Visual Place Recognition via Robust ℓ_2-Norm Distance Based Holism and Landmark Integration

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

Abstract

Visual place recognition is essential for large-scale simultaneous localization and mapping (SLAM). Long-term robot operations across different time of the days, months, and seasons introduce new challenges from significant environment appearance variations. In this paper, we propose a novel method to learn a location representation that can integrate the semantic landmarks of a place with its holistic representation. To promote the robustness of our new model against the drastic appearance variations due to long-term visual changes, we formulate our objective to use non-squared ℓ_2-norm distances, which leads to a difficult optimization problem that minimizes the ratio of the $\ell_{2,1}$-norms of matrices. To solve our objective, we derive a new efficient iterative algorithm, whose convergence is rigorously guaranteed by theory. In addition, because our solution is strictly orthogonal, the learned location representations can have better place recognition capabilities. We evaluate the proposed method using two large-scale benchmark data sets, the CMU-VL and Nordland data sets. Experimental results have validated the effectiveness of our new method in long-term visual place recognition applications.

During visual simultaneous localization and mapping (SLAM), especially in large-scale loopy environments, place recognition is essential to detect and close loops to reduced uncertainty for constructing maps and to improve localization accuracy (Lowry et al. 2016). Recently, driven by several critical applications of SLAM, such as self-driving, long-term place recognition has attracted increased attentions to improve outdoor localization when a robot (or an autonomous vehicle) needs to operate in a dynamic environment for long periods of time. Besides traditional perception difficulties, long-term place recognition introduces additional challenges mainly caused by significant visual appearance changes of the environment over time (Sünderhauf and Protzel 2011; Zhang, Han, and Wang 2016; Latif et al. 2017; Han et al. 2018). For example, when an autonomous vehicle drives through the same place at different time of the day (*e.g.*, noon *vs.* midnight), the same place can look very different due to dramatic illumination changes. Similarly, at different time and seasons, the changing weather (*e.g.*, sunny *vs.* snowy) and vegetation (*e.g.*, trees with or without leaves) during long-term autonomy cause the same challenge to place recognition.

Over the past several years, a number of methods have been proposed to address this critical problem of long-term place recognition (Lowry et al. 2016). From the perspective of features, these methods are based either on local (*e.g.*, SIFT) or global (*e.g.*, HOG or deep features) features; from the perspective of localization cues, these methods can be generally grouped into two categories, based on either holistic layouts or landmarks, respectively. The holistic layout of the environment is typically represented using global features (Han et al. 2017; Wu and Rehg 2011) that are learned or manually constructed to encode long-term changes. Very recently, several techniques take advantages of semantic landmarks (*e.g.*, traffic lights and buildings) within the environment as an intermediate representation to address long-term place recognition (Yuan, Chan, and Lee 2011; Sunderhauf et al. 2015). Although these methods using either holistic layouts or landmarks have shown promising place recognition capability, the research problem of how to integrate these two localization cues in a principled way has not yet been well addressed.

In this paper, we present a novel method to learn location representations. It first learns a projection from the semantic landmarks in a place image. Then it projects the holistic image representation of the same place into the learned subspace. As a result, the learned location representation simultaneously captures the information from both the semantic landmarks of the place and its holistic characterization. In the proposed objective to learn the projection from the landmarks of a place image, we aim to preserve both global and local consistencies of the landmarks in the projected subspace, which leads to an optimization problem that minimizes the ratio of the matrix traces. By taking into account the drastic visual variations at the same place over long period of time, we further develop the proposed objective by replacing the squared ℓ_2-norm distances used in most traditional learning models by the non-squared ℓ_2-norm distances, such that the robustness of the learned location representations against outlying visual characterizations is promoted (Gao 2008; Wright et al. 2009; Nie et al. 2011; Wang, Nie, and Huang 2013b; Nie et al. 2013; Wang, Nie, and Huang 2014b; Liu and Wang 2015; Liu et al. 2017; 2018; Liu and Wang 2018). The contributions of this paper

can be summarized as follows:

- We propose a new representation learning method to integrate semantic landmarks and holistic layouts, which is more descriptive and robust for place recognition. To explicitly address significant appearance changes over long periods of time, we formulate a new objective function to minimize the ratio of the $\ell_{2,1}$-norms of matrices.
- To solve the challenging optimization problem, we derive an efficient iterative algorithm with theoretically guaranteed convergence. It is worth noting that, the solution obtained by our algorithm is strictly orthogonal. As a result, our solution has a better data representation capability that leads to improved long-term place recognition results.

Problem Formulation and Our New Method

In this section, we propose to learn the location representation that is robust to visual appearance variations in long-term place recognition by integrating the holistic information and the semantic landmarks at a place.

We first introduce the notations used in the following of our paper. Given a matrix $\mathbf{M} = [m_{ij}] \in \Re^{d\times n}$, its Frobenius norm is denoted as $\|\mathbf{M}\|_{\mathrm{F}}$ and its $\ell_{2,1}$-norm is defined as $\|\mathbf{M}\|_{2,1} = \sum_i \left\|\mathbf{m}^i\right\|_2 = \sum_i(\sqrt{\sum_j m_{ij}^2})$. If $\mathbf{M}$ is a square matrix, its trace is defined as $\mathbf{tr}\left(\mathbf{M}\right) = \sum_{i=1}^n m_{ii}$.

Suppose that we have a set of training images for varied scenarios (*e.g.*, different seasons, months, angles, *etc.*). We denote each image in the training set as $\mathcal{X} = \{\mathbf{x}, \mathbf{X}\}$, where $\mathbf{x} \in \Re^d$ is the holistic representation of the image and $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \Re^{d\times n}$ denotes the n semantic landmarks in the same image. Here, $\mathbf{x}_i \in \Re^d$ denotes the feature vector of one semantic landmark.

Given the training images $\{\mathcal{X}\}$, our goal is to learn an integrated representation of $\mathbf{y} = f\left(\mathcal{X}\right)$ for each image $\mathcal{X}$, which captures both holistic and landmark information in the image and then can be used for place recognition in the test. Ideally, different images of the same place should have similar representations, although they could be captured from various scenarios over a long duration. To achieve this, we look for a projection to map similar places into similar representations in a new subspace. Meanwhile, we also hope that different scenarios can still be discriminated. Because semantic landmarks have shown promising performance to improve long-term place recognition, we learn the projection $g(\cdot)$ from landmarks within the scenes. Then the learned projection will be used to project the holistic representation of an image to obtain its integrated representation by computing $\mathbf{y} = f\left(\mathcal{X}\right) = h\left(g(\mathbf{X}), \mathbf{x}\right)$, which is expected to capture both holistic and landmark information. To begin with, we learn the projection $g(\cdot)$ to retain both global and local consistencies of semantic landmarks.

Learn to Integrate Holism and Landmarks

Our goal is to learn a projection from the original feature space to a subspace while preserving as much information as possible. To achieve this, we use the Globally and Locally consistent Unsupervised Projection (GLUP) method proposed in our previous work (Wang, Nie, and Huang 2014a). This method learns a linear projection $\mathbf{W} \in \Re^{d\times r}$ from the landmarks $\mathbf{X}$ in a training image, which will project the holistic representation $\mathbf{x}$ of the same image in 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 geometrical structures of the input data will be preserved after the projection. We present some details of the GLUP method (Wang, Nie, and Huang 2014a) as follows.

Mathematically, suppose we have $\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i$, we can learn the projection $\mathbf{W}$ by maximizing the following objective to retain as much information as possible following principal component analysis (Jolliffe 1986):

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

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, we remove the constant factor $\frac{1}{n}$ for brevity. Maximizing $\mathcal{J}_{\text{Global}}$ enforces geometrical structures of data in projected subspace to be close to that in the original space. Thus $\mathbf{W}$ helps preserve geometrical data distribution when we perform the projection.

On the other hand, landmarks with close semantics should be close to each other in the projected subspace. To this end, we minimize the local variance around every landmark in the learned subspace for better local consistency. We use $\mathcal{N}_i$ to denote K-nearest neighborhood of $\mathbf{x}_i$ to define its locality and use $\bar{\mathbf{x}}_i = \frac{1}{K+1}\sum_{\mathbf{x}_j\in\mathcal{N}_i\cup\{\mathbf{x}_i\}}\mathbf{x}_j$ to represent the mean vector of the neighborhood of $\mathbf{x}_i$. Then we can achieve the local consistency by minimizing the following objective:

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

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

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

Again, the constant factor $\frac{1}{K+1}$ is omitted here for brevity.

In light of the two objectives above that capture the global and local consistencies separately, we now turn to learn the projection by simultaneously capturing the consistencies of both by using following GLUP objective (Wang, Nie, and Huang 2014a):

$$\begin{aligned}\min_{\mathbf{W}^T\mathbf{W}=\mathbf{I}} &\frac{\mathbf{tr}\left(\mathbf{W}^T\mathbf{S}_L\mathbf{W}\right)}{\mathbf{tr}\left(\mathbf{W}^T\mathbf{S}_G\mathbf{W}\right)} \\ &= \frac{\sum_{i=1}^n \sum_{\mathbf{x}_j\in\mathcal{N}_i\cup\{\mathbf{x}_i\}} \left\|\mathbf{W}^T\left(\mathbf{x}_j - \bar{\mathbf{x}}_i\right)\right\|_2^2}{\sum_{i=1}^n \left\|\mathbf{W}^T\left(\mathbf{x}_i - \bar{\mathbf{x}}\right)\right\|_2^2} \\ &= \frac{\left\|\mathbf{W}^T\mathbf{A}\right\|_{\mathrm{F}}^2}{\left\|\mathbf{W}^T\mathbf{B}\right\|_{\mathrm{F}}^2} \ ,\end{aligned} \tag{4}$$

where each column of $\mathbf{B}$ is one $(\mathbf{x}_i - \bar{\mathbf{x}})$ and each column of $\mathbf{A}$ is one $(\mathbf{x}_j - \bar{\mathbf{x}}_i)$.

Despite its success in a variety of real-world applications, the GLUP objective in Eq. (4) uses squared ℓ_2-norm distances, which is notoriously know to be sensitive to outlying features and samples. However, in long-term autonomy, the scene at the same location could change drastically in different scenarios, which motivate us to further develop the GLUP objective in Eq. (4) to make it more robust against significant scene changes as follows.

First, we notice that the following equality holds when the constraint of $\mathbf{W}^T\mathbf{W} = \mathbf{I}$ is given:

$$\begin{aligned} &\left\|\mathbf{b}_i - \mathbf{W}\mathbf{W}^T\mathbf{b}_i\right\|_2^2 \\ &= \left(\mathbf{b}_i - \mathbf{W}\mathbf{W}^T\mathbf{b}_i\right)^T\left(\mathbf{b}_i - \mathbf{W}\mathbf{W}^T\mathbf{b}_i\right) \\ &= \mathbf{b}_i^T\mathbf{b}_i - 2\mathbf{b}_i^T\mathbf{W}\mathbf{W}^T\mathbf{b}_i + \mathbf{b}_i^T\mathbf{W}\mathbf{W}^T\mathbf{b}_i \\ &= \|\mathbf{b}_i\|_2^2 - \left\|\mathbf{W}^T\mathbf{b}_i\right\|_2^2 \ , \end{aligned} \tag{5}$$

by which we can write the objective in Eq. (4) as following:

$$\begin{aligned} &\min_{\mathbf{W}^T\mathbf{W}=\mathbf{I}} \frac{\left\|\mathbf{W}^T\mathbf{A}\right\|_{\mathrm{F}}^2}{\|\mathbf{B}\|_{\mathrm{F}}^2 - \|\mathbf{B} - \mathbf{W}\mathbf{W}^T\mathbf{B}\|_{\mathrm{F}}^2} \\ &= \frac{\sum_{i=1}^s \left\|\mathbf{W}^T\mathbf{a}_i\right\|_2^2}{\sum_{i=1}^d \|\mathbf{b}_i\|_2^2 - \sum_{i=1}^d \|\mathbf{b}_i - \mathbf{W}\mathbf{W}^T\mathbf{b}_i\|_2^2} \ . \end{aligned} \tag{6}$$

Then, motivated by prior works that to promote the robustness of our model against sample (data) outliers using not squared ℓ_2-norm distances (Gao 2008; Wright et al. 2009; Nie et al. 2011; 2013; Wang, Nie, and Huang 2013b; 2014b; Liu and Wang 2015; Liu et al. 2017; 2018; Liu and Wang 2018), we develop our new objective as following:

$$\begin{aligned} &\min_{\mathbf{W}^T\mathbf{W}=\mathbf{I}} \frac{\sum_{i=1}^s \left\|\mathbf{W}^T\mathbf{a}_i\right\|_2}{\sum_{i=1}^d \|\mathbf{b}_i\|_2 - \sum_{i=1}^d \|\mathbf{b}_i - \mathbf{W}\mathbf{W}^T\mathbf{b}_i\|_2} \\ &= \frac{\left\|\left(\mathbf{W}^T\mathbf{A}\right)^T\right\|_{2,1}}{\left\|(\mathbf{B})^T\right\|_{2,1} - \left\|\left(\mathbf{B} - \mathbf{W}\mathbf{W}^T\mathbf{B}\right)^T\right\|_{2,1}} \ . \end{aligned} \tag{7}$$

In Eq. (7), we replace the squared errors in Eq. (6) by not squared errors for better robustness. Here we note that we proposed a similar objective to Eq. (7) in our previous work (Han, Wang, and Zhang 2018) that uses the ℓ_1-norm distances in its objective, which can tolerate the compromise between feature outliers and data outliers. However, in the real-world applications, in contrast to visual occlusions or image corruptions that happen rarely, long-term place recognition uses images with drastic scene variances thereby data outliers by nature. Thus, in this paper we propose our new objective in Eq. (7) to focus on alleviating the impacts of outlying data samples.

Visual Place Recognition via Integrated Image Representations

Upon solving the optimization problem in Eq. (7) (using optimization algorithm in the next section), we can obtain the new representation of an input image by computing $\mathbf{y} = \mathbf{W}^T\mathbf{x}$, such that the holistic information is integrated with the semantic landmarks.

After obtaining the learned representation which integrates landmark and holistic information, we can calculate the matching scores by using cosine similarity between query image and each template image in the projected subspace (Naseer et al. 2014; 2015; Han et al. 2017), and then determine whether two locations are matched by comparing the score with a user-defined threshold. Compared with existing long-term place recognition methods that use either holistic information or semantic landmarks only, our new method is more advantageous since it learns an integrated representation that can capture both insights. Due to the non-squared ℓ_2-norm distances, our method is more robust against drastic visual appearance changes caused by outliers or long-term appearance variations, which hence improves the accuracy of place recognition. It is worth noting that, though in this work we use image-based place recognition, our proposed method for integrated representation learning can be readily applied in more sophisticated long-term place recognition methods, such as sequence-based matching.

Optimization Algorithm

Our new objective in Eq. (7) can be seen as a special case of the following general optimization problem:

$$\min_{v\in\mathcal{C}} \frac{f(v)}{g(v)} \ , \quad \text{where } g(v) \geq 0 \ \ (\forall\, v \in \mathcal{C}) \ \ , \tag{8}$$

which has been studied in our previous works (Wang, Nie, and Huang 2014a; 2014b). The algorithm to solve Eq. (8) was also presented, which is summarized in Algorithm 1 [1] (Wang, Nie, and Huang 2014a; 2014b). Theorem 1 and Theorem 2 guarantees the convergence of Algorithm 1 and that it converges fast.

Algorithm 1: The algorithm to solve problem (8) (Wang, Nie, and Huang 2014a, Algorithm 1)(Wang, Nie, and Huang 2014b, Algorithm 2).

1. Set $t = 1$ and initialize $v \in \mathcal{C}$;
while *not converge* **do**
  2. Calculate $\lambda_t = \frac{f(v_t)}{g(v_t)}$;
  3. Update v_{t+1} by solving the following problem:

$$v_{t+1} = \operatorname*{argmin}_{v\in\mathcal{C}} f(v) - \lambda_t g(v) \ . \tag{9}$$

  4. $t = t + 1$.;
end

Theorem 1. *(Wang, Nie, and Huang 2014a; 2014b, Theorem 2) Algorithm 1 decreases the objective value of the problem in Eq.* (8) *in each iteration till converges.*

[1] It can be verified that the denominator in Eq (7) is always greater than 0 which satisfies the constraint in Eq. (8).

Theorem 2. *(Wang, Nie, and Huang 2014a; 2014b, Theorem 3) Algorithm 1 is a Newton's method to find the global solution of Eq.* (8)*.*

Now we turn to solve our objective in Eq. (7). According to Step 3 of Algorithm 1, we need to solve the following optimization problem in every iteration (Here we drop the subscript of λ for brevity.):

$$\min_{\mathbf{W}^T\mathbf{W}=\mathbf{I}} \left\|\left(\mathbf{W}^T\mathbf{A}\right)^T\right\|_{2,1} - \lambda\left(\left\|(\mathbf{B})^T\right\|_{2,1} - \left\|\left(\mathbf{B}-\mathbf{W}\mathbf{W}^T\mathbf{B}\right)^T\right\|_{2,1}\right) , \tag{10}$$

which is equivalent to the following optimization problem:

$$\min_{\mathbf{W}^T\mathbf{W}=\mathbf{I}} \left\|\left(\mathbf{W}^T\mathbf{A}\right)^T\right\|_{2,1} + \lambda\left\|\left(\mathbf{B}-\mathbf{W}\mathbf{W}^T\mathbf{B}\right)^T\right\|_{2,1} , \tag{11}$$

because $\left\|(\mathbf{B})^T\right\|_{2,1}$ is a constant for a given training data set.

In the following, we derive the solution algorithm of the optimization problem in Eq. (11) using the Alternating Direction Method of Multipliers (ADMM) (Boyd et al. 2011). ADMM was originally proposed for convex problems and was extended to nonseparable, nonconvex problems. Given the following constrained optimization problem:

$$\min_{\mathbf{X},\mathbf{Z}} f(\mathbf{X},\mathbf{Z}), \qquad s.t. \quad h(\mathbf{X},\mathbf{Z})=0 , \tag{12}$$

ADMM gives the solution through the updating procedures described in Algorithm 2 (Boyd et al. 2011).

Algorithm 2: ADMM Method to solve Eq. (12).

```
Initialize μ > 0 and set ρ > 1;
while not converge do
    1. Update X by solving X^{k+1} =
       arg min_X (f(X, Z^k) + (μ/2)‖h(X, Z^k) + (1/μ)Y^k‖_F^2);
    2. Update Z by solving
       Z^{k+1} = arg min_Z (f(X^{k+1}, Z) +
       (μ/2)‖h(X^{k+1}, Z) + (1/μ)Y^k‖_F^2);
    3. Update Y by
       Y^{k+1} = Y^k + μh(X^{k+1}, Z^{k+1});
    4. Update μ by μ = ρμ;
end
```

We first rewrite Eq. (11) as the following equivalent optimization problem:

$$\begin{aligned}\min_{\mathbf{W},\mathbf{F},\mathbf{G},\mathbf{H}} \ & \|\mathbf{F}\|_{2,1} + \lambda\|\mathbf{G}\|_{2,1} \ , \\ s.t. \ & \mathbf{F}=\left(\mathbf{W}^T\mathbf{A}\right)^T, \mathbf{G}=\left(\mathbf{B}-\mathbf{W}\mathbf{W}^T\mathbf{B}\right)^T, \\ & \mathbf{H}=\mathbf{W}, \mathbf{H}^T\mathbf{H}=\mathbf{I} ,\end{aligned} \tag{13}$$

in which the orthonormal constraint on $\mathbf{W}$ is implicitly imposed through to the constraints of $\mathbf{H}=\mathbf{W}$ and $\mathbf{H}^T\mathbf{H}=\mathbf{I}$.

According to Algorithm 2, we need to solve the following optimization problem:

$$\begin{aligned}\min_{\mathbf{W},\mathbf{F},\mathbf{G},\mathbf{H},\mathbf{\Lambda},\mathbf{\Sigma},\mathbf{\Theta}} \ & \|\mathbf{F}\|_{2,1} + \lambda\|\mathbf{G}\|_{2,1} \\ & + \frac{\mu}{2}\left\|\mathbf{F}-\left(\mathbf{W}^T\mathbf{A}\right)^T+\frac{1}{\mu}\mathbf{\Lambda}\right\|_{\mathrm{F}}^2 \\ & + \frac{\mu}{2}\left\|\mathbf{G}-\left(\mathbf{B}-\mathbf{H}\mathbf{W}^T\mathbf{B}\right)^T+\frac{1}{\mu}\mathbf{\Sigma}\right\|_{\mathrm{F}}^2 \\ & + \frac{\mu}{2}\left\|\mathbf{W}-\mathbf{H}+\frac{1}{\mu}\mathbf{\Theta}\right\|_{\mathrm{F}}^2 \\ s.t. \ & \mathbf{H}^T\mathbf{H}=\mathbf{I} ,\end{aligned} \tag{14}$$

where $\mathbf{\Lambda}\in\Re^{s\times r}$ is the Lagrangian multiplier for the constraint of $\mathbf{F}=\left(\mathbf{W}^T\mathbf{A}\right)^T$, $\mathbf{\Sigma}\in\Re^{s\times d}$ is the Lagrangian multiplier for the constraint of $\mathbf{G}=\left(\mathbf{B}-\mathbf{W}\mathbf{W}^T\mathbf{B}\right)^T$, and $\mathbf{\Theta}\in\Re^{p\times r}$ is the Lagrangian multiplier for the constraint of $\mathbf{H}=\mathbf{W}$.

Step 1. Initialization.

Step 2. We solve $\mathbf{F}$, when we fix the other variables:

$$\min_{\mathbf{F}} \ \|\mathbf{F}\|_{2,1} + \frac{\mu}{2}\|\mathbf{F}-\mathbf{M}\|_{\mathrm{F}}^2 , \tag{15}$$

where we denote $\mathbf{M}=\left(\mathbf{W}^T\mathbf{A}\right)^T-\frac{1}{\mu}\mathbf{\Lambda}$ for brevity.

The optimization problem in Eq. (15) can be decoupled row by row to solve the following subproblems:

$$\min_{\mathbf{f}^i} \ \frac{1}{\mu}\left\|\mathbf{f}^i\right\|_2+\frac{1}{2}\left\|\mathbf{f}^i-\mathbf{m}^i\right\|_2^2 . \tag{16}$$

The solution of Eq. (16) can be derived as:

$$\mathbf{f}^i=\begin{cases}\left(1-\frac{1}{\mu\|\mathbf{m}^i\|_2}\right)\mathbf{m}^i, & \left\|\mathbf{m}^i\right\|_2>1/\mu , \\ \mathbf{0} & \left\|\mathbf{m}^i\right\|_2\le 1/\mu .\end{cases} \tag{17}$$

Step 3. We solve $\mathbf{G}$, when we fix the other variables:

$$\min_{\mathbf{G}} \ \lambda\|\mathbf{G}\|_{2,1}+\frac{\mu}{2}\|\mathbf{F}-\mathbf{N}\|_{\mathrm{F}}^2 , \tag{18}$$

where we denote $\mathbf{N}=\left(\mathbf{B}-\mathbf{H}\mathbf{W}^T\mathbf{B}\right)^T-\frac{1}{\mu}\mathbf{\Sigma}$ for brevity.

Similarly, the optimization problem in Eq. (18) can be decoupled row by row to solve the following subproblems:

$$\min_{\mathbf{g}^i} \ \frac{\lambda}{\mu}\left\|\mathbf{g}^i\right\|_2+\frac{1}{2}\left\|\mathbf{g}^i-\mathbf{n}^i\right\|_2^2 . \tag{19}$$

Thus, the solution of Eq. (19) can be derived as:

$$\mathbf{g}^i=\begin{cases}\left(1-\frac{\lambda}{\mu\|\mathbf{n}^i\|_2}\right)\mathbf{n}^i, & \left\|\mathbf{n}^i\right\|_2>\lambda/\mu , \\ \mathbf{0} & \left\|\mathbf{n}^i\right\|_2\le\lambda/\mu .\end{cases} \tag{20}$$

Step 4. We solve $\mathbf{H}$, when we fix the other variables:

$$\max_{\mathbf{H}} \ \mathbf{tr}\left(\mathbf{H}^T\mathbf{Z}\right) \quad s.t. \quad \mathbf{H}^T\mathbf{H}=\mathbf{I} , \tag{21}$$

where we denote $\mathbf{Z}=(\mathbf{B}^T-\mathbf{G}-\frac{1}{\mu}\mathbf{\Theta})^T\mathbf{B}^T\mathbf{W}+\mathbf{W}+\frac{1}{\mu}\mathbf{\Theta}$ for brevity. According to (Wang, Nie, and Huang 2013a,

Theorem 1), the problem in Eq. (21) can be solved by computing the SVD of $\mathbf{Z}$: if $svd(\mathbf{Z}) = \mathbf{U}\mathbf{A}\mathbf{V}^T$, the solution of Eq. (21) is given by $\mathbf{U}\mathbf{V}^T$.

Step 5. We solve $\mathbf{W}$, when we fix the other variables:

$$\min_{\mathbf{W}} \left\| \mathbf{F} - \left(\mathbf{W}^T\mathbf{A}\right)^T + \frac{1}{\mu}\mathbf{\Lambda} \right\|_{\mathrm{F}}^2 + \left\| \mathbf{W} - \mathbf{H} + \frac{1}{\mu}\mathbf{\Theta} \right\|_{\mathrm{F}}^2 + \left\| \mathbf{G} - \left(\mathbf{B} - \mathbf{H}\mathbf{W}^T\mathbf{B}\right)^T + \frac{1}{\mu}\mathbf{\Sigma} \right\|_{\mathrm{F}}^2 . \quad (22)$$

Because there is no constraint in Eq. (22), we can solve it by taking the derivative of it w.r.t. $\mathbf{W}$ and setting the derivative to be equal to 0, which leads to the following solution:

$$\mathbf{W} = \left(\mathbf{A}\mathbf{A}^T + \mathbf{B}\mathbf{B}^T + \mathbf{I}\right)^{-1}\mathbf{Q} . \quad (23)$$

where $\mathbf{Q} = \mathbf{A}\left(\mathbf{F} + \frac{1}{\mu}\mathbf{\Lambda}\right) + \mathbf{B}(\mathbf{B}^T - \mathbf{G} - \frac{1}{\mu}\mathbf{\Theta})\mathbf{H} + \left(\mathbf{H} - \frac{1}{\mu}\mathbf{\Theta}\right)$.

Step 6. Update $\mathbf{\Lambda}$, $\mathbf{\Sigma}$, $\mathbf{\Theta}$ and μ as step 3 and 4 in Algorithm 2.

Experimental Results

In this section, we evaluate the proposed method for long-term visual place recognition by experimenting with two large-scale public data sets: the CMU-VL data set with different month scenarios and the Nordland data set with different season scenarios.

In our experiments, different feature extraction methods are used for extracting visual features from the video frames including: (1) color features (Lee, Kim, and Myung 2013), (2) Local Binary Patterns (LBP) visual features (Qiao, Cappelle, and Ruichek 2015), (3) Speeded Up Robust Features (SURF) (Badino, Huber, and Kanade 2012), (4) Deep features learned by Convolutional Neural Network (CNN) (Sunderhauf et al. 2015), and (5) Histogram of Oriented Gradients (HOG) features (Naseer et al. 2014). Compared to these raw visual featurs, our method can improve the representation capability by incorporating landmark relationships and preserving their global and local consistencies.

We test the performance of our method by conducting both quantitative and qualitative evaluations. Baseline and recent methods, including the BRIEF-GIST (Sünderhauf and Protzel 2011), Normalized Gradients (NormG) of grayscale images (used in SeqSLAM (Milford and Wyeth 2012)), and the methods that simply based upon color, LBP, SURF, CNN, HOG features, are compared in our experiments.

Study of the Hyperparameter K of Our New Method

To begin with our experiments, we first study the impacts of the hyperparameter K of our method. We first experiment with the Nordland data set. Since K denotes the number of nearest neighbors of a semantic landmark $\mathbf{x}_i$, we vary it from 2 to 5 and report the objective value in Eq. (7) over iterations in the top panel of Fig. 1. We can see that, the objective value converges fastest (from 1.05 to 0.0027) when $K = 2$, while

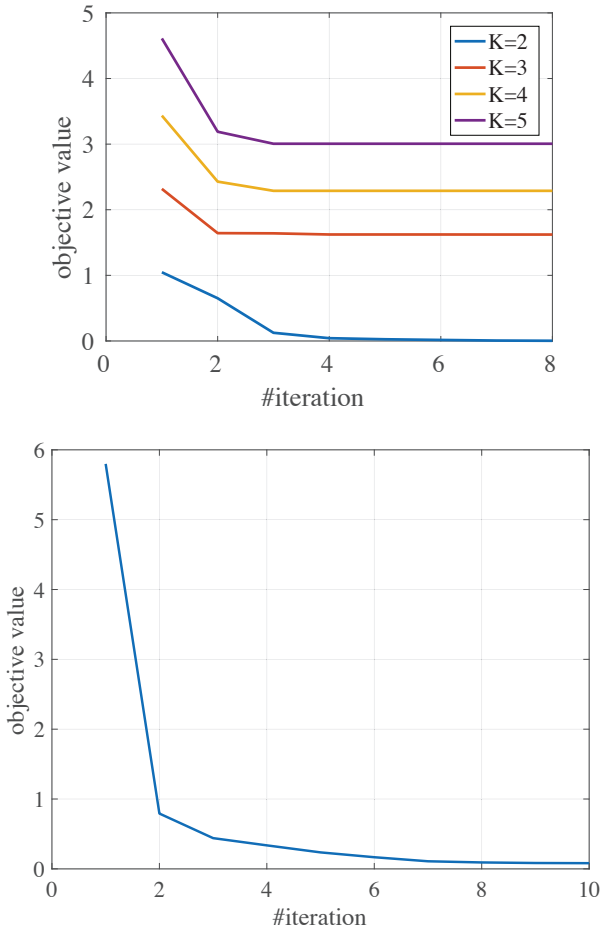

Figure 1: Top: Objective value changes with different K value on the Nordland data set. Bottom: Objective value updates with $K = 2$ on the CMU-VL data set.

the objective value converges slower when K is selected as other values. The same observation can be seen on the CUM-VL data set. Thus, we empirically select $K = 2$ in all our following experiments.

Study of the Convergence of Our New Method

The proposed algorithm to solve our new objective is an iterative algorithm. Thus, we study its convergence empirically. Fig. 1 shows our algorithm does converge on the both experimental data sets. In addition, the bottom panel of Fig. 1 shows the variation of the objective value of our method over iterations on the CMU-VL data set. We can see that 1): the objective value monotonically decreases over iterations and 2): the objective value converges very fast (from 5.78 to 0.04). These observations clearly demonstrate the correctness and effectiveness of our new method.

Study of the Orthogonality of the Solutions of Our New Method

In our algorithm, $\mathbf{W}$ is optimized by approximating $\mathbf{H}$, where $\mathbf{H} = \mathbf{U}\mathbf{V}^T$ is strictly orthogonal. Thus the opti-

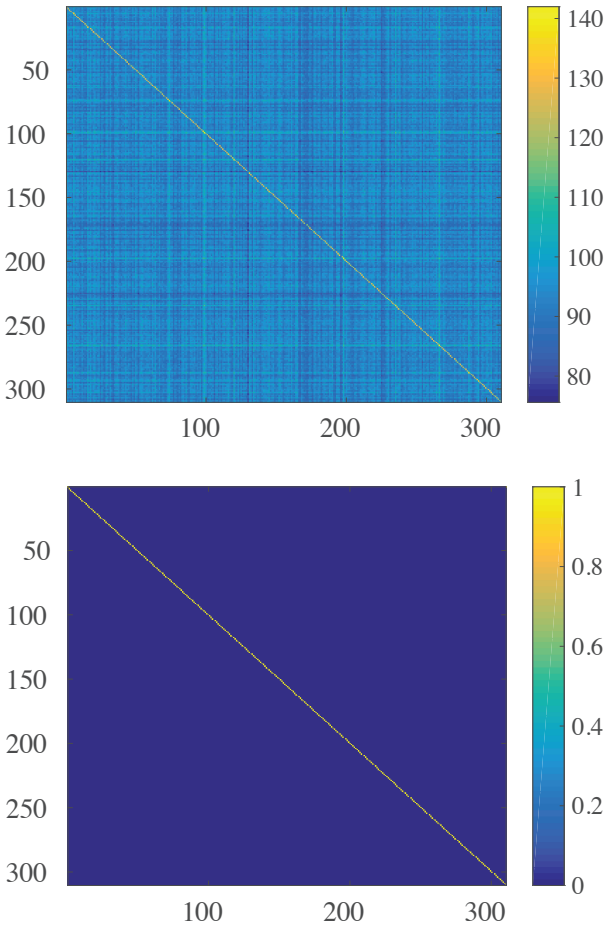

Figure 2: The heatmaps of $\mathbf{W}^T\mathbf{W}$ by a competing method (top) and our new method (bottom).

mized $\mathbf{W}$ should also be orthogonal. Although many existing methods tried to solve the orthogonal constraint solution, most of them indeed cannot get strictly orthogonal solution, such as (Xiang, Nie, and Zhang 2008; Wang, Nie, and Huang 2013b). Fig. 2 shows the heatmaps of $\mathbf{W}^T\mathbf{W}$ obtained by (Xiang, Nie, and Zhang 2008) and by our new method. We can see that our method can get a strictly orthogonal $\mathbf{W}$ while the counterpart fails.

Results on the CMU-VL (Different Months) and Nordland (Different Seasons) Datasets

The CMU Visual Localization (CMU-VL) data set (Badino, Huber, and Kanade 2012) was recorded by two cameras installed on a vehicle traveling the same route five times in different months, while the Nordland data set (Sünderhauf, Neubert, and Protzel 2013) was obtained from a ten-hour long train journey in different seasons.

There are many challenges in visual place recognition in both data sets, such as new buildings and dynamic objects (cars, pedestrians, *etc*.). Views change due to possible route deviations, and above all, as well as long-term appearance variations due to vegetation, illumination and weather changes in different seasons or months. To train our proposed model, the same semantic objects and landmarks recorded in different scenarios are utilized to learn the optimal projection matrix $\mathbf{W}$, such as stop signs, trees, houses, *etc*. in CMU-VL and railroad tracks, trees, houses, among others in Nordland.

For quantitative evaluation and comparison, we use precision-recall curves as a metric following (Sunderhauf et al. 2015; Zhang, Han, and Wang 2016), where high area under the curve means both high recall (relates to a low false positive rate) and high precision (relates to a low false negative rate). Inspired by the conclusion drawn by (Han et al. 2017) that HOG features perform the best among other types of raw visual features, we extract HOG features from landmarks as the input to generate the integrated representation by the proposed method. As is shown in Fig. 4 and Fig. 5, our new method outperforms the previous HOG-based features, which again demonstrates that the holism-landmark integration by our method can improve representation capability.

The qualitative results obtained by our method are reported in Fig. 3. The first half shows the examples of matched frames between October (top row) and December (bottom row) of the CMU-VL data set and the bottom half shows the matched frames between Winter (top row) and Spring (bottom row) of the Nordland data set, respectively. The matched frames are determined by the maximum similarity score between two compared frames. From Fig. 3 we see that scenes in same location exhibit significant appearance variations in various scenarios, yet our new method can still recognize correctly.

Conclusion

In this paper, we proposed a novel method to combine the holistic information with landmarks of a place to construct an integrative representation, which can help improve long-term visual place recognition. Our method is implemented by minimizing the ratio objective with the $\ell_{2,1}$-norm of matrices that enforces both global and local consistency of the input data in the projected subspace while staying robustness to sample outliers. To solve the challenging ratio objective, we proposed a new optimization algorithm which can theoretically guarantee a sequence decreasing solution. We conducted experiments on two large-scale public benchmark data sets collected for long-term place recognition and promising results demonstrated the effectiveness and superiority of our proposed method.

Acknowledgments

All correspondence should be addressed to: Hua Wang (huawangcs@gmail.com). This work was partially supported by National Science Foundation under Grant NSF-IIS 1652943. This research was also partially supported by Army Research Office (ARO) under Grant W911NF-17-1-0447, U.S. Air Force Academy (USAFA) under Grant FA7000-18-2-0016, and the Distributed and Collaborative Intelligent Systems and Technology (DCIST) CRA under Grant W911NF-17-2-0181.

Figure 3: The matched scenes by our proposed method on CMU-VL (upper) and Nordland (lower) data set.

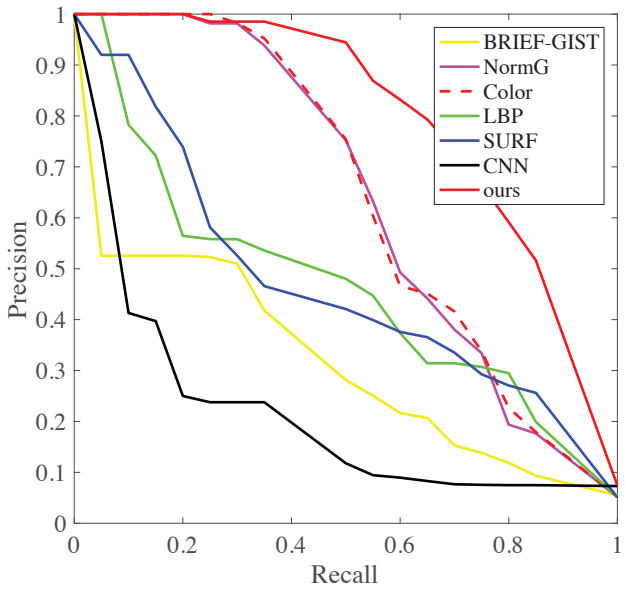

Figure 4: The Precision-Recall curve of our proposed method compared with other methods on CMU-VL data set.

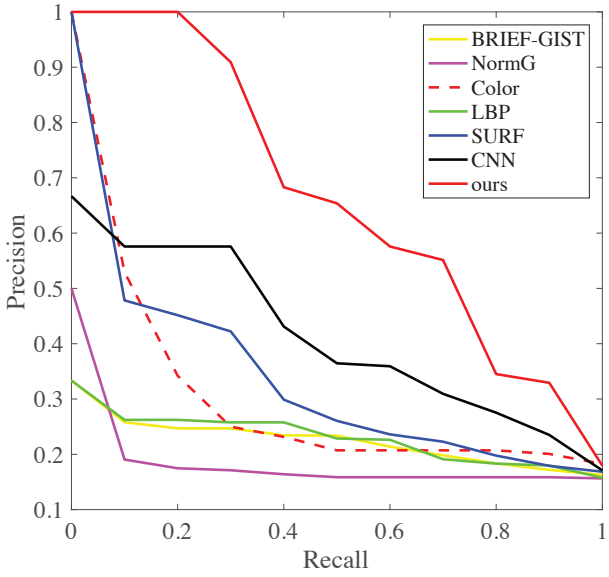

Figure 5: The Precision-Recall curve of our proposed method compared with other methods on Nordland data set.

References

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

Boyd, S.; Parikh, N.; Chu, E.; Peleato, B.; Eckstein, J.; et al. 2011. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning* 3(1):1–122.

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.

Han, F.; El Beleidy, S.; Wang, H.; Ye, C.; and Zhang, H. 2018. Learning of holism-landmark graph embedding for place recognition in long-term autonomy. *IEEE Robotics and Automation Letters* 3(4):3669–3676.

Han, F.; Wang, H.; and Zhang, H. 2018. Learning of integrated holism-landmark representations for long-term loop closure detection. In *AAAI*.

Jolliffe, I. T. 1986. Principal component analysis and factor analysis. In *Principal component analysis*. Springer. 115–128.

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.

Liu, K., and Wang, H. 2015. Robust multi-relational clustering via ℓ_1-norm symmetric nonnegative matrix factorization. In *ACL*, volume 2, 397–401.

Liu, K., and Wang, H. 2018. High-order co-clustering via strictly orthogonal and symmetric ℓ_1-norm nonnegative matrix tri-factorization. In *IJCAI*, 2454–2460.

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.

Liu, K.; Wang, H.; Nie, F.; and Zhang, H. 2018. Learning multi-instance enriched image representations via non-greedy ratio maximization of the ℓ_1-norm distances. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7727–7735.

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.

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.

Nie, F.; Huang, H.; Ding, C.; Luo, D.; and Wang, H. 2011. Robust principal component analysis with non-greedy ℓ_1-norm maximization. In *IJCAI*.

Nie, F.; Wang, H.; Huang, H.; and Ding, C. H. 2013. Early active learning via robust representation and structured sparsity. In *IJCAI*, 1572–1578.

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.

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.; and Huang, H. 2013a. Multi-view clustering and feature learning via structured sparsity. In *International conference on machine learning*, 352–360.

Wang, H.; Nie, F.; and Huang, H. 2013b. Robust and discriminative self-taught learning. In *International Conference on Machine Learning*, 298–306.

Wang, H.; Nie, F.; and Huang, H. 2014a. Globally and locally consistent unsupervised projection. In *AAAI*, 1328–1333.

Wang, H.; Nie, F.; and Huang, H. 2014b. Robust distance metric learning via simultaneous ℓ_1-norm minimization and maximization. In *ICML*, 1836–1844.

Wright, J.; Ganesh, A.; Rao, S.; Peng, Y.; and Ma, Y. 2009. Robust principal component analysis: Exact recovery of corrupted. *NIPS* 116.

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.

Xiang, S.; Nie, F.; and Zhang, C. 2008. Learning a mahalanobis distance metric for data clustering and classification. *Pattern Recognition* 41(12):3600–3612.

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*.