Learning of Holism-Landmark Graph Embedding for Place Recognition in Long-Term Autonomy

Fei Han, Saad El Beleidy, Hua Wang, *Senior Member, IEEE*, Cang Ye, *Senior Member, IEEE*, and Hao Zhang, *Member, IEEE*

***Abstract*—Place recognition plays an important role to perform loop closure detection of large-scale, long-term simultaneous localization and mapping in loopy environments. The *long-term* place recognition problem is challenging because the environment appearance exhibits significant long-term variations across various times of the day, months, and seasons. In this letter, we introduce a novel place representation approach that simultaneously integrates semantic landmarks and holistic information to achieve place recognition in long-term autonomy. First, a graph is constructed for each place. The graph nodes encode all landmarks and the holistic image of the place scene recorded in different scenarios. The edges connecting the nodes indicate that these nodes represent the same landmark or place, even though places and landmarks encoded by the nodes may exhibit different appearances in the long-term periods. Then, a graph embedding is learned to preserve the locality in the feature descriptor space, i.e., finding a projection such that the same landmark and place have the identical representation in the new projected descriptor space, no matter in what scenarios they are recorded. We formulate the embedding learning as an optimization problem and implement a new solver that provides a theoretical convergence guarantee. Extensive evaluations are conducted using large-scale benchmark datasets of place recognition in long-term autonomy, which has shown our approach's promising performance.**

***Index Terms*—Visual learning, recognition, SLAM, localization.**

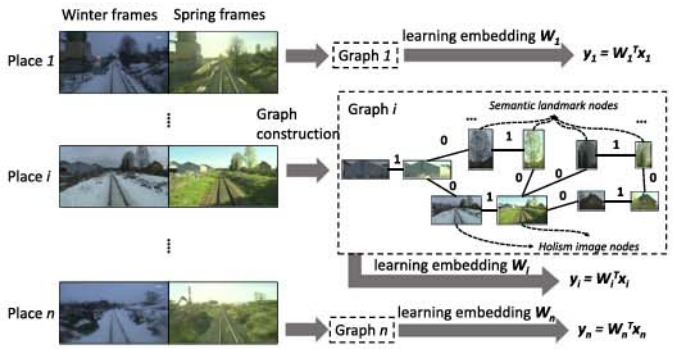

Fig. 1. Overview of the proposed HALGE approach to learn a representation for each place that simultaneously integrates semantic landmarks and holistic cues across various environmental scenarios for improved long-term place recognition. Each place recorded in multiple scenarios is constructed by a graph. The graph nodes represent landmarks and holism images in all scenarios. The edges denote relationships between nodes: nodes representing the same landmark or holism image have an edge with weight 1, otherwise 0. After constructing the graph, we learn an embedding of the graph, and obtain the projection matrix $\mathbf{W}_i$ for each place i, such that the same landmark and holism image have the identical representation in the projected space. During testing, we apply the learned projection to each holistic query image to compute its representation and find the best matching place.

I. Introduction

PLACE recognition attracted significant attention in robotics over the past decades because of its important role to perform loop closing for SLAM [1]–[3]. The purpose of place recognition is to identify whether the current query place is the same as one of the previously visited places; if so, which one it is.

More recently, motivated by the increasing interest in long-term autonomy within the robotics community and autonomous driving industry, *long-term* place recognition has become a rapidly growing research field, aiming at improving accuracy and reliability of outdoor localization during long-term operations [3]–[5].

Besides the problems in the conventional place recognition (e.g., viewpoint change and occlusion), it becomes even more challenging when operated in the long-term autonomy scenario. The same place can look drastically different in various times of the day, months and seasons, caused by changes of illumination (e.g., noon v.s. evening), vegetation conditions (e.g., green leaves versus fallen leaves), weather (e.g., covered by snow versus no snow), among other factors. In addition, multiple places may have the similar appearance, also known as the perceptual aliasing issue, which is another challenge that makes long-term place recognition difficult.

Due to its significance, there are many works that investigate the long-term place recognition problem [6]–[8]. A large group of previous methods perform scene matching based on representations extracted from local [9], [10] or holistic [11]–[13] cues of the scene. Local semantic landmarks in the scene are also an important indication to represent the place. Several methods using semantic landmarks were introduced recently [14], [15]. However, these landmark-based approaches cannot

Manuscript received February 24, 2018; accepted June 19, 2018. Date of publication July 16, 2018; date of current version August 2, 2018. This letter was recommended for publication by Associate Editor N. Hawes and Editor D. Lee upon evaluation of the reviewers' comments. This work was supported in part by Army Research Office under Grant W911NF-17-1-0447 and in part by National Science Foundation under Grants NSF-IIS 1423591 and NSF-IIS 1652943. *(Corresponding author: Hao Zhang.)*

F. Han, S. El Beleidy, H. Wang, and H. Zhang are with the Department of Computer Science, Colorado School of Mines, Golden, CO 80401 USA (e-mail: fhan@mines.edu; selbeleidy@mines.edu; huawangcs@gmail.com; hzhang@mines.edu).

C. Ye is with the Department of Computer Science, Virginia Commonwealth University, Richmond, VA 23284 USA (e-mail: cye@vcu.edu).

Digital Object Identifier 10.1109/LRA.2018.2856274

preserve the relationship between landmarks. Also, they cannot integrate both landmark and holistic scene information to construct a representation that is discriminative to identify different places while robust to long-term appearance changes; that is, previous techniques are only able to encode either holism cues or landmarks.

In this letter, we introduce a novel approach for representing each place as a graph, where its nodes represent all semantic landmarks in the scene recorded from different scenarios, and its edges represent the relationship between nodes, as shown in Fig. 1. Two nodes have a connecting edge only when they represent the same landmark. The holism image is treated as a special landmark of the scene and constructed in the same graph of the place. This enables our proposed approach to integrate both semantic landmarks and holistic cues for place representation. The final representation is then computed by embedding the constructed graph into the representation of the place. After formulating the embedding learning into a regularized optimization problem, we find a projection from the original feature space into a low-dimensional projected space such that the graph structure is preserved. That is, the locality is also preserved, i.e., the same landmark and place have the identical representation in the new projected space, no matter from which long-term scenario they are recorded. Given our approach's advantage of fusing holistic cues and landmarks to perform long-term place recognition, we name it *Holism-And-Landmark Graph Embedding* (HALGE).

The contribution of this letter is threefold:

- We propose the new HALGE approach to learn a representation that integrates both holistic cues and semantic landmarks to represent places in long-term autonomy, which is discriminative to identify different places while robust to long-term environmental changes.
- We propose to construct a graph representation for each place and learn its embedding through a novel optimization formulation that preserves the graph structure. We also develop an efficient iterative algorithm to solve the formulated optimization problem, whose convergence is theoretically guaranteed by rigorous proofs.
- Extensive evaluation is performed on public benchmark datasets to evaluate HALGE's performance on recognizing places in long-term scenarios across various months and seasons, which have demonstrated the performance improvement resulted from our approach.

The remainder of the letter is organized as follows. The previous related research on place recognition is presented in Section II. Our HALGE approach is discussed in detail in Section III. After presenting experimental results in Section IV, we conclude our letter in Section V.

II. Related Work

Place recognition (a.k.a loop closure detection) plays an important role in SLAM since it performs data association when a robot revisits any previous seen places. Without place recognition, SLAM cannot obtain globally consistent maps due to the accumulated mapping error caused by sensing drift. In this section, we provide a brief review of previous representation methods for place recognition, especially long-term place recognition.

A. Place Representations

It is known that the appearance of a place can be significantly different caused by changes of illumination, weather, and vegetation conditions. The main methodology to address this critical long-term place recognition problem is to look for a representation that is robust to significant appearance variations [3].

It has been demonstrated that representations based upon keypoints, such as SIFT and SURF, cannot perform well in general when there are long-term perceptual changes of the environment, and holistic representations (e.g., based on HOG [11], GIST [12], CNN [13]) are necessary to encode places with long-term changes [7], [16]. There are methods to fuse multiple modalities to make the representation more robust in long-term place recognition [7], [17]. For example, our previous work [7] autonomously learns representative feature modalities that are shared in various scenarios and fuses these modalities for the long-term place representation. In addition, there are learning-based descriptor enrich approaches to further improve the long-term place recognition, e.g., descriptor space projection [18], and metric learning [19]. All these methods encode the place with holistic information but not capturing the meaningful landmarks. For example, a crossing with stop signs can be easily distinguished from another one without stop signs, though these two crossings look very similar. However, most of the existing holism-based methods cannot encode this useful information.

Recently, several methods [14], [15] address long-term place recognition problem by utilizing semantic landmarks (e.g., stop signs and houses) in the scene as an intermediate representation based upon the fact that the existence of those landmarks are consistent under different environmental conditions, making the resulted representation invariant to appearance changes. For example, [14] uses features based on convolution neural network as robust landmark descriptors to represent and recognize places, without the need of any environment-specific training for the place recognition purpose. As an extension of [14], [15] enforces consistency of the relative positions between landmarks in different views, which shows significant impact on the place recognition performance. However, these methods ignore the rich holistic information in the scene and cannot abstract the discriminative representation for the current place.

In this letter, we propose a new method to integrate both semantic landmarks and holistic cues to extract a robust representation for long-term place recognition. We construct a graph for each place that include both landmarks and holistic images recorded in various scenarios. The edges in the graph models the relationships between different nodes, that is, only those nodes representing the same landmark are connected. Provided the constructed graph, we enable to learn an embedding for the final place representation that is robust to strong appearance variations, since the same place has the consistent representation in the learned embedding space even though they are different in the original raw feature space.

B. Image Matching for Long-Term Place Recognition

Besides place representation, scene matching is important to determine the long-term place recognition performance, which can be divided into two categories: image-based and sequence-based matching.

Many place recognition methods are achieved by matching the query image and the template images in the database. One of the most popular technique is based on pairwise similarity score due to its efficiency. It calculates the similarity scores of the query image and each of the template image based on a certain distance metric. The template image with the highest matching score is selected as the matching template [20], [21]. Another alternative image matching approach is based on nearest neighbor search. A search tree is built to find the most similar template image with respect to the query image [9]. For example, RTAB-MAP [22] implements the KD tree [23] to achieve fast nearest neighbor search for the loop closure detection in SLAM. More recently, sparse regression is also applied in place recognition problem to solve the matching ambiguity problem [12], [24]. [12] formulates the place recognition into a convex optimization problem with ℓ_1-norm based regularization in order to find only a small set of matched templates. However, it has been demonstrated that place recognition based on single image suffers from the perceptual aliasing problem and cannot perform well in long-term place recognition.

In order to integrate rich information, especially temporal information, there are also many methods use a set of consecutive images to decrease the influence by perceptual aliasing and thus achieve better performance in life-long place recognition [1], [20], [21], [25], [26]. Similar to image-based scene matching, pairwise sequence similarity scoring is also applied for the place recognition [20], [21], [26]. Besides that, other sequence-based matching methods have also been reported. For example, [11] formulates the visual robot localization problem across seasons as a minimum cost flow task to exploit the sequence information effectively. [27] applies the Hidden Markov Models (HMMs) to find the paired query sequences in a set of templates. Also, the Conditional Random Fields (CRF) is also used in solving the sequence-based scene matching problem [28].

However, most of the existing image-based and sequence-based methods do not incorporate the rich information in the scene, which include both semantic landmarks and holistic cues. In this letter, we propose a new method that enables to construct a representation graph that introduces both of them simultaneously. In order to address the long-term place recognition problem, a projection is then learned through graph embedding such that the same landmark and place have the identical representation in the projected subspace.

III. Holism-Landmark Graph Embedding

In this section, we introduce our novel HALGE approach to learn the representation for each place, which is constructed by observations in different scenarios in a long period of time. It has not been investigated in any of the previous research for long-term place recognition. We formulate the problem into a novel graph embedding problem using a new optimization formulation. In addition, we also developed a new algorithm to solve the formulated optimization problem, with the theoretical convergence guarantee.

Notation: Throughout this letter, we write matrices as bold uppercase letters and vectors as bold lowercase letters. Given a matrix $\mathbf{M} = [m_{ij}] \in \Re^{n\times m}$, its i-th row and j-th column are denoted as $\mathbf{m}^i$ and $\mathbf{m}_j$, respectively. The trace of $\mathbf{M}$ is defined as $\mathrm{tr}(\mathbf{M}) = \sum_i m_{ii}$.

A. Problem Formulation

To solve the critical long-term place recognition problem, observed images recorded in different scenarios (e.g., different times of the day, months, or seasons) are used for place representation. Different from most of the previous methods that use the holistic cue to extract the place representation, we utilize both holism and semantic landmarks in the scene to represent the place, which is organized as a graph. Formally, we perform the long-term place recognition in a set of input scenes, and we study each input place in all of these given scenes $\mathcal{X} = \{\mathbf{X}_h, \mathbf{X}_l, \mathbf{S}_l\}$, where $\mathbf{X}_h \in \Re^{d\times s}$ denotes the holistic representations in s different scenes, $\mathbf{X}_l = [\mathbf{x}_1, \ldots, \mathbf{x}_m] \in \Re^{d\times m}$ denotes a collection of m semantic landmarks in different scenes, and $\mathbf{S}_l \in \Re^{m\times m}$ is adjacency matrix of the graph constructed by the landmarks such that s_{ij} assesses the similarity between $\mathbf{x}_i$ and $\mathbf{x}_j$, respectively. In our formulation, $s_{ij} = 1$ when node i and j represent the same landmark; Otherwise $s_{ij} = 0$. Here $\mathbf{x}_i \in \Re^d$ is the feature vector of a semantic landmark of the input image recorded in a specific scenario, as shown in Fig. 1. In this letter, we treat the entire image as a special landmark, that is, the feature vectors of the holism image $\mathbf{X}_h$ in s scenarios can be seen as additional landmarks and augmented in $\mathbf{X}$. The landmarks and holistic images apply the same feature descriptor thus have the same dimension d. Therefore, the input scene can be rewritten as $\mathcal{X} = \{\mathbf{X}, \mathbf{S}\}$, where $\mathbf{X} = [\mathbf{X}_h, \mathbf{X}_l] \in \Re^{d\times n}$ and $\mathbf{S} \in \Re^{n\times n}$ is the adjacency matrix of the augmented graph, where $n = m + s$. The node in the graph is either a semantic landmark or the holistic image in one specific scenario, and the graph is constructed for each place, as shown in Fig. 1.

Our goal is to learn from this graph $\mathcal{X}$ a representation of $\mathbf{y} = f(\mathcal{X})$ that captures both landmark and holistic cues conveyed by this graph, which will be then used for place recognition as described later in Section III-B. Intuitively, in order to solve the long-term place recognition problem, we want to enforce the same holism image and landmark in different scenarios with the identical representation, while different holism images or landmarks have different representations. Since different nodes representing the same landmark/place are connected in the constructed graph, we are interested in finding a new space *w.r.t.* the original raw feature space such that the distances of nodes that are neighbors in the graph are minimized in the projected new space. In other works, we preserve the *locality* (neighborhood relationship) between nodes in the graph in the projected new space.

In order to preserve the locality relationships, we need to learn a projection *w.r.t.* the raw feature space of each node such that the connected nodes will have the same representation in the

projected space, which can be formulated to solve the following problem to find the optimal space:

$$\min_{\mathbf{W}^T\mathbf{X}\mathbf{D}\mathbf{X}^T\mathbf{W}=\mathbf{I}} \sum_{i,j=1}^{n} s_{ij} \left\|\mathbf{W}^T\mathbf{x}_i - \mathbf{W}^T\mathbf{x}_j\right\|_2^p, \tag{1}$$

where

$\mathbf{D}$ is a diagonal matrix with the i-th diagonal element as $\sum_j s_{ij}$, and p is a hyper-parameter and satisfying $0 < p \leq 2$.

It should be pointed out that the squared distances when $p = 2$ do not tolerate large value of distance, thus makes the distances in the embedding subspace tend to be even, i.e., not too large but also not too small. Therefore, the squared distances would makes the method unable to find the optimal subspace such that most of the distances of local data pairs are minimized but a few of them are large. We will find the optimal configuration of p in the experimental section in detail.

Although the motivation of Eq. (1) is clear, it is a non-smooth objective and difficult to be solved efficiently. Thus, in the next section, we will introduce an iterative algorithm to solve the problem (1). We will show that the original weight matrix $\mathbf{W}$ would be adaptively re-weighted to capture clearer cluster structures after each iteration.

B. Place Recognition

In the training process, semantic landmarks in the scene can be manually segmented or detected by state-of-the-art segmentation approaches [29]. After constructing the graph using landmarks and holism image in different scenarios, we learn the projection matrix $\mathbf{W}_i$ that preserves the locality relationships by solving the optimization problem in Eq. (1) using Algorithm 1. Finally, the representation for the current place i is calculated by $\mathbf{y}_i = \mathbf{W}_i^T\mathbf{x}_i$, where $\mathbf{x}_i$ is the raw feature representation of the holistic image of place i. $\mathbf{W}_i$ is learned per place i. Since two places may have different data distributions (e.g., different lighting conditions due to shining and shadows), their weight matrices can be different.

In the testing phase, given a new query observation represented by the high-dimensional raw feature vector $\mathbf{x}_q \in \Re^{d\times 1}$ of the holistic image, we project it into a low-dimensional space $\mathbf{y}_{qi} = \mathbf{W}_i^T\mathbf{x}_q \in \Re^{r\times 1}, i = 1, \ldots, n$ using the projection matrix $\mathbf{W}_i$ learned by each place i in the training process. We then calculate a matching score between the query image and each template image in the projected low-dimensional feature space. Following [7], [11], [30], we choose the cosine similarity in this letter. We can determine whether two places are matched by comparing the matching score with a pre-defined threshold. Semantic landmarks in the scene are only required in the training procedure, while not needed for query images in the testing phase, making our HALGE approach highly efficient in real world place recognition tasks.

Due to the graph construction and embedding learning, HALGE is able to deal with strong appearance changes caused by illumination, weather and vegetation variations, thereby improving the robustness of feature matching for long-term place recognition. In addition, although the image-based scene matching method is applied in this work, our HALGE approach can be well integrated with more sophisticated place recognition methods such as sequence-based or manifold-based matching.

Algorithm 1: The algorithm to solve Eq. (1).

Input: Training data $\mathbf{X} \in \Re^{d\times n}$. The original weight matrix $\mathbf{S} \in \Re^{n\times n}$. $\mathbf{D}$ is a diagonal matrix with the i-th diagonal element as $\sum_j s_{ij}$.

1 Initialize $\mathbf{W} \in \Re^{d\times r}$ such that $\mathbf{W}^T\mathbf{X}\mathbf{D}\mathbf{X}^T\mathbf{W} = \mathbf{I}$;

2 **While** *not converge* **do**

3 1. Calculate $\tilde{\mathbf{L}} = \tilde{\mathbf{D}} - \tilde{\mathbf{S}}$, where $\tilde{s}_{ij} = \frac{p}{2}s_{ij}\|\mathbf{W}^T x_i - \mathbf{W}^T x_j\|_2^{p-2}$, $\tilde{\mathbf{D}}$ is a diagonal matrix with the i-th diagonal element as $\sum_j \tilde{s}_{ij}$;

4 2. Update $\mathbf{W}$. The columns of the updated $\mathbf{W}$ are the first r eigenvectors of $(\mathbf{X}\mathbf{D}\mathbf{X}^T)^{-1}\mathbf{X}\tilde{\mathbf{L}}\mathbf{X}^T$ corresponding to the first r smallest eigenvalues ;

Output: $\mathbf{W} \in \Re^{d\times r}$.

C. Optimization Algorithm

Although the proposed objective in Eq. (1) is clearly motivated, it is difficult to solve due to the p-th power of the ℓ_2-norm distances. In this subsection, we will derive an iterative solution algorithm with rigorously proved convergence.

The Lagrangian function of the optimization problem in Eq. (1) is

$$\begin{aligned}\mathcal{L}(\mathbf{W}) = &\sum_{i,j=1}^{n} s_{ij} \left\|\mathbf{W}^T\mathbf{x}_i - \mathbf{W}^T\mathbf{x}_j\right\|_2^p \\ &- \mathrm{tr}\left(\mathbf{\Lambda}(\mathbf{W}^T\mathbf{X}\mathbf{D}\mathbf{X}^T\mathbf{X} - \mathbf{I})\right).\end{aligned} \tag{2}$$

Here we define a Laplacian matrix $\tilde{\mathbf{L}} = \tilde{\mathbf{D}} - \tilde{\mathbf{S}}$, where $\tilde{\mathbf{S}}$ is a re-weighted weight matrix defined by

$$\tilde{s}_{ij} = \frac{p}{2}s_{ij} \left\|\mathbf{W}^T\mathbf{x}_i - \mathbf{W}^T\mathbf{x}_j\right\|_2^{p-2} \tag{3}$$

and $\tilde{\mathbf{D}}$ is a diagonal matrix with the i-th diagonal element as $\sum_j \tilde{s}_{ij}$. Taking the derivative of $\mathcal{L}(\mathbf{W})$ with respect to $\mathbf{W}$, and setting it to zero, we have:

$$\frac{\partial\mathcal{L}(\mathbf{W})}{\partial\mathbf{W}} = \mathbf{X}\tilde{\mathbf{L}}\mathbf{X}^T\mathbf{W} - \mathbf{X}\mathbf{D}\mathbf{X}^T\mathbf{W}\mathbf{\Lambda} = 0, \tag{4}$$

which indicates that the solution $\mathbf{W}$ is the eigenvectors of $(\mathbf{X}\mathbf{D}\mathbf{X}^T)^{-1}\mathbf{X}\tilde{\mathbf{L}}\mathbf{X}^T$. Note that $(\mathbf{X}\mathbf{D}\mathbf{X}^T)^{-1}\mathbf{X}\tilde{\mathbf{L}}\mathbf{X}^T$ is dependent on $\mathbf{W}$. Thus we can derive an iterative algorithm as listed in Algorithm 1 to obtain the solution $\mathbf{W}$ such that the KKT conditions in Eq. (4) is satisfied.

In every iteration of Algorithm 1, $\tilde{\mathbf{L}}$ is calculated with the current solution of $\mathbf{W}$, then $\mathbf{W}$ is updated according to the currently calculated $\tilde{\mathbf{L}}$. This iteration procedure repeats until it converges. From the algorithm we can see that the original weight matrix $\mathbf{S}$ is adaptively re-weighted to minimize the objective in Eq. (1) during iterations.

In the following, we have the convergence theorem of of Algorithm 1.

TABLE I
SCENARIOS CONSIDERED IN EXPERIMENTS ON THE CMU-VL DATASET

Scenario	Month	Description
#1	Mid Sep.	Sunny, abundant green vegetation, vertical shadows
#2	Early Nov.	Sunny, reduced colored vegetation, fallen leaves
#3	Late Nov.	Sunny, strong slanted shadows
#4	Late Dec.	Cloudy, lot of snow on ground
#5	Early Mar.	Partially cloudy, some shadows

Theorem 1: The Algorithm 1 will monotonically decrease the objective in Eq. (1) in each iteration, and converge to a local optimum of the problem.

Proof: See appendix. ■

IV. EXPERIMENTS

In this section, we evaluate and analyze our proposed HALGE method for long-term place recognition using two large-scale public benchmark datasets: CMU-VL dataset and Nordland dataset. Our prior work [7] demonstrated that Histogram of Gradients (HOG) and Convolution Neural Network (CNN) based features are two good features (in comparison to Color, SURF, NormG features, etc) for the long-term place recognition problem. Without loss of generality, we choose HOG as the raw feature to represent the place and apply our HALGE on it. Any global features can be applied in our HALGE method though we use HOG in our two experiments. In addition, multimodal features can also be utilized to further improve the performance by applying our HALGE method. In our experiments, only stable and static objects were selected to construct the graph for each place, e.g., houses, tress, traffic signs, etc. Since all these semantic landmarks were only required in the training procedure, we manually segmented those objects.

A. Evaluation in Scenarios of Different Months

The CMU Visual Localization (CMU-VL) dataset [31] is a public dataset that was recorded using a vehicle with mounted two cameras operating on a 8.8 km route under a variety of scenarios in different months throughout the whole year. Two cameras were mounted on the roof of the vehicle and oriented to left and right respectively. GPS data were also measured to be used as the ground truth of the recorded places. The environmental conditions in the CMU-VL dataset vary a lot across different months of the year (e.g., sunny, cloudy, snowing, green vegetation, reduced colored vegetation, etc), which makes it very challenging to recognize the same place in the long-term span.

The scenarios considered in this experiment via CMU-VL dataset are shown in Table I and Fig. 2. The five videos recorded in these five scenarios are used to train the HALGE model and obtain the projection matrix $\mathbf{W}$ in Eq. (1) for each place. We synchronized the videos recorded in different scenarios according to the GPS data. Without loss of generality, the video recorded in the fourth scenario (late December) is used as the template data, while the video recorded in the fifth scenario (early March) is utilized as the query data. We need to find the best matching frames in late December provided the frames recorded in early March in order to evaluate the performance of our proposed HALGE method.

Fig. 2. Three example places in five different scenarios in the experiment over the CMU-VL dataset. The description of each scenario can be found in Table I.

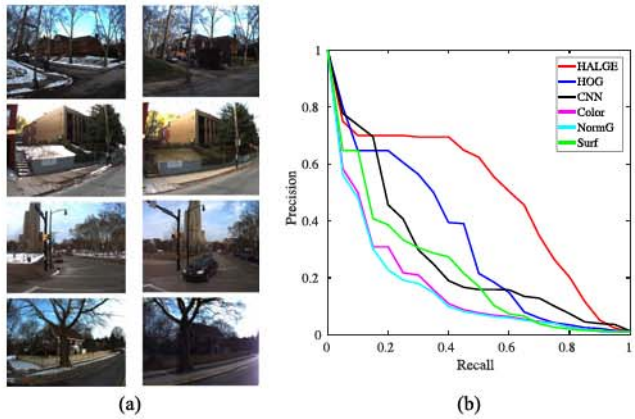

Fig. 3. Evaluation of the HALGE approach on the CMU-VL dataset across different months.

We used frames recorded in December and March that were strictly synchronized by GPS for training. The other frames for testing. The long-term place recognition evaluation results over the CMU-VL dataset are illustrated in Fig. 3, in which the qualitative evaluation is illustrated in Fig. 3(a). The left columns show the query images recorded in December, and the right columns illustrate the images recorded in March that have the maximum matching score with the template images. It can be observed from Fig. 3(a) that the proposed HALGE approach can find the correct place matches even when there are significant appearance variations caused by strong lightening and vegetation changes in different months.

We also quantitatively evaluate our HALGE method in Fig. 3(b) using the precision-recall curves. It is known that better performance can be indicated by a larger area under the precision-recall curve (i.e., the curve is closer to the up-right corner). From Fig. 3(b) we observe that our HALGE approach achieves better performance than that based on raw HOG features, demonstrating the effectiveness and superior performance of our proposed HALGE approach since the HOG-based representation learned by our HALGE method outperforms that based on raw HOG. In addition, we also compare our HALGE approach with previous image-based methods in Fig. 3(b). It is observed that HALGE outperforms approaches by many popularly used features, including color

Fig. 4. Three example places in four seasons in the experiment from the Nordland dataset.

features [7], NormG features [20], SURF features (with bag of words (BoW) encoding) [9], CNN features [13].

B. Evaluation in Scenarios of Different Seasons

We also evaluate the performance of our HALGE method over the Nordland dataset. Nordland dataset [6] is another public benchmark dataset that records the scenes in a ten-hour long trip by train traveling around 3000 KM in Nordland. Visual data in four seasons are recorded and aligned frame by frame in the dataset. The video has a 1920×1080 resolution and 25 FPS.

There are significant appearance variations in the Nordland dataset, which are caused by various weather, vegetation and illumination conditions in four seasons. For example, there is almost full snow coverage on the ground in winter. Moreover, the journey passes through many wild places with similar appearances, making the dataset with strong perceptual aliasing problem. All these difficulties make the Nordland one of the most challenging dataset for long-term visual place recognition. In this experiment, the videos are downsampled with 640×360 resolution and 5 FPS. Frames 1001-6000 are used for training, and the others are used for testing.

The scenarios considered in this experiment via Nordland dataset are shown in Fig. 4. Four videos recorded in four seasons are used to train the HALGE model and obtain the projection matrix $\mathbf{W}$ in Eq. (1) for each place. The videos recorded in four seasons have already been strictly synchronized in Nordland dataset, and we utilized these synchronized frame information as the ground truth. Without loss of generality, the video recorded in winter is used as the template data, while the video recorded in spring is utilized as the query data. The objective is to find the best matching frames in winter provided the frames recorded in spring.

We still use part of the video frames recorded in winter and spring for training, and the remaining for testing. The long-term place recognition evaluation results via the Nordland dataset are illustrated in Fig. 5, in which the qualitative evaluation is illustrated in Fig. 5(a). The left columns show the query images recorded in winter, and the right columns illustrate the images recorded in spring that have the maximum matching score with the template images. It can be observed from Fig. 5(a) that the proposed HALGE approach can find the correct place matches even when there are significant appearance variations caused by strong weather, lightening and vegetation changes in different seasons.

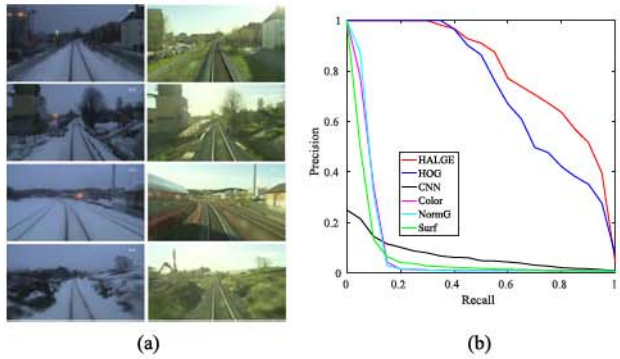

Fig. 5. Evaluation of the HALGE approach on the Nordland dataset across different seasons.

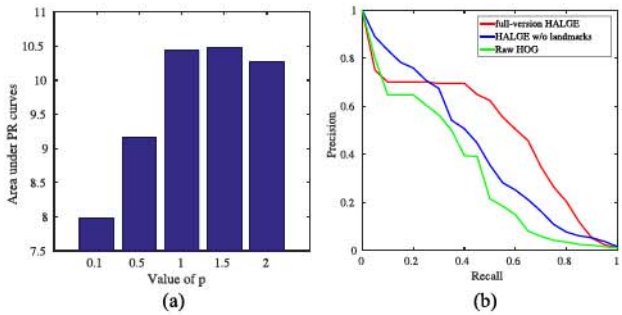

Fig. 6. Parameter analysis and discussion of our HALGE approach using the CMU-VL dataset. Fig. 6(a) shows the area under precision-recall curves with respect to different values of p in Eq. (1). HALGE achieves the best performance when $p = 1.5$. Fig. 6(b) illustrates the importance of semantic landmarks constructed in the graph.

Similar to the previous experiments via the CMU-VL dataset, we also quantitatively evaluate our HALGE method via Nordland dataset in Fig. 5(b) using the precision-recall curves. From Fig. 5(b) we observe that our HALGE approach achieves better performance than that based on raw HOG features, demonstrating the effectiveness and superior performance of our proposed HALGE approach since the HOG-based representation learned by our HALGE method outperforms that based on raw HOG. In addition, we also compare our HALGE approach with previous image-based methods in Fig. 5(b). It is observed that HALGE outperforms previous approaches using popularly used features, including color features [7], NormG features [20], SURF features (with bag of words (BoW) encoding) [9], CNN features [13].

C. Discussion

The main parameters of the HALGE method are discussed and analyzed in this subsection. Without loss of generality, the experimental results via the CMU-VL dataset is chosen to evaluate the effects of the parameter selection in our HALGE method, which are illustrated in Fig. 6.

The HALGE's performance will be affected by the hyper-parameter p in Eq. (1). In Fig. 6(a), we compare the HALGE approach with different values of p using the challenging CMU-VL dataset. It is observed from 6(a) that the best performance is achieved when $p = 1.5$. When $p = 2$, the HALGE

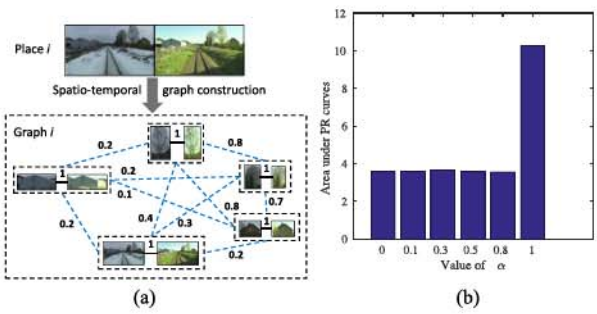

Fig. 7. Analysis of spatial relationships of objects in the scene. Fig. 7(a) shows the constructed graph with consideration of both spatial and temporal relationships between data points. Solid black edges represent temporal relationships, and dashed blue edges denote spatial relationships according to the pixel distance between two nodes in the image. Fig. 7(b) illustrates the impact of spatial relationships.

method reduces to the simple locality preserving projection (LPP) model. We also evaluate the importance of semantic landmarks in the graph construction. Fig. 6(b) demonstrates that the full-version HALGE with both semantic landmarks and holistic image achieves the best performance.

Our HALGE approach considers the appearance relationships between different nodes (i.e., landmarks and holism image) recorded in different scenarios, we call it temporal relationships since they are recorded in different times of the day, month, and seasons. It is also interesting to incorporate spatial relationships between nodes. That is, the geometry structure of different landmarks in the scene can be another important cue for the place representation. The resulting graph integrating both spatial and temporal relationships is illustrated in Fig. 7(a), where the edges representing spatial relationships are weighted by the normalized distance between two nodes in pixel. In Fig. 7(a), dashed blue lines represent the spatial relationships, and solid black lines denote the temporal relationships.

In order to evaluate the effect of spatial relationships, we compare the place recognition performance via the CMU-VL dataset with different weight of sub-graph considering only spatial relationships between nodes. Formally, the adjacent matrix the graph representation of a place $\mathbf{S} = \alpha \mathbf{S}_t + (1-\alpha)\mathbf{S}_s$, where $\mathbf{S}_t$ represents the sub-graph with only temporal relationships, and $\mathbf{S}_s$ denotes that with only spatial relationships. The place recognition performance with respect to different values of α is presented in Fig. 7(b), where we can observe that the best performance is achieved when $\alpha = 1$. It means the spatial sub-graph cannot help in the long-term place recognition, since most of the adjacent landmarks are totally different, connecting them by their geometry distance will degrade the final performance.

Our HALGE method is a general representation learning framework. The raw feature engineering is not the focus of HALGE. In our experimental evaluations, the same HOG descriptor is applied based on the prior knowledge that it performs well in both CMU-VL and Nordland datasets [7]. The performance can be further improved if other advanced features (either single feature or multimodal features) are applied in our HALGE method. In addition, a sequence of image frames represented by our HALGE representation can also further improve the long-term place recognition performance since sequence representation can decrease the negative effect of long-term appearance change and improve the accuracy of life-long place recognition [7]. Since our HALGE method only applies a projection matrix to the raw feature to find the best match, making it highly efficient to be implemented in real world applications. In our experiment using HOG descriptor, the recognition speed can reach up to 1.2e4 Hz (excluding the time for HOG extraction) by using a workstation with 3.7GHz Intel i7 and 16GB memory, without any GPU acceleration.

V. Conclusion

In this letter, we introduce the novel HALGE approach under the optimization framework that integrates information from both semantic landmarks and holistic cues to construct a comprehensive representation for long-term place recognition. HALGE constructs a graph of each place to represent landmarks and holistic images from different scenarios and model their relationships. Given the graph, a projection is then learned through graph embedding, which can preserve the graph structure, such that the same place and landmark have the identical representation in the projected space. We formulate graph embedding under an optimization framework, and design a solver that possesses a guarantee to converge to the optima theoretically. To evaluate HALGE, we conduct experiments based upon two large-scale public datasets that are widely used to benchmark long-term place recognition techniques. The promising experimental results have validated the performance improvement resulted from the HALGE approach.

Appendix
Proof of Theorem 1

Before proving Theorem 1, we have two lemmas.

Lemma 1: For any scalar x, when $0 < p \leq 2$, we have $2|x|^p - px^2 + p - 2 \leq 0$.

Lemma 2: For any nonzero vectors $\mathbf{v}$ and $\mathbf{v}_0$, when $0 < p \leq 2$, the following inequality holds:

$$\|\mathbf{v}\|_2^p - \frac{p}{2}\|\mathbf{v}_0\|_2^{p-2}\|\mathbf{v}\|_2^2 \leq \|\mathbf{v}_0\|_2^p - \frac{p}{2}\|\mathbf{v}_0\|_2^{p-2}\|\mathbf{v}_0\|_2^2. \quad (5)$$

Now we prove Theorem 1.

Theorem: The Algorithm 1 will monotonically decrease the objective in Eq. (1) in each iteration, and converge to a local optimum of the problem.

Proof: Suppose the updated $\mathbf{W}$ is $\tilde{\mathbf{W}}$. According to the step 2 in the Algorithm 1, we know that

$$\tilde{\mathbf{W}} = \arg\min_{\mathbf{W}^T\mathbf{X}\mathbf{D}\mathbf{X}^T\mathbf{W}=\mathbf{I}} \mathrm{tr}(\mathbf{W}^T\mathbf{X}\tilde{\mathbf{L}}\mathbf{X}^T\mathbf{W}) = \arg\min_{\mathbf{W}^T\mathbf{X}\mathbf{D}\mathbf{X}^T\mathbf{W}=\mathbf{I}} \sum_{i,j=1}^{n} \tilde{s}_{ij}\left\|\mathbf{W}^T\mathbf{x}_i - \mathbf{W}^T\mathbf{x}_j\right\|_2^2. \quad (6)$$

Note that $\tilde{s}_{ij} = \frac{p}{2} s_{ij} \left\| \mathbf{W}^T \mathbf{x}_i - \mathbf{W}^T \mathbf{x}_j \right\|_2^{p-2}$, we have

$$\sum_{i,j=1}^{n} \frac{p}{2} s_{ij} \left\| \mathbf{W}^T \mathbf{x}_i - \mathbf{W}^T \mathbf{x}_j \right\|_2^{p-2} \left\| \tilde{\mathbf{W}}^T \mathbf{x}_i - \tilde{\mathbf{W}}^T \mathbf{x}_j \right\|_2^2 \leq \sum_{i,j=1}^{n} \frac{p}{2} s_{ij} \left\| \mathbf{W}^T \mathbf{x}_i - \mathbf{W}^T \mathbf{x}_j \right\|_2^{p-2} \left\| \mathbf{W}^T \mathbf{x}_i - \mathbf{W}^T \mathbf{x}_j \right\|_2^2. \tag{7}$$

According to Lemma 2, we have

$$\sum_{i,j=1}^{n} s_{ij} \left\| \tilde{\mathbf{W}}^T \mathbf{x}_i - \tilde{\mathbf{W}}^T \mathbf{x}_j \right\|_2^p - \sum_{i,j=1}^{n} \frac{p}{2} s_{ij} \left\| \mathbf{W}^T \mathbf{x}_i - \mathbf{W}^T \mathbf{x}_j \right\|_2^{p-2} \left\| \tilde{\mathbf{W}}^T \mathbf{x}_i - \tilde{\mathbf{W}}^T \mathbf{x}_j \right\|_2^2 \leq \sum_{i,j=1}^{n} s_{ij} \left\| \mathbf{W}^T \mathbf{x}_i - \mathbf{W}^T \mathbf{x}_j \right\|_2^p - \sum_{i,j=1}^{n} \frac{p}{2} s_{ij} \left\| \mathbf{W}^T \mathbf{x}_i - \mathbf{W}^T \mathbf{x}_j \right\|_2^{p-2} \left\| \tilde{\mathbf{W}}^T \mathbf{x}_i - \tilde{\mathbf{W}}^T \mathbf{x}_j \right\|_2^2. \tag{8}$$

Summing Eq. (7) and Eq. (8) in the two sides, we have

$$\sum_{i,j=1}^{n} s_{ij} \left\| \tilde{\mathbf{W}}^T \mathbf{x}_i - \tilde{\mathbf{W}}^T \mathbf{x}_j \right\|_2^p \leq \sum_{i,j=1}^{n} s_{ij} \left\| \mathbf{W}^T \mathbf{x}_i - \mathbf{W}^T \mathbf{x}_j \right\|_2^p. \tag{9}$$

Thus the Algorithm 1 monotonically decreases the objective in Eq. (1) in each iteration until the algorithm converges. In the convergence, the equality in Eq. (9) holds, thus $\mathbf{W}$ and $\tilde{\mathbf{L}}$ will satisfy Eq. (4), the KKT condition of the problem in Eq. (1). Therefore, the Algorithm 1 will converge to a local optimum of the problem in Eq. (1). ■

REFERENCES

[1] H. Zhang, F. Han, and H. Wang, "Robust multimodal sequence-based loop closure detection via structured sparsity," in *Robotics: Science and Systems*. Cambridge, MA, USA: MIT Press, 2016.

[2] F. Han, H. Wang, G. Huang, and H. Zhang, "Sequence-based sparse optimization methods for long-term loop closure detection in visual slam," *Auton. Robots*, pp. 1–13, 2018.

[3] S. Lowry, N. Sünderhauf, P. Newman, J. J. Leonard, D. Cox, P. Corke, and M. J. Milford, "Visual place recognition: A survey," *IEEE Trans. Robot.*, vol. 32, no. 1, pp. 1–19, Feb. 2016.

[4] M. Labbe and F. Michaud, "Appearance-based loop closure detection for online large-scale and long-term operation," *IEEE Trans. Robot.*, vol. 29, no. 3, pp. 734–745, Jun. 2013.

[5] H. Grimmett *et al.*, "Integrating metric and semantic maps for vision-only automated parking," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2015, pp. 2159–2166.

[6] N. Sünderhauf, P. Neubert, and P. Protzel, "Are we there yet? Challenging SeqSLAM on a 3000 km journey across all four seasons," in *Proc. Workshop IEEE Int. Conf. Robot. Autom.*, 2013.

[7] 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 Robot. Autom. Lett.*, vol. 2, no. 2, pp. 1172–1179, Apr. 2017.

[8] C. Linegar, W. Churchill, and P. Newman, "Made to measure: Bespoke landmarks for 24-hour, all-weather localisation with a camera," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2016, pp. 787–794.

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

[10] R. Mur-Artal and J. D. Tardós, "Fast relocalisation and loop closing in keyframe-based SLAM," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2014, pp. 846–853.

[11] T. Naseer, L. Spinello, W. Burgard, and C. Stachniss, "Robust visual robot localization across seasons using network flows," in *Proc. AAAI Conf. Artif. Intell.*, 2014, pp. 2564–2570.

[12] Y. Latif, G. Huang, J. J. Leonard, and J. Neira, "An online sparsity-cognizant loop-closure algorithm for visual navigation," in *Proc. Robot. Sci. Syst.*, 2014.

[13] N. Sünderhauf, S. Shirazi, F. Dayoub, B. Upcroft, and M. Milford, "On the performance of convnet features for place recognition," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2015, pp. 4297–4304.

[14] N. Sunderhauf *et al.*, "Place recognition with ConvNet landmarks: Viewpoint-robust, condition-robust, training-free," in *Proc. Robot. Sci. Syst.*, 2015.

[15] P. Panphattarasap and A. Calway, "Visual place recognition using landmark distribution descriptors," in *Proc. Asian Conf. Comput. Vis.*, 2016, pp. 487–502.

[16] J. Wu and J. M. Rehg, "CENTRIST: A visual descriptor for scene categorization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 8, pp. 1489–1501, Aug. 2011.

[17] A. Pronobis, O. M. Mozos, B. Caputo, and P. Jensfelt, "Multi-modal semantic place classification," *Int. J. Robot. Res.*, vol. 29, no. 2/3, pp. 298–320, 2010.

[18] N. Carlevaris-Bianco and R. M. Eustice, "Learning visual feature descriptors for dynamic lighting conditions," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2014, pp. 2769–2776.

[19] Z. Chen, S. Lowry, A. Jacobson, Z. Ge, and M. Milford, "Distance metric learning for feature-agnostic place recognition," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2015, pp. 2556–2563.

[20] M. J. Milford and G. F. Wyeth, "SeqSLAM: Visual route-based navigation for sunny summer days and stormy winter nights," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2012, pp. 1643–1649.

[21] M. J. Milford, G. F. Wyeth, and D. Prasser, "RatSLAM: A hippocampal model for simultaneous localization and mapping," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2004, pp. 403–408.

[22] M. Labbé and F. Michaud, "Online global loop closure detection for large-scale multi-session graph-based slam," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2014, pp. 2661–2666.

[23] J. L. Bentley, "Multidimensional binary search trees used for associative searching," *Commun. ACM*, vol. 18, no. 9, pp. 509–517, 1975.

[24] X. Yang, F. Han, H. Wang, and H. Zhang, "Enforcing template representability and temporal consistency for adaptive sparse tracking," in *Proc. Int. Joint Conf. Artif. Intell.*, 2016, pp. 3522–3529.

[25] 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 *Proc. IEEE Int. Conf. Robot. Automat.*, 2015, pp. 6328–6335.

[26] E. Johns and G.-Z. Yang, "Feature co-occurrence maps: Appearance-based localisation throughout the day," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2013, pp. 3212–3218.

[27] P. Hansen and B. Browning, "Visual place recognition using HMM sequence matching," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2014, pp. 4549–4555.

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

[29] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[30] T. Naseer, M. Ruhnke, C. Stachniss, L. Spinello, and W. Burgard, "Robust visual SLAM across seasons," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2015, pp. 2529–2535.

[31] H. Badino, D. Huber, and T. Kanade, "Real-time topometric localization," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2012, pp. 1635–1642.