Robust Multimodal Sequence-Based Loop Closure Detection via Structured Sparsity

Hao Zhang, Fei Han, and Hua Wang
Department of Electrical Engineering and Computer Sceience
Colorado School of Mines, Golden, Colorado 80401
hzhang@mines.edu, fhan@mines.edu, huawangcs@gmail.com

Abstract—**Loop closure detection is an essential component for simultaneously localization and mapping in a variety of robotics applications. One of the most challenging problems is to perform long-term place recognition with strong perceptual aliasing and appearance variations due to changes of illumination, vegetation, weather, etc. To address this challenge, we propose a novel Robust Multimodal Sequence-based (ROMS) method for long-term loop closure detection, by formulating image sequence matching as an optimization problem regularized by structured sparsity-inducing norms. Our method is able to model the sparsity nature of place recognition, i.e., the current location should match only a small subset of previously visited places, as well as to model underlying structures of image sequences and incorporate multiple feature modalities to construct a discriminative scene representation. In addition, a new optimization algorithm is developed to efficiently solve the formulated problem, which has a theoretical guarantee to converge to the global optimal solution. To evaluate the ROMS algorithm, extensive experiments are performed using large-scale benchmark datasets, including St Lucia, CMU-VL, and Nordland datasets. Experimental results have validated that our algorithm outperforms previous loop closure detection methods, and obtains the state-of-the-art performance on long-term place recognition.**

I. Introduction

Simultaneous Localization and Mapping (SLAM) has been an active research area in robotics for several decades, which is an essential component required by an autonomous robot to navigate through real-world environments in numerous critical applications, including search and rescue [19], service robotics [8, 41], and planetary exploration [12]. *Loop closure detection*, i.e., to identify a previously visited location and find the match from existing map templates using image matching techniques, is necessary in all SLAM systems, because loop closing is able to reduce the ambiguity and uncertainty in estimated maps and robot poses, thereby can significantly improve the accuracy of robot mapping and localization.

Given its importance, the problem of loop closure detection has been extensively investigated over the last decade in visual SLAM [1, 6, 23, 24]. Most previous techniques applied global [2, 24, 26] or local [1, 7, 22, 29] visual features to represent new observations and the scene templates of previously visited places; then a newly observed scene is matched with a scene template using a similarity score or a nearest neighbor search [2, 6]. However, loop closure detection problems become very challenging for the methods to solve robustly when two places look similar, which is often referred to as perceptual aliasing [14], or when the same location exhibits significant appearance changes during a long-term SLAM operation (e.g., in different illumination, weather, and vegetation conditions across months or seasons [18, 26]), often called long-term place recognition.

To address these challenges, several loop closure detection methods [4, 16, 35] fuzed the information from various sensing modalities and devices (such as RGB-D cameras and LiDAR), to construct a more comprehensive and discriminative representation of scene templates and observations. The increasing onboard computational power on mobile robots, especially the accessibility of GPUs, allows for real-time multimodal feature extraction. Another promising direction of research to address the challenges is based on the idea of integrating temporal information and consistency by matching sequences of templates and observed frames, instead of individual images. Sequence-based loop closure detection methods demonstrated a high robustness to perceptual aliasing as well as substantial perceptual changes due to weather, daylight, and season changes in long-term loop closure detection [2, 17, 18, 26].

In this research, we implement a novel *RObust Multimodal Sequence-based* (ROMS) loop closure detection algorithm, by integrating both spatial (via multimodal features) and temporal (via sequence-based matching) information to improve scene representation power and place recognition performance. Our algorithm is inspired by the insight that loop closuring events are inherently sparse [24], i.e., the current sequence of frames matches only a small subset (if any) of the template sequences of scenes previously visited by a robot. A convex optimization problem is formulated to robustly and efficiently solve the loop closure detection task, and structured sparsity-inducing norms are developed to model the spatiotemporal relationship of both scene template and query sequences of various length.

Our contributions are threefold:

- We propose a novel ROMS loop closure detection method that is able to model the sparsity nature of place recognition in SLAM, capture the underlying structure of both template and query sequences, and integrate multimodal features to build highly discriminative representations.
- We introduce a new optimization algorithm to efficiently solve the proposed sequence-based loop closure detection problem, which is theoretically guaranteed to find the best solution to the problem.
- We present a novel paradigm to formulate the sequence-based loop closure detection problem as an optimization task regularized by structured sparsity-inducing norms.

II. Related Work

Existing SLAM methods can be broadly divided into three groups based on extended Kalman filters, particle filters, and graph optimization paradigms [40]. Loop closure detection is an integrated component of all visual SLAM techniques, which uses visual features to recognize revisited locations [25].

A. Visual Features for Scene Representation

A large number of visual features are developed and applied by SLAM methods to represent the scenes observed by robots during navigation. These features can be generally categorized into two classes: global and local features [36, 44].

Global features extract information from the whole image, and a feature vector is often formed based on feature statistics (e.g., histograms). These global features can encode raw image pixels, shape signatures and color information. For example, GIST features [24], built from responses of steerable filters at different orientations and scales, were applied to perform place recognition [37]. The Local Difference Binary (LDB) features were used to represent scenes by directly computing a binary string using simple intensity and gradient differences of image grid cells [2]. The SeqSLAM approach [26] utilized the sum of absolute differences between contrast low-resolution images as global features to perform sequence-based place recognition. Deep features based on convolutional neural networks (CNNs) were adopted to match image sequences [31]. Global features can encode whole image information and no dictionary-based quantization is required, which showed promising performance for long-term place recognition [2, 26, 27, 30, 33].

On the other hand, local features utilize a detector to locate points of interest (e.g., corners) in an image and a descriptor to capture local information of a patch centered at each interest point. Place recognition based on local features typically uses the Bag-of-Words (BoW) model as a quantization technique to construct a feature vector. For example, this model was applied to the Scale-Invariant Feature Transform (SIFT) features to detect loops from 2D images [1]. FAB-MAP [6, 7] utilized the Speeded Up Robust Features (SURF) for visual loop closure detection. Both local features were also applied by the RTAB-Map SLAM [21, 22]. A bag of binary words based on BRIEF and FAST features were used to perform fast place recognition [9]. Recently, ORB features showed promising performance of loop closure identification [28, 29]. The BoW representation based on local visual features are discriminative and (partially) invariant to scale, orientation, affine distortion and illumination changes, thus are widely used in SLAM for place recognition.

The proposed ROMS loop closure detection algorithm is a general multimodal approach that can utilize a combination of global and/or local features to construct a more comprehensive spatial representation of scenes.

B. Image Matching for Place Recognition

Given a query observation and the scene templates of previously visited locations (represented as feature vectors), image matching aims at determining the most similar templates to the query observation, thereby recognizing the revisits.

Most of the place recognition methods are based on image-to-image matching, which localize the most similar individual image that best matches the current frame obtained by a robot. The existing image-to-image matching methods in the SLAM literature can be generally categorized into three groups, based on pairwise similarity scoring, nearest neighbor search, and sparse optimization. Early methods compute a similarity score of the query image and each template based on certain distance metrics and select the template with the maximum similarity score [5, 14]. Matching techniques based on nearest neighbor search typically construct a search tree to efficiently locate the most similar scene template to the query image. For example, the Chow Liu tree was used by the FAB-MAP SLAM [6, 7]. The KD tree was implemented using FLANN to perform fast nearest neighbor search in the RTAB-MAP [21, 22] and some other methods [2, 21] for efficient image-to-image matching. Very recently, methods based on sparsity-inducing norms were introduced to decide the globally most similar template to the query image [24] (details in Section III-A). These image-to-image matching methods typically suffer from the perceptual aliasing problem, due to the limited information carried by a single image [2, 26]. In addition, approaches based on nearest neighbor search or sparse optimization are typically incapable to address sequence-based loop closuring, because they cannot satisfy the constraint that the selected group of the most similar templates are temporally adjacent.

It has been demonstrated that integrating information from a sequence of frames can significantly improve place recognition accuracy and decrease the effect of perceptual aliasing [2, 17, 18, 26, 27]. The majority of sequence-based matching techniques, including RatSLAM [27], SeqSLAM [26], Cooc-Map [18], among others [17, 20], compute sequence similarity using all possible pairings of images within the template and query sequences to create a similarity matrix, and then select the local template sequence with a statistically high score from this matrix. Other sequence-based matching methods were also proposed. For example, this problem is formulated in [30] as a minimum cost flow task in a data association graph to exploit sequence information. Hidden Markov Models (HMMs) [15] and Conditional Random Fields (CRFs) [4] were also applied to align a pair of template and query sequences. However, all previous sequence-based methods are not capable to model the sparsity nature of place recognition for loop closuring. Also, previous approaches only used a local similarity score without considering global constraints to model the interrelationship of the sequences. The proposed ROMS loop closure detection method addresses these issues and is theoretically guaranteed to find the best solution.

III. ROMS Loop Closure Detection

In this section, we introduce the formulation of loop closure detection from the sparse convex optimization point of view. Then, our novel multimodal algorithm is introduced to detect loop closure from a sequence of frames based on heterogenous features, named *RObust Multimodal Sequence-based* (ROMS) loop closure recognition. A new optimization algorithm is also proposed to efficiently solve this problem. Theoretical analysis of the algorithm is provided.

Notation. In this paper, matrices are written using boldface, capital letters, and vectors are represented 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:

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

A. Formulation of Image-to-Image Matching as Sparse Convex Optimization for Loop Closure Detection

Given a collection of image templates from the mapped area $\mathbf{D} = [\mathbf{d}_1, \mathbf{d}_2, \cdots, \mathbf{d}_n] \in \mathbb{R}^{m \times n}$, and a feature vector extracted from the current image $\mathbf{b} \in \mathbb{R}^m$, loop closure detection can be formulated as a convex optimization problem using sparsity-inducing norms, as presented by Latif et al. [24]:

$$\min_{\mathbf{a}} \|\mathbf{D}\mathbf{a} - \mathbf{b}\|_2 + \lambda \|\mathbf{a}\|_1 \ , \tag{2}$$

where $\lambda > 0$ is a trade-off parameter, and $\mathbf{a} \in \mathbb{R}^n$ indicates the weights of all image templates to encode $\mathbf{b}$. A larger value of a_i means the image template $\mathbf{d}_i$ is more similar to the current image $\mathbf{b}$ and can better represent it.

The first term in Eq. 2 is a convex loss function to measure the error of using the templates to explain the current image. The second term is a regularization used to prevent overfitting or introduce additional information to model design objectives. By applying the ℓ_1-norm as a regularization term in Eq. 2, we can enforce the sparsity of $\mathbf{a}$, and seek an explanation of the query image $\mathbf{b}$ that uses the fewest templates from the mapped region. A loop is recognized if an image template has a high similarity (i.e., with a large weight) to the current frame $\mathbf{b}$. If no matches are found within $\mathbf{D}$, then $\mathbf{a}$ is dense, which assigns a small weight to a large portion of the image templates in $\mathbf{D}$. As validated in [24], loop closure detection methods based on sparse convex optimization are able to obtain very promising performance to detect revisited locations.

B. Multimodal Sequence-Based Loop Closure Detection

Our objective is to solve the loop closure detection problem in challenging environments through incorporating a temporal sequence of image frames for place recognition and a set of heterogenous visual features to capture comprehensive image information. Formally, we have a set of templates that encode scenes from the mapped area $\mathbf{D} = [\mathbf{d}_1, \mathbf{d}_2, \cdots, \mathbf{d}_n] \in \mathbb{R}^{m \times n}$, which has rich information structures. Each template contains a set of heterogenous features extracted from different sources $\mathbf{d}_i = [(\mathbf{d}_i^1)^\top, (\mathbf{d}_i^2)^\top, \cdots, (\mathbf{d}_i^r)^\top]^\top \in \mathbb{R}^m$, where $\mathbf{d}_i^j \in \mathbb{R}^{m_j}$ is the feature vector of length m_j that is extracted from the j-th feature modality and $m = \sum_{j=1}^r m_j$. In addition, the feature templates $[\mathbf{d}_1, \mathbf{d}_2, \cdots, \mathbf{d}_n]$ are divided into k separate groups, i.e., $\mathbf{D} = [\mathbf{D}^1, \mathbf{D}^2, \cdots, \mathbf{D}^k]$, where each group $\mathbf{D}^j$ denotes the j-th sequence that contains n_j images acquired in a short time interval and used together for sequence-based matching, where $n = \sum_{j=1}^k n_j$. Given a query observation of the current scene, which contains a sequence of s image frames encoded by their multimodal feature vectors $\mathbf{B} = [\mathbf{b}_1, \mathbf{b}_2, \cdots, \mathbf{b}_s] \in \mathbb{R}^{m \times s}$, solving the loop closure detection problem from the perspective of sparse optimization is to learn a set of weight vectors, $\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_s]$, which can be expanded as:

$$\mathbf{A} = \begin{bmatrix} \mathbf{a}_1^1 & \mathbf{a}_2^1 & \dots & \mathbf{a}_s^1 \\ \mathbf{a}_1^2 & \mathbf{a}_2^2 & \dots & \mathbf{a}_s^1 \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{a}_1^k & \mathbf{a}_2^k & \dots & \mathbf{a}_s^k \end{bmatrix} \in \mathbb{R}^{n \times s} \ , \tag{3}$$

where each component weight vector $\mathbf{a}_p^q \in \mathbb{R}^{n_q}$ represents the weights of the templates in the q-th group $\mathbf{D}^q$ with respect to the p-th query image $\mathbf{b}_p$, which indicates the similarity of the templates in $\mathbf{D}^q$ and $\mathbf{b}_p$.

Since we want each frame $\mathbf{b}$ in the observation relies on the fewest number of templates for place recognition, following [24], an intuitive objective function to solve the problem is:

$$\min_{\mathbf{A}} \sum_{i=1}^{s} \left(\|\mathbf{D}\mathbf{a}_i - \mathbf{b}_i\|_2 + \lambda \|\mathbf{a}_i\|_1 \right) , \tag{4}$$

which minimizes the error of applying $\mathbf{D}$ to explain each $\mathbf{b}_i$ in the query observation, and at the same time enforces sparsity of the used scene templates by using the ℓ_1-norm to regularize each $\mathbf{a}_i$ in $\mathbf{A}$ ($1 \leq i \leq s$). We concisely rewrite Eq. 4 utilizing the following traditional Lasso model [42]:

$$\min_{\mathbf{A}} \|(\mathbf{D}\mathbf{A} - \mathbf{B})^\top\|_{2,1} + \lambda \|\mathbf{A}\|_1, \tag{5}$$

where $\|\mathbf{A}\|_1 = \sum_{i=1}^s \|\mathbf{a}_i\|_1$.

However, this formulation suffers from two critical issues. First, the Lasso model in Eq. 5 is equivalent to independently applying Lasso to each $\mathbf{b}$ and ignores the relationship among the frames in the observation $\mathbf{B}$. Since the frames in the same observation are obtained within a short time period, the visual content of these image frames is similar; thus the frames are correlated and should be explained by the same subset of the templates. Second, the model in Eq. 5 ignores the underlying group structure of the scene templates (each group containing a sequence of templates acquired in previous time), and thus is incapable of matching between sequences, i.e., the selected scene templates with large weights are typically not temporally adjacent or from the same template group. Both issues must be addressed to accurately model the sequence-based loop closure detection problem.

To model the correlation among the frames in an observation $\mathbf{B}$, the $\ell_{2,1}$-norm is proposed as follows:

$$\min_{\mathbf{A}} \| \left(\mathbf{D}\mathbf{A} - \mathbf{B} \right)^\top \|_{2,1} + \lambda \|\mathbf{A}\|_{2,1}. \tag{6}$$

The $\ell_{2,1}$-norm is an advanced technique that addresses both the frame correlation and sparsity issues, by enforcing an ℓ_2-norm across frames (i.e., all frames in $\mathbf{B}$ have a similar weight for a same template) and an ℓ_1-norm across templates (i.e., selected templates are sparse), as illustrated in Fig. 1.

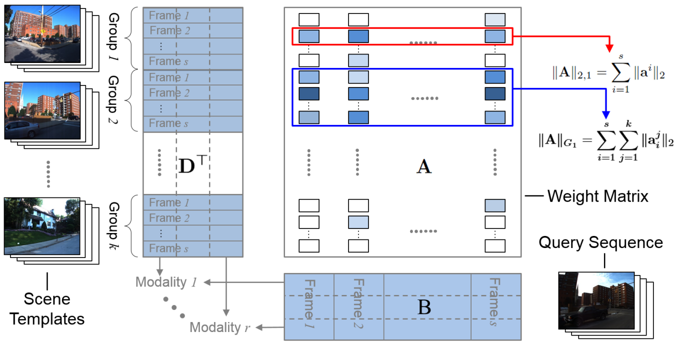


Fig. 1. Illustration of the proposed ROMS algorithm. We model the grouping structure of the scene templates using the G_1-norm regularization ($\|\mathbf{A}\|_{G_1}$), and enforce the query sequence of images to jointly match the same templates using the $\ell_{2,1}$-norm regularization ($\|\mathbf{A}\|_{2,1}$).

To model the grouping structure among the templates in $\mathbf{D}$ and realize sequence-based matching, which was not addressed in previous loop closure detection techniques based on nearest-neighbor search or sparsity optimization, we propose to further regulate the weight matrix $\mathbf{A}$ by adding a new regulation term named the group ℓ_1-norm (G_1-norm) to Eq. 6, which is an ℓ_1 sum of the ℓ_2-norms of group-specific weight vectors:

$$\|\mathbf{A}\|_{G_1} = \sum_{i=1}^{s}\sum_{j=1}^{k} \|\mathbf{a}_i^j\|_2 \,. \tag{7}$$

Because the G_1-norm uses ℓ_2-norm within each group and the ℓ_1-norm between groups, it enforces sparsity between different groups, i.e., if a group of templates are not representative for the observation $\mathbf{B}$, the weights of the templates in this group are assigned with zeros (in ideal case, usually they are very small values); otherwise, their weights are large. The ℓ_2-norm applied on each group enables that the templates within the same group have similar weight values. We illustrate the effect of the G_1-norm regulation in Fig. 1.

To sum up, the final objective function is formulated as:

$$\min_{\mathbf{A}} \|(\mathbf{D}\mathbf{A}-\mathbf{B})^\top\|_{2,1} + \lambda_1\|\mathbf{A}\|_{2,1} + \lambda_2\|\mathbf{A}\|_{G_1} . \tag{8}$$

Through combining the $\ell_{2,1}$-norm with the G_1-norm, a small number of scene templates (can be none) in non-representative groups can also learn a large weight. The combined regularizer can address sequence misalignment challenges, by activating individual templates that are highly similar to the observation but not in the most representative template group. Comparing to traditional regression that utilizes a squared loss (e.g., the Frobenius norm) as the loss function, in our new objective in Eq. 8, the loss term encoded by the $\ell_{2,1}$-norm is an absolute loss, which can significantly improve the robustness of loop closure detection, by reducing the effect of outliers caused by occlusions and dynamic objects (e.g., pedestrians and cars).

After obtaining the optimal $\mathbf{A}$ in Eq. 8, a revisited location (i.e., a loop) is recognized, if one group of scene templates $\mathbf{D}^j$ have large weights, i.e., $\sum_{i=1}^{s} \|\mathbf{a}_i^j\|_1/s \geq \tau$, where τ is close to 1, meaning $\mathbf{D}^j$ well matches the query sequence $\mathbf{B}$. After the query sequence $\mathbf{B}$ is processed, the scene templates $\mathbf{D} = [\mathbf{D^1}, \mathbf{D^2}, \cdots, \mathbf{D^k}]$ are updated as $\mathbf{D} = [\mathbf{D}, \mathbf{B}]$.

C. Optimization Algorithm and Analysis

Although the optimization problem in Eq. 8 is convex, since the objective function contains three non-smooth terms, it is challenging to be solved. We derive a new efficient algorithm to solve this optimization problem, and provide a theoretical analysis to prove that the proposed algorithm converges to the global optimal solution.

Taking the derivative of Eq. 8 with respect to $\mathbf{A}$ and setting it to zero, we obtain[1]:

$$\mathbf{D}^\top\mathbf{D}\mathbf{A}\mathbf{U} - \mathbf{D}^\top\mathbf{B}\mathbf{U} + \lambda_1\mathbf{V}\mathbf{A} + \lambda_2\mathbf{W}^i\mathbf{A} = \mathbf{0} \,, \tag{9}$$

where $\mathbf{U}$ is a diagonal matrix with the i-th diagonal element as $u_{ii} = \frac{1}{2\|\mathbf{b}_i-\mathbf{D}\mathbf{a}_i}\|_2$, $\mathbf{V}$ is a diagonal matrix with the i-th element as $\frac{1}{2\|\mathbf{a}^i\|_2}$, and $\mathbf{W}^i$ $(1 \leq i \leq s)$ is a block diagonal matrix with the j-th diagonal block as $\frac{1}{2\|\mathbf{a}_i^j\|_2}\mathbf{I}_j$, where $\mathbf{I}_j$ $(1 \leq j \leq k)$ is an identity matrix of size n_j for each template group. Thus, for each i, we have:

$$u_{ii}\mathbf{D}^\top\mathbf{D}\mathbf{a}_i - u_{ii}\mathbf{D}^\top\mathbf{b}_i + \lambda_1\mathbf{V}\mathbf{a}_i + \lambda_2\mathbf{W}^i\mathbf{a}_i = \mathbf{0} \,. \tag{10}$$

Then, we calculate $\mathbf{a}_i$ as follows:

$$\mathbf{a}_i = u_{ii}\left(u_{ii}\mathbf{D}^\top\mathbf{D} + \lambda_1\mathbf{V} + \lambda_2\mathbf{W}^i\right)^{-1}\mathbf{D}^\top\mathbf{b}_i \,, \tag{11}$$

where we can efficiently compute $\mathbf{a}_i$ through solving the linear equation $u_{ii}(\mathbf{D}^\top\mathbf{D} + \lambda_1\mathbf{V} + \lambda_2\mathbf{W}^i)a_i = u_i i\mathbf{D}^\top\mathbf{b}_i$, without computing the computationally expensive matrix inversion.

Note that $\mathbf{U}$, $\mathbf{V}$, and $\mathbf{W}$ in Eq. 11 depend on $\mathbf{A}$ and thus are also unknown variables. We propose an iterative algorithm to solve this problem, which is presented in Algorithm 1.

In the following, we analyze the algorithm convergence and prove that Algorithm 1 converges to the global optimum. First, we present a lemma from Nie et al. [32]:

Lemma 1. *For any vector* $\tilde{\mathbf{v}}$ *and* $\mathbf{v}$*, the following inequality holds:* $\|\tilde{\mathbf{v}}\|_2 - \frac{\|\tilde{\mathbf{v}}\|_2^2}{2\|\mathbf{v}\|_2} \leq \|\mathbf{v}\|_2 - \frac{\|\mathbf{v}\|_2^2}{2\|\mathbf{v}\|_2}$.

Then, we prove the convergence of our Algorithm 1 in the following theorem.

Theorem 1. *Algorithm 1 monotonically decreases the objective value of the problem in Eq. 8 in each iteration.*

[1] When $\mathbf{D}\mathbf{a}_i - \mathbf{b}_i = \mathbf{0}$, Eq. 8 is not differentiable. Following [13, 43], we can regularize the i-the diagonal element of the matrix $\mathbf{U}$ using $u_{ii} = \frac{1}{2\sqrt{\|\mathbf{D}\mathbf{a}_i-\mathbf{b}_i\|_2^2+\zeta}}$. Similarly, when $\mathbf{a}^i = \mathbf{0}$, the i-th diagonal element of the matrix $\mathbf{V}$ can be regularized using $\frac{1}{2\sqrt{\|\mathbf{a}^i\|_2^2+\zeta}}$. When $\mathbf{a}_i^j = \mathbf{0}$, we employ the same small perturbation to regularize the j-th diagonal block of $\mathbf{W}^i$ as $\frac{1}{2\sqrt{\|\mathbf{a}_i^j\|_2^2+\zeta}}\mathbf{I}_j$. Then, the derived algorithm can be proved to minimize the following function: $\sum_{i=1}^{s}\sqrt{\|\mathbf{D}\mathbf{a}_i-\mathbf{b}_i\|_2^2+\zeta} + \lambda_1\sum_{i=1}^{n}\sqrt{\|\mathbf{a}^i\|_2^2+\zeta} + \lambda_2\sum_{i=1}^{s}\sum_{j=1}^{k}\sqrt{\|\mathbf{a}_i^j\|_2^2+\zeta}$. It is easy to verify that this new problem is reduced to the problem in Eq. 8, when $\zeta \to 0$.

Algorithm 1: An efficient algorithm to solve the optimization problem in Eq. 8.

```
Input  : The scene templates $\mathbf{D} \in \mathbb{R}^{m \times n}$,
         the query sequence of frames $\mathbf{b} \in \mathbb{R}^{m \times s}$.
Output : The weight matrix $\mathbf{A} \in \mathbb{R}^{n \times s}$.
1: Initialize $\mathbf{A} \in \mathbb{R}^{n \times s}$;
2: while not converge do
3:     Calculate the diagonal matrix $\mathbf{U}$ with the $i$-th diagonal element as $u_{ii} = \frac{1}{2\|\mathbf{b}_i - \mathbf{D}\mathbf{a}_i\|_2}$;
4:     Calculate the diagonal matrix $\mathbf{V}$ with the $i$-th diagonal element as $\frac{1}{2\|\mathbf{a}^i\|_2}$;
5:     Calculate the block diagonal matrix $\mathbf{W}^i$ $(1 \leq i \leq s)$ with the $j$-th diagonal block as $\frac{1}{2\|\mathbf{a}_i^j\|_2}\mathbf{I}_j$;
6:     For each $\mathbf{a}_i$ $(1 \leq i \leq s)$, calculate
       $\mathbf{a}_i = u_{ii}\left(u_{ii}\mathbf{D}^\top\mathbf{D} + \lambda_1\mathbf{V} + \lambda_2\mathbf{W}^i\right)^{-1}\mathbf{D}^\top\mathbf{b}_i$;
7: end
8: return $\mathbf{A} \in \mathbb{R}^{n \times s}$.
```

Proof: Assume the update of $\mathbf{A}$ is $\tilde{\mathbf{A}}$. According to Step 6 in Algorithm 1, we know that:

$$\begin{aligned}\tilde{\mathbf{A}} &= \arg\min_{\mathbf{A}} Tr((\mathbf{D}\mathbf{A}-\mathbf{B})\mathbf{U}(\mathbf{D}\mathbf{A}-\mathbf{B})^\top) \\ &\quad +\lambda_1 Tr(\mathbf{A}^\top\mathbf{V}\mathbf{A}) + \lambda_2 \sum_{i=1}^{s} Tr(\mathbf{a}_i^\top\mathbf{W}^i\mathbf{a}_i)\,, \end{aligned} \tag{12}$$

where $Tr(\cdot)$ is the trace of a matrix. Thus, we can derive

$$\begin{aligned} & Tr((\mathbf{D}\tilde{\mathbf{A}}-\mathbf{B})\mathbf{U}(\mathbf{D}\tilde{\mathbf{A}}-\mathbf{B})^\top) \\ &\quad +\lambda_1 Tr(\tilde{\mathbf{A}}^\top\mathbf{V}\tilde{\mathbf{A}}) + \lambda_2 \sum_{i=1}^{s} Tr(\tilde{\mathbf{a}}_i^\top\mathbf{W}^i\tilde{\mathbf{a}}_i) \\ \leq\; & Tr((\mathbf{D}\mathbf{A}-\mathbf{B})\mathbf{U}(\mathbf{D}\mathbf{A}-\mathbf{B})^\top) \\ &\quad +\lambda_1 Tr(\mathbf{A}^\top\mathbf{V}\mathbf{A}) + \lambda_2 \sum_{i=1}^{s} Tr(\mathbf{a}_i^\top\mathbf{W}^i\mathbf{a}_i) \end{aligned} \tag{13}$$

According to the definition of $\mathbf{U}$, $\mathbf{V}$, and $\mathbf{W}$, we have

$$\begin{aligned} & \sum_{i=1}^{s}\left(\frac{\|\mathbf{D}\tilde{\mathbf{a}}_i-\mathbf{b}_i\|_2^2}{2\|\mathbf{D}\mathbf{a}_i-\mathbf{b}_i\|_2} + \lambda_1\frac{\|\tilde{\mathbf{a}}\|_2^2}{2\|\mathbf{a}\|_2} + \lambda_2\sum_{j=1}^{k}\frac{\|\tilde{\mathbf{a}}_i^j\|_2^2}{2\|\mathbf{a}_i^j\|_2}\right) \\ \leq\; & \sum_{i=1}^{s}\left(\frac{\|\mathbf{D}\mathbf{a}_i-\mathbf{b}_i\|_2^2}{2\|\mathbf{D}\mathbf{a}_i-\mathbf{b}_i\|_2} + \lambda_1\frac{\|\mathbf{a}\|_2^2}{2\|\mathbf{a}\|_2} + \lambda_2\sum_{j=1}^{k}\frac{\|\mathbf{a}_i^j\|_2^2}{2\|\mathbf{a}_i^j\|_2}\right) \end{aligned} \tag{14}$$

According to Lemma 1, we can obtain the following inequalities:

$$\begin{aligned} & \sum_{i=1}^{s}\left(\|\mathbf{D}\tilde{\mathbf{a}}_i-\mathbf{b}_i\|_2 - \frac{\|\mathbf{D}\tilde{\mathbf{a}}_i-\mathbf{b}_i\|_2^2}{2\|\mathbf{D}\mathbf{a}_i-\mathbf{b}_i\|_2}\right) \\ \leq\; & \sum_{i=1}^{s}\left(\|\mathbf{D}\mathbf{a}_i-\mathbf{b}_i\|_2 - \frac{\|\mathbf{D}\mathbf{a}_i-\mathbf{b}_i\|_2^2}{2\|\mathbf{D}\mathbf{a}_i-\mathbf{b}_i\|_2}\right) \end{aligned}$$

$$\sum_{i=1}^{s}\left(\|\tilde{\mathbf{a}}\|_2 - \lambda_1\frac{\|\tilde{\mathbf{a}}\|_2^2}{2\|\mathbf{a}\|_2}\right) \leq \sum_{i=1}^{s}\left(\|\mathbf{a}\|_2 - \lambda_1\frac{\|\mathbf{a}\|_2^2}{2\|\mathbf{a}\|_2}\right) \tag{15}$$

$$\sum_{i=1}^{s}\sum_{j=1}^{k}\left(\|\tilde{\mathbf{a}}_i^j\|_2 - \frac{\|\tilde{\mathbf{a}}_i^j\|_2^2}{2\|\mathbf{a}_i^j\|_2}\right) \leq \sum_{i=1}^{s}\sum_{j=1}^{k}\left(\|\mathbf{a}_i^j\|_2 - \frac{\|\mathbf{a}_i^j\|_2^2}{2\|\mathbf{a}_i^j\|_2}\right)$$

Computing the summation of the three equations in Eq. 15 on both sides (weighted by λs), we obtain:

$$\begin{aligned} & \sum_{i=1}^{s}\|(\mathbf{D}\tilde{\mathbf{a}}_i-\mathbf{b}_i)^\top\|_2 + \lambda_1\|\tilde{\mathbf{a}}\|_2 + \lambda_2\|\tilde{\mathbf{a}}\|_2 \\ \leq\; & \sum_{i=1}^{s}\|(\mathbf{D}\mathbf{a}_i-\mathbf{b}_i)^\top\|_2 + \lambda_1\|\mathbf{a}\|_2 + \lambda_2\|\mathbf{a}\|_2 \end{aligned} \tag{16}$$

Therefore, Algorithm 1 monotonically decreases the objective value in each iteration. ■

Since the optimization problem in Eq. 8 is convex, Algorithm 1 converges to the global optimal solution fast. In each iteration of our algorithm, computing Steps 3–5 is trivial. We compute Step 6 by solving a system of linear equations with a quadratic complexity.

IV. Experimental Results

To assess the performance of our ROMS algorithm on place recognition for loop closure detection, we conducted extensive experiments. This section discusses our implementations, and presents and analyzes the experimental results.

A. Experiment Setup

Three large-scale public benchmark datasets were used for validation in different conditions during various time spans. A summary of the dataset statistics is presented in Table I. Four types of visual features were employed in our experiments for all datasets, including LDB features [2] applied on 64 × 64 downsampled images, GIST features [24] applied on 320×240 downsampled images, CNN-based deep features [34] applied on 227 × 227 downsampled images, and ORB local features [29] extracted from 320 × 240 downsampled images. These features are concatenated into a final vector to represent scene templates and query observations.

TABLE I
Statistics and scenarios of the public benchmark datasets used for algorithm validation in our experiments

Dataset	Sequence	Image Statistics	Scenario
St Lucia [11]	10 × 12 km	$10\times \sim 22,000$ frames 640 × 480 at 15 FPS	Different times of the day
CMU-VL [3]	5 × 8 km	$5\times \sim 13,000$ frames 1024 × 768 at 15 FPS	Different months
Nordland [38]	4 × 728 km	$4\times \sim 900,000$ frames 1920 × 1080 at 25 FPS	Different seasons

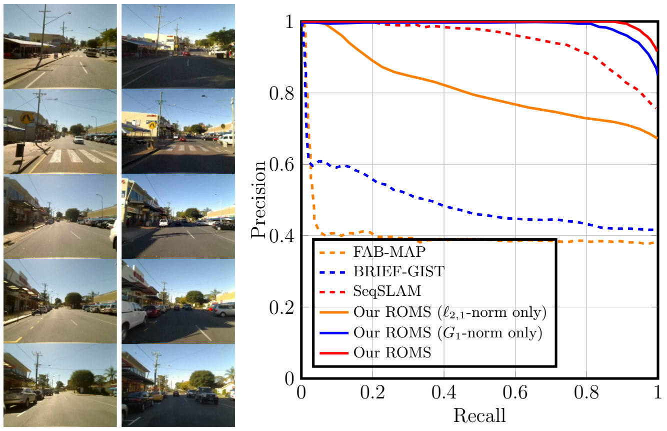
We implement three versions of the proposed ROMS loop closure detection method. First, we set λ_2 in Eq. 8 to 0, which only employs the $\ell_{2,1}$-norm and thereby only considers frame consistency in the query observation. Second, we set λ_1 in Eq. 8 equal to 0, which only uses the G_1-norm to match between sequences without considering frame correlations. Finally, the full version of the proposed ROMS algorithm is implemented, which both models frame consistency and performs sequence matching. The current implementation was programmed using a mixture of unoptimized Matlab and C++ on a Linux laptop with an i7 3.0 GHz GPU, 16G memory and 2G GPU. Similar to other state-of-the-art methods [31, 39], the implementation in this current stage is not able to perform large-scale long-term loop closure detection in real time. A key limiting factor is that the runtime is proportional to the number of previously visited places. Utilizing memory management techniques [22], combined with an optimized implementation, can potentially overcome this challenge to achieve real-time performance. In these experiments, we qualitatively and quantitatively evaluate our algorithms, and compare them with several state-of-the-art methods, including BRIEF-GIST [37], FAB-MAP [7] (using the OpenFABMAP v2.0 implementation [10]), and SeqSLAM [26] (using the OpenSeqSLAM implementation [38]).

B. Results on the St Lucia Dataset (Various Times of the Day)

The *St Lucia dataset* [11] was collected by a single camera installed on a car in the suburban area of St Lucia in Australia at various times over several days during a two-week period. Each data instance includes a video of 20-25 minutes. GPS data was also recorded, which is used in the experiment as the ground truth for place recognition. The dataset contains several challenges including appearance variations due to illumination changes at different times of a day, dynamic objects including pedestrians and vehicles, and viewpoint variations due to slight route deviations. The dataset statistics is shown in Table I.

Loop closure detection results over the St Lucia dataset are illustrated in Fig. 2. The quantitative performance is evaluated using a standard precision-recall curve, as shown in Fig. 2(b). The high precision and recall values (close to 1) indicate that our ROMS methods with G_1-norms obtain high performance and well match morning and afternoon video sequences. The ROMS method only using the G_1-norm regulation outperforms the implementation only using the $\ell_{2,1}$-norm regulation, which underscores the importance of grouping effects and sequence-based matching. When combined both norms together, the full version of the ROMS algorithm obtains the best performance, which indicates that promoting consistency of the frames in the query sequence is also beneficial. To qualitatively evaluate the experimental results, an intuitive example of the sequence-based matching is presented in Fig. 2(a). We show the template image (left column of Fig. 2(a)) that has the maximum weight for a query image (right column of Fig. 2(a)) within a sequence containing 75 frames. This qualitative results demonstrate that the proposed ROMS algorithm works well with the presence of dynamic objects and other vision challenges including camera motions and illumination changes.

Comparisons with some of the main state-of-the-art methods are also graphically presented in Fig. 2(b). It is observed that for long-term loop closure detection, sequence-based methods, such as our ROMS algorithms with G_1-norms and SeqSLAM, outperform the methods based on individual image matching, including FAB-MAP and BRIEF-GIST, due to the significant appearance variations of the same location at different times. In addition, our sequence-based ROMS methods (i.e., with the G_1-norm) obtain superior performance over SeqSLAM, which is mainly resulted from the ability of our ROMS algorithm to detect the global optimal match, comparing to depending on a local similarity score for place recognition. The quantitative comparison of the evaluated sequence-based approaches, using the metric of recall at 100% precision, is summarized in Table II, which indicates the percentage of loop closures that can be recognized without any false positives [7]. We do not include the methods based on individual image matching in this table, because they generally obtain a zero-percent recall at a perfect precision, as illustrated in Fig. 2(b). As indicated by Table II, our ROMS loop closure detection algorithm achieves the best recall of 65.31% with a perfect precision.



(a) Matched sequences

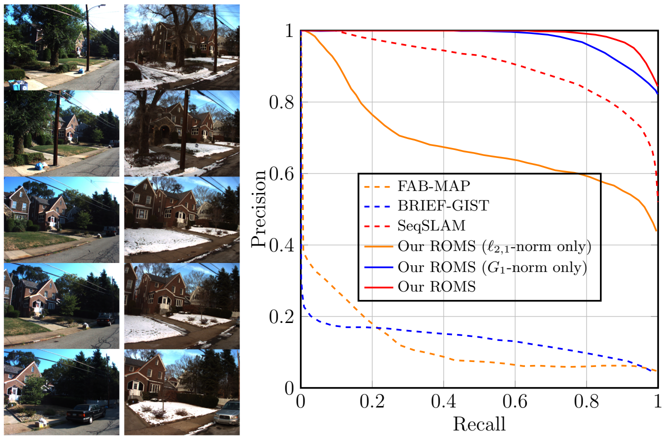
(b) Precision-recall curves

Fig. 2. Experimental results over the St Lucia dataset. Fig. 2(a) presents an example showing the matched template and query sequences recorded at 15:45 on 08/18/2009 and 10:00 on 09/10/2009, respectively. Fig. 2(b) illustrates the precision-recall curves that indicate the performance of our ROMS algorithms. Quantitative comparisons with some of the main state-of-the-art loop closure detection methods are shown in Fig. 2(b). The figures are best seen in color.

C. Results on the CMU-VL Dataset (Different Months)

The *CMU Visual Localization (VL) dataset* [3] was gathered using two cameras installed on a car that traveled the same route five times in Pittsburgh areas in the USA during different months in varying climatological, environmental and weather conditions. GPS information is also available, which is used as the ground truth for algorithm evaluation. This dataset contains seasonal changes caused by vegetation, snow, and illumination variations, as well as urban scene changes due to constructions and dynamic objects. The visual data from the left camera is used in this set of experiments.

The qualitative and quantitative testing results obtained by our ROMS algorithms on the CMU-VL dataset are graphically shown in Fig. 3. Each of the scene template groups and query sequences include 75 frames obtained every five seconds. The qualitative results in Fig. 3(a) show the template images (left column) with the maximum weight for each query image (right



(a) Matched sequences

(b) Precision-recall curves

Fig. 3. Experimental results over the CMU-VL dataset. Fig. 3(a) presents an example demonstrating the matched template and query sequences recorded in October and December, respectively. Fig. 3(b) illustrates the precision-recall curves and compares our methods with several previous loop closure detection approaches. The figures are best viewed in color.

TABLE II

COMPARISON OF USED SEQUENCE-BASED LOOP CLOSURE DETECTION METHODS USING THE METRIC OF RECALL (%) AT 100% PRECISION. APPROACHES BASED ON SINGLE IMAGE MATCHING ARE NOT INCLUDED HERE BECAUSE THEY GENERALLY OBTAIN A ZERO VALUE.

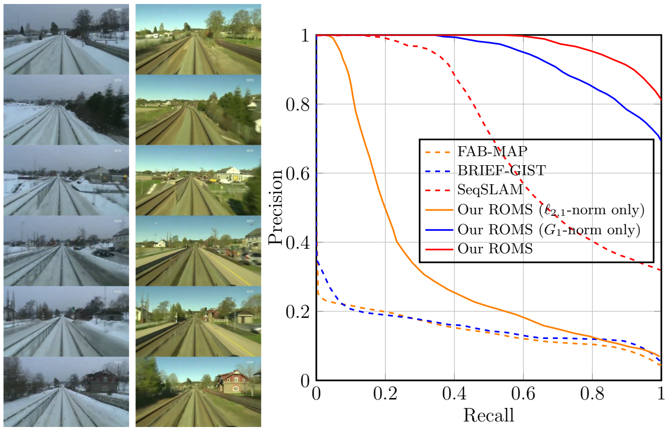
Methods	St Lucia	CMU-VL	Nordland
SeqSLAM [26, 38]	32.25	12.83	16.26
ROMS ($\ell_{2,1}$-norm only)	31.81	2.54	4.83
ROMS (G_1-norm only)	52.55	50.17	36.92
Our ROMS algorithm	65.31	66.47	57.36

column) in an observed sequence. It is clearly observed that our ROMS method is able to well match scene sequences and recognize same locations across different months that exhibit significant weather, vegetation, and illumination changes. The quantitative experimental results in Fig. 3(b) indicate that the ROMS methods with G_1-norm regulations obtain much better performance than the version using only the $\ell_{2,1}$-norm, which is the same phenomenon observed in the experiment using the St Lucia dataset. The reason is the ROMS method using only $\ell_{2,1}$-norm regulations actually matches a sequence of observed images to a set of independent scene templates, i.e., the group structure of the scene templates is not considered. On the other hand, the ROMS methods using G_1-norm regulations perform sequence-based matching, by using the G_1-norm to model the underlying structure of the scene templates. This underscores the importance of sequence-based matching for long-term loop closure detection across months. By integrating both sparsity-inducing norms, the full version of our algorithm achieves very promising performance as shown in Fig. 3 and Table II.

Fig. 3(b) also illustrates comparisons of our ROMS methods with several previous loop closure detection approaches, which shows the same conclusion as in the St Lucia experiment that sequence-based loop closure detection approaches significantly outperform methods based on single image matching for long-term place recognition. In addition, we observe that the ROMS algorithm only using $\ell_{2,1}$-norms as the regularization (i.e., not sequence-sequence matching) still performs much better than traditional approaches based on image-image matching. This is because although the group structure of the scene templates is not modeled, the ROMS algorithm with only the $\ell_{2,1}$-norm considers a sequence of currently observed frames to match a small set of independent templates, which essentially performs the optimal sequence-image matching. The comparison in Fig. 3(b) also demonstrates that even the optimal sequence-image matching approach (i.e., our ROMS algorithm using only the $\ell_{2,1}$-norm) cannot perform as good as sequence-based methods (e.g., SeqSLAM and ROMS with G_1-norms).

D. Results on the Nordland Dataset (Different Seasons)

The *Nordland dataset* [38] contains visual data from a ten-hour long journey of a train traveling around 3000 km, which was recorded in four seasons from the viewpoint of the train's front cart. GPS data was also collected, which is employed as the ground truth for algorithm evaluation. Because the dataset was recorded across different seasons, the visual data contains strong scene appearance variations in different climatological, environmental and illumination conditions. In addition, since multiple places of the wilderness during the trip exhibit similar appearances, this dataset contains strong perceptual aliasing. The visual information is also very limited in several locations such as tunnels. The difficulties make the Nordland dataset one of the most channelling datasets for loop closure detection.

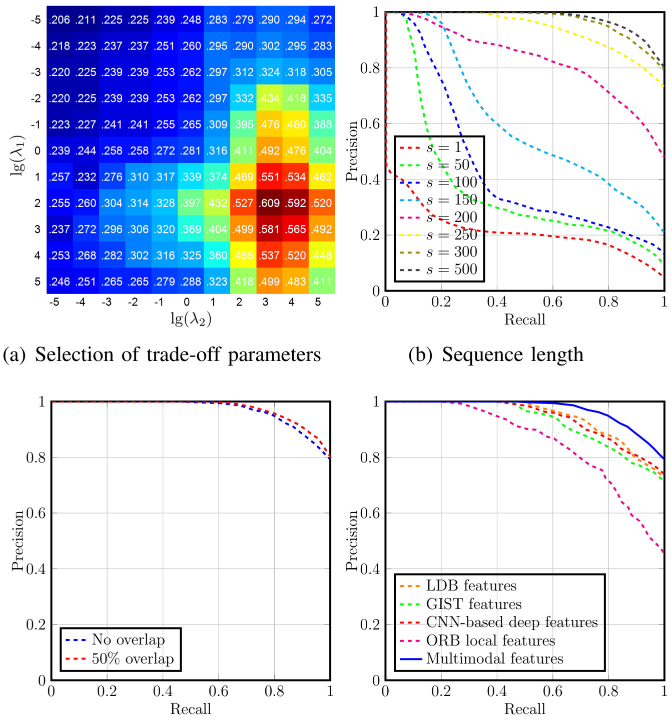


(a) Matched sequences

(b) Precision-recall curves

Fig. 4. Experimental results over the Nordland dataset. Fig. 4(a) presents an example illustrating the matched scene template and query sequences recorded in Winter and Spring, respectively. Fig. 4(b) shows the precision-recall curves and compares our methods with previous loop closure detection methods. The figures are best viewed in color.

Fig. 4 presents the experimental results obtained by matching the spring data to the winter video in the Nordland dataset. Fig. 4(a) demonstrates several example images from one of the matched scene template (left column) and query (right column) sequences, each including 300 frames. Fig. 4(a) validates that our ROMS algorithm can accurately match image sequences that contain dramatic appearance changes across seasons. Fig. 4(b) illustrates the quantitative result obtained by our ROMS algorithms and the comparison with the previous techniques. We can observe similar phenomena that show the state-of-the-art performance of our sequence-based loop closure detection algorithm, which is also supported by its highest recall value at a perfect precision as compared in Table II.



(a) Selection of trade-off parameters

(b) Sequence length

(c) Sequence overlap

(d) Representation modality

Fig. 5. Performance analysis of our ROMS algorithm with respect to varying parameters and algorithm setups using the Nordland dataset. These figures are best viewed in color.

E. Discussion and Parameter Analysis

We discuss and analyze the characteristics and key parameters of the ROMS algorithm, using experimental results of the first hour of the winter and spring visual data in the Nordland dataset as an example, as demonstrated in Fig. 5.

The effect of the trade-off parameters used by our problem formulation in Eq. 8 is illustrated in Fig. 5(a), using recall at 100% perception as an evaluation metric. When $\lambda_1 = 10^2$ and $\lambda_2 = 10^3$, our ROMS approach obtains the best performance. This validates both sparsity-inducing norms are necessary, and the G_1-norm regulation that enables sequence-based matching is more important. When λ_1 and λ_2 take very large values, the performance decreases, because the loss function that models the sequence matching error is almost ignored. When λ_1 and λ_2 take very small values, the algorithm cannot well enforce sequence-based matching and frame consistency of the query sequence, thus resulting in performance decrease. Specifically, when $\lambda_1 = \lambda_2 = 0$, i.e., no global sparsity is considered, the algorithm only minimizes the error of using template groups to explain query sequences, which is similar to the methods based on similarity scores (e.g., SeqSLAM). Similar phenomena are also observed on other datasets in the experiments.

The temporal length of the image sequences is another key parameter that affects the performance of sequence-based loop closure detection techniques. We illustrate the precision-recall curves obtained by our ROMS methods with varying sequence lengths in Fig. 5(b). In general, a longer sequence results in a better performance. On the other hand, when the sequence is longer than 250 (for the Nordland dataset), the improvement is limited. Similar observations are obtained using other datasets. In suburban environments, we notice a sequence length of five seconds (i.e., 75 images for St Lucia and CMU-VL datasets) can result in promising performance. In natural environments with stronger perceptual aliasing, a longer image sequence that includes more information is needed, as demonstrated in Fig. 5(b) using the Nordland dataset. The camera's frame rate and movement speed also need to be considered when determining the number of frames used in the sequences.

Effects of different algorithm setups are also analyzed. For example, the sliding window technique can be flexibly used by our ROMS algorithm through overlapping a number of frames in the sequences. However, we observe in the experiments that the approaches using different sizes of overlaps obtain almost identical performance, as shown by the Nordland example in Fig. 5(c). This is mainly because the highly similar templates outside of the selected group can be activated (and vise versa) by the $\ell_{2,1}$-norm to address the sequence misalignment issue. In addition, we analyze algorithm performance variations with respect to different modality settings. The experimental results over the Nordland dataset are demonstrated in Fig. 5(d). It is observed that the global features (i.e., LDB, GIST, and CNN-based deep features) applied on downsampled images perform better than local features, and are more descriptive to deal with significant scene changes across seasons. The experiment also illustrates that using an ensemble of features can improve the performance of sequence-based image matching.

V. Conclusion

We propose a novel robust multimodal sequence-based loop closure detection algorithm that formulates sequence matching as an optimization problem regularized by structured sparsity-inducing norms. Our ROMS algorithm captures the sparsity nature of loop closure detection, models the grouping structure of template and query sequences, and incorporates multimodal features. A new optimization algorithm is also implemented to efficiently solve the formulated problem, which guarantees to obtain the global optimal solution to the problem. To evaluate the performance of the ROMS method, extensive experiments are performed based on three large-scale benchmark datasets. Qualitative results have validated that our algorithm is able to robustly perform long-term place recognition under significant scene variations across different times of the day, months and seasons. Quantitative evaluation results have also demonstrated that our ROMS algorithm outperforms previous techniques and obtains the state-of-the-art place recognition performance.

References

[1] Adrien Angeli, David Filliat, Stéphane Doncieux, and Jean-Arcady Meyer. Fast and incremental method for loop-closure detection using bags of visual words. *IEEE Transactions on Robotics*, 24(5):1027–1037, 2008.

[2] Roberto Arroyo, Pablo F Alcantarilla, Luis M Bergasa, and Eduardo Romera. Towards life-long visual localization using an efficient matching of binary sequences from images. In *IEEE International Conference on Robotics and Automation*, 2015.

[3] Hernán Badino, Daniel Huber, and Takeo Kanade. Real-time topometric localization. In *IEEE International Conference on Robotics and Automation*, 2012.

[4] César Cadena, Dorian Gálvez-López, Juan D Tardós, and José Neira. Robust place recognition with stereo sequences. *IEEE Transactions on Robotics*, 28(4):871–885, 2012.

[5] Cheng Chen and Han Wang. Appearance-based topological Bayesian inference for loop-closing detection in a cross-country environment. *The International Journal of Robotics Research*, 25(10):953–983, 2006.

[6] Mark Cummins and Paul Newman. FAB-MAP: Probabilistic localization and mapping in the space of appearance. *The International Journal of Robotics Research*, 27(6):647–665, 2008.

[7] Mark Cummins and Paul Newman. Highly scalable appearance-only SLAM-FAB-MAP 2.0. In *Robotics: Science and Systems*, 2009.

[8] Carlos Estrada, José Neira, and Juan D Tardós. Hierarchical SLAM: Real-time accurate mapping of large environments. *IEEE Transactions on Robotics*, 21(4): 588–596, 2005.

[9] Dorian Gálvez-López and Juan D Tardós. Bags of binary words for fast place recognition in image sequences. *IEEE Transactions on Robotics*, 28(5):1188–1197, 2012.

[10] Arren Glover, William Maddern, Michael Warren, Stephanie Reid, Michael Milford, and Gordon Wyeth. OpenFABMAP: An open source toolbox for appearance-based loop closure detection. In *IEEE International Conference on Robotics and Automation*, 2012.

[11] Arren J Glover, William P Maddern, Michael J Milford, and Gordon F Wyeth. FAB-MAP + RatSLAM: appearance-based SLAM for multiple times of day. In *IEEE International Conference on Robotics and Automation*, 2010.

[12] Steven B Goldberg, Mark W Maimone, and Lany Matthies. Stereo vision and rover navigation software for planetary exploration. In *IEEE Aerospace Conference Proceedings*, 2002.

[13] Irina F Gorodnitsky and Bhaskar D Rao. Sparse signal reconstruction from limited data using FOCUSS: A re-weighted minimum norm algorithm. *IEEE Transactions on Signal Processing*, 45(3):600–616, 1997.

[14] Jens-Steffen Gutmann and Kurt Konolige. Incremental mapping of large cyclic environments. In *IEEE International Symposium on Computational Intelligence in Robotics and Automation*, 1999.

[15] Paul Hansen and Brett Browning. Visual place recognition using HMM sequence matching. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2014.

[16] Peter Henry, Michael Krainin, Evan Herbst, Xiaofeng Ren, and Dieter Fox. RGB-D mapping: Using Kinect-style depth cameras for dense 3D modeling of indoor environments. *The International Journal of Robotics Research*, 31(5):647–663, 2012.

[17] Kin Leong Ho and Paul Newman. Detecting loop closure with scene sequences. *International Journal of Computer Vision*, 74(3):261–286, 2007.

[18] Edward Johns and Guang-Zhong Yang. Feature co-occurrence maps: Appearance-based localisation throughout the day. In *IEEE International Conference on Robotics and Automation*, 2013.

[19] Alexander Kleiner and Christian Dornhege. Real-time localization and elevation mapping within urban search and rescue scenarios. *Journal of Field Robotics*, 24(8-9): 723–745, 2007.

[20] Manfred Klopschitz, Christopher Zach, Arnold Irschara, and Dieter Schmalstieg. Generalized detection and merging of loop closures for video sequences. In *3D Data Processing, Visualization, and Transmission*, 2008.

[21] Mathieu Labbe and Francois Michaud. Appearance-based loop closure detection for online large-scale and long-term operation. *IEEE Transactions on Robotics*, 29 (3):734–745, 2013.

[22] Mathieu Labbe and François Michaud. Online global loop closure detection for large-scale multi-session graph-based SLAM. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2014.

[23] Yasir Latif, César Cadena, and José Neira. Robust loop closing over time for pose graph SLAM. *The International Journal of Robotics Research*, pages 1611–1626, 2013.

[24] Yasir Latif, Guoquan Huang, John Leonard, and José Neira. An Online Sparsity-Cognizant Loop-Closure Algorithm for Visual Navigation. In *Robotics: Science and Systems Conference*, 2014.

[25] Stephanie Lowry, Niko Sunderhauf, Paul Newman, John J Leonard, David Cox, Peter Corke, and Michael J Milford. Visual Place Recognition: A Survey. *IEEE Transactions on Robotics*, in press, 2016.

[26] Michael J Milford and Gordon F Wyeth. SeqSLAM: Visual route-based navigation for sunny summer days and stormy winter nights. In *IEEE International Conference on Robotics and Automation*, 2012.

[27] Michael J Milford, Gordon F Wyeth, and DF Rasser. RatSLAM: a hippocampal model for simultaneous localization and mapping. In *IEEE International Conference on Robotics and Automation*, 2004.

[28] Raúl Mur-Artal and Juan D Tardós. Fast relocalisation and loop closing in keyframe-based SLAM. In *IEEE*

International Conference on Robotics and Automation, 2014.

[29] Raul Mur-Artal, Jose Maria Martinez Montiel, and Juan D Tardos. ORB-SLAM: a versatile and accurate monocular SLAM system. *IEEE Transactions on Robotics*, 31(5):1147–1163, 2015.

[30] Tayyab Naseer, Luciano Spinello, Wolfram Burgard, and Cyrill Stachniss. Robust visual robot localization across seasons using network flows. In *AAAI Conference on Artificial Intelligence*, 2014.

[31] Tayyab Naseer, Michael Ruhnke, Cyrill Stachniss, Luciano Spinello, and Wolfram Burgard. Robust visual SLAM across seasons. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2015.

[32] Feiping Nie, Heng Huang, Xiao Cai, and Chris H Ding. Efficient and robust feature selection via joint $\ell_{2,1}$-norms minimization. In *Advances in Neural Information Processing Systems*, 2010.

[33] Edward Pepperell, Peter Corke, and Michael J Milford. All-environment visual place recognition with SMART. In *IEEE International Conference on Robotics and Automation*, 2014.

[34] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, 2015.

[35] João Machado Santos, Micael S Couceiro, David Portugal, and Rui P Rocha. A sensor fusion layer to cope with reduced visibility in SLAM. *Journal of Intelligent & Robotic Systems*, 80(3):401–422, 2015.

[36] Elena S Stumm, Christopher Mei, and Simon Lacroix. Building location models for visual place recognition. *The International Journal of Robotics Research*, 35(4): 334–356, 2015.

[37] Niko Sünderhauf and Peter Protzel. BRIEF-Gist – Closing the loop by simple means. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2011.

[38] Niko Sünderhauf, Peer Neubert, and Peter Protzel. Are we there yet? Challenging SeqSLAM on a 3000 km journey across all four seasons. In *Workshop on IEEE International Conference on Robotics and Automation*, 2013.

[39] Niko Sunderhauf, Sareh Shirazi, Adam Jacobson, Feras Dayoub, Edward Pepperell, Ben Upcroft, and Michael Milford. Place recognition with ConvNet landmarks: Viewpoint-robust, condition-robust, training-free. In *Robotics: Science and Systems*, 2015.

[40] Sebastian Thrun and John J Leonard. Simultaneous localization and mapping. In *Springer handbook of robotics*, pages 871–889. Springer, 2008.

[41] Sebastian Thrun, Wolfram Burgard, and Dieter Fox. A real-time algorithm for mobile robot mapping with applications to multi-robot and 3D mapping. In *IEEE International Conference on Robotics and Automation*, 2000.

[42] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[43] Hua Wang, Feiping Nie, and Heng Huang. Multi-view clustering and feature learning via structured sparsity. In *International Conference on Machine Learning*, 2013.

[44] Hao Zhang, Christopher Reardon, and Lynne E Parker. Real-time multiple human perception with color-depth cameras on a mobile robot. *Cybernetics, IEEE Transactions on*, 43(5):1429–1441, 2013.