collected by two cameras oriented to left and right separately. The ground truth for place recognition evaluation is gathered through a GPS. The challenges in this dataset are on environmental conditions, like the variations of vegetation (green and fall leaves), weather (snow, cloudy and sunny), which make the dataset very challenging.

The quantitative results are shown in Figure 4(a). We can observe that our method outperforms the other state-of-the art methods under the precision-recall evaluation metric. We



(a) Precision-recall curves (b) Matched places

Fig. 4: Experimental results over the CMU-VL dataset. Figure 4(a) demonstrates the quantitative results on precision-recall curves. Figure 4(b) presents qualitative results of place matches between the query and template images, which are recorded in December (left) and October (right), respectively. The figures are best viewed in color.

obtained the largest area ratio of 0.6274 listed in Table II. The qualitative results are presented in Figure 4(b), which contain three pairs of matched places recorded in October and December separately. The environment condition changes a lot from query image to template image caused by the variation of vegetation, weather and dynamic objects. And also, some landmarks are disappeared due to the occlusion of trees and some new landmarks are added. Our method can still well address these challenges by correctly find the matches of landmarks between query and template images. Because of the correct matches of landmarks, the spatial relationship of landmarks can be well preserved so that we can obtain a high similarity.

D. Discussion

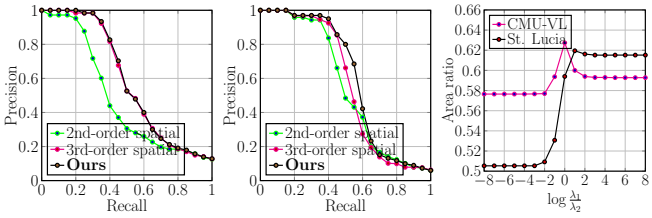
Without losing generality, the importance of different spatial relationships and the main hyperparameters of our proposed approach will be studied on both of the datasets.

1) The Importance of Different Spatial Relationships: The performance on St. Lucia dataset obtained from partial and complete of our approach that uses different spatial relationships are demonstrated in Figure 5(a). We can see that the angular relationship is more important than the distance relationship. If we only use distance relationship, based on the area under the precision-recall curve, we obtain the score of 0.5054 and the score obtained from only utilizing angular relationships is 0.6152. The combination of distance and angular relationships can achieve a score of 0.6249 which is slightly higher than the score obtained from using angular relationships.

On dataset CMU-VL, we similarly evaluate the importance of each spatial relationship and the results are presented in Figure 5(b). We can also see that the angular relationship is more important than the distance relationship for long-term place recognition. Using distance relationships obtains a score of 0.5765 and using angular relationships gets a score of 0.5928. Combining them obtains a score of 0.6274. We can conclude that the angular relationship is more important than the distance relationship. Since angles of a triangle are invariant to scale change, and angular representation is more robust to spatial deformation.

2) Hyperparameter Analysis: The cross-validation results on both datasets obtained from our method are shown in Figure 5(c). We analyze the performance variation of our method with different hyperparameter λ_1 and λ_2.

Similar to the quantitative analysis, we use the ratio between the area under the precision-recall curve and whole area space as the single value evaluation metric. Since the final matching score is influenced by the ratio between λ_1 and λ_2, we evaluate their ratio in range $[10^{-8}, 10^{8}]$. Based on the results shown in Figure 5(c), the peak of the curve on St. Lucia dataset is when $\frac{\lambda_1}{\lambda_2} = 10$, which can obtain the best performance. And also the best performance can be obtained on CMU-VL dataset is when $\frac{\lambda_1}{\lambda_2} = 1$. In addition, we can observe that the performance is better when $\frac{\lambda_1}{\lambda_2}$ is larger and the performance is worse when the ratio is smaller. This phenomenon demonstrates that the angular relationship is more important than the distance relationship, which is consistent with our analysis in Figure 5(a) and (b).



(a) St. Lucia (b) CMU-VL (c) Hyperparameters

Fig. 5: Analysis of the proposed approach over both datasets. Figures 5(a) and (b) compare methods using different spatial relationships over the St. Lucia and CMU-VL datasets, respectively. Figure 5(c) presents the results of hyperparameter analysis, which shows performance changes of our approach given different hyperparameter ratios using both datasets.

V. Conclusion

We propose the novel worst-case graph matching approach that integrates spatial relationships of landmarks with appearance cues to perform long-term place recognition. Our approach employs graph representations to encode appearances and spatial relationships of landmarks in order to represent places. Then, our approach formulates place recognition as a worst-case graph matching problem, which maximizes the spatial similarity of the landmarks with the worst appearance similarity in order to address challenges caused by appearing and disappearing landmarks. In addition, the matching score of two places is directly computed by our approach without requiring further matching procedures. Experimental results on two public benchmark datasets have shown that our approach obtains promising long-term place recognition performance, with a changing number of landmarks.

REFERENCES

[1] 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.

[2] 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.

[3] J. Engel, T. Schöps, and D. Cremers, "LSD-SLAM: large-scale direct monocular SLAM," in *European conference on computer vision*, 2014.

[4] I. Ulrich and I. Nourbakhsh, "Appearance-based place recognition for topological localization," in *IEEE International Conference on Robotics and Automation*, 2000.

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

[6] S. Lowry and M. J. Milford, "Supervised and unsupervised linear learning techniques for visual place recognition in changing environments," *IEEE Transactions on Robotics*, vol. 32, no. 3, pp. 600–613, 2016.

[7] R. Mur-Artal and J. D. Tardós, "Visual-inertial monocular SLAM with map reuse," *IEEE Robotics and Automation Letters*, vol. 2, no. 2, pp. 796–803, 2017.

[8] C. Kerl, J. Sturm, and D. Cremers, "Dense visual SLAM for RGB-D cameras," in *IEEE International Conference on Intelligent Robots and Systems*, 2013.

[9] H. Zhou, D. Zou, L. Pei, R. Ying, P. Liu, and W. Yu, "StructSLAM: visual SLAM with building structure lines," *IEEE Transactions on Vehicular Technology*, vol. 64, no. 4, pp. 1364–1375, 2015.

[10] X. Gao and T. Zhang, "Unsupervised learning to detect loops using deep neural networks for visual SLAM system," *Autonomous robots*, vol. 41, no. 1, pp. 1–18, 2017.

[11] H. Strasdat, J. M. Montiel, and A. J. Davison, "Visual SLAM: Why filter?" *Image and Vision Computing*, vol. 30, no. 2, pp. 65–77, 2012.

[12] B. M. Kitt, J. Rehder, A. D. Chambers, M. Schonbein, H. Lategahn, and S. Singh, "Monocular visual odometry using a planar road model to solve scale ambiguity," 2011.

[13] P. Neubert, N. Sünderhauf, and P. Protzel, "Superpixel-based appearance change prediction for long-term navigation across seasons," *Robotics and Autonomous Systems*, vol. 69, pp. 15–27, 2015.

[14] S. Yang, S. A. Scherer, X. Yi, and A. Zell, "Multi-camera visual SLAM for autonomous navigation of micro aerial vehicles," *Robotics and Autonomous Systems*, vol. 93, pp. 116–134, 2017.

[15] H. Lategahn, A. Geiger, and B. Kitt, "Visual SLAM for autonomous ground vehicles," in *IEEE International Conference on Robotics and Automation*, 2011.

[16] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "NetVLAD: CNN architecture for weakly supervised place recognition," in *IEEE conference on computer vision and pattern recognition*, 2016.

[17] 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.

[18] N. Sünderhauf, S. Shirazi, F. Dayoub, B. Upcroft, and M. Milford, "On the performance of convnet features for place recognition," in *IEEE international conference on intelligent robots and systems*, 2015.

[19] F. Endres, J. Hess, N. Engelhard, J. Sturm, D. Cremers, and W. Burgard, "An evaluation of the RGB-D SLAM system." in *IEEE International Conference on Robotics and Automation*, 2012.

[20] T. Naseer, L. Spinello, W. Burgard, and C. Stachniss, "Robust visual robot localization across seasons using network flows," in *AAAI Conference on Artificial Intelligence*, 2014.

[21] Y. Latif, G. Huang, J. J. Leonard, and J. Neira, "An online sparsity-cognizant loop-closure algorithm for visual navigation." in *Robotics: Science and Systems*, 2014.

[22] F. Han, X. Yang, Y. Deng, M. Rentschler, D. Yang, and H. Zhang, "SRAL: shared representative appearance learning for long-term visual place recognition," *IEEE Robotics and Automation Letters*, vol. 2, no. 2, pp. 1172–1179, 2017.

[23] F. Han, S. El Beleidy, H. Wang, C. Ye, and H. Zhang, "Learning of holism-landmark graph embedding for place recognition in long-term autonomy," *IEEE Robotics and Automation Letters*, vol. 3, no. 4, pp. 3669–3676, 2018.

[24] S. Siva and H. Zhang, "Omnidirectional multisensory perception fusion for long-term place recognition," in *IEEE International Conference on Robotics and Automation*, 2018.

[25] P. Panphattarasap and A. Calway, "Visual place recognition using landmark distribution descriptors," in *Asian Conference on Computer Vision*, 2016.

[26] Z. Chen, F. Maffra, I. Sa, and M. Chli, "Only look once, mining distinctive landmarks from ConvNet for visual place recognition," in *IEEE International Conference on Intelligent Robots and Systems*, 2017.

[27] A. Pronobis, O. Martinez Mozos, B. Caputo, and P. Jensfelt, "Multi-modal semantic place classification," *The International Journal of Robotics Research*, vol. 29, no. 2-3, pp. 298–320, 2010.

[28] A. Babenko and V. Lempitsky, "Aggregating local deep features for image retrieval," in *IEEE international conference on computer vision*, 2015.

[29] C. L. Zitnick and P. Dollár, "Edge boxes: locating object proposals from edges," in *European conference on computer vision*, 2014.

[30] Y. Hou, H. Zhang, and S. Zhou, "Evaluation of object proposals and convnet features for landmark-based visual place recognition," *Journal of Intelligent and Robotic Systems*, vol. 92, no. 3-4, pp. 505–520, 2018.

[31] K. Liu, H. Wang, F. Han, and H. Zhang, "Visual place recognition via robust l2-norm distance based holism and landmark integration," in *AAAI Conference on Artificial Intelligence*, 2019.

[32] F. Han, H. Wang, G. Huang, and H. Zhang, "Sequence-based sparse optimization methods for long-term loop closure detection in visual SLAM," *Autonomous Robots*, vol. 42, no. 7, pp. 1323–1335, 2018.

[33] K. L. Ho and P. Newman, "Loop closure detection in SLAM by combining visual and spatial appearance," *Robotics and Autonomous Systems*, vol. 54, no. 9, pp. 740–749, 2006.

[34] P. Newman, D. Cole, and K. Ho, "Outdoor SLAM using visual appearance and laser ranging," in *IEEE International Conference on Robotics and Automation*, 2006.

[35] T. Naseer, M. Ruhnke, C. Stachniss, L. Spinello, and W. Burgard, "Robust visual SLAM across seasons," in *IEEE International Conference on Intelligent Robots and Systems*, 2015.

[36] M. J. Milford, G. F. Wyeth, and D. Prasser, "RatSLAM: a hippocampal model for simultaneous localization and mapping," in *IEEE International Conference on Robotics and Automation*, 2004.

[37] M. Cummins and P. Newman, "FAB-MAP: probabilistic localization and mapping in the space of appearance," *The International Journal of Robotics Research*, vol. 27, no. 6, pp. 647–665, 2008.

[38] H. Zhang, F. Han, and H. Wang, "Robust multimodal sequence-based loop closure detection via structured sparsity." in *Robotics: Science and systems*, 2016.

[39] E. Johns and G.-Z. Yang, "Feature co-occurrence maps: Appearance-based localisation throughout the day," in *IEEE International Conference on Robotics and Automation*, 2013.

[40] S. M. Siam and H. Zhang, "Fast-SeqSLAM: a fast appearance based place recognition algorithm," in *IEEE International Conference on Robotics and Automation*, 2017.

[41] P. Hansen and B. Browning, "Visual place recognition using HMM sequence matching," in *IEEE International Conference on Intelligent Robots and Systems*, 2014.

[42] C. Cadena, D. Gálvez-López, J. D. Tardós, and J. Neira, "Robust place recognition with stereo sequences," *IEEE Transactions on Robotics*, vol. 28, no. 4, pp. 871–885, 2012.

[43] S. Rabanser, O. Shchur, and S. Günnemann, "Introduction to tensor decompositions and their applications in machine learning," *Machine Learning*, vol. 98, no. 1-2, pp. 1–5, 2015.

[44] H. E. Romeijn and D. R. Morales, "A class of greedy algorithms for the generalized assignment problem," *Discrete Applied Mathematics*, vol. 103, no. 1-3, pp. 209–235, 2000.

[45] A. J. Glover, W. P. Maddern, M. J. Milford, and G. F. Wyeth, "FAB-MAP+ RatSLAM: Appearance-based SLAM for multiple times of day," in *IEEE International Conference on Robotics and Automation*, 2010.

[46] H. Badino, D. Huber, and T. Kanade, "Real-time topometric localization," in *IEEE International Conference on Robotics and Automation*, 2012.

[47] D. Lee, H. Kim, and H. Myung, "2D image feature-based real-time RGB-D 3D SLAM," in *Robot Intelligence Technology and Applications*, 2013.

[48] N. Sünderhauf and P. Protzel, "Brief-Gist-Closing the loop by simple means," in *IEEE International Conference on Intelligent Robots and Systems*, 2011.