The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

Learning Integrated Holism-Landmark Representations for Long-Term Loop Closure Detection

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

Abstract

Loop closure detection is a critical component of large-scale simultaneous localization and mapping (SLAM) in loopy environments. This capability is challenging to achieve in long-term SLAM, when the environment appearance exhibits significant long-term variations across various time of the day, months, and even seasons. In this paper, we introduce a novel formulation to learn an integrated long-term representation based upon both holistic and landmark information, which integrates two previous insights under a unified framework: (1) holistic representations outperform keypoint-based representations, and (2) landmarks as an intermediate representation provide informative cues to detect challenging locations. Our new approach learns the representation by projecting input visual data into a low-dimensional space, which preserves both the global consistency (to minimize representation error) and the local consistency (to preserve landmarks' pairwise relationship) of the input data. To solve the formulated optimization problem, a new algorithm is developed with theoretically guaranteed convergence. Extensive experiments have been conducted using two large-scale public benchmark data sets, in which the promising performances have demonstrated the effectiveness of the proposed approach.

Introduction

Loop closure detection, also referred to as place recognition, has been an active research field over the past decades in simultaneous localization and mapping (SLAM) (Sünderhauf and Protzel 2011; Zhang, Han, and Wang 2016; Latif et al. 2017; Zhang, Lilly, and Vela 2016) and structure from motion (Lynen, Bosse, and Siegwart 2016; Cao and Snavely 2014). The purpose of loop closure detection is to identify a previously visited location, where the matched locations can be used later to eliminate or reduce the uncertainty and ambiguity in the constructed map, thus improving positioning and mapping performance.

In recent several years, motivated by autonomous driving, *long-term* place recognition has been attracting a significant attention to improve accuracy and reliability of outdoor localization during long-term operations (Lowry et al. 2016; Labbe and Michaud 2013). Consider the scenario when a self-driving vehicle operates during the whole year. In this long-term autonomy, the same location can look very differently at various time of the day (*e.g.*, caused by illumination changes) and different months and seasons because of long-term appearance changes such as vegetation changes (*e.g.*, trees with or without leaves) and long-term weather changes (*e.g.*, a place covered by snow in winter or not in summer). The long-term appearance change is widely recognized as one of the biggest challenge to enable long-term loop closure detection.

Given its significance, the problem of long-term loop closure detection has been actively studied (Lowry et al. 2016; Sünderhauf, Neubert, and Protzel 2013; Han et al. 2017; Linegar, Churchill, and Newman 2016). A well received insight obtained in the previous research is that representations based upon keypoints, such as SIFT (Lowe 2004), generally cannot work well when the environment exhibits long-term

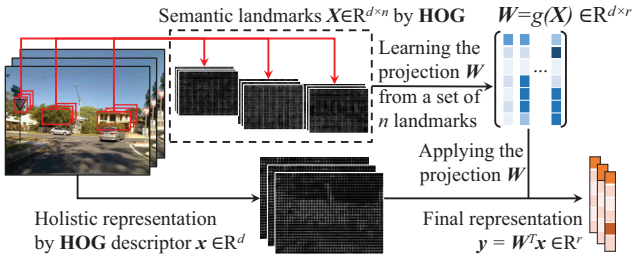


Figure 1: Overview of the proposed method to learn a representation that integrates holistic information and landmark relationship for improved long-term loop closure detection. Our new representation is constructed by unsupervisedly learning a projection that encodes both global and local consistencies, where the global consistency is used to preserve similar distribution of data points during the projection, and the local consistency is developed to preserve the relationship of landmarks. In this illustration, the red bounding boxes denote the *landmarks*, which are represented by a collection of landmark descriptors $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]$. The *holistic* feature vector of the entire frame is described by another holistic descriptor $\mathbf{x}$. A projection parameterized by $\mathbf{W}$ is learned from $\mathbf{X}$ to project $\mathbf{x}$ into a lower-dimensional subspace by computing $\mathbf{y} = \mathbf{W}^T\mathbf{x}$, which thereby integrates both holistic cues and landmarks of an input frame.

perceptual changes, and *holistic representations* (*e.g.*, based on HOG (Felzenszwalb, McAllester, and Ramanan 2008)) are necessary to encode long-term changes (Han et al. 2017; Wu and Rehg 2011). More recently, several studies (Yuan, Chan, and Lee 2011; Sunderhauf et al. 2015) investigated long-term loop closure detection by utilizing semantic *landmarks* (*e.g.*, buildings and trees) in the scene as an intermediate representation. Despite the promising performance, previous landmark-based methods (Yuan, Chan, and Lee 2011; Sunderhauf et al. 2015) cannot preserve the relationship between the landmarks. In addition, no method in literature thus far can integrate the both insights (*i.e.*, holism-based and landmark-based representations) for loop closure detection with long-term appearance changes; that is, previous techniques on building holistic representations are generally not able to encode semantic landmarks, and vise versa.

In this paper, we introduce a novel approach to unsupervisedly learn a long-term representation that integrates both landmark and holistic information to improve the encoding power to address long-term perceptual changes for loop closure detection in long-term autonomy. Our new approach learns a projection from the semantic landmarks within the scene to a low-dimensional representation space, which preserves both the global consistency (*i.e.*, data distribution in the projected subspace is similar to the raw data distribution) and the local consistency (*i.e.*, pairwise distance within a set of neighboring landmarks in the raw data is also preserved in the projected space) of an input frame. Then, this projection is used to project the holistic cues obtained from the entire scene to the newly learned representation space, thus to incorporate holistic information with landmarks embedded in the projection to construct an integrated holism-landmark based representation. Since our approach incorporates the insights of landmark and holism — two valuable insights to improve long-term loop closure detection, as highlighted in Figure 1, we name it *Holism-And-Landmark Integration* (HALI).

The contributions of this paper are threefold:

- We propose the new HALI approach to learn a representation in an unsupervised fashion that integrates both landmark and holistic cues, which provides a more descriptive representation of long-term perceptual variations for loop closure detection in long-term autonomy.
- We introduce a new formulation for representation construction, which projects raw data into a low-dimensional space, while preserving a similar data distribution in the learned projected space (*i.e.*, global consistence) and simultaneously preserving landmark relationship (*i.e.*, local consistency) during the projection.
- We implement an efficient iterative algorithm to solve the formulated optimization problem, whose convergence is theoretically guaranteed by the rigorous proofs.

Extensive experiments are performed over two public long-term loop closure detection datasets to evaluate HALI's performance on loop closure detection across different months and seasons respectively, which have demonstrated the improved performance resulted from our new approach.

The HALI Approach

In this section, we discuss the formulation of our new HALI approach to integrate the holism-landmark insights to improve long-term loop closure detection, which is achieved through learning a projection that incorporates both global consistency (*i.e.*, preserving similar data distribution) and local consistency (*i.e.*, preserving landmark relationship) of an input frame.

General notations used in the paper are defined as follows. In this paper, we write matrices as bold uppercase letters and vectors as bold lowercase letters. Given a sqaure matrix $\mathbf{M} = [m_{ij}] \in \Re^{n\times n}$, its trace is defined as $\mathbf{tr}(\mathbf{M}) = \sum_{i=1}^{n} m_{ii}$. Given a positive constant $p \geq 1$, the ℓ_p-norm of the vector $\mathbf{v} \in \Re^d$ is defined as $\|\mathbf{v}\|_p = \left(\sum_{i=1}^{d} (|v_i|^p)\right)^{\frac{1}{p}}$.

Problem Formalization

Our goal is to learn a representation of places/locations that is robust to appearance variations in the long-term duration, which we achieve by our new formulation that integrates both holistic information and semantic landmarks.

Formally, we assume that we have a collection of training images $\{\mathcal{X}\}$ recorded in different scenarios (*e.g.*, different time of a day, month, and season). Each image can be denoted as $\mathcal{X} = \{\mathbf{x}, \mathbf{X}\}$, where $\mathbf{x} \in \Re^d$ denotes its holistic representation and $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbb{R}^{d\times n}$ denotes a collection of n sematic landmarks within this image, respectively. Here $\mathbf{x}_i \in \Re^d$ is the feature vector of a semantic landmark within the image (shown as a red box in Figure 1).

Given all the training images $\{\mathcal{X}\}$, our goal is to learn each image $\mathcal{X}$ an integrated representation of $\mathbf{y} = f(\mathcal{X})$ that captures both landmark and holistic information conveyed by the image, which will be then used for matching locations during testing. Ideally, the same place should have the identical representation, even if those perception data is captured in different scenarios in long-term autonomy, which motivates us to find a projection that maps similar places into similar representations while still keeping their capability to discriminate different scenarios. Since landmarks possess semantic meanings that have shown promising performance in the literature to improve loop closure detection, we propose to learn the projection $g(\cdot)$ from the n semantic landmarks $\{\mathbf{x}_i\}_{i=1}^{n}$ within the scenes captured in different scenarios. After that, the learned projection is applied to the holistic representation of each image to obtain the final integrated representation $\mathbf{y} = f(\mathcal{X}) = h(g(\mathbf{X}), \mathbf{x})$, which thereby captures both landmark and holistic cues of each image. To implement this formulation, we propose to learn the projection $g(\cdot)$ that encodes both global and local consistencies, which is detailed in the following subsection.

Representation Learning through Projection

We focus on developing an objective function to learn a representation projection $\mathbf{W} = g(\cdot)$ that preserves the global and local consistencies of semantic landmarks.

The first objective is to learn a projection from the input feature space to a subspace that preserves as much information as possible. In practice, since directly processing high-

dimensional input features (*e.g.*, extracted from camera observations) often suffers from the "curse of dimensionality", we expect the projected subspace has a lower dimensionality. As a result, learning a low-dimensional subspace while maintaining geometrical structures of the original input (*i.e.*, distribution of the projected data in the subspace is similar to the original data distribution) is desired for practical use. To achieve this objective, we propose to learn a linear projection $\mathbf{W} \in \Re^{d\times r}$ from the landmarks $\mathbf{X}$ of a training image, which maps the holistic representation $\mathbf{x}$ in the high d-dimensional space into a vector $\mathbf{y}$ in a lower r-dimensional space by computing $\mathbf{y} = \mathbf{W}^T\mathbf{x}$ where $r < d$, such that the overall distribution of the input data in the projected spaced $\Re^r$ is preserved.

Formally, given the mean vector of the input feature vectors $\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i$, we compute the projection $\mathbf{W}$ by maximizing the following objective:

$$\mathcal{J}_{\text{Global}}\left(\mathbf{W}\right) = \mathbf{tr}\left(\mathbf{W}^T\mathbf{S}_G\mathbf{W}\right) = \sum_{i=1}^{n}\left\|\mathbf{W}^T\left(\mathbf{x}_i - \bar{\mathbf{x}}\right)\right\|_2^2 \ , \quad s.t.\ \ \mathbf{W}^T\mathbf{W} = \mathbf{I} \ , \tag{1}$$

where $\mathbf{S}_G = \sum_{i=1}^{n}\left(\mathbf{x}_i - \bar{\mathbf{x}}\right)\left(\mathbf{x}_i - \bar{\mathbf{x}}\right)^T$ measures the covariance of $\mathbf{X}$. Here, the constant factor $\frac{1}{n}$ is removed for brevity. Because maximizing $\mathcal{J}_{\text{Global}}$ enforces that the distribution of the data points in the projected subspace to be same as that in the original space, the learned projection $\mathbf{W}$ is globally consistent, which preserves geometrical data distribution during the projection. Similar objectives are often seen in traditional dimension reduction techniques, such as Principal Component Analysis (PCA)

Although the objective function of $\mathcal{J}_{\text{Global}}$ in Eq. (1) preserves data distribution when computing the projection, it is incapable of incorporating the relationship of landmarks in the projected subspace. We propose to enable this capability by considering the local consistency of landmarks. Ideally, the landmarks with similar semantics in the learned subspace should be close to each other. Thus, beyond maximizing $\mathcal{J}_{\text{Global}}$ to enforce similar data distribution in the projected subspace, we also want to minimize the local variance around every landmark in the learned subspace via projection (*i.e.*, local consistency). Formally, given that we use K-nearest neighbors to define the locality of each $\mathbf{x}_i$, which is written as $\mathcal{N}_i$ (where K is the number of neighbors contained in $\mathcal{N}_i$), and represent the mean vector of the neighbors of $\mathbf{x}_i$ as $\bar{\mathbf{x}}_i = \frac{1}{K+1}\sum_{\mathbf{x}_j\in\mathcal{N}_i\cup\{\mathbf{x}_i\}}\mathbf{x}_j$, we want to achieve the overall local consistency by minimizing the following objective (Wang, Nie, and Huang 2015):

$$\begin{aligned}\mathcal{J}_{\text{Local}}\left(\mathbf{W}\right) &= \mathbf{tr}\left(\mathbf{W}^T\mathbf{S}_L\mathbf{W}\right) \\ &= \sum_{i=1}^{n}\sum_{\mathbf{x}_j\in\mathcal{N}_i\cup\{\mathbf{x}_i\}}\left\|\mathbf{W}^T\left(\mathbf{x}_j - \bar{\mathbf{x}}_i\right)\right\|_2^2 \ , \\ s.t.\ \ &\mathbf{W}^T\mathbf{W} = \mathbf{I} \ ,\end{aligned} \tag{2}$$

where $\mathbf{S}_L = \sum_{i=1}^{n}\mathbf{S}_{Li}$ and

$$\mathbf{S}_{Li} = \sum_{\mathbf{x}_j\in\mathcal{N}_i\cup\{\mathbf{x}_i\}}\left(\mathbf{x}_j - \bar{\mathbf{x}}_i\right)\left(\mathbf{x}_j - \bar{\mathbf{x}}_i\right)^T \ . \tag{3}$$

Obviously, $\mathbf{S}_{Li}$ measures the local variance around $\mathbf{x}_i$, which we propose to minimize following our previous work (Wang, Nie, and Huang 2015). Similarly, the constant factor $\frac{1}{K+1}$ is omitted here for brevity. In Eq. (2) we did not use the objective of locality preserving projection (He and Niyogi 2004) to capture the locality of the input data due to its notorious performance sensitivity with respect to the parameters to construct graphs. The problem can be remedied in supervised and semi-supervised learning tasks through cross-validation, which, though, is usually not feasible in unsupervised representation learning since labels of the training images are not available in general. As a result, our objective in Eq. (2) is more advantageous in that it has less parameter and easier to fine tune (Wang, Nie, and Huang 2015).

Armed with the above two objectives that separately capture the global and local consistencies via learning a projection, we consider to learn the projection by simultaneously capturing the both consistencies. Among several possible ways to combine the two objectives, we formulate our new objective using the distance ratio that maximizes:

$$\mathcal{J}\left(\mathbf{W}\right) = \frac{\sum_{i=1}^{n}\left\|\mathbf{W}^T\left(\mathbf{x}_i - \bar{\mathbf{x}}\right)\right\|_1}{\sum_{i=1}^{n}\sum_{\mathbf{x}_j\in\mathcal{N}_i\cup\{\mathbf{x}_i\}}\left\|\mathbf{W}^T\left(\mathbf{x}_j - \bar{\mathbf{x}}_i\right)\right\|_1} \ , \quad s.t.\ \ \mathbf{W}^T\mathbf{W} = \mathbf{I} \ . \tag{4}$$

Note that, in long-term autonomy, the scene at the same location could change drastically at different time. This motivates us to use the ℓ_1-norm distances in the proposed objective in Eq. (4), rather than traditional squared ℓ_2-norm distances, to promote its robustness against both outlier data instances and outlier features (Gao 2008; Wright et al. 2009; Wang et al. 2013; Wang, Nie, and Huang 2014).

Upon solving the optimization problem in Eq. (4) (using Algorithm 2 detailed in the next section), the learned $\mathbf{W}$ not only preserves the global distribution of the input data in the learned subspace, but also preserves the geometric relationship of the landmarks. Given the holistic feature vector $\mathbf{x}$ of an input image, our approach computes its new representation by computing $\mathbf{y} = \mathbf{W}^T\mathbf{x}$ that integrates landmark information with holistic features, where the landmark relationship is embedded in the learned $\mathbf{W}$.

Place Recognition

Based on the learned representation that integrates landmark and holistic information, we then calculate a matching score between the query observation and each template image in the projected low-dimensional subspace. Following (Naseer et al. 2014; 2015; Han et al. 2017), we use the cosine similarity. Finally, we can determine whether two locations are matched by comparing the score with a user-defined threshold. Different to existing loop closure detection methods that use either holistic information or semantic landmarks, our HALI approach is superior in that it automatically learns an integrated representation that can capture the both insights. Our approach is more robust to strong appearance changes caused by long-term appearance variations and outliers, hence improving the accuracy of long-term place matching. It is worth noting that, although image-based place matching is used in this work, our new ap-

proach for representation learning can be well integrated with more sophisticated loop closure detection methods such as sequence-based matching.

Our Optimization Algorithm

The proposed objective in Eq. (4) maximizes the ratio of the summations of a number of ℓ_1-norm distances, which is difficult to efficiently solve in general. Thus, we derive an efficient iterative solution algorithm and prove its convergence in this subsection. As a theoretical contribution, the proposed solution algorithm is non-greedy in nature.

Solving a General Optimization Problem

We first study the following general optimization problem and derive an efficient algorithm to solve it.

$$\max_{v\in\mathcal{C}} \frac{f(v)}{g(v)} \quad where \;\; g(v)\geq 0 \;\; (\forall v\in\mathcal{C}) \;\; . \tag{5}$$

To solve the above optimization problem, we propose a simple, yet efficient, iterative algorithm as summarized in Algorithm 1, whose convergence can be proved by Theorem 1.

Algorithm 1: Solve the general optimization problem in Eq. (5).

1: Randomly initialize $v^{(0)}\in\Omega$ and set $t=1$
2: Calculate $\lambda^{(t)}=\frac{f\left(v^{(t-1)}\right)}{g\left(v^{(t-1)}\right)}$
3: Find a $v^{(t)}\in\Omega$ satisfying $f\left(v^{(t)}\right)-\lambda^{(t)}g\left(v^{(t)}\right)\geq 0$
4: $t=t+1$, and goto Step 2 until convergence
5: **return** v

Theorem 1. *Algorithm 1 increases the objective in each iteration until converges.*

Proof. Because $\forall v\in\mathcal{C}\;\; g(v)>0$, according to Step 3 of Algorithm 1, we can derive $\frac{f(\mathbf{x}^{(t)})}{g(\mathbf{x}^{(t)})}\geq\lambda^{(t)}$. Step 2 of Algorithm 1 defines that $\lambda^{(t)}=\frac{f(\mathbf{x}^{(t-1)})}{g(\mathbf{x}^{(t-1)})}$. Thus, we have $\frac{f(\mathbf{x}^{(t)})}{g(\mathbf{x}^{(t)})}\geq\frac{f(\mathbf{x}^{(t-1)})}{g(\mathbf{x}^{(t-1)})}$, which completes the proof. □

Our Algorithm to Solve Eq. (4)

Because our new objective in Eq. (4) is a special case of the general optimization problem in Eq. (5), we can derive Algorithm 2 to solve Eq. (4), whose convergence is rigorously guaranteed by Algorithm 1 and Theorem 1.

Now we need solve the problem in Eq. (4) in Algorithm 2, for which we first introduce the following two lemmas.

Lemma 1. *(Liu et al. 2017, Theorem 1) For any vector $\boldsymbol{\xi}=[\xi_1,\cdots,\xi_m]^T\in\Re^m$, we have $\|\boldsymbol{\xi}\|_1=\max_{\boldsymbol{\eta}\in\Re^m}(\mathrm{sign}(\boldsymbol{\eta}))^T\boldsymbol{\xi}$, where the maximum value is attained if and only if $\boldsymbol{\eta}=a\times\boldsymbol{\xi}$, where $a>0$ is a scalar.*

Lemma 2. *(Jenatton, Obozinski, and Bach 2010, Lemma 3.1) For any vector $\boldsymbol{\xi}=[\xi_1,\cdots,\xi_m]^T\in\Re^m$, we have $\|\boldsymbol{\xi}\|_1=\min_{\boldsymbol{\eta}\in\Re^m_+}\frac{1}{2}\sum_{i=1}^m\frac{\xi_i^2}{\eta_i}+\frac{1}{2}\|\boldsymbol{\eta}\|_1$, where the minimum value is attained if and only if $\eta_j=|\xi_j|, j\in\{1,2,\cdots,m\}$.*

Algorithm 2: Solve the proposed objective in Eq. (4).

1: Randomly initialize $\mathbf{W}^{(0)}$ satisfying $\left(\mathbf{W}^{(0)}\right)^T\mathbf{W}^{(0)}=\mathbf{I}$ and set $t=1$
2: Calculate

$$\lambda^{(t)}=\frac{\sum_{i=1}^n\left\|\left(\mathbf{W}^{(t-1)}\right)^T(\mathbf{x}_i-\bar{\mathbf{x}})\right\|_1}{\sum_{i=1}^n\sum_{\mathbf{x}_j\in\mathcal{N}_i\cup\{\mathbf{x}_i\}}\left\|\left(\mathbf{W}^{(t-1)}\right)^T(\mathbf{x}_j-\bar{\mathbf{x}}_i))\right\|_1} \tag{6}$$

3: Find a $\mathbf{W}^{(t)}$ satisfying

$$\mathcal{Q}\left(\mathbf{W}^{(t)}\right)=\mathcal{J}_{\text{Global}}\left(\mathbf{W}^{(t)}\right)-\lambda^{(t)}\mathcal{J}_{\text{Local}}\left(\mathbf{W}^{(t)}\right)\geq 0 \tag{7}$$

by Algorithm 3
5: $t=t+1$, and goto Step 2 until convergence
6: **return** $\mathbf{W}$

According Lemma 1 and Lemma 2, to solve the problem in Eq. (4) we introduce the following function:

$$\begin{aligned}&\mathcal{L}\left(\mathbf{W}^{(t)},\mathbf{W}^{(t-1)}\right)\\&=\mathcal{F}\left(\mathbf{W}^{(t)},\mathbf{W}^{(t-1)}\right)-\lambda^{(t)}\mathcal{G}\left(\mathbf{W}^{(t)},\mathbf{W}^{(t-1)}\right)\;\;,\end{aligned} \tag{8}$$

where

$$\begin{aligned}&\mathcal{F}\left(\mathbf{W}^{(t)},\mathbf{W}^{(t-1)}\right)\\&=\sum_{m=1}^r(\mathbf{w}_m^{(t)})^T\mathbf{B}\,\mathrm{sign}\left(\mathbf{B}^T\left(\mathbf{w}_m^{(t-1)}\right)\right)\;\;,\end{aligned} \tag{9}$$

$$\begin{aligned}&\mathcal{G}\left(\mathbf{W}^{(t)},\mathbf{W}^{(t-1)}\right)\\&=\frac{1}{2}\sum_{m=1}^r\left(\mathbf{w}_m^{(t)}\right)^T\mathbf{A}_m\mathbf{w}_m^{(t)}+\left(\mathbf{w}_m^{(t-1)}\right)^T\mathbf{A}_m\mathbf{w}_m^{(t-1)}\;\;,\end{aligned} \tag{10}$$

where we denote $\mathbf{w}_m^{(t)}$ and $\mathbf{w}_m^{(t-1)}$ as the m-th column of matrices $\mathbf{W}^{(t)}$ and $\mathbf{W}^{(t-1)}$, respectively, and define

$$\mathbf{B}=[\bar{\mathbf{x}}_1-\bar{\mathbf{x}},\bar{\mathbf{x}}_2-\bar{\mathbf{x}},\cdots,\bar{\mathbf{x}}_n-\bar{\mathbf{x}}]\;\;, \tag{11}$$

$$\mathbf{A}_m=\sum_{i=1}^n\sum_{\mathbf{x}_j\in\mathcal{N}_i\cup\{\mathbf{x}_i\}}\frac{(\mathbf{x}_j-\bar{\mathbf{x}}_i)(\mathbf{x}_j-\bar{\mathbf{x}}_i)^T}{\left|\left(\mathbf{w}_m^{(t-1)}\right)^T(\mathbf{x}_j-\bar{\mathbf{x}}_i)\right|}\;\;. \tag{12}$$

In Eq. (9), $\mathrm{sign}(x)$ is the sign function.

Then we prove the following theorem.

Theorem 2. *For any $\mathbf{W}^{(t)}\in\Re^{d\times r}$ and $\mathbf{W}^{(t-1)}\in\Re^{d\times r}$, we have $\mathcal{L}(\mathbf{W}^{(t)},\mathbf{W}^{(t-1)})\leq\mathcal{Q}(\mathbf{W}^{(t)})$. The equality holds on if and only if $\mathbf{W}^{(t)}=\mathbf{W}^{(t-1)}$.*

Proof. According to Lemma 1, we can derive:

$$\begin{aligned}
&\mathcal{J}_{\text{Global}}\left(\mathbf{W}^{(t)}\right) = \sum_{i=1}^{n}\left\|\left(\mathbf{W}^{(t)}\right)^T(\mathbf{x}_i-\bar{\mathbf{x}})\right\|_1 \\
&= \sum_{m=1}^{r}\sum_{i=1}^{n}\left\|\left(\mathbf{W}^{(t)}\right)^T(\mathbf{x}_i-\bar{\mathbf{x}})\right\|_1 \\
&\geq \sum_{m=1}^{r}\sum_{i=1}^{n}\text{sign}\left((\mathbf{w}_m^{(t-1)})^T(\mathbf{x}_i-\bar{\mathbf{x}})\right)\left(\left(\mathbf{w}_m^{(t)}\right)^T(\mathbf{x}_i-\bar{\mathbf{x}}\right) \\
&= \sum_{m=1}^{r}\left(\mathbf{w}_m^{(t)}\right)^T\mathbf{B}\,\text{sign}\left(\mathbf{B}^T\left(\mathbf{w}_m^{(t-1)}\right)\right) \\
&= \mathcal{F}\left(\mathbf{W}^{(t)},\mathbf{W}^{(t-1)}\right) \ . \qquad (13)
\end{aligned}$$

According to Lemma 2, we can derive:

$$\begin{aligned}
&\sum_{i=1}^{n}\sum_{\mathbf{x}_j\in\mathcal{N}_i\cup\{\mathbf{x}_i\}}\frac{1}{2}\frac{\boldsymbol{\xi}^T(\mathbf{x}_j-\bar{\mathbf{x}}_i)(\mathbf{x}_j-\bar{\mathbf{x}}_i)^T\boldsymbol{\xi}}{\boldsymbol{\xi}^T(\mathbf{x}_j-\bar{\mathbf{x}}_i)} \\
&\qquad + \frac{1}{2}\left\|\boldsymbol{\xi}^T(\mathbf{x}_j-\bar{\mathbf{x}}_i)\right\|_1 \\
&\leq \sum_{i=1}^{n}\sum_{\mathbf{x}_j\in\mathcal{N}_i\cup\{\mathbf{x}_i\}}\frac{1}{2}\frac{\boldsymbol{\xi}^T(\mathbf{x}_j-\bar{\mathbf{x}}_i)(\mathbf{x}_j-\bar{\mathbf{x}}_i)^T\boldsymbol{\xi}}{\boldsymbol{\eta}^T(\mathbf{x}_j-\bar{\mathbf{x}}_i)} \\
&\qquad + \frac{1}{2}\left\|\boldsymbol{\eta}^T(\mathbf{x}_j-\bar{\mathbf{x}}_i)\right\|_1 \ ,
\end{aligned} \qquad (14)$$

which indicates that:

$$\begin{aligned}
&\mathcal{J}_{\text{Local}}\left(\mathbf{W}^{(t)}\right) = \sum_{i=1}^{n}\sum_{\mathbf{x}_j\in\mathcal{N}_i\cup\{\mathbf{x}_i\}}\left\|\left(\mathbf{W}^{(t)}\right)^T(\mathbf{x}_j-\bar{\mathbf{x}}_i)\right\|_1 \\
&= \sum_{m=1}^{r}\sum_{i=1}^{n}\sum_{\mathbf{x}_j\in\mathcal{N}_i\cup\{\mathbf{x}_i\}}\frac{1}{2}\frac{\left(\mathbf{w}_m^{(t)}\right)^T(\mathbf{x}_j-\bar{\mathbf{x}}_i)(\mathbf{x}_j-\bar{\mathbf{x}}_i)^T\mathbf{w}_m^{(t)}}{\left|\left(\mathbf{w}_m^{(t)}\right)^T(\mathbf{x}_j-\bar{\mathbf{x}}_i)\right|} \\
&\quad + \frac{1}{2}\left\|\left(\mathbf{w}_m^{(t)}\right)^T(\mathbf{x}_j-\bar{\mathbf{x}}_i)\right\|_1 \\
&\leq \sum_{m=1}^{r}\sum_{i=1}^{n}\sum_{\mathbf{x}_j\in\mathcal{N}_i\cup\{\mathbf{x}_i\}}\frac{1}{2}\frac{\left(\mathbf{w}_m^{(t)}\right)^T(\mathbf{x}_j-\bar{\mathbf{x}}_i)(\mathbf{x}_j-\bar{\mathbf{x}}_i)^T\mathbf{w}_m^{(t)}}{\left|\left(\mathbf{w}_m^{(t-1)}\right)^T(\mathbf{x}_j-\bar{\mathbf{x}}_i)\right|} \\
&\quad + \frac{1}{2}\left\|\left(\mathbf{w}_m^{(t-1)}\right)^T(\mathbf{x}_j-\bar{\mathbf{x}}_i)\right\|_1 \\
&= \frac{1}{2}\sum_{m=1}^{r}\left(\mathbf{w}_m^{(t)}\right)^T\mathbf{A}_m\mathbf{w}_m^{(t)} + \left(\mathbf{w}_m^{(t-1)}\right)^T\mathbf{A}_m\mathbf{w}_m^{(t-1)} \\
&= \mathcal{G}\left(\mathbf{W}^{(t)},\mathbf{W}^{(t-1)}\right) \ . \qquad (15)
\end{aligned}$$

According to the inequalities in Eq. (13) and Eq. (15), we easily can derive:

$$\begin{aligned}
&\mathcal{L}\left(\mathbf{W}^{(t)},\mathbf{W}^{(t-1)}\right) \qquad (16)\\
&= \mathcal{F}\left(\mathbf{W}^{(t)},\mathbf{W}^{(t-1)}\right) - \lambda^{(t)}\mathcal{G}\left(\mathbf{W}^{(t)},\mathbf{W}^{(t-1)}\right) \\
&\leq \mathcal{J}_{\text{Global}}\left(\mathbf{W}^{(t)}\right) - \lambda^{(t)}\mathcal{J}_{\text{Local}}\left(\mathbf{W}^{(t)}\right) = \mathcal{Q}\left(\mathbf{W}^{(t)}\right) \ ,
\end{aligned}$$

which completes the proof. □

Substituting $\mathbf{W}^{(t)} = \mathbf{W}^{(t-1)}$ into the function $\mathcal{L}\left(\mathbf{W}^{(t)},\mathbf{W}^{(t-1)}\right)$, we have:

$$\mathcal{L}\left(\mathbf{W}^{(t-1)},\mathbf{W}^{(t-1)}\right) = \mathcal{Q}\left(\mathbf{W}^{(t-1)}\right) = 0 \ . \qquad (17)$$

In the t-th iteration in solving the objective function in Eq. (4), the optimal solution $\mathbf{W}^\star$ satisfies

$$\mathcal{L}(\mathbf{W}^\star,\mathbf{W}^{k-1}) \geq \mathcal{L}(\mathbf{W}^{k-1},\mathbf{W}^{k-1}) = 0 \ . \qquad (18)$$

Then, we have:

$$\begin{aligned}
\mathcal{Q}(\mathbf{W}^\star) &\geq \mathcal{L}\left(\mathbf{W}^\star,\mathbf{W}^{(t-1)}\right) \\
&\geq \mathcal{L}(\mathbf{W}^{(t-1)},\mathbf{W}^{(t-1)}) \\
&= \mathcal{Q}(\mathbf{W}^{(t-1)}) = 0 \ .
\end{aligned} \qquad (19)$$

Theorem 2 and Eq. (19) indicate that the solution of the problem in Eq. (4) can be transformed to solve the problem of $\mathcal{L}\left(\mathbf{W}^{(t)},\mathbf{W}^{(t-1)}\right) \geq 0$, which can be solved by the projected subgradient method with Armigo line search (Sun and Yuan 2006). Thus we compute the subgradient of $\mathcal{L}\left(\mathbf{W}^{(t)},\mathbf{W}^{(t-1)}\right)$ at $\mathbf{W}^{(t)}$ as:

$$\begin{aligned}
\partial L(\mathbf{W}^{(t)},\mathbf{W}^{(t-1)}) &= \mathbf{B}\,\text{sign}\left(\mathbf{B}^T\mathbf{W}^{(t-1)}\right) \\
&\quad - \lambda^k\left[\mathbf{A}_1\mathbf{w}_1^{(t)},\mathbf{A}_2\mathbf{w}_2^{(t)},\cdots,\mathbf{A}_p\mathbf{w}_p^{(t)}\right].
\end{aligned} \qquad (20)$$

Note that for a given matrix $\mathbf{W}^{(t)}$, the operator $\mathcal{P}(\mathbf{W}^{(t)}) = \mathbf{W}^{(t)}\left(\left(\mathbf{W}^{(t)}\right)^T\mathbf{W}^{(t)}\right)^{-\frac{1}{2}}$ can project $\mathbf{W}^{(t)}$ onto an orthogonal cone. This guarantees the orthogonality constraint of the projection matrix, *i.e.*, $\left(\mathbf{W}^{(t)}\right)^T\mathbf{W}^{(t)} = \mathbf{I}$. Algorithm 3 summarizes the solution to the problem in Eq. (4).

Algorithm 3: Solve the optimization problem in Eq. (4).

Input : $\mathbf{W}^{(t-1)}$ and the parameter $0 < \beta < 1$

1: Calculate $\lambda^{(t)} = \sum_{i=1}^{n}\left\|(\mathbf{W}^{(t-1)})^T(\mathbf{x}_i-\bar{\mathbf{x}})\right\|_1 / \sum_{i=1}^{n}\sum_{\mathbf{x}_j\in\mathcal{N}_i\cup\{\mathbf{x}_i\}}\left\|(\mathbf{W}^{(t-1)})^T(\mathbf{x}_j-\bar{\mathbf{x}}_i))\right\|_1$, thus the subgradient $G^{(t-1)} = \partial L(\mathbf{W}^{(t-1)},\mathbf{W}^{(t-1)})$ and set $m = 1$

2: Calculate $\mathbf{W}^{(t)} = \mathcal{P}(\mathbf{W}^{(t-1)} + \beta^m G^{(t-1)})$

3: Calculate $\mathcal{Q}(\mathbf{W}^{(t)})$ by Eq. (4). If $\mathcal{Q}(\mathbf{W}^{(t)}) \geq 0$, then goto Step 4, otherwise $m = m + 1$ and goto Step 2

4: **return** $\mathbf{W}^{(t)}$

Experimental Results

Extensive experiments are performed to validate and evaluate the performance of our HALI approach over long-term loop closure detection, using two large-scale public datasets recorded in different long-term situations, including: CMU-VL (scenarios in different months) and Nordland (scenarios in different seasons) data sets.

In our experiments, a variety of feature extraction techniques are implemented to extract visual features from input frames, including: (1) color features (Lee, Kim, and Myung 2013) applied on downsampled visual frames, (2) BRIEF-GIST features (Latif et al. 2014) applied on downsampled frames, (3) Histogram of Oriented Gradients (HOG) features (Naseer et al. 2014) computed over downsampled images, (4) Local Binary Patterns (LBP) visual features (Qiao, Cappelle, and Ruichek 2015) applied on downsampled images, (5) Speeded Up Robust Features (SURF) (Badino, Huber, and Kanade 2012) applied on downsampled images, and (6) Deep features learned by Convolutional Neural Network (CNN) (Sunderhauf et al. 2015) applied on downsampled images. Given these raw holistic features, our HALI approach improves their representation power by incorporating landmark relationships and enforcing their local and global consistency.

Both qualitative and quantitative evaluations are conducted to evaluate the performance of HALI. In addition, several baseline and recent methods are compared in each experiment, including the BRIEF-GIST (Sünderhauf and Protzel 2011), Normalized Gradients (NormG) of grayscale images (used in SeqSLAM (Milford and Wyeth 2012)), and techniques based only upon color, LBP, HOG, SURF, CNN features. To demonstrated the performance improvement truly resulted from HALI, the simple image-based matching is intentionally implemented for location matching.

Throughout the experiments, the hyperparameter $K = 2$ was used, which is resulted from our sensitivity analysis for the hyperparameter value selection (presented at the end of this section). The projection matrix $\mathbf{W}$ is learned on patches containing landmarks or semantic objects coming from a separate held-back subset of the datasets during the training phase; then the learned projection is applied during the testing phase over a separate, previously unseen testing instances for validation and evaluation. For quantitative evaluation and comparison, following previous studies (Sunderhauf et al. 2015; Zhang, Han, and Wang 2016), we use precision-recall curves as a metric, which shows the trade-off between precision and recall for different threshold. A high area under the curve represents both high recall and high precision, where high precision relates to a low false positive rate, and high recall relates to a low false negative rate.

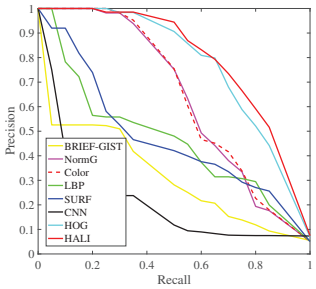
Results on CMU-VL Dataset (Different Months)

The CMU Visual Localization (CMU-VL) dataset (Badino, Huber, and Kanade 2012) was recorded from two monocular cameras installed on a vehicle that traveled the same route five times around Pittsburgh areas across different months with a variety of weather, environmental, and climatological conditions. The length of the single route is around 8 KM. There are five videos in the dataset, and each of them consists of around $13,000$ frames. The resolution of each frame in these video is 1024×768 and the frame rate is 15 FPS. The GPS information during the travels was also recorded, which is defined by the dataset author as the ground truth of localization.

We observe several challenges from the CMU-VL dataset,

(a) Examples of matched locations during testing

(b) Landmarks for training



(c) Precision-recall curves

Figure 2: Experiments on the CMU-VL dataset. Figure 2(a) shows several matched locations recorded in October (top) and December (bottom), respectively. Figure 2(b) illustrates several exemplary patches that contain landmarks used for training. Figure 2(c) presents the precision-recall curves that evaluate the performances of our HALI approach, as well the baseline and recent competing techniques. The figures are best viewed in color.

including urban scene changes due to constructions and dynamic objects, viewpoint changes caused by slight route deviations, and more importantly, long-term appearance variations due to weather, illumination, and vegetation changes in different months, as shown in Figure 2(a). To train our HALI approach, the same landmark or semantic objects recorded in different months are used to learn the optimal projection matrix $\mathbf{W}$. Figure 2(b) illustrates several exemplary patches containing the landmarks used for training over the CMU-VL dataset. In the CMU-VL experiment, objects commonly seen in urban areas were selected, such as stop signs, trees, houses, among others.

The qualitative results obtained by HALI on the CMU-VL dataset are presented in Figure 2(a), which illustrates several examples of matched frames recorded in October (in the top row) and December (in the bottom row) in the testing dataset. The matched frames were selected as the two images with the maximum similarity score that is computed by our approach. From the figure, we can observe the scenes in the same location in December and October exhibit significant appearance variations caused by different snow coverage and vegetation. In addition, we can observe that HALI is able to well address the long-term appearance change challenge and accurately match scene images to recognize same places across different months.

The quantitative results obtained by our approach in terms of precision-recall curves, and its comparisons with baseline and previous methods are illustrated in Figure 2(c). We observe from these results that representations based on HOG

features outperform representations based upon other types of holistic visual features, which is consistent with the conclusion drawn by (Han et al. 2017). Based upon this observation, we extract HOG features from landmarks as the input to our approach to create the proposed integrated representation. As shown in Figure 2(c), the resulted HALI approach outperforms the previous technique based on HOG features, which demonstrates the improved representation capability resulted from holism-landmark integration by our approach. Similarly, we performed comparisons using other visual features and observed similar performance improvement that is obtained by our HALI approach. In general, for any type of features that can encode each landmark into a fixed-length vector, HALI is able to adopt the vector as the input to construct a more descriptive representation through integrating the relationship of landmarks with the features.

Results on Nordland Dataset (Different Seasons)

The large-scale Nordland dataset (Sünderhauf, Neubert, and Protzel 2013) was collected in four different seasons from a ten-hour long journey of a train. The length of the route is 728 KM. This dataset contains four videos, and each of them includes around $900,000$ frames. The video has a resolution of 1920×1080 and the frame rate is 25 FPS. Though there is also GPS information together with the dataset, the image frames are perfectly aligned by the dataset author, which can be used as the ground truth of localization matching.

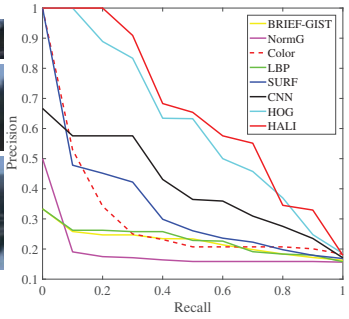
The scenes in the dataset exhibit significant long-term appearance changes due to various illumination, weather, and vegetation conditions in different seasons. For example, the ground is almost completely covered by snow during winter time in the dataset, while the ground remains green covered with grass and trees, as shown in Figure 3(a). In addition, a great number of locations in the wilderness on the trip exhibit similar appearances, which means there are strong perceptual aliasing issues in the Nordland dataset. These difficulties make the Nordland dataset one of the most challenging datasets for long-term loop closure detection.

In our experiments on the Nordland dataset, we select various landmarks and semantic objects to train our approach, such as railroad tracks, trees, houses, among others. Figure 3(b) shows several exemplary patches of landmarks used for learning the projection. The qualitative results obtained by HALI are illustrated in Figure 3(a), which presents several matched frames in Winter (in the top row) and Spring (in the bottom row), respectively. We observe that given the significant long-term appearance variation in Winter and Spring, which is mainly resulted from snow and different vegetation, HALI is able to well recognize same places across different seasons.

The quantitative results over the Nordland dataset are presented in Figure 3(c), which shows the precision-recall curve of our approach and the comparisons with various methods. Similar to our experiment on the CMU-VL dataset, we evaluate which holistic features can lead to good performance, and we obtain a consistent observation that the HOG features (which encodes shape information) works better than other holistic features. Then, we compare this baseline approach with our HALI approach that uses HOG as the feature extraction methods over the landmarks. As illustrated by Figure 3(c), the HALI approach further improves the performance over the baseline method that uses HOG only by modeling landmark relationship.

(a) Examples of matched locations during testing

(b) Training landmarks



(c) Precision-recall curves

Figure 3: Experiments on the Nordland dataset. Figure 3(a) illustrates several pairs of scenes recorded in Spring (top) and Winter (bottom), respectively. Figure 3(b) shows several examples of landmarks used to learn the projection matrix in our approach. Figure 3(c) demonstrates the precision-recall curves that indicate the quantitative performance of our approach. Comparisons with various techniques are also shown in Figure 3(c). The figures are best viewed in color.

Conclusion

In this paper, we introduce a novel approach under the optimization framework that integrates holistic information with landmarks to construct a unified representation to improve long-term loop closure detection. Our HALI approach is implemented through designing a new objective that enforces both global and local consistency of the input data points in the projected subspace, where the global consistency is designed to preserve similar distribution of data points during the project, and the local consistency is designed to preserve the relationship of the landmarks. To solve the challenging objective function in our formulation, we also implement a new optimization solver which possesses a theoretical guarantee to converge to the global optimal solution. Experiments are conducted based on two large-scale public benchmark datasets collected for long-term place recognition. Promising results have demonstrated performance improvement resulted from the proposed approach.

Acknowledgments

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

References

Badino, H.; Huber, D.; and Kanade, T. 2012. Real-time topometric localization. In *IEEE International Conference on Robotics and Automation*.

Cao, S., and Snavely, N. 2014. Minimal scene descriptions from structure from motion models. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Felzenszwalb, P.; McAllester, D.; and Ramanan, D. 2008. A discriminatively trained, multiscale, deformable part model. In *IEEE Conference on Computer Vision and Pattern Recognition*, 1–8.

Gao, J. 2008. Robust ℓ_1 principal component analysis and its bayesian variational inference. *Neural Computation* 20(2):555–572.

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

He, X., and Niyogi, P. 2004. Locality preserving projections. In *Advances in Neural Information Processing Systems*.

Jenatton, R.; Obozinski, G.; and Bach, F. 2010. Structured sparse principal component analysis. In *International Conference on Artificial Intelligence and Statistics*.

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

Latif, Y.; Huang, G.; Leonard, J.; and Neira, J. 2014. An online sparsity-cognizant loop-closure algorithm for visual navigation. In *Robotics: Science and Systems*.

Latif, Y.; Huang, G.; Leonard, J.; and Neira, J. 2017. Sparse optimization for robust and efficient loop closing. *Robotics and Autonomous Systems* 93:13–26.

Lee, D.; Kim, H.; and Myung, H. 2013. 2D image feature-based real-time RGB-D 3D SLAM. In *Robot Intelligence Technology and Applications*. 485–492.

Linegar, C.; Churchill, W.; and Newman, P. 2016. Made to measure: Bespoke landmarks for 24-hour, all-weather localisation with a camera. In *IEEE International Conference on Robotics and Automation*.

Liu, Y.; Gao, Q.; Miao, S.; Gao, X.; Nie, F.; and Li, Y. 2017. A non-greedy algorithm for ℓ_1-norm lda. *IEEE Transactions on Image Processing* 26(2):684–695.

Lowe, D. G. 2004. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision* 60(2):91–110.

Lowry, S.; Sünderhauf, N.; Newman, P.; Leonard, J. J.; Cox, D.; Corke, P.; and Milford, M. J. 2016. Visual place recognition: A survey. *IEEE Transactions on Robotics* 32(1):1–19.

Lynen, S.; Bosse, M.; and Siegwart, R. 2016. Trajectory-based place-recognition for efficient large scale localization. *International Journal of Computer Vision* 1–16.

Milford, M. J., and 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*.

Naseer, T.; Spinello, L.; Burgard, W.; and Stachniss, C. 2014. Robust visual robot localization across seasons using network flows. In *AAAI Conference on Artificial Intelligence*.

Naseer, T.; Ruhnke, M.; Stachniss, C.; Spinello, L.; and Burgard, W. 2015. Robust visual SLAM across seasons. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2529–2535.

Qiao, Y.; Cappelle, C.; and Ruichek, Y. 2015. Place recognition based visual localization using LBP feature and SVM. In *Advances in Artificial Intelligence and Its Applications*. 393–404.

Sun, W., and Yuan, Y.-X. 2006. *Optimization theory and methods: nonlinear programming*, volume 1. Springer Science & Business Media.

Sünderhauf, N., and Protzel, P. 2011. BRIEF-Gist – Closing the loop by simple means. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*.

Sunderhauf, N.; Shirazi, S.; Jacobson, A.; Dayoub, F.; Pepperell, E.; Upcroft, B.; and Milford, M. 2015. Place recognition with ConvNet landmarks: Viewpoint-robust, condition-robust, training-free. In *Robotics: Science and Systems*.

Sünderhauf, N.; Neubert, P.; and 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*.

Wang, H.; Nie, F.; Cai, W.; and Huang, H. 2013. Semi-supervised robust dictionary learning via efficient $\ell_{2,0^+}$-norms minimization. In *Proceedings of the IEEE International Conference on Computer Vision*, 1145–1152.

Wang, H.; Nie, F.; and Huang, H. 2014. Robust distance metric learning via simultaneous ℓ_1-norm minimization and maximization. In *International Conference on Machine Learning*, 1836–1844.

Wang, H.; Nie, F.; and Huang, H. 2015. Learning robust locality preserving projection via p-order minimization. In *AAAI*, 3059–3065.

Wright, J.; Ganesh, A.; Rao, S.; Peng, Y.; and Ma, Y. 2009. Robust principal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization. In *Advances in Neural Information Processing Systems*, 2080–2088.

Wu, J., and Rehg, J. M. 2011. CENTRIST: A visual descriptor for scene categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33(8):1489–1501.

Yuan, L.; Chan, K. C.; and Lee, C. G. 2011. Robust semantic place recognition with vocabulary tree and landmark detection. In *Workshop of IEEE/RSJ International Conference on Intelligent Robots and Systems*.

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

Zhang, G.; Lilly, M. J.; and Vela, P. A. 2016. Learning binary features online from motion dynamics for incremental loop-closure detection and place recognition. In *IEEE International Conference on Robotics and Automation*, 765–772.