SRAL: Shared Representative Appearance Learning for Long-Term Visual Place Recognition

Fei Han, Xue Yang, Yiming Deng, Mark Rentschler, Dejun Yang, and Hao Zhang

***Abstract*—Place recognition, or loop closure detection, is an essential component to address the problem of visual simultaneous localization and mapping (SLAM). Long-term navigation of robots in outdoor environments introduces new challenges to enable life-long SLAM, including the strong appearance change resulting from vegetation, weather, and illumination variations across various times of the day, different days, months, or even seasons. In this paper, we propose a new shared representative appearance learning (SRAL) approach to address long-term visual place recognition. Different from previous methods using a single feature modality or a concatenation of multiple features, our SRAL method autonomously learns representative features that are shared in all scene scenarios, and then fuses the features together to represent the long-term appearance of environments observed by a robot during life-long navigation. By formulating SRAL as a regularized optimization problem, we use structured sparsity-inducing norms to model interrelationships of feature modalities. In addition, an optimization algorithm is developed to efficiently solve the formulated optimization problem, which holds a theoretical convergence guarantee. Extensive empirical study was performed to evaluate the SRAL method using large-scale benchmark datasets, including St Lucia, CMU-VL, and Nordland datasets. Experimental results have shown that our SRAL method obtains superior performance for life-long place recognition using individual images, outperforms previous single image-based methods, and is capable of estimating the importance of feature modalities.**

***Index Terms*—Loop closure detection, long-term place recognition, simultaneous localization and mapping (SLAM), visual learning.**

I. Introduction

VISUAL place recognition (also referred to as loop closure detection) during a long time period is an essential capability required by a robot to perform life-long visual Simultaneous Localization and Mapping (SLAM) [1], which thus attracted an increasing attention in the robotics community [1]–[3]. The goal of place recognition is to identify the place that is previously visited by a mobile robot, which enables the robot to localize itself in an environment during navigation and correct incremental pose drifts during SLAM. However, visual place recognition is a challenging problem due to robot movement, occlusion, and dynamic objects (e.g., cars and pedestrians). In addition, multiple places can have similar appearance, which results in the so-called perceptual aliasing issue. In particular, long-term visual place recognition also introduces additional difficulties. Specifically, the same place can look quite differently during daytime versus night, or across different days, months, or seasons, which is named the *Long-term Appearance Change* (LAC) problem.

In the past several years, long-term place recognition has been intensively investigated [1], [4]–[6], due to its importance to life-long robot navigation. For example, FAB-MAP [7] is one of the first methods to address long-term place recognition for loop closure during visual SLAM, which detects revisited places by matching individual images according to their visual appearance, and was demonstrated on a route of over 1000 KM [3]. Recently, techniques based on sequence-based image matching demonstrated promising performance on

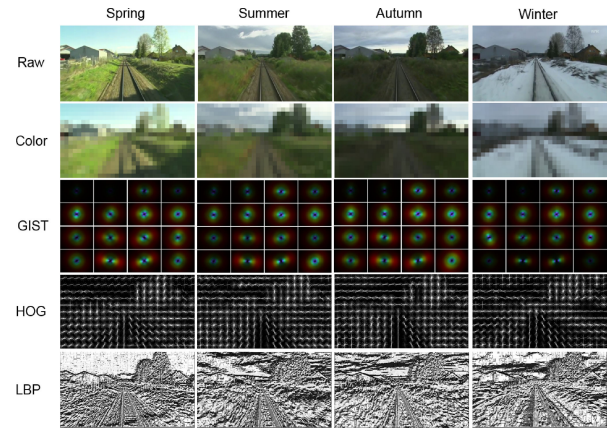

Fig. 1. Motivating examples of long-term visual place recognition. The same place can show significant appearance changes resulting from different weather, vegetation, and illumination conditions (from the Nordland dataset) across different times of the day, months, or seasons. Our SRAL approach automatically identifies visual features (e.g., Color, GIST, HOG, and LBP) that are not only representative but also shared among all scene scenarios, and then effectively integrates them together to perform more accurate image matching for robust long-term visual place recognition.

Manuscript received September 9, 2016; accepted January 12, 2017. Date of publication February 1, 2017; date of current version March 1, 2017. This paper was recommended for publication by Associate Editor Dr. N. Sunderhauf and Editor C. Stachniss upon evaluation of the reviewers' comments.

F. Han, X. Yang, D. Yang, and H. Zhang are with the Division of Computer Science, Colorado School of Mines, Golden, CO 80401 USA (e-mail: fhan@mines.edu; xueyang@mines.edu; djyang@mines.edu; hzhang@mines.edu).

Y. Deng is with the Department of Electrical and Computer Engineering, Michigan State University, East Lansing, MI 48824 USA (e-mail: dengyimi@msu.edu).

M. Rentschler is with the Department of Mechanical Engineering, University of Colorado Boulder, CO 80309 USA (e-mail: mark.rentschler@colorado.edu).

This paper has supplementary downloadable material available at http://ieeexplore.ieee.org, provided by the authors. The Supplemental Material contains the detailed proof of Algorithm 1 in the main paper. This material is 180 kB in size.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LRA.2017.2662061

long-term visual place recognition, including SeqSLAM [8], ABLE-M [9], among others [10], [11], which use a temporal sequence of images for place matching. However, previous visual place recognition techniques either used a single visual feature to represent place appearance, or combine a small number of features by simple concatenation [2], [12]. The problem of how to automatically estimate the importance of heterogeneous features and integrate multimodal features together for robust place recognition is not well understood, especially across a long-term time duration (e.g., different seasons as shown in Fig. 1).

To build a descriptive representation for long-term visual place recognition, we introduce a novel approach, named the *Shared Representative Appearance Learning* (SRAL), which is able to automatically learn representative features that are shared by all scene scenarios, and fuse multimodal features together to encode the long-term appearance of environments observed by a robot during life-long visual navigation. The proposed SRAL approach models the *key insight* that, for long-term place recognition, the best features should be not only representative (so they can be accurately matched) but also shared among all scene scenarios (so the same features can be extracted in all scenarios). If visual features are only discriminative for a specific scenario (e.g., winter), they are less useful to match images from different scenarios, as they cannot well encode the environment in other scenarios. For the first time, the SRAL method automatically learns shared representative representations from multimodal features that can encode the long-term appearance in all scene scenarios with strong perceptual aliasing and the LAC challenge.

The contributions of the paper are threefold:

1) We introduce a new representation based on the shared representative features to encode long-term appearance changes, and propose a new formulate to build the representation as a multi-task multimodal feature learning problem;
2) We propose a novel SRAL approach under the regularized sparse optimization framework to automatically identify the shared representative features and integrate heterogeneous multimodal features for long-term visual place recognition;
3) A new optimization algorithm is implemented to solve the formulated problem, which possesses a theoretical guarantee to converge to the optimal solution.

To facilitate researchers to evaluate and compare their own features using SRAL, the code that implements our approach is available as open source at: http://hcr.mines.edu/code/SRAL.html

The rest of this paper is organized as follows. We describe related research on long-term place recognition in Section II. Section III introduces the SRAL approach for long-term representation learning. Experimental results are discussed in Section IV. Finally, we conclude our paper in Section V.

II. Related Work

Existing visual SLAM methods can be broadly categorized into three groups based on graph optimization, particle filter, and Extended Kalman Filter. Based on the relationship of visual inputs and mapping outputs, SLAM can be categorized into 2D SLAM (using 2D images to generate 2D maps) [13], monocular SLAM (using 2D image sequences to generate 3D maps) [14], [15], and 3D SLAM (using 3D point clouds or RGB-D images to build 3D maps) [16], [17]. Place recognition is an integrated component for loop closuring in all visual SLAM methods, which employs visual features to identify locations previously visited by robots.

A. Image Matching for Long-Term Place Recognition

Most previous place recognition methods use image-based matching [10], [18], [19], which can be broadly divided into two groups, based on pairwise similarity scoring and nearest neighbor search. Methods based on pairwise similarity scores calculate a similarity score of the current scene image and each template based on a certain distance metric and select the template image that has the maximum score [20], [21]. On the other hand, techniques based on the nearest neighbor search typically build a search tree to efficiently identify the most similar template to the current scene. For example, the FAB-MAP SLAM [3], [7] uses the Chow Liu tree to locate the most similar template; RTAB-MAP [16] uses the KD tree to perform fast nearest neighbor search; and other search trees are also used in different place recognition methods for efficient image-to-image matching [9]. Manifold-based methods were also proposed for scene recognition [1], [22], which apply dense correspondence optimization to balance scene manifold and appearance information for image matching. However, due to the limited information provided by a single image, previous image-based matching methods usually suffer from the LAC and perceptual aliasing issues, and are not robust to perform life-long place recognition.

To integrate more information, several recent methods use a sequence of image frames to decrease the negative effect of LAC and perceptual aliasing and improve the accuracy of life-long place recognition [8], [9]. Most sequence-based place recognition techniques are based on a similarity score computed from a pair of template and query image sequences. For example, SeqSLAM [8], one of the earliest sequence-based life-long place recognition methods, computes a summation of image-based similarity scores to match image sequences. The recent ABLE-M technique [9] concatenates binary features extracted from each image in a sequence as a representation to match between sequences. Such similarity scores are also utilized by other sequence-based image matching techniques to perform life-long place recognition for robot navigation [23], which calculate a sequence-based similarity score of the query and all template sequences to construct a similarity matrix, and then choose the template sequence with a statistically high score as a match. To model the temporal relationship among individual frames in the sequence, data association methods based on Hidden Markov Models (HMMs) [11], Conditional Random Fields (CRFs) [24], and graph flow [25] were also proposed to align a pair of template and query sequences. Recently, an unsupervised sequence-based matching method [12] was proposed by formu-

lating the task as a regularized optimization problem, where multiple feature modalities were simply concatenated to build scene representations.

However, the previous methods, usually based on a single type of features or feature concatenation, are not able to effectively integrate the rich information offered by multiple modalities of visual features. We address this problem by introducing a novel multimodal fusion and feature importance learning method, which can work with image-based matching methods for long-term place recognition.

B. Visual Features for Place Representation

A large number of visual features are proposed in previous long-term place recognition techniques to encode the scenes observed by robots during navigation, which can be broadly classified into two categories: local and global features.

Local features apply a detector to detect points of interest (e.g., corners) in an image and a descriptor to encode local information around each detected point of interest. Then, the Bag-of-Words (BoW) model is typically employed by place recognition methods to build a feature vector for each image. The Scale-Invariant Feature Transform (SIFT) feature is used along with a BoW model to detect revisited locations from 2D images [26]. FAB-MAP [3], [7] applies the Speeded Up Robust Features (SURF) for visual place recognition. Both features are also used by the RTAB-MAP SLAM [16]. The binary BRIEF and FAST features are applied to build a BoW model to perform fast loop closure detection [27]. The ORB feature is also used for place recognition [28]. The local visual features are usually discriminative to differentiate each scene scenario (e.g., daytime and night, or different seasons). However, they are often not shared by different scenarios and thus less representative to encode all scene scenarios for image matching in long-term place recognition.

Global features extract information from the whole image, and a feature vector is often formed based on concatenation or feature statistics (e.g., histograms). These global features can encode raw image pixels, shape signatures, and other information. GIST features [29], constructed from the response of steerable filters at different orientations and scales, were employed to perform place recognition [2]. Local Difference Binary (LDB) features were applied to represent scenes by directly using intensity and gradient differences of image grid cells to calculate a binary string [9]. SeqSLAM [8] used the sum of absolute differences of low-resolution images as global features to conduct sequence-based place recognition. Convolutional Neural Networks (CNNs) were also utilized to extract deep features to match image sequences [19], [30]–[32]. Other global features were also proposed to perform life-long place recognition and loop-closure detection [25], [33]. These global features can encode whole image information and no dictionary-based quantization is required. They also showed promising performance for long-term place recognition [1]. However, the problem of identifying more discriminative global features and effectively fuzing them together were not well studied in previous methods based on global features, which is addressed in this paper.

III. Shared Representative Appearance Learning for Long-Term Place Recognition

We propose the concept of learning and fusing heterogenous multimodal features that are not only representative but also shared by different scenarios of the same location/place across a long time span, which was not well investigated in the previous research to address long-term place recognition. We introduce our novel formulation and SRAL approach to build shared representative representations from the perspective of regularized sparse optimization. A new optimization algorithm is also developed to solve the formulated problem, which holds a theoretical guarantee to converge to the global optimal solution.

Notation: In this paper, matrices are written as boldface, capital letters, and vectors are written as boldface lowercase letters. Given a matrix $\mathbf{M} = \{m_{ij}\} \in \mathbb{R}^{n\times m}$, we refer to its i-th row and j-th column as $\mathbf{m}^i$ and $\mathbf{m}_j$, respectively. The ℓ_1-norm of a vector $\mathbf{v} \in \mathbb{R}^n$ is defined as $\|\mathbf{v}\|_1 = \sum_{i=1}^n |v_i|$. The ℓ_2-norm of $\mathbf{v}$ is defined as $\|\mathbf{v}\|_2 = \sqrt{\mathbf{v}^\top \mathbf{v}}$. The $\ell_{2,1}$-norm of the matrix $\mathbf{M}$ is defined as:

$$\|\mathbf{M}\|_{2,1} = \sum_{i=1}^{n} \sqrt{\sum_{j=1}^{m} m_{ij}^2} = \sum_{i=1}^{n} \|\mathbf{m}^i\|_2 \,, \tag{1}$$

and the Frobenius norm is defined as:

$$\|\mathbf{M}\|_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} m_{ij}^2} \,. \tag{2}$$

A. Problem Formulation

Given a collection of images acquired during long-term visual navigation in different scenarios (e.g., different times of the day, months and seasons), the heterogeneous multimodal feature vectors extracted from the image set are represented as $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbb{R}^{d\times n}$, and their corresponding scene scenarios are denoted as $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_n]^\top \in \mathbb{R}^{n\times c}$, where $\mathbf{x}_i = [(\mathbf{x}_i^1)^\top, \ldots, (\mathbf{x}_i^m)^\top]^\top \in \mathbb{R}^d$ is a vector of multimodal features from the i-th image, which consists of m modalities such that $d = \sum_{j=1}^m d_j$, and d_j is the size of the j-th feature modality. $\mathbf{y}_i \in \mathbb{Z}^c$ is the indicator vector of scene scenarios with y_{ij} indicating whether the i-th image belongs to the j-th scenario. Then, shared representative appearance learning is formulated as a regularized sparse optimization problem:

$$\min_{\mathbf{W}} \|\mathbf{X}^\top \mathbf{W} - \mathbf{Y}\|_F^2 + \lambda \|\mathbf{W}\|_{2,1}, \tag{3}$$

where $\mathbf{W} \in \mathbb{R}^{d\times c}$ is the weight matrix with w_{ij} representing the importance of the i-th visual feature with respect to the j-th scene scenario, and $\lambda \geq 0$ is a trade-off hyperparameter. The first term represented by the squared Frobenius norm in (3) is the loss function that models the squared error of using the weighted features to identify a scene scenario. The second term based on the $\ell_{2,1}$-norm is a regularization term that enforces the sparsity of the rows of $\mathbf{W}$ and the grouping effect of the weights in each row. The sparsity enforced by this regularization allows our method to find discriminative features with larger weights (i.e., non-discriminative features have a weight that is close to 0); the grouping effect enforces the discriminative features have

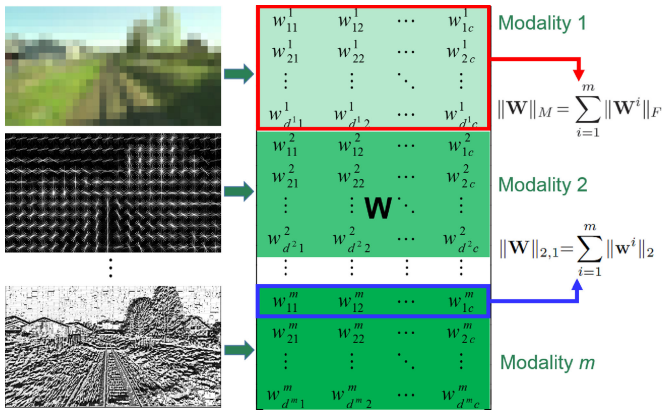

Fig. 2. Illustration of the structured sparsity-inducing norms used in SRAL. Given the feature weight matrix $\mathbf{W} = [\mathbf{W}^1; \mathbf{W}^2; \cdots; \mathbf{W}^m] \in \Re^{d \times c}$, we model the interrelationship of the feature modalities using the new M-norm regularizer that enforces grouping effect of all features in each modality and sparsity among modalities, thus allowing our SRAL approach to identify the shared representative modalities across different scenarios. We also use the ℓ_{21}-norm regularization to identify individual representative shared features. This figure is best viewed in color.

similar weights for all scene scenarios, since when the weights of a feature are equal for all scenarios, $\|\mathbf{m}^i\|_2$ in the $\ell_{2,1}$-norm in (1) obtains the minimum value. Thus, this regularization term models our insight of identifying *shared representative features* to build representations that can represent long-term changes for life-long place recognition. This insight cannot be captured by traditional regularization terms (e.g. the ℓ_1-norm), which are not able to model grouping effect while enforcing sparsity.

Because different types of visual features typically capture different attributes of the scene scenarios (e.g., color, shape, etc.), when multiple modalities of features are used together, it is important to encode the underlying structure introduced by multiple modalities. To realize this capability, we define a novel regularization term that enforces the grouping effect of the features belonging to the same modality, as well as the sparsity between modalities to identify *shared discriminative modalities*, as follows:

$$\|\mathbf{W}\|_M = \sum_{i=1}^{m} \sqrt{\sum_{p=1}^{d_i} \sum_{q=1}^{c} (w_{pq}^i)^2} = \sum_{i=1}^{m} \|\mathbf{W}^i\|_F \,, \tag{4}$$

where $\mathbf{W}^i \in \mathbb{R}^{d_i \times c}$ is the weight matrix for the i-th modality. Then, the final problem formulation of shared representative appearance learning becomes:

$$\min_{\mathbf{W}} \|\mathbf{X}^\top \mathbf{W} - \mathbf{Y}\|_F^2 + \lambda_1 \|\mathbf{W}\|_{2,1} + \lambda_2 \|\mathbf{W}\|_M \,, \tag{5}$$

which models the underlying structure of feature modalities and recognizes shared representative features and modalities (shown in Fig. 2) to encode long-term appearance changes.

B. Multimodal Place Recognition

After solving the optimization problem in (5) using the algorithm detailed in Section III-C, we can obtain the optimal weight matrix $\mathbf{W}^* = [\mathbf{W}^{1*}; \mathbf{W}^{2*}; \cdots; \mathbf{W}^{m*}] \in \Re^{d \times c}$. Then, the overall weight $\omega_i, i = 1, \cdots, m$ for the i-th feature modality can be computed by $\omega_i = \|\mathbf{W}^{i*}\|_F$.

Given a new observation represented by a set of heterogeneous features $\mathbf{x} = [\mathbf{x}^1; \mathbf{x}^2; \cdots; \mathbf{x}^m] \in \Re^d$, we calculate a matching score, $s_i, i = 1, \cdots, m$, between the query observation and each template image using each feature modality separately. Then, the final matching score is computed using the following fusion method:

$$s = \sum_{i=1}^{m} \bar{\omega}_i s_i, i = 1, \cdots, m, \tag{6}$$

where $\bar{\omega}_i$ denotes the normalized weight for the i-th modality. Finally, we can determine whether two locations are matched by comparing the score with a user-defined threshold.

Different from most existing place recognition methods that utilize a single feature modality or are based on simple feature concatenation [12], [30], an advantage of our SRAL approach is its capability to automatically learn representative modalities that are shared by different scenarios with strong appearance changes caused by illumination, weather and vegetation variations, thereby improving the robustness of feature matching for long-term place recognition. Also, the learned weights provide clues for analysis and future integration of various feature modalities for long-term place recognition. In addition, although the image-based matching is applied in this work, our approach can be well integrated with more sophisticated matching methods such as sequence-based or manifold-based matching. When multiple features are extracted, the proposed approach can be applied to each frame within a sequence or each manifold in an image.

C. Optimization Algorithm and Analysis

Because the objective in (5) includes two non-smooth regularization terms (i.e., the $\ell_{2,1}$-norm and the M-norm), it is difficult to solve in general. To this end, we implement a new iterative algorithm to solve the optimization problem in (5) with the non-smooth regularization terms.

Taking the derivative of the objective with respect to $\mathbf{w}_i$, $(1 \le i \le c)$, and setting it to a zero vector, we obtain:

$$\mathbf{X}\mathbf{X}^\top \mathbf{w}_i - \mathbf{X}\mathbf{y}_i + \gamma_1 \mathbf{D}\mathbf{w}_i + \gamma_2 \tilde{\mathbf{D}}\mathbf{w}_i = \mathbf{0}, \tag{7}$$

where $\mathbf{D}$ is a diagonal matrix with the i-th diagonal element as $\frac{1}{2\|\mathbf{w}^i\|_2}$, $\tilde{\mathbf{D}}$ is also a diagonal matrix with the i-th diagonal block as $\frac{1}{2\|\mathbf{W}^i\|_F}\mathbf{I}_i$ ($\mathbf{I}_i$ is an identity matrix of size d_i), and $\mathbf{W}^i$ is the i-th row segment of $\mathbf{W}$ that includes the weights of the i-th feature modality. Thus, we have

$$\mathbf{w}_i = (\mathbf{X}\mathbf{X}^\top + \gamma_1 \mathbf{D} + \gamma_2 \tilde{\mathbf{D}})^{-1} \mathbf{X}\mathbf{y}_i. \tag{8}$$

Because $\mathbf{D}$ and $\tilde{\mathbf{D}}$ are dependent on the weight matrix $\mathbf{W}$, they are also unknown parameters needed to be estimated. An iterative algorithm is proposed to solve this problem, which is described in Algorithm 1.

Algorithm 1 decreases the objective value during each iteration, and we conclude the convergence property of this algorithm by the following Theorem 1.

Algorithm 1: An efficient algorithm to solve the optimization problem in Eq. (5).

Input : $\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_n] \in \Re^{d \times n}$ and $\mathbf{Y} = [\mathbf{y}_1, \cdots, \mathbf{y}_n]^\top \in \Re^{n \times c}$

Let $t = 1$. Initialize $\mathbf{W}(t)$ by solving $\min_{\mathbf{W}} \|\mathbf{X}^\top \mathbf{W} - \mathbf{Y}\|_F^2$.

while *not converge* **do**

  Calculate the block diagonal matrix $\mathbf{D}(t+1)$, where the i-th diagonal element of $\mathbf{D}(t+1)$ is $\frac{1}{2\|\mathbf{w}^i(t)\|_2}\mathbf{I}_i$.

  Calculate the block diagonal matrix $\tilde{\mathbf{D}}(t+1)$, where the i-th diagonal block of $\tilde{\mathbf{D}}(t+1)$ is $\frac{1}{2\|\mathbf{W}^i(t)\|_F}\mathbf{I}_i$.

  For each $\mathbf{w}_i (1 \leq i \leq c)$, $\mathbf{w}_i(t+1) = \left(\mathbf{X}\mathbf{X}^\top + \gamma_1 \mathbf{D}(t+1) + \gamma_2 \tilde{\mathbf{D}}(t+1)\right)^{-1} \mathbf{X}\mathbf{y}_i$.

  $t = t + 1$.

end

Output: $\mathbf{W} = \mathbf{W}(t) \in \Re^{d \times c}$

Theorem 1: Algorithm 1 monotonically decreases the objective value of the problem in (5) in each iteration.

Proof: In the supplementary material. ■

Because the problem in (5) is convex, the optimization algorithm converges to the global optimal solution. In addition, since the solution has a closed form in each iteration, Algorithm 1 converges very fast. In terms of running time in each iteration, computing Steps (3), (4) and (6) is trivial; Step (5) can be implemented through solving a system of linear equations with a quadratic complexity, but not inverting matrices. Dimension reduction methods, such as the coresets for SVD [34] can be applied to further improve the algorithm efficiency.

IV. Experimental Results

We perform extensive experiments to evaluate our SRAL approach's performance on long-term place recognition, using three large-scale benchmark datasets recorded in different conditions across a variety of time spans,

Six different types of visual features were implemented in our experiments for long-term place recognition, including (1) color histograms [35] applied on 32×24 downsampled visual frames, (2) GIST features [29] applied on 320×240 downsampled frames, (3) Histogram of Oriented Gradients (HOG) features [25] computed over 320×240 downsampled images, (4) Local Binary Patterns (LBP) visual features [36] applied on 320×240 downsampled images, (5) Speeded Up Robust Features (SURF) [37] applied on 320×240 downsampled images, and (6) Deep features learned by Convolutional Neural Network (CNN) [19] applied on 320×240 downsampled images. Instead of simply concatenating the visual features into a bigger vector to generate scene representations, our SRAL approach learns a weight of each feature modality and fuses all modalities that are more robust to the LAC challenge in the problem of long-term place recognition.

Qualitative and quantitative evaluations are performed to evaluate our SRAL approach. Furthermore, we compare with several baseline and recent methods during each experiment, including the multiple feature concatenation method, BRIEF-GIST [2], Normalized Gradients (NormG) of grayscale images (used in SeqSLAM [8]), and techniques based only on color, LBP, HOG, SURF, or CNN features. We implemented those feature modalities in the same image-based matching framework to make sure the performance improvement of our SRAL is truly resulted from multimodal fusion, but not from the other procedures (e.g., sequence-based matching). Throughout all experiments, the hyperparameter values $\lambda_1 = 0.001$ and $\lambda_2 = 0.01$ are applied, with the sensitivity analysis for hyper-parameter selection described at the end of this section. In all experiments, the weight matrix is learned on a separate held-back subset of the datasets, and the testing results (precision-recall curves) are obtained by the proposed approach on this subset that is not used in the training process. The separation of training and testing makes sure our evaluation obtains real performance without overfitting.

A. St Lucia Dataset (Different Times of the Day)

St Lucia dataset [38] was recorded using a single camera that was installed on a car in suburban areas of St Lucia in Australia at different times over several days during a two-week period. The length of the whole driving route is around 12 KM. This dataset contains 10 videos, each video lasting 20-25 minutes and consisting of around 22000 frames. The video resolution is 640×480 and the frame rate is 15 frames per second (FPS). GPS data was also documented. The St Lucia dataset shows long-term appearance changes at various times of the day. Frames 1-6000 of the video were used for training.

Several challenges can be observed in this dataset, including appearance variations due to illumination changes at different times of the day, dynamic objects such as pedestrians and vehicles, and viewpoint variations caused by slight route deviations. In this experiment, the GPS information is used as the ground truth for location matching during evaluation.

The long-term visual place recognition results over the St Lucia dataset are illustrated in Fig. 3, in which the qualitative matching results are demonstrated in Fig. 3(a). We show the template images from afternoon scenes (in the top row) that obtain the maximum matching score with the query images from morning scenes (in the bottom row). It is observed that the proposed SRAL approach is able to match locations with significant appearance variations caused by dynamic objects, as well as visual sensing challenges such as camera motions and illumination changes at different times of the day.

The quantitative performance is evaluated using the standard metric of the precision-recall curve, as demonstrated in Fig. 3(b), with a bigger area under the curve (i.e., the curve is closer to the up-right corner) indicating better performance. Fig. 3(b) illustrates that our SRAL approach with both $\ell_{2,1}$-norm and M-norm obtains high performance and accurately matches locations observed in the afternoon with the scenes observed in the morning. Comparisons with previous image-based and baseline methods are also presented in Fig. 3(b). It is observed that SRAL outperforms previous techniques based on single or a small number

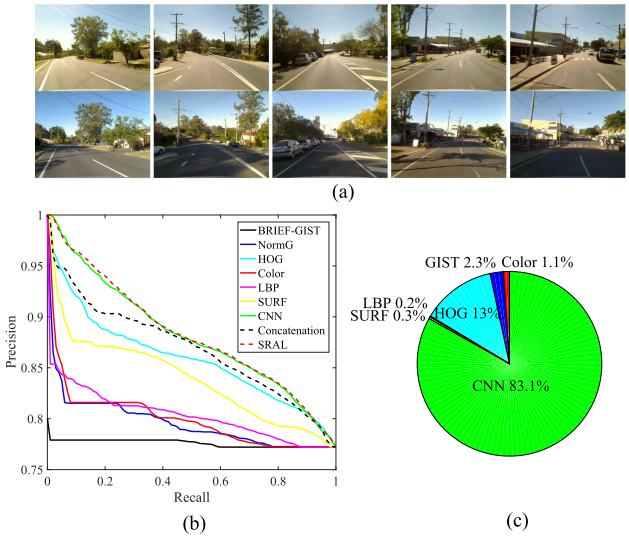

Fig. 3. Experimental results over the St Lucia dataset. (a) presents an example demonstrating the matched images recorded at 15:45 on 08/18/2009 and 10:00 on 09/10/2009, respectively. (b) illustrates the precision-recall curves that indicate the matching performance of our SRAL approach. Quantitative comparisons with baseline and recent methods are also depicted in (b). (c) illustrates the weight distribution of feature modalities automatically learned by the proposed SRAL approach. The figures are best viewed in color. (a) Examples of matched locations, (b) Precision-recall curves, (c) Feature weight.

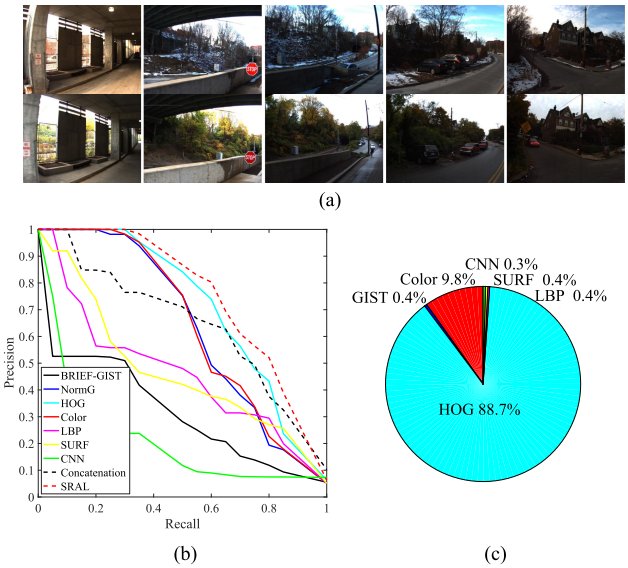

Fig. 4. Experimental results over the CMU-VL dataset. 4(a) presents an example showing the matched images recorded in December and October, respectively. (b) shows the precision-recall curves that demonstrate the performance of our SRAL approach. Quantitative comparisons with baseline and recent methods are also demonstrated in (b). (c) illustrates the normalized weights of feature modalities learned by the SRAL approach. These figures are best viewed in color. (a) Examples of matched locations, (b) Precision-recall curves, (c) Feature weights.

of features (including BRIEF-GIST [2], SeqSLAM [8], SLAM based on color [35] and SURF [37], localization based on LBP [36] and HOG [25], and CNN-based place recognition [19]), by fusing multimodal features to build a representation that is more robust to long-term appearance variations at different times of the day. The SRAL method also performs better than the concatenation method by explicitly learning the weights of feature modalities.

The modality weights of the Color, GIST, HOG, LBP, SURF and CNN features, which are automatically learned by our approach to represent feature importance, are graphically presented in Fig. 3(c). CNN features have the biggest weight among all used visual feature modalities, which shows the ability of CNN to address long-term appearance changes to some extent. When a feature modality is not representative to represent a scene or not shared among different scenarios (e.g., morning & afternoon in this experiment), its normalized weight is very close to 0, for example, the LBP and SURF modalities in Fig. 3(c) with a normalized weight of 0.2% and 0.3%, respectively. These weights are used to integrate feature modalities in our SRAL approach. Also, if the weight of a modality is low (e.g., the LBP feature), then we don't need to compute it during online applications, thus greatly saving computing resources.

B. CMU-VL Dataset (Different Months)

CMU visual localization dataset (CMU-VL) [37] was recorded from two monocular cameras installed on a car that traveled the same route five times around Pittsburgh areas in different months with various weather, climatological, and environmental conditions. The length of the route is around 8 KM. The CMU-VL dataset contains 5 videos, each video consisting of around 13000 frames. The video resolution is 1024×768 and frame rate is 15 FPS. GPS information is also recorded.

This CMU-VL dataset includes the challenge of long-term appearance variation resulted from weather, illumination, and vegetation changes in different months, as well as from urban scene changes due to constructions and dynamic objects. The visual data obtained from the left camera is used in this set of experiments. The images from the first 350 frames were used for training, and frames 351-764 for testing.

The qualitative and quantitative testing results obtained by our SRAL approach on the CMU-VL dataset are graphically shown in Fig. 4. The qualitative results in Fig. 4(a) illustrate the matched image from December (in the top row) with the maximum score for each query image from October (in the bottom row). The scenes in the same location in December and October exhibit obvious appearance variations caused by snow and different vegetation. Our SRAL method is able to well address this LAC challenge and accurately match scene images to recognize same places across different months.

Fig. 4(b) demonstrates the precision-recall curves obtained by the SRAL approaches, and presents the performance comparison of our method with baseline and previous techniques. We observe that the proposed SRAL method outperforms the conventional feature fusion method based on concatenation. The normalized weights of feature modalities are illustrated in Fig. 4(c). The experimental results illustrate that the CNN-based deep features do not always obtain the best performance among all individual feature modalities. In this experiment on the CMU-VL dataset, the HOG feature modality dominates, demonstrating the significance of shape features to recognize places with long-term appearance variations.

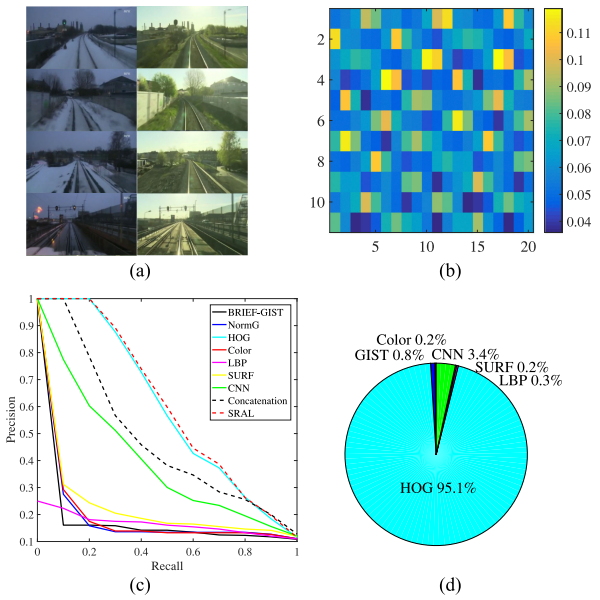

Fig. 5. Experimental results over the Nordland dataset. (a) illustrates an example showing the matched images recorded during winter and spring, respectively. (b) shows the weight of each cell based on HOG features. (c) illustrates the precision-recall curves that indicate the performance of our SRAL approach. Quantitative comparisons with baseline and recent techniques are also demonstrated in (c). (d) illustrates the normalized weights of feature modalities learned by our SRAL approach. The figures are best viewed in color. (a) Examples of matched locations, (b) Cell weights using HOG features, (c) Precision-recall curves, (d) Feature weights.

C. Nordland Dataset (Different Seasons)

Nordland dataset [4] contains visual data from a ten-hour long journey of a train traveling around 3000 KM, which was gathered in four seasons from the viewpoint of the train's front cart. The length of the route is 728 KM. The Nordland dataset contains 4 videos, each video including around 900000 frames. The video resolution is 1920×1080 and the frame rate is 25 FPS. This dataset also includes GPS information. Frames 1001-6000 of the downsampled video were used for training, and frames 1-900 for testing.

The scenes in this dataset exhibit significant appearance variations caused by various weather, vegetation, and illumination conditions in different seasons. In particular in winter, the ground is almost completely covered by snow. Moreover, since multiple places in the wilderness during the trip exhibit similar appearances, this dataset contains strong perceptual aliasing. These difficulties make the dataset one of the most challenging datasets for long-term visual place recognition.

Fig. 5(a) presents the qualitative results obtained by matching the winter data (top row) to the spring images (bottom row) in the Nordland dataset, which show that the SRAL approach can accurately match places with dramatic visual appearance changes across seasons. The precision-recall curves obtained by the proposed SRAL approach as well as from baseline and previous methods are presented and compared in Fig. 5(c). The figure shows that our SRAL approach obtain better results than methods using only a single feature modality, indicating the importance of feature integration for long-term place recognition.

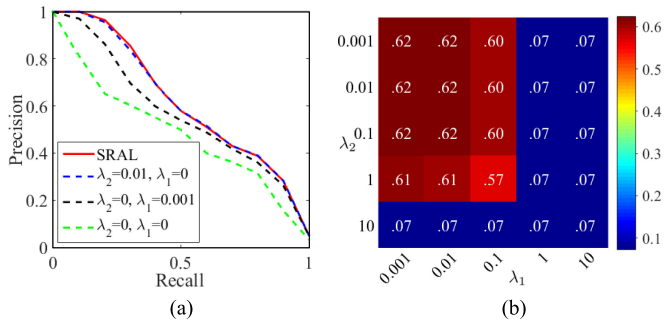

Fig. 6. Performance analysis of our SRAL method with respect to hyperparameters over the Nordland dataset. The figures are best viewed in color. (a) Precision-recall curves. (b) Sensitivity analysis.

Moreover, the SRAL approach significantly outperforms the traditional concatenation-based fusion method, indicating the significance of learning feature weights to fuse multimodal heterogeneous features. Normalized weights of the feature modalities are shown in Fig. 5(d), with an observation that HOG is the most important modality among the used features in this scenario. Moreover, as shown in Fig. 5(b), the features along the horizon have greater weights on average, indicating these areas are more informative and contribute more to place recognition. Our SRAL method achieves greater improvement over other techniques including feature concatenation in the experiments using the Nordland dataset, which is the most challenging used dataset as it includes appearance variations across the longest periods of time (i.e., in different seasons). This emphasizes the our approach's significance when dealing with more challenging environments, which is the purpose of this work. On the other hand, if the environment appearance does not vary much or remains the same (e.g., in short periods of time as in other used datasets), the relatively small appearance variation could be encoded by traditional methods (e.g., concatenation), thus they may obtain similar performance. However, theoretically, our approach is guaranteed to outperform traditional methods based on simple feature concatenation, as they are a special case of our SRAL approach when all learned feature weights have the same value.

D. Discussion

Same as all techniques based on regularized optimization (e.g., Support Vector Machines), the SRAL's performance is affected by hyperparameter values. In Fig. 6(a), we compare the full SRAL approach with the degenerated versions with either or both hyperparameters having a zero value, using the most challenging Nordland dataset with appearance changes across seasons. The precision-recall curves demonstrate that that the full SRAL approach using both M-norm and $\ell_{2,1}$-norm regulations performs the best, but with similar performance to the method that only adopts the M-norm ($\lambda_1 = 0$). Also, both significantly outperform the version only using the $\ell_{2,1}$-norm ($\lambda_2 = 0$). This phenomenon generally indicates that our novel M-norm, which enforces the sparsity among the feature modalities, is more critical than the conventional $\ell_{2,1}$-norm to learn the *shared representative* appearance. We further perform sensitivity analysis to evaluate performance variations with respect to the hyperpa-

rameters with non-zero values. The result, using the area size below precision-recall curves as a metric, over the Nordland dataset is demonstrated in Fig. 6(b), which indicates that there exists a region in the hyperparameter space that results in better performance over other regions in the hyperparameter space.

To demonstrate the efficiency of the SRAL approach, we performed an experiment over the Nordland dataset using a laptop (Intel i5 2.7 GHz CPU and 8 GB memory) without any GPU acceleration by Matlab implementation. It required 10.6 minutes to obtain the weight matrix in the training process with 5000 images in the training set. In testing, each query image can be matched by an average of 0.44 milliseconds. This indicates the promise of our SRAL approach to identify places in real-time applications.

V. Conclusion

In this paper, we introduce a new concept of learning long-term appearance representations from multimodal features, which are not only representative but also shared by all scene scenarios to address the LAC problem and robustly perform long-term place recognition. We propose the SRAL approach to formulate our concept as a regularized sparse optimization problem, and develop a new optimization algorithm to solve the problem with the theoretical convergence guarantee. The SRAL approach is capable to integrate heterogeneous feature modalities by automatically learning modality weights. Experiments results on three benchmarkdatasets have validated that SRAL can estimate modality importance and robustly perform long-term visual place recognition with significant appearance variations across different days, months and seasons. Quantitative comparisons have also shown that SRAL outperforms the previous and baseline methods.

References

[1] S. Lowry *et al.*, "Visual place recognition: A survey," *IEEE Trans. Robot*, vol. 32, no. 1, pp. 1–19, Feb. 2016.

[2] N. Sünderhauf and P. Protzel, "BRIEF-Gist – Closing the loop by simple means," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2011, pp. 1234–1241.

[3] M. Cummins and P. Newman, "Highly scalable appearance-only SLAM-FAB-MAP 2.0," in *Proc. Robot.: Sci. Syst.*, 2009.

[4] 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, p. 2013.

[5] W. Churchill and P. Newman, "Experience-based navigation for long-term localisation," *Int. J Robot. Res.*, vol. 32, no. 14, pp. 1645–1661, 2013.

[6] 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. Autom.*, 2016, pp. 787–794.

[7] 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.

[8] 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. Autom.*, 2012, pp. 1643–1649.

[9] R. Arroyo, P. Alcantarilla, L. Bergasa, and E. Romera, "Towards life-long visual localization using an efficient matching of binary sequences from images," in *Proc. IEEE Int. Conf. Robot. Autom.*, pp. 6328–6335, 2015.

[10] 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.*, pp. 2529–2535, 2015.

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

[12] H. Zhang, F. Han, and H. Wang, "Robust multimodal sequence-based loop closure detection via structured sparsity," in *Proc. Robot.: Sci. Syst.*, 2016.

[13] M. Kaess, H. Johannsson, R. Roberts, V. Ila, J. J. Leonard, and F. Dellaert, "iSAM2: Incremental smoothing and mapping using the Bayes tree," *Int. J. Robot. Res.*, vol. 31, no. 2, pp. 216–235, 2012.

[14] J. Engel, T. Schöps, and D. Cremers, "LSD-SLAM: Large-scale direct monocular SLAM," in *Proc. Eur. Conf. Comput. Vis.*, pp. 834–849, 2014.

[15] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardos, "ORB-SLAM: A versatile and accurate monocular SLAM system," *IEEE Trans. Robot*, vol. 31, no. 5, pp. 1147–1163, Oct. 2015.

[16] M. Labbe 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.

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

[18] E. Stumm, C. Mei, S. Lacroix, J. Nieto, M. Hutter, and R. Siegwart, "Robust visual place recognition with graph kernels," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 4535–4544.

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

[20] C. Chen and H. Wang, "Appearance-based topological Bayesian inference for loop-closing detection in a cross-country environment," *Int. J. Robot. Res.*, vol. 25, no. 10, pp. 953–983, 2006.

[21] J.-S. Gutmann and K. Konolige, "Incremental mapping of large cyclic environments," in *Proc. IEEE Int. Symp. Comput. Intell. Robot. Autom.*, 1999, pp. 318–325.

[22] D. Song and D. Tao, "Biologically inspired feature manifold for scene classification," *IEEE Trans. Image Process.*, vol. 19, no. 1, pp. 174–184, Jan. 2010.

[23] M. Klopschitz, C. Zach, A. Irschara, and D. Schmalstieg, "Generalized detection and merging of loop closures for video sequences," *Proc. 3D Data Process., Visual. Transm.*, 2008.

[24] 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.

[25] 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.

[26] A. Angeli, D. Filliat, S. Doncieux, and J.-A. Meyer, "Fast and incremental method for loop-closure detection using bags of visual words," *IEEE Trans. Robot*, vol. 24, no. 5, pp. 1027–1037, Oct. 2008.

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

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

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

[30] 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.

[31] Y. Jia *et al.*, "Caffe: Convolutional architecture for fast feature embedding," in *Proc. ACM Int. Conf. Multimedia*, 2014, pp. 675–678.

[32] Z. Chen, O. Lam, A. Jacobson, and M. Milford, "Convolutional neural network-based place recognition," in *Proc. Australiaian Conf. Robot. Autom.*, 2014.

[33] E. Pepperell, P. Corke, and M. J. Milford, "All-environment visual place recognition with SMART," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2014, pp. 1612–1618.

[34] D. Feldman, M. Volkov, and D. Rus, "Dimensionality reduction of massive sparse datasets using coresets," in *Proc. Annu. Conf. Neural Inform. Process. Syst.*, 2016, pp. 2766–2774.

[35] D. Lee, H. Kim, and H. Myung, "2D image feature-based real-time RGB-D 3D SLAM," in *Proc. Robot Intell. Technol. Appl.*, 2013, pp. 485–492.

[36] Y. Qiao, C. Cappelle, and Y. Ruichek, "Place recognition based visual localization using LBP feature and SVM," in *Proc. Adv. Artif. Intell. Appl.*, 2015, pp. 393–404.

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

[38] 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 *Proc. IEEE Int. Conf. Robot. Autom.*, 2010, pp. 3507–3512.