Autonomous Robots
https://doi.org/10.1007/s10514-018-9736-3

Sequence-based sparse optimization methods for long-term loop closure detection in visual SLAM

Fei Han[1] · Hua Wang[1] · Guoquan Huang[2] · Hao Zhang[1]

Received: 15 February 2017 / Accepted: 2 April 2018

Abstract
Loop closure detection is one of the most important module in Simultaneously Localization and Mapping (SLAM) because it enables to find the global topology among different places. A loop closure is detected when the current place is recognized to match the previous visited places. When the SLAM is executed throughout a long-term period, there will be additional challenges for the loop closure detection. The illumination, weather, and vegetation conditions can often change significantly during the life-long SLAM, resulting in the critical strong perceptual aliasing and appearance variation problems in loop closure detection. In order to address this problem, we propose a new Robust Multimodal Sequence-based (ROMS) method for robust loop closure detection in long-term visual SLAM. A sequence of images is used as the representation of places in our ROMS method, where each image in the sequence is encoded by multiple feature modalites so that different places can be recognized discriminatively. We formulate the robust place recognition problem as a convex optimization problem with structured sparsity regularization due to the fact that only a small set of template places can match the query place. In addition, we also develop a new algorithm to solve the formulated optimization problem efficiently, which guarantees to converge to the global optima theoretically. Our ROMS method is evaluated through extensive experiments on three large-scale benchmark datasets, which record scenes ranging from different times of the day, months, and seasons. Experimental results demonstrate that our ROMS method outperforms the existing loop closure detection methods in long-term SLAM, and achieves the state-of-the-art performance.

Keywords Long-term place recognition · Loop closure detection · Visual SLAM · Long-term autonomy

1 Introduction

Autonomous robots as well as self-driving cars must have the ability to navigate themselves in unstructured and unknown environments. To achieve this goal, a map of the navigating area has to be built so that autonomous robots and self-driving cars are able to localize themselves in it. This is one of the fundamental problems in the robotics community known as Simultaneous Localization and Mapping (SLAM), which has been applied in many applications (Kleiner and Dornhege 2007; Estrada et al. 2005; Thrun et al. 2000; Goldberg et al. 2002). Unlike odometry which interprets the environment as an infinite map and suffers from the drifting problem, SLAM incorporates the loop closure detection module and is able to understand the global real topology of the environment (Cadena et al. 2016). Loop closure detection can improve improve the accuracy of the maps and robot moving trajectories.

Numerous results on loop closure detection have been reported over the last decade due to its importance in visual SLAM (Angeli et al. 2008; Cummins and Newman 2008; Latif et al. 2013, 2014). Visual features of images are applied to represent different places in visual SLAM, including local

This is one of several papers published in *Autonomous Robots* comprising the "Special Issue on Robotics Science and Systems".

✉ Fei Han
fhan@mines.edu

Hua Wang
huawangcs@gmail.com

Guoquan Huang
ghuang@udel.edu

Hao Zhang
hzhang@mines.edu

[1] Department of Computer Science, Colorado School of Mines, Golden, CO 80401, USA

[2] Department of Mechanical Engineering, University of Delaware, Newark, DE 19716, USA

Published online: 21 April 2018

features (Angeli et al. 2008; Cummins and Newman 2009; Labbe and Michaud 2014; Mur-Artal et al. 2015) and global features (Arroyo et al. 2015; Latif et al. 2014; Milford and Wyeth 2012). In most existing approaches, a new query scene is matched with a scene template in the database using scene matching methods, i.e. maximum similarity score or nearest neighbor search (Arroyo et al. 2015; Cummins and Newman 2008). However, most of those previous techniques cannot perform well when two places have the similar appearance, which is also known as the perceptual aliasing problem (Gutmann and Konolige 1999). Moreover, the appearance of a place can be extremely changed in different times of the day, month, or seasons due to changes of illumination, weather, and vegetation conditions (Johns and Yang 2013; Milford and Wyeth 2012). This is called long-term loop closure detection problem in visual SLAM. When a SLAM system needs to update the map during the long-term operation, this long-term loop closure detection is essential (Labbe and Michaud 2013; Han et al. 2017). The traditional loop closure detectors tend to fail when considering this long-term issue (Zhang et al. 2016; Han et al. 2017).

There are various methods reported to address these challenges. One direction is to fuze data from various feature modalities and/or sensors (i.e. depth, radar, and LiDAR, etc) to construct a more discriminative and robust representation of places (Cadena et al. 2012; Henry et al. 2012; Santos et al. 2015). Instead of representing places using single image, a sequence of consecutive images are proposed to be the representation of one place, which integrates much more information than the single image-based representation method (temporal and more spatial information). It has been demonstrated that sequence-based loop closure detection methods are more robust when dealing with both perceptual aliasing problems and appearance changes caused by light, weather, and vegetation changes in long-term loop closure detection (Arroyo et al. 2015; Ho and Newman 2007; Johns and Yang 2013; Milford and Wyeth 2012).

In this paper, we propose a novel *RObust Multimodal Sequence-based* (ROMS) method for loop closure detection in long-term visual SLAM. It integrates both spatial and temporal information via multimodal features and sequence-based scene recognition, respectively. In our ROMS method, a sequence of images is used for the representation of a single place, where each image in the sequence is encoded by multimodal features. Inspired by (Latif et al. 2014) that the query sequence of frames only matches a small subset of template sequences stored in the database, the loop closure detection problem becomes to find the most representative sequence in templates that matches the query sequence. By this idea, we propose a new sparse optimization formulation as well as an efficient algorithm to solve the loop closure detection problem in long-term visual SLAM.

The contribution of the paper is threefold:

1. We propose a novel ROMS method for loop closure detection that integrates the insights of the place sparsity and the sequence-based place recognition framework, which allows to robustly model the long-term variability of places for loop closing in visual SLAM.
2. We develop and implement a new paradigm to formulate robust sequence-based loop closure detection as a regularized optimization problem based on structured sparsity-inducing norms.
3. We introduce an efficient algorithm to solve the formulated non-smooth convex optimization problem, by which the theoretical convergence to the global optima is guaranteed.

The rest of the paper is organized as follows. We first review state-of-the-art approaches addressing loop closure detection in Sect. 2. Then, the proposed ROMS method with structured sparsity regularization is presented in Sect. 3. Experimental results are illustrated in Sect. 4. Finally, the conclusion is drawn in Sect. 5.

2 Related work

SLAM addresses the robot navigation problem in unknown environments, which can provide accurate environment models and robot pose estimates.

We can broadly categorize existing SLAM methods into three groups based on extended Kalman filters, particle filters, and graph optimization paradigms (Thrun and Leonard 2008). Loop closure detection is an integrated component of all visual SLAM techniques, which uses visual features to recognize revisited locations (Lowry et al. 2016). The global topology of the environment will be updated when a loop closure is detected (Cadena et al. 2016). In this section, we provide a brief review of visual features and image matching methods used in visual SLAM.

2.1 Visual features for scene representation

In the computer vision and robotics community, scenes observed by robots during navigation are usually represented by visual features. Many visual features are developed and applied in SLAM systems during the past decade, which can be generally divided into two classes: global and local features.

Global features extract information from the entire 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 (Latif et al. 2014), built from

responses of steerable filters at different orientations and scales, were applied to perform place recognition (Sünderhauf and Protzel 2011). 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 (Arroyo et al. 2015). The SeqSLAM approach (Milford and Wyeth 2012) 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 (Naseer et al. 2015). Global features can encode whole image information and no dictionary-based quantization is required, which showed promising performance for long-term place recognition (Arroyo et al. 2015; Milford and Wyeth 2012; Milford et al. 2004; Naseer et al. 2014; Pepperell et al. 2014).

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. The Bag-of-Words (BoW) model is often used as a quantization technique for local features in order to construct a feature vector in place recognition applications. For example, this model was applied to the Scale-Invariant Feature Transform (SIFT) features to detect loops from 2D images (Angeli et al. 2008). FAB-MAP (Cummins and Newman 2008, 2009) utilized the Speeded Up Robust Features (SURF) for visual loop closure detection. Both local features were also applied by the RTAB-Map SLAM (Labbe and Michaud 2013, 2014). A bag of binary words based on BRIEF and FAST features were used to perform fast place recognition (Gálvez-López and Tardós 2012). Recently, ORB features showed promising performance of loop closure identification (Mur-Artal and Tardós 2014; Mur-Artal et al. 2015). 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.

Different from most existing methods that use only one kind of feature modality as the place representation, our 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.

2.2 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 (Chen and Wang 2006; Gutmann and Konolige 1999). 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 (Cummins and Newman 2008, 2009). The KD tree was implemented using FLANN to perform fast nearest neighbor search in the RTAB-MAP (Labbe and Michaud 2013, 2014) and some other methods (Arroyo et al. 2015; Labbe and Michaud 2013) 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 (Latif et al. 2014) (details in Sect. 3.1). These image-to-image matching methods typically suffer from the perceptual aliasing problem, due to the limited information carried by a single image (Arroyo et al. 2015; Milford and Wyeth 2012). 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 (Arroyo et al. 2015; Ho and Newman 2007; Johns and Yang 2013; Milford and Wyeth 2012; Milford et al. 2004). The majority of sequence-based matching techniques, inclsuding RatSLAM (Milford et al. 2004), SeqSLAM (Milford and Wyeth 2012), Cooc-Map (Johns and Yang 2013), among others (Ho and Newman 2007; Klopschitz et al. 2008), 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 (Naseer et al. 2014) as a minimum cost flow task in a data association graph to exploit sequence information. Hidden Markov Models (HMMs) (Hansen and Browning 2014) and Conditional Random Fields (CRFs) (Cadena et al. 2012) 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.

3 ROMS loop closure detection

In this section, the formulation of loop closure detection from the sparse convex optimization point of view is introduced. Then, we present the novel multimodal algorithm 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 also provided.

Notation Throughout the 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 ith row and jth 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:

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

3.1 Formulation of image-to-image matching as sparse convex optimization for loop closure detection

Given a collection of image templates from the mapped area $\mathbf{D} = [\mathbf{d}_1, \mathbf{d}_2, \ldots, \mathbf{d}_n] \in \mathbb{R}^{m\times n}$, and a feature vector extracted from the current image $\mathbf{b} \in \mathbb{R}^m$, loop closure detection problem can be formulated as a convex optimization problem with sparsity regularization, as presented by Latif et al. (2014):

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

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 loss function based on ℓ_2-norm 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 encode structure in the model for 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 Latif et al. (2014), loop closure detection methods based on sparse convex optimization are able to obtain very promising performance to detect revisited locations and close the loop in SLAM.

3.2 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 $\mathbf{D} = [\mathbf{d}_1, \mathbf{d}_2, \ldots, \mathbf{d}_n] \in \mathbb{R}^{m\times n}$, which has rich information structures. Each template contains a set of heterogenous features extracted from different sources $\mathbf{d}_i = [(\mathbf{d}_i^1)^\top, (\mathbf{d}_i^2)^\top, \ldots, (\mathbf{d}_i^r)^\top]^\top \in \mathbb{R}^m$, where $\mathbf{d}_i^j \in \mathbb{R}^{m_j}$ is the feature vector of length m_j that is extracted from the jth feature modality and $m = \sum_{j=1}^r m_j$. In addition, the feature templates $[\mathbf{d}_1, \mathbf{d}_2, \ldots, \mathbf{d}_n]$ are divided into k separate groups, i.e., $\mathbf{D} = [\mathbf{D}^1, \mathbf{D}^2, \ldots, \mathbf{D}^k]$, where each group $\mathbf{D}^j$ denotes the jth 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 $\mathbf{B} = [\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_s] \in \mathbb{R}^{m\times s}$, solving the loop closure detection problem from the perspective of sparse optimization is to learn a set of weight vectors, $\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_s]$, which can be expended as:

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

where each component weight vector $\mathbf{a}_p^q \in \mathbb{R}^{n_q}$ represents the weights of the templates in the qth group $\mathbf{D}^q$ with respect to the pth 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 Latif et al. (2014), an intuitive objective function to solve the problem is:

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

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 (Tibshirani 1996):

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

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:

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

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.

To model the grouping structure among the templates in $\mathbf{D}$ and achieve sequence-based matching, which was not addressed in previous loop closure detection techniques based on nearest-neighbor search or sparsity optimization,

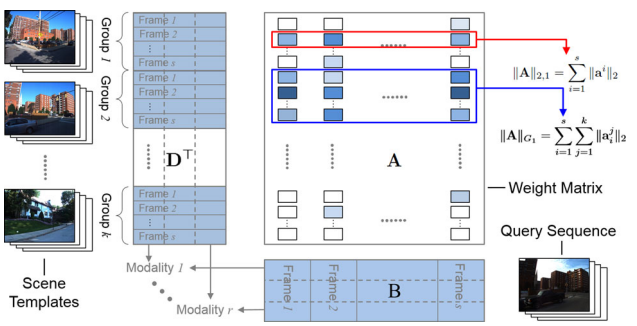


Fig. 1 Illustration of the proposed ROMS algorithm. We model the grouping structure of the scene templates using the G_1-norm regularization ($\|\mathbf{A}\|_{G_1}$), and enforce the query sequence of images to jointly match the same templates using the $\ell_{2,1}$-norm regularization ($\|\mathbf{A}\|_{2,1}$)

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:

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

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 $\mathbf{D} = [\mathbf{D^1}, \mathbf{D^2}, \ldots, \mathbf{D^k}]$ are updated as $\mathbf{D} = [\mathbf{D}, \mathbf{B}]$.[1]

3.3 Optimization algorithm and analysis

Although the optimization problem in Eq. 8 is convex, it is very challenging for us to solve it efficiently since the

[1] The template groups can be designed to have overlaps, e.g., using sliding window techniques. However, in the experiments, we found that groups with or without overlaps result in almost identical performance, as demonstrated by the example in Fig. 5c, since our method can activate highly similar scene templates outside of the selected group (and vise versa) to solve the sequence misalignment issue.

objective function contains non-smooth terms. We can formulate the problem as a second-order cone programming (SOCP) or semidefinite programming (SDP) problem, which can be solved by some existing methods, i.e., interior point method or the bundle method. However, solving those problems is expensive in computation, which limits its application in robust loop closure detection of visual SLAM.

Addressing this problem, we derive a new efficient algorithm to solve the optimization problem in Eq. 8, 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[2]:

$$\mathbf{D}^\top\mathbf{DAU} - \mathbf{D}^\top\mathbf{BU} + \lambda_1\mathbf{VA} + \lambda_2\mathbf{W}^i\mathbf{A} = \mathbf{0}\,, \tag{9}$$

where $\mathbf{U}$ is a diagonal matrix with the ith 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 ith element as $\frac{1}{2\|\mathbf{a}^i\|_2}$, and $\mathbf{W}^i$ $(1 \le i \le s)$ is a block diagonal matrix with the jth diagonal block as $\frac{1}{2\|\mathbf{a}_i^j\|_2}\mathbf{I}_j$, where $\mathbf{I}_j$ $(1 \le j \le k)$ 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. Accordingly, 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. (2010):

[2] When $\mathbf{D}\mathbf{a}_i - \mathbf{b}_i = \mathbf{0}$, Eq. 8 is not differentiable. Following Gorodnitsky and Rao (1997) and Wang et al. (2013), 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 ith 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 jth 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 ith 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 ith diagonal element as $\frac{1}{2\|\mathbf{a}^i\|_2}$;
5: Calculate the block diagonal matrix $\mathbf{W}^i$ $(1 \le i \le s)$ with the jth diagonal block as $\frac{1}{2\|\mathbf{a}_i^j\|_2}\mathbf{I}_j$;
6: For each $\mathbf{a}_i$ $(1 \le i \le s)$, 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** $\mathbf{A} \in \mathbb{R}^{n\times s}$.

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

Proof in "Appendix". □

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

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

Proof in "Appendix". □

Since the optimization problem in Eq. 8 is convex, Algorithm 1 converges to the global optima. In each iteration of our algorithm, computing Steps 3–5 is trivial. We could compute Step 6 by solving a system of linear equations with a quadratic complexity without the matrix inverse operation. The complexity of Algorithm 1 can be further improved by dimension reduction techniques, e.g. PCA (Li et al. 2015) and its derivations (Han et al. 2018).

4 Experimental results

Extensive experiments were conducted to evaluate the performance of our ROMS algorithm on place recognition for loop closure detection. In this section, our implementations are discussed firstly. Then, the experimental results using three public benchmark datasets are presented and analyzed.

4.1 Experiment setup

Three large-scale public benchmark datasets were used for validation in different conditions with various time spans. A summary of the used dataset statistics is presented in Table

1. Four types of visual features were employed in our experiments for all three datasets, including LDB features (Arroyo et al. 2015) applied on 64 × 64 downsampled images, GIST features (Latif et al. 2014) applied on 320 × 240 downsampled images, CNN-based deep features (Ren et al. 2015) applied on 227 × 227 downsampled images, and ORB local features (Mur-Artal et al. 2015) extracted from 320 × 240 downsampled images. These features are concatenated into a final vector to represent scene templates and query observations.

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 (Naseer et al. 2015; Sünderhauf et al. 2015), 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 (Labbe and Michaud 2014), 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 techniques, including BRIEF-GIST (Sünderhauf and Protzel 2011), FAB-MAP (Cummins and Newman 2009) [using the implementation from OpenFABMAP v2.0 (Glover et al. 2012)], and SeqSLAM (Milford and Wyeth 2012) [using the OpenSeqSLAM implementation (Sünderhauf et al. 2013)].

4.2 Results on the St Lucia dataset (various times of the day)

The *St Lucia dataset* (Glover et al. 2010) 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 2-week period. Each data instance includes a video of 20–25 min. 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 1.

Loop closure detection results over the St Lucia dataset are illustrated in Fig. 2. In Fig. 2a, an example sequence of multimodal features for the place representation is illustrated. The quantitative performance is evaluated using a standard precision-recall curve, as shown in Fig. 2c. 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. 2b. We show the template image (left column of Fig. 2b) that has the maximum weight for a query image (right column of Fig. 2b) 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.

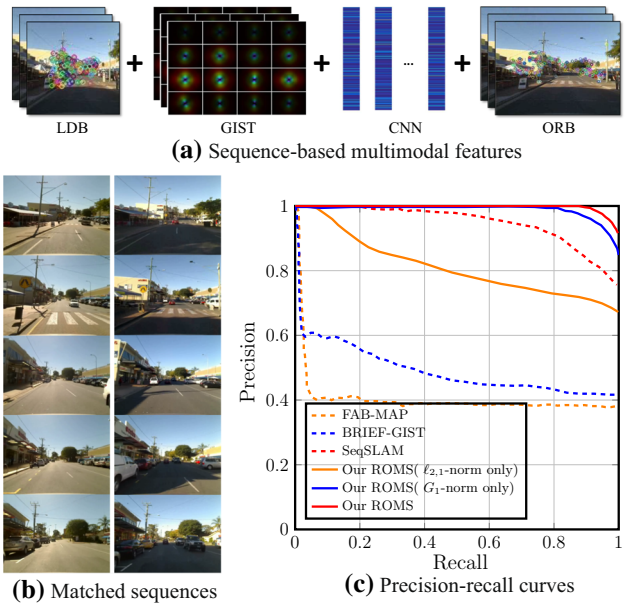
Comparisons with some of the main state-of-the-art methods are also graphically presented in Fig. 2c. It is observed that for long-term loop closure detection, sequence-based methods, such as our ROMS algorithms with G_1-norms and SeqSLAM, outperform the methods based on individual image matching, including 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 2, which indicates the percentage of loop closures that can be recognized without any false positives (Cummins and Newman 2009). 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. 2c. As indicated by Table 2, our ROMS loop closure detection algorithm achieves the best recall of 65.31% with a perfect precision.

4.3 Results on the CMU-VL dataset (different months)

The *CMU Visual Localization (VL) dataset* (Badino et al. 2012) was gathered using two cameras installed on a car that traveled the same route five times in Pittsburgh areas in the

Table 1 Statistics and scenarios of the public benchmark datasets used for algorithm validation in our experiments

Dataset	Sequence	Image statistics	Scenario
St Lucia (Glover et al. 2010)	10 × 12 km	10× ∼ 22,000 frames, 640 × 480 at 15 FPS	Different times of the day
CMU-VL (Badino et al. 2012)	5 × 8 km	5× ∼ 13,000 frames, 1024 × 768 at 15 FPS	Different months
Nordland (Sünderhauf et al. 2013)	4 × 728 km	4× ∼ 900,000 frames, 1920 × 1080 at 25 FPS	Different seasons



(a) Sequence-based multimodal features

(b) Matched sequences

(c) Precision-recall curves

Fig. 2 Experimental results over the St Lucia dataset. **a** illustrates an example sequence of multimodal features used for the place representation. **b** 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. **c** 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 (**c**). The figures are best seen in color (Color figure online)

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 5 s. An example sequence of multimodal features for the place representation is illustrated in Fig. 3a. The qualitative results in Fig. 3b show the template images (left column) with the maximum weight for each query image (right 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. 3c 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 2.

Figure 3c 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. 3c 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).

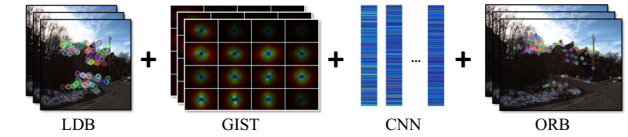
4.4 Results on the Nordland dataset (different seasons)

The *Nordland dataset* (Sünderhauf et al. 2013) contains visual data from a ten-hour long journey of a train traveling around 3000 km, which was recorded in four seasons from

Table 2 Comparison of used sequence-based loop closure detection methods using the metric of recall (%) at 100% precision

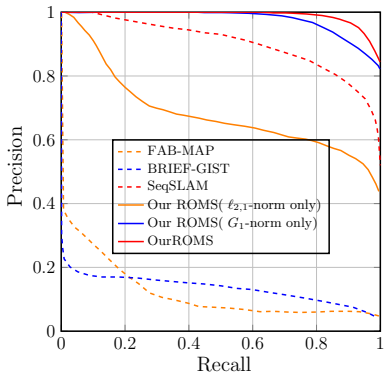
Methods	St Lucia	CMU-VL	Nordland
SeqSLAM (Milford and Wyeth 2012; Sünderhauf et al. 2013)	32.25	12.83	16.26
ROMS ($\ell_{2,1}$-norm only)	31.81	2.54	4.83
ROMS (G_1-norm only)	52.55	50.17	36.92
Our ROMS algorithm	65.31	66.47	57.36

Approaches based on single image matching are not included here because they generally obtain a zero value



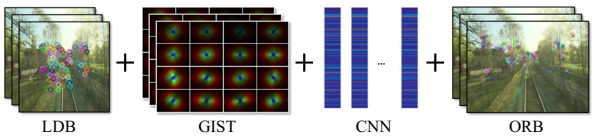
(a) Sequence-based multimodal features

(b) Matched sequences



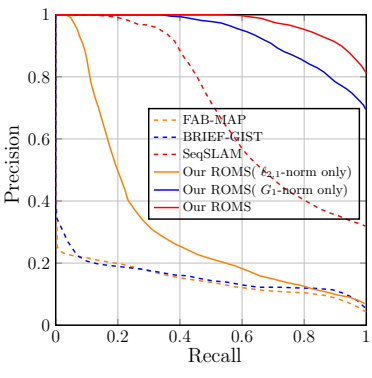
(c) Precision-recall curves

Fig. 3 Experimental results over the CMU-VL dataset. **a** illustrates an example sequence of multimodal features used for the place representation. **b** presents an example demonstrating the matched template and query sequences recorded in October and December, respectively. **c** illustrates the precision-recall curves and compares our methods with several previous loop closure detection approaches. The figures are best viewed in color (Color figure online)



(a) Sequence-based multimodal features

(b) Matched sequences



(c) Precision-recall curves

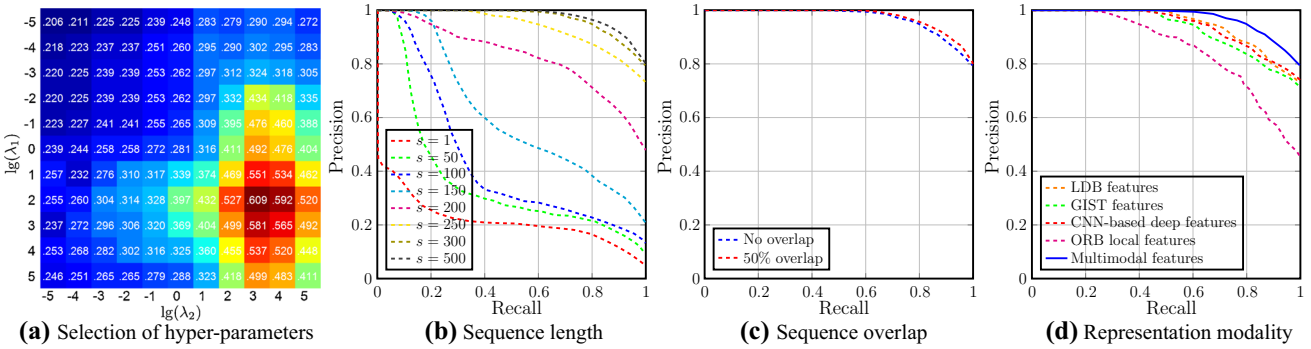
Fig. 4 Experimental results over the Nordland dataset. **a** illustrates an example sequence of multimodal features used for the place representation. **b** presents an example illustrating the matched scene template and query sequences recorded in Winter and Spring, respectively. **c** shows the precision-recall curves and compares our methods with previous loop closure detection methods. The figures are best viewed in color (Color figure online)

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.

Figure 4 presents the experimental results obtained by matching the spring data to the winter video in the Nordland dataset. Figure 2a illustrates an example sequence of multimodal features for the place representation. Figure 4b demonstrates several example images from one of the matched scene template (left column) and query (right column) sequences, each including 300 frames. Figure 4b validates that our ROMS algorithm can accurately match image sequences that contain dramatic appearance changes across seasons. Figure 4c 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 2.

4.5 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.



(a) Selection of hyper-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 (Color figure online)

The effect of the trade-off parameters used by our problem formulation in Eq. 8 is illustrated in Fig. 5a, 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. 5b. 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 5 s (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. 5b 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. 5c. 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. 5d. 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. An interesting future work can be focused on multi-feature learning to study the optimal combination of the features for long-term loop closure detection (Han et al. 2017).

5 Conclusion

In this paper, a novel robust loop closure detection method for long-term visual SLAM is proposed, which is formulated as a convex optimization problem with structured sparsity regularization. Our ROMS method enables to model the sparsity nature of place recognition, where only a small set of template sequences can be matched to the query sequence. Besides that, it is also able to model the grouping structure of template and query sequences, and incorporate multimodal features for discriminative scene representations. In order to solve the formulated non-smooth optimization problem efficiently, a new algorithm with the capability to converge to the global optima is also developed. The proposed ROMS method is finally evaluated by extensive experiments using 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.

Acknowledgements This work was partially supported by ARO W911NF-17-1-0447, NSF-IIS 1423591, and NSF-IIS 1652943.

Appendix Proof of Lemma 1:

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

Proof Obviously, the inequality $-(\|\tilde{\mathbf{a}}\|_2 - \|\mathbf{a}\|_2)^2 \leq 0$ holds. Thus, we have:

$$-(\|\tilde{\mathbf{v}}\|_2 - \|\mathbf{v}\|_2)^2 \leq 0 \Rightarrow 2\|\tilde{\mathbf{v}}\|_2\|\mathbf{v}\|_2 - \|\tilde{\mathbf{v}}\|_2^2 \leq \|\mathbf{v}\|_2^2$$
$$\Rightarrow \|\tilde{\mathbf{v}}\|_2 - \frac{\|\tilde{\mathbf{v}}\|_2^2}{2\|\mathbf{v}\|_2} \leq \|\mathbf{v}\|_2 - \frac{\|\mathbf{v}\|_2^2}{2\|\mathbf{v}\|_2}$$

This completes the proof. □

Proof of Theorem 1:

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

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

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

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

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

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

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

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

$$\begin{aligned} &\sum_{i=1}^{s} \left(\|\mathbf{D}\tilde{\mathbf{a}}_i - \mathbf{b}_i\|_2 - \frac{\|\mathbf{D}\tilde{\mathbf{a}}_i - \mathbf{b}_i\|_2^2}{2\|\mathbf{D}\mathbf{a}_i - \mathbf{b}_i\|_2} \right) \leq \sum_{i=1}^{s} \left(\|\mathbf{D}\mathbf{a}_i - \mathbf{b}_i\|_2 - \frac{\|\mathbf{D}\mathbf{a}_i - \mathbf{b}_i\|_2^2}{2\|\mathbf{D}\mathbf{a}_i - \mathbf{b}_i\|_2} \right) \\ &\sum_{i=1}^{s} \left(\|\tilde{\mathbf{a}}\|_2 - \lambda_1 \frac{\|\tilde{\mathbf{a}}\|_2^2}{2\|\mathbf{a}\|_2} \right) \leq \sum_{i=1}^{s} \left(\|\mathbf{a}\|_2 - \lambda_1 \frac{\|\mathbf{a}\|_2^2}{2\|\mathbf{a}\|_2} \right) \\ &\sum_{i=1}^{s} \sum_{j=1}^{k} \left(\|\tilde{\mathbf{a}}_i^j\|_2 - \frac{\|\tilde{\mathbf{a}}_i^j\|_2^2}{2\|\mathbf{a}_i^j\|_2} \right) \leq \sum_{i=1}^{s} \sum_{j=1}^{k} \left(\|\mathbf{a}_i^j\|_2 - \frac{\|\mathbf{a}_i^j\|_2^2}{2\|\mathbf{a}_i^j\|_2} \right) \end{aligned} \tag{15}$$

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. □

References

Angeli, A., Filliat, D., Doncieux, S., & Meyer, J. A. (2008). Fast and incremental method for loop-closure detection using bags of visual words. *IEEE Transactions on Robotics*, *24*(5), 1027–1037.

Arroyo, R., Alcantarilla, P., Bergasa, L., & Romera, E. (2015). Towards life-long visual localization using an efficient matching of binary sequences from images. In *IEEE international conference on robotics and automation*.

Badino, H., Huber, D., & Kanade, T. (2012). Real-time topometric localization. In *IEEE international conference on robotics and automation*.

Cadena, C., Carlone, L., Carrillo, H., Latif, Y., Scaramuzza, D., Neira, J., et al. (2016). Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age. *IEEE Transactions on Robotics*, *32*(6), 1309–1332.

Cadena, C., Gálvez-López, D., Tardós, J. D., & Neira, J. (2012). Robust place recognition with stereo sequences. *IEEE Transactions on Robotics*, *28*(4), 871–885.

Chen, C., & Wang, H. (2006). Appearance-based topological Bayesian inference for loop-closing detection in a cross-country environment. *The International Journal of Robotics Research*, *25*(10), 953–983.

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

Cummins, M., & Newman, P. (2009). Highly scalable appearance-only SLAM-FAB-MAP 2.0. In *Robotics: Science and systems*.

Estrada, C., Neira, J., & Tardós, J. D. (2005). Hierarchical SLAM: Real-time accurate mapping of large environments. *IEEE Transactions on Robotics*, *21*(4), 588–596.

Gálvez-López, D., & Tardós, J. D. (2012). Bags of binary words for fast place recognition in image sequences. *IEEE Transactions on Robotics*, *28*(5), 1188–1197.

Glover, A. J., Maddern, W. P., Milford M. J., & Wyeth, G. F. (2010). FAB-MAP + RatSLAM: Appearance-based SLAM for multiple times of day. In *IEEE international conference on robotics and automation*.

Glover, A., Maddern, W., Warren, M., Reid, S., Milford, M., & Wyeth, G. (2012). OpenFABMAP: An open source toolbox for appearance-based loop closure detection. In *IEEE international conference on robotics and automation*.

Goldberg, S. B., Maimone, M. W., & Matthies, L. (2002). Stereo vision and rover navigation software for planetary exploration. In *IEEE aerospace conference proceedings*.

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

Gutmann, J. S., & Konolige, K. (1999). Incremental mapping of large cyclic environments. In *IEEE international symposium on computational intelligence in robotics and automation*.

Han, F., Wang, H., & Zhang, H. (2018). Learning of integrated holism-landmark representations for long-term loop closure detection. In *AAAI conference on artificial intelligence*.

Han, F., Yang, X., Deng, Y., Rentschler, M., Yang, D., & Zhang, H. (2017). SRAL: Shared representative appearance learning for long-term visual place recognition. *IEEE Robotics and Automation Letters*, *2*(2), 1172–1179.

Hansen, P., & Browning, B. (2014). Visual place recognition using HMM sequence matching. In *IEEE/RSJ international conference on intelligent robots and systems*.

Henry, P., Krainin, M., Herbst, E., Ren, X., & Fox, D. (2012). 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.

Ho, K. L., & Newman, P. (2007). Detecting loop closure with scene sequences. *International Journal of Computer Vision*, *74*(3), 261–286.

Johns, E., & Yang, G. Z. (2013). Feature co-occurrence maps: Appearance-based localisation throughout the day. In *IEEE international conference on robotics and automation*.

Kleiner, A., & Dornhege, C. (2007). Real-time localization and elevation mapping within urban search and rescue scenarios. *Journal of Field Robotics*, *24*(8–9), 723–745.

Klopschitz, M., Zach, C., Irschara, A., & Schmalstieg, D. (2008). Generalized detection and merging of loop closures for video sequences. In *3D data processing, visualization, and transmission*.

Labbe, M., & Michaud, F. (2013). Appearance-based loop closure detection for online large-scale and long-term operation. *IEEE Transactions on Robotics*, *29*(3), 734–745.

Labbe, M., & Michaud, F. (2014). Online global loop closure detection for large-scale multi-session graph-based SLAM. In *IEEE/RSJ international conference on intelligent robots and systems*.

Latif, Y., Cadena, C., & Neira, J. (2013). Robust loop closing over time for pose graph SLAM. *The International Journal of Robotics Research*, *32*, 1611–1626.

Latif, Y., Huang, G., Leonard, J., & Neira, J. (2014). An online sparsity-cognizant loop-closure algorithm for visual navigation. In *Robotics: Science and systems conference*.

Li, S., Huang, H., Zhang, Y., & Liu, M. (2015). An efficient multi-scale convolutional neural network for image classification based on PCA. In *International conference on real-time computing and robotics*.

Lowry, S., Sünderhauf, N., Newman, P., Leonard, J. J., Cox, D., Corke, P., et al. (2016). Visual place recognition: A survey. *IEEE Transactions on Robotics*, *32*, 1.

Milford, M. J., & Wyeth, G. F. (2012). SeqSLAM: Visual route-based navigation for sunny summer days and stormy winter nights. In *IEEE international conference on robotics and automation*.

Milford, M. J., Wyeth, G. F., & Rasser, D. (2004). RatSLAM: A hippocampal model for simultaneous localization and mapping. In *IEEE international conference on robotics and automation*.

Mur-Artal, R., Montiel, J. M. M., & Tardos, J. D. (2015). ORB-SLAM: A versatile and accurate monocular SLAM system. *IEEE Transactions on Robotics*, *31*(5), 1147–1163.

Mur-Artal, R., & Tardós, J. D. (2014). Fast relocalisation and loop closing in keyframe-based SLAM. In *IEEE international conference on robotics and automation*.

Naseer, T., Ruhnke, M., Stachniss, C., Spinello, L., & Burgard, W. (2015). Robust visual SLAM across seasons. In *IEEE/RSJ international conference on intelligent robots and systems*.

Naseer, T., Spinello, L., Burgard, W., & Stachniss, C. (2014). Robust visual robot localization across seasons using network flows. In *AAAI conference on artificial intelligence*.

Nie, F., Huang, H., Cai, X., & Ding, C. H. (2010). Efficient and robust feature selection via joint $\ell_{2,1}$-norms minimization. In *Advances in neural information processing systems*.

Pepperell, E., Corke, P., & Milford, M. J. (2014). All-environment visual place recognition with SMART. In *IEEE international conference on robotics and automation*.

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*.

Santos, J. M., Couceiro, M. S., Portugal, D., & Rocha, R. P. (2015). A sensor fusion layer to cope with reduced visibility in SLAM. *Journal of Intelligent & Robotic Systems*, *80*(3), 401–422.

Sünderhauf, N., Neubert, P., & Protzel, P. (2013). 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*.

Sünderhauf, N., & Protzel, P. (2011). BRIEF-Gist—closing the loop by simple means. In *IEEE/RSJ international conference on intelligent robots and systems*.

Sünderhauf, N., Shirazi, S., Jacobson, A., Dayoub, F., Pepperell, E., Upcroft, B., & Milford, M. (2015). ConvNet landmarks: Viewpoint-robust, condition-robust, training-free. In *Robotics: Science and systems*.

Thrun, S., Burgard, W., & Fox, D. (2000). A real-time algorithm for mobile robot mapping with applications to multi-robot and 3D mapping. In *IEEE international conference on robotics and automation*.

Thrun, S., & Leonard, J. J. (2008). Simultaneous localization and mapping. In B. Siciliano & O. Khatib (Eds.), *Springer handbook of robotics* (pp. 871–889). Berlin: Springer.

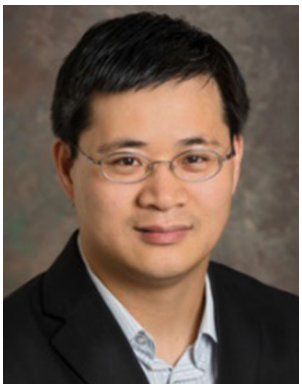
Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B (Methodological)*, *58*, 267–288.

Wang, H., Nie, F., & Huang, H. (2013). Multi-view clustering and feature learning via structured sparsity. In *International conference on machine learning*.

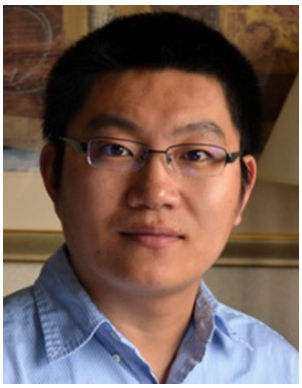
Zhang, H., Han, F., & Wang, H. (2016). Robust multimodal sequence-based loop closure detection via structured sparsity. In *Robotics: Science and systems*.

Fei Han received the Ph.D. degree in Mechanical Engineering from the City University of Hong Kong in 2014, and the B.E. degree in Automation from the University of Science and Technology of China in 2009. He is currently pursuing another Ph.D. degree in Computer Science at Colorado School of Mines. His research interests include computer vision and machine learning in robotics, decision making, human-robot teaming and nonlinear control systems.

Hua Wang is an Assistant Professor at the Department of Computer Science at Colorado School of Mines. He received the Ph.D. degree in Computer Science from the University of Texas at Arlington in 2012. Before that, he received the Bachelor's degree from Tsinghua University, China in 1999 and the Master's degree from Nanyang Technological University, Singapore in 2003. His research interests include machine learning and data mining, as well as their applications in robotics, bioinformatics, health informatics, medical image analysis, computer vision and cheminformatics.



Guoquan Huang is an Assistant Professor in Mechanical Engineering at the University of Delaware. He received the B.E. degree in automation (Electrical Engineering) from the University of Science and Technology, Beijing, China, in 2002, and the M.Sc. and Ph.D. degrees in computer science (robotics) from the University of Minnesota, Twin Cities, in 2009 and 2012, respectively. His research interests include robotics, computer vision and robot learning, with special emphasis on probabilistic perception, estimation, and control of autonomous ground, aerial, and underwater vehicles.



Hao Zhang received the Ph.D. degree in Computer Science from the University of Tennessee, Knoxville in 2014, the M.S. in Electrical Engineering from the Chinese Academy of Sciences in 2009, and the B.S. in Electrical Engineering from the University of Science and Technology of China in 2006. He is currently an Assistant Professor in the Department of and Computer Science at Colorado School of Mines. His research interests include human-robot teaming, robot learning and adaptation, multisensory perception, and robot decision making.