# Long-Term Loop Closure Detection through Visual-Spatial Information Preserving Multi-Order Graph Matching

**Peng Gao, Hao Zhang**
Human-Centered Robotics Laboratory, Colorado School of Mines, Golden, CO 80401
{gaopeng, hzhang}@mines.edu

## Abstract

Loop closure detection is a fundamental problem for simultaneous localization and mapping (SLAM) in robotics. Most of the previous methods only consider one type of information, based on either visual appearances or spatial relationships of landmarks. In this paper, we introduce a novel visual-spatial information preserving multi-order graph matching approach for long-term loop closure detection. Our approach constructs a graph representation of a place from an input image to integrate visual-spatial information, including visual appearances of the landmarks and the background environment, as well as the second and third-order spatial relationships between two and three landmarks, respectively. Furthermore, we introduce a new formulation that formulates loop closure detection as a multi-order graph matching problem to compute a similarity score directly from the graph representations of the query and template images, instead of performing conventional vector-based image matching. We evaluate the proposed multi-order graph matching approach based on two public long-term loop closure detection benchmark datasets, including the St. Lucia and CMU-VL datasets. Experimental results have shown that our approach is effective for long-term loop closure detection and it outperforms the previous state-of-the-art methods.

## Introduction

Loop closure detection (also referred to as place recognition) is a fundamental challenge for visual simultaneous localization and mapping (SLAM) in robotics, which has become an active research field over the past decades. The goal of loop closure detection is to determine whether the robot's current location has been previously visited by matching the current robot's observation with previous experiences. Loop closure detection is a necessary component for all SLAM techniques to reduce ambiguity and accumulated errors in constructed maps, thus significantly improving a robot's localization and mapping accuracy (Durrant-Whyte and Bailey 2006).

Over the past several years, long-term loop closure detection has attracted an increasing attention for visual SLAM in various real-world long-term autonomy applications, such as autonomous driving, with the goal to identify previously visited locations during long-term robot operations (e.g., at different times of a day, or over months and seasons). Long-term changes in the environment makes long-term loop closure detection a challenging problem to solve. For example, when an autonomous vehicle operates over various seasons, the same place in the environment can look quite differently due to long-term changes, including illumination (e.g., noon versus midnight), weather (rainy versus sunny), and vegetation changes.



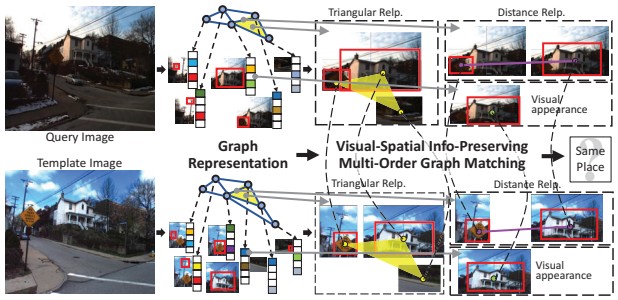



Figure 1: Overview of the proposed visual-spatial information preserving multi-order graph matching approach. It creates graph representations from the query and template images to integrate information of both visual appearances and multi-order spatial relationships, and it formulates loop closure detection as a new multi-order graph matching problem to perform place matching directly from the graph representations, instead of relying on vector-based place representation and matching.

Given the importance of long-term loop closure detection, a number of methods were developed to address this problem (Sünderhauf, Neubert, and Protzel 2013; Linegar, Churchill, and Newman 2016). A category of these methods focused on using visual appearance information to represent and match places, e.g., based on local features (Cummins and Newman 2008; Mur-Artal and Tardós 2014) and visual holistic scenes (Naseer et al. 2014; Latif et al. 2014). Another category of these previous methods considered spatial relationships of landmarks for loop closure detection (Ho and Newman 2006; Panphattarasap and Calway 2016). In spite of their promising performance, the problem of how to integrate visual appearance cues and spatial relationships

of landmarks in a principled way has not yet been well addressed for long-term loop closure detection.

We propose a novel *visual-spatial information preserving multi-order graph matching* method for long-term loop closure detection in this paper. The proposed approach divides an input image into regions that contain landmarks and non-landmark background regions. These regions are represented as graph nodes and visual features are extracted from the regions to encode the visual appearances of the nodes. Then, our approach computes distances between two nodes and angles among three nodes to encode the second and third-order spatial relationships, respectively. Thus, the constructed representation integrates both visual and spatial information of the place from an input image. Given graph representations of a pair of query and template images, our approach formulates loop closure detection as a multi-order graph matching problem, which computes a similarity score between the pair of graph representations that encode both visual and spatial cues from the query and template images.

The main contribution of this paper focuses on the proposal of the novel visual-spatial information preserving multi-order graph matching method for long-term loop closure detection. Specifically, we first implement one of the first graph representations that integrates both visual appearances and multi-order spatial relationships of the image regions encoding landmarks and the background environment. Second, we introduce a novel formulation that formulates loop closure detection as a multi-order graph matching problem to compute a similarity score directly from the graphs of query and template images, instead of performing vector-based place matching used in almost all previous approaches. Third, we develop an effective optimization algorithm to solve the formulated non-convex optimization problem in order to obtain the best match of multi-order graphs.

## Related Work

### Representations for Loop Closure Detection

Constructing a robust representation of places is essential for long-term loop closure detection (Williams et al. 2009; Lowry et al. 2015). Existing approaches for representation construction can be divided into two major categories, based on visual appearance information or spatial relationships of landmarks.

Approaches based upon visual appearances used local, global, or a combination of both types of visual features to build a representation of places for loop closure detection. Representations based on local features were shown less effective to represent long-term environment changes (Naseer et al. 2014). Thus, most methods based on visual cues used global features, such as GIST (Latif et al. 2014), HOG (Naseer et al. 2014), and CNN (Sünderhauf et al. 2015), to construct representations of the holistic scene in the robot view. Besides using a single type of features, several approaches integrated multiple types of features to encode places (Pronobis et al. 2010; Han et al. 2017).

The other category of approaches use the spatial relationship of landmarks in the environment to perform loop closure detection, such as using co-visibility matrix to encode the neighborhood relationship of consistently co-observed landmarks in a sequence (Stumm et al. 2015), encoding the connection between semantic regions by graphs (Gawel et al. 2018). In addition, random walk (Gawel et al. 2018), graph kernel (Stumm et al. 2016), graph embedding (Han, Wang, and Zhang 2018; Liu et al. 2019) techniques are widely applied to embed the graph-based structural information into the linear vector space. Deep learning techniques are also commonly used to encoding the spatial structure of landmarks into a vector-based descriptor. For example, the CNN-based landmark spatial descriptor (Chen et al. 2017; Panphattarasap and Calway 2016), bag of semantic words to encode the layout of 3D map (Schönberger et al. 2018). Instead of using vector-based descriptors, associations between two scans are also studied based on second order spatial relationship of key-points by CRF-matching (Ramos, Kadous, and Fox 2009) and based on high-order potentials through hypergraph matching (Nguyen, Gautier, and Hein 2015).

Existing methods generally utilize one type of cues only, i.e., based on visual appearances of the holistic environment, or based on spatial relationships of landmarks. The problem of fusing visual appearance of holistic environment and spatial relationship of landmarks in a principled way has not yet been well studied. In addition, most methods implicitly encode structural information in linear vector space or hidden states, which have low interpretability and are hard to analyze the importance of each cue. Our proposed approach aims to address this key research problem.

### Matching Methods for Loop Closure Detection

After obtaining a representation of places, existing methods generally apply a separate matching procedure that matches a query observation to templates of previously visited places for loop closure detection. Given a vector-based representation, matching methods typically calculate a similarity score between query and template image based upon an Euclidian or cosine distance (Newman, Cole, and Ho 2006; Naseer et al. 2015). Given the association between graphs, separate procedure using the number of associations is applied to calculate the matching similarity (Ramos, Fox, and Durrant-Whyte 2007).

The previous methods are generally consider representation construction and place recognition as two separate procedures, i.e., after optimizing vector-based representations of input images, a separate matching technique is applied on the representation vectors to perform place matching. We introduce a formulation based on multi-order graph representation and matching that formulates loop closure detection under a unified optimization framework.

## The Proposed Approach

**Notation.** We denote matrices and tensors (i.e., 3D matrices) by bold capital letters, e.g., $\mathbf{M} = \{m_{ij}\} \in \mathbb{R}^{n \times n'}$ and $\mathbf{T} = \{t_{ijk}\} \in \mathbb{R}^{n \times n' \times n''}$, respectively. Vectors are denoted by bold lowercase letters. In addition, we denote the vectorized form of the matrix $\mathbf{M} \in \mathbb{R}^{n \times n'}$ using $\mathbf{m} \in \mathbb{R}^{nn'}$ that is a concatenation of the columns of $\mathbf{M}$ into a vector.

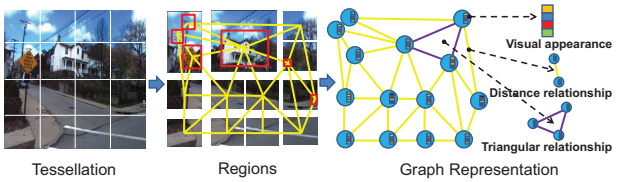



Figure 2: Given an image, our approach creates a graph representation that encodes the visual appearances of landmarks and the background environment, as well as the second and third-order spatial relationships of the nodes.

## Problem Formulation

**Graph Representation.** We represent places with a graph representation that fuses both visual and spatial information from input images (Figure 2). Specifically, given an image with observed landmarks, we tessellate the image into a grid of rectangular cells. Each cell contains either none, partial, or complete landmarks. Cells with a part of the same landmark are merged to generate a region that contains a complete landmark. All generated regions that include a landmark are denoted by the set $\mathcal{S}_l$. The remaining cells containing no landmarks are denoted by $\mathcal{S}_n$. Then, the observed image can be divided into a set of regions $\mathcal{S} = \{\mathcal{S}_l, \mathcal{S}_r\}$.

Given the region set $\mathcal{S}$, we represent a place with a graph $\mathcal{G} = (\mathcal{V}, \mathcal{H}, \mathcal{E}, \mathcal{C})$. The nodes $\mathcal{V} = \{\mathbf{v}_i, i = 1, \dots, n\}$ represents the locations of all regions in $\mathcal{S}$, with $\mathbf{v}_i$ encoding the location of the $i$-th region. The node set can be further divided into two subsets $\mathcal{V} = \{\mathcal{V}_l, \mathcal{V}_r\}$ to represent the landmark regions and non-landmark regions, respectively. We define a feature set $\mathcal{C} = \{\mathbf{c}_i, i = 1, \dots, n\}$ to include visual features that encode the appearances of the regions in $\mathcal{S}$, where $\mathbf{c}_i \in \mathbb{R}^d$ is the feature vector of length $d$ that is extracted from the $i$-th region. The feature set $\mathcal{C}$ can be also divided into two subsets $\mathcal{C} = \{\mathcal{C}_l, \mathcal{C}_r\}$ to include features extracted from landmark regions and non-landmark background regions, respectively. $\mathcal{C}_r$ and $\mathcal{C}_l$ together describe the visual appearance of the holistic environment as the features in $\mathcal{C} = \{\mathcal{C}_l, \mathcal{C}_r\}$ cover all pixels of the whole image. The set $\mathcal{E} = \{e_{i,j}, i, j = 1, 2, \dots, n, i \neq j\}$ represents the distance between a pair of nodes, where $e_{i,j}$ represents the distance of the $i$-th and the $j$-th nodes in $\mathcal{V}$. Since a distance describes the relationship between two nodes, the set $\mathcal{E}$ encodes the second-order spatial relationship of the nodes. The set $\mathcal{H} = \{h_{i,j,k} = [\theta_i, \theta_j, \theta_k], i, j, k = 1, \dots, n, i \neq j \neq k\}$ encodes the triangular relationship among three nodes, where $h_{i,j,k} = [\theta_i, \theta_j, \theta_k]$ denotes three angles of the triangle constructed by the $i$-th, $j$-th and $k$-th nodes in $\mathcal{V}$. Since the angle vector $[\theta_i, \theta_j, \theta_k]$ describes the relationship among three nodes, the set $\mathcal{H}$ encodes the third-order spatial relationship.

**Graph-based Place Matching.** Based upon the graph representation, we propose the new formulation that formulates loop closure detection as a multi-order graph matching problem. Formally, given a query image and a template image, we build their graph-based representations $\mathcal{G} = (\mathcal{V}, \mathcal{H}, \mathcal{E}, \mathcal{C})$ and $\mathcal{G}' = (\mathcal{V}', \mathcal{H}', \mathcal{E}', \mathcal{C}')$, respective. The two graphs $\mathcal{G}$ and $\mathcal{G}'$ can contain a different number of nodes (i.e., $n$ can be different from $n'$), because the query and template images may contain a different number of landmarks. Then, loop closure detection is performed by matching the graphs $\mathcal{G}$ and $\mathcal{G}'$.

In order to preserve the information of both visual appearances and spatial relationships of the nodes when matching the graphs, we model three types of similarities in a unified multi-order graph matching framework. First, we compute a vectorized matrix $\mathbf{b} = \{b_{ii'}\} \in \mathbb{R}^{nn'}$ to represent feature similarities, where $b_{ii'}$ represents the similarity between the feature vectors $\mathbf{c}_i \in \mathcal{C}$ and $\mathbf{c}'_i \in \mathcal{C}'$, which can be computed using a dot product of two feature vectors:

$$b_{ii'} = \frac{\mathbf{c}_i \cdot \mathbf{c}'_i}{\|\mathbf{c}_i\| \|\mathbf{c}'_i\|} \tag{1}$$

Since $\mathcal{C}$ and $\mathcal{C}'$ encodes visual appearances of the landmarks and the background environment, Eq. (1) represents the similarity of visual appearances of the landmarks and the background. Because Eq. (1) uses one node from each graph, we refer to this similarity as the first-order similarity.

Second, we calculate the distance similarity matrix $\mathbf{A} = \{a_{ii',jj'}\} \in \mathbb{R}^{nn' \times nn'}$, where $a_{ii',jj'}$ represents the similarity of the edge $e_{i,j} \in \mathcal{E}$ and the edge $e'_{i',j'} \in \mathcal{E}'$, which can be computed by:

$$a_{ii',jj'} = \exp\left(-\frac{1}{\beta}(e_{i,j} - e_{i',j'})^2\right) \tag{2}$$

This similarity is computed using the nonlinear exponential function, with $\beta$ as the hyper parameter, to transfer any non-negative input to an output value between 0 and 1. Because Eq. (2) computes the similarity of the second-order spatial relationships that involve two nodes from each graph, this similarity is referred to as the second-order similarity.

Third, we also compute the angular similarity tensor $\mathbf{T} = \{t_{ii',jj',kk'}\} \in \mathbb{R}^{nn' \times nn' \times nn'}$, where $t_{ii',jj',kk'}$ denotes the similarity between $h_{i,j,k} = [\theta_i, \theta_j, \theta_k] \in \mathcal{H}$ and $h'_{i',j',k'} = [\theta'_{i'}, \theta'_{j'}, \theta'_{k'}] \in \mathcal{H}'$, which can be computed by:

$$t_{ii',jj',kk'} = \exp\left(-\frac{1}{\gamma} \sum_{u \in i,j,k; v \in i',j',k'} |\cos(\theta_u) - \cos(\theta_v)|\right) \tag{3}$$

where $\gamma$ is a hyperparameter of the exponential normalization function. Since Eq. (3) calculates the similarity of the third-order spatial relationships, we refer to this similarity as the their-order similarity.

Given the definitions of the three similarities that are computed from graph representations of the query and template images, loop closure detection is formulated as multi-order graph matching by solving the optimization problem:

$$\begin{aligned} \max_{\mathbf{X}} \; & \lambda_3 \sum_{ii'=1}^{nn'} \sum_{jj'=1}^{nn'} \sum_{kk'=1}^{nn'} t_{ii',jj',kk'} x_{ii'} x_{jj'} x_{kk'} \\ & + \lambda_2 \sum_{ii'}^{nn'} \sum_{jj'}^{nn'} a_{ii',jj'} x_{ii'} x_{jj'} + \lambda_1 \sum_{ii'}^{nn'} b_{ii'} x_{ii'} \\ \text{s.t.} \quad & \mathbf{X}\mathbf{1}_{n' \times 1} \leq \mathbf{1}_{n \times 1}, \mathbf{X}^\top \mathbf{1}_{n \times 1} \leq \mathbf{1}_{n' \times 1} \end{aligned} \tag{4}$$

where $\mathbf{X} \in \{0,1\}^{n\times n'}$ is correspondence matrix, with $x_{ii'} = 1$ denoting that the $i$-th node in $\mathcal{V}$ and the $i'$-th node in $\mathcal{V}'$ are matched, and $\mathbf{1}$ is a all 1 vector.

The first term in Eq. (4) represents the accumulated third-order triangular relationship of two graphs given the correspondence matrix $\mathbf{X}$, which sums up all third-order similarities $t_{ii',jj',kk'}$ of two triangles $h_{i,j,k} \in \mathcal{H}$ and $h'_{i',j',k'} \in \mathcal{H}'$. The second term represents the accumulated second-order relationship of two graphs, which sums up all second-order similarities $a_{ii',jj'}$ of two edges $e_{i,j} \in \mathcal{E}$ and $e'_{i',j'} \in \mathcal{E}'$. The third term denotes the accumulated similarity on visual appearances of two graphs, which sums up all visual appearance similarities of the feature vectors $\mathbf{c}_i \in \mathcal{C}$ and $\mathbf{c}'_i \in \mathcal{C}'$. $\lambda_1 \geq 0$, $\lambda_2 \geq 0$ and $\lambda_3 \geq 0$ are hyperparameters to control the importance of the first, second and third-order similarities, which satisfy $\lambda_1 + \lambda_2 + \lambda_3 = 1$. The constraints in Eq. (4) are designed to enforce the one-to-one correspondence: each row or column in $\mathbf{X}$ can at most have one element equal to 1, and all others are equal to 0.

We can rewrite Eq. (4) into a concise matrix form:

$$\begin{aligned}\max_{\mathbf{X}}\ & \lambda_3 \mathbf{T} \otimes_1 \mathbf{x} \otimes_2 \mathbf{x} \otimes_3 \mathbf{x} \\ & + \lambda_2 \mathbf{x}^\top \mathbf{A}\mathbf{x} + \lambda_1 \mathbf{b}^\top \mathbf{x} \\ \text{s.t.}\quad & \mathbf{X}\mathbf{1}_{n'\times 1} \leq \mathbf{1}_{n\times 1}, \mathbf{X}^\top \mathbf{1}_{n\times 1} \leq \mathbf{1}_{n'\times 1} \end{aligned} \tag{5}$$

where $\mathbf{x} = \{x_{ii'}\} \in \{0,1\}^{nn'}$ is the vectorized correspondence matrix $\mathbf{X}$, $\otimes$ denotes a tensor product, and $\otimes_l, l = 1,2,3$ denotes multiplication between $\mathbf{x}$ and the mode-$l$ matricization of $\mathbf{T}$ (Rabanser, Shchur, and Günnemann 2015).

After optimizing $\mathbf{X}$ to decide the correspondence between the nodes in two graphs, the value of the objective function in Eq. (4) denotes the similarity of the two graphs built from the query and template images. Different from conventional methods that convert each image into a vector representation and match images based on vector matching, our approach computes a similarity score directly from graph representations of the query and template images through multi-order graph matching.

**Addressing Scale Changes.** Linear perspective variations (i.e., objects look smaller when they are further away from a camera) cause scale variations of objects in images, which are not considered in the definitions of the visual and spatial similarities used in the formulation in Eq. (5). Furthermore, landmark regions (e.g., buildings and stop signs) are usually more informative than regions of background environments for place matching. Accordingly, we propose to address the scale changes of the landmarks in order to improve loop closure detection. Mathematically, given the $i$-th region $s_i \in \mathcal{S}$, we first normalize the sizes of all landmark regions by:

$$w_i = \frac{\Omega(s_i)}{\sum_{s_j \in \mathcal{S}_l} \Omega(s_j)} \mathbb{1}(s_i \in \mathcal{S}_l) \tag{6}$$

where $\Omega(s_i)$ represents the area size of $s_i$ and $\mathbb{1}(\cdot)$ denotes an indicator function, which is equal to 1 when $s_i \in \mathcal{S}_l$ (i.e., $s_i$ is a landmark region), otherwise 0. Eq. (6) computes the normalized sizes for all landmark regions, and assigns a zero value for all regions from the background environment using $\mathbb{1}(\cdot)$ to distinguish landmark and background regions.

Then, in order to address scale variations of the landmark regions, we introduce the weight matrix $\mathbf{O} = \{o_{ii',jj',kk'} \in \mathbb{R}\}^{nn'\times nn'\times nn'}$ to adjust the third-order spatial similarities $\mathbf{T}$, the weight matrix $\mathbf{P} = \{p_{ii',jj'} \in \mathbb{R}\}^{nn'\times nn'}$ to adjust the second-order spatial similarities $\mathbf{A}$, and the matrix $\mathbf{q} = \{q_{ii'} \in R\}^{nn'}$ to adjust the visual appearance similarities $\mathbf{b}$, which are defined as follows:

$$o_{ii',jj',kk'} = \exp\left(-\frac{1}{\sigma} \sum_{u\in i,j,k; v\in i',j',k'} |w_u - w_v|\right) \tag{7}$$

$$p_{ii',jj'} = \exp\left(-\frac{1}{\sigma} \sum_{u\in i,j; v\in i',j'} |w_u - w_v|\right) \tag{8}$$

$$q_{ii'} = \exp\left(-\frac{1}{\sigma} |w_i - w_{i'}|\right) \tag{9}$$

where $\sigma$ is a hyperparameter of the exponential normalization function. The weights are designed to address the scale change of the landmarks. For example, if two pairs of landmark regions have similar scales (i.e., $s_i \in \mathcal{S}_l$ and $s'_{i'} \in \mathcal{S}'_l$ as well as $s_j \in \mathcal{S}_l$ and $s'_{j'} \in \mathcal{S}'_l$ exhibit similar scales), then $p_{ii',jj'}$ has a large value. If the two pairs of landmark regions show different scales, $p_{ii',jj'}$ has a smaller value. When the pairs contain regions from the background, $p_{ii',jj'}$ also takes small values. $o_{ii',jj',kk'}$ and $q_{ii'}$

After integrating the weight matrices to address the scale changes of landmark regions, we formulate loop closure detection as follows:

$$\begin{aligned}\max_{\mathbf{X}}\ & \lambda_3 \sum_{ii'=1}^{nn'} \sum_{jj'=1}^{nn'} \sum_{kk'=1}^{nn'} o_{ii',jj',kk'} t_{ii',jj',kk'} x_{ii'} x_{jj'} x_{kk'} \\ & + \lambda_2 \sum_{ii'}^{nn'} \sum_{jj'}^{nn'} p_{ii',jj'} a_{ii',jj'} x_{ii'} x_{jj'} \\ & + \lambda_1 \sum_{ii'}^{nn'} q_{ii'} b_{ii'} x_{ii'} \\ \text{s.t.}\quad & \mathbf{X}\mathbf{1}_{n'\times 1} \leq \mathbf{1}_{n\times 1}, \mathbf{X}^\top \mathbf{1}_{n\times 1} \leq \mathbf{1}_{n'\times 1} \end{aligned} \tag{10}$$

We can rewrite the formulation into a concise matrix form:

$$\begin{aligned}\max_{\mathbf{X}}\ & \lambda_3 \mathbf{O} \circ \mathbf{T} \otimes_1 \mathbf{x} \otimes_2 \mathbf{x} \otimes_3 \mathbf{x} \\ & + \lambda_2 \mathbf{x}^\top \mathbf{P} \circ \mathbf{A}\mathbf{x} + \lambda_1 (\mathbf{q} \circ \mathbf{b})^\top \mathbf{x} \\ \text{s.t.}\quad & \mathbf{X}\mathbf{1}_{n'\times 1} \leq \mathbf{1}_{n\times 1}, \mathbf{X}^\top \mathbf{1}_{n\times 1} \leq \mathbf{1}_{n'\times 1} \end{aligned} \tag{11}$$

where $\circ$ denotes the entrywise product.

**Addressing Varying Number of Nodes.** After solving the optimization problem in Eq. (11) to obtain the optimal $\mathbf{X}^*$, the matching score between the query and template images can be computed as the value of the objective function:

$$\begin{aligned} S = & \lambda_3 \mathbf{O} \circ \mathbf{T} \otimes_1 \mathbf{x}^* \otimes_2 \mathbf{x}^* \otimes_3 \mathbf{x}^* \\ & + \lambda_2 (\mathbf{x}^*)^\top \mathbf{P} \circ \mathbf{A}\mathbf{x}^* + \lambda_1 (\mathbf{q} \circ \mathbf{b})^\top \mathbf{x}^* \end{aligned} \tag{12}$$

However, graph representations of input images often contain a varying number of nodes, and this matching score also changes its value proportionally to the number of nodes. In the following, we show that the matching score is bounded.

**Theorem 1.** *The objective function of the optimization problem in Eq. (11) is bounded by* $[0, \sum_{i=1}^{3} \frac{\lambda_i r!}{(r-i)!}]$, *where* $r = \min\{n, n'\}$.

*Proof.* Assume that two graphs $\mathcal{G}$, $\mathcal{G}'$ have $n$ and $n'$ nodes, and $r = \min\{n, n'\}$. Computing the third-order similarity depends on $r$ which is the number of matches between $\mathcal{G}$ and $\mathcal{G}'$. Then, the number of the third-order similarities is equal to the permutation of picking three pairs of nodes from the final matches, which can be calculated by $\frac{r!}{(r-3)!}$. Since $t_{ii',jj',kk'} \in [0, 1]$ and $o_{ii',jj',kk'} \in [0, 1]$, we obtain:

$$0 \leq \mathbf{O} \circ \mathbf{T} \otimes_1 \mathbf{x} \otimes_2 \mathbf{x} \otimes_3 \mathbf{x} \leq \frac{r!}{(r-3)!} \tag{13}$$

Similarly, the number of the second-order similarities is equal to the permutation of selecting two pairs of nodes from the final matches, and the number of the appearance similarities is equal to the permutation of selecting one pair of node from the final matches. So we obtain:

$$0 \leq \mathbf{x}^\top \mathbf{P} \circ \mathbf{A}\mathbf{x} \leq \frac{r!}{(r-2)!} \tag{14}$$

$$0 \leq (\mathbf{q} \circ \mathbf{b})^\top \mathbf{x} \leq \frac{r!}{(r-1)!} \tag{15}$$

Adding these three equations on both size weighted by the hyperparameters, we obtain:

$$\begin{aligned} 0 \leq\ & \lambda_3 \mathbf{O} \circ \mathbf{T} \otimes_1 \mathbf{x} \otimes_2 \mathbf{x} \otimes_3 \mathbf{x} \\ & + \lambda_2 \mathbf{x}^\top \mathbf{P} \circ \mathbf{A}\mathbf{x} + \lambda_1 (\mathbf{q} \circ \mathbf{b})^\top \mathbf{x} \\ \leq & \sum_{i=1}^{3} \frac{\lambda_i r!}{(r-i)!} \end{aligned} \tag{16}$$

Thus, the objective function of the optimization problem in Eq. (11) is bounded by $[0, \sum_{i=1}^{3} \frac{\lambda_i r!}{(r-i)!}]$. □

In order to enable our approach to match graphs that have different numbers of nodes, we calculate the final matching score by $S\left(\sum_{i=1}^{3} \frac{\lambda_i r!}{(r-i)!}\right)^{-1}$, which utilizes the upper bound to normalize the score to take values always between $[0, 1]$ for any number of nodes. Then, we determine whether two places are matched by comparing the normalized matching score with a threshold. Different from existing methods based upon vector-based matching, our approach formulates loop closure detection as a multi-order graph matching problem to compute a matching score directly from graphs.

## Non-Convex Optimization Solver

The optimization problem in Eq.(11) is challenging to solve, since it is a non-convex problem with no closed-form solution (Duchenne et al. 2011). Accordingly, we implement a new solver based upon the general random re-weighted walk framework (Lee, Jungmin and Cho, Minsu and Lee, Kyoung Mu 2011), which is presented in Algorithm 1.

After defining $\mathbf{M} = \lambda_3 \mathbf{O} \circ \mathbf{T}$, $\mathbf{N} = \lambda_2 \mathbf{P} \circ \mathbf{A}$ and $\mathbf{z} = \lambda_1 \mathbf{q} \circ \mathbf{b}$ in Step 1, Step 2 rewrites Eq. (11) as:

$$\max_{\mathbf{x}} \mathbf{M} \otimes_1 \mathbf{x} \otimes_2 \mathbf{x} \otimes_3 \mathbf{x} + \mathbf{x}^\top \mathbf{N} \mathbf{x} + \mathbf{z}^\top \mathbf{x} \tag{17}$$

**Algorithm 1:** An algorithm to solve the formulated non-convex optimization problem in Eq. (11).

**Input :** $\mathbf{O}$ and $\mathbf{T} \in \mathbb{R}^{nn' \times nn' \times nn'}$, $\mathbf{P}$ and $\mathbf{A} \in \mathbb{R}^{nn' \times nn'}$, and $\mathbf{q}$ and $\mathbf{b} \in \mathbb{R}^{nn'}$
**Output:** $\mathbf{X} = \in \{0, 1\}^{n \times n'}$

1: Initialize the vectorized matrix $\mathbf{x} \in \{0, 1\}^{nn'}$;
2: Compute $\mathbf{M} = \mathbf{O} \circ \mathbf{T}$, $\mathbf{N} = \mathbf{P} \circ \mathbf{A}$ and $\mathbf{z} = \mathbf{q} \circ \mathbf{b}$;
3: Compute $\mathbf{M}'$ $\mathbf{N}'$ and $\mathbf{z}'$ according to Eq. (18), Eq. (19), and Eq. (20), respectively;
4: **while** *not converge* **do**
5: Update $\mathbf{x}$ by Eq. (21);
6: Compute the jump vector $\mathbf{j}$ by Eq. (22);
7: Normalize $\mathbf{j}$ using the bistochastic normalization;
8: Update $\mathbf{x}$ with reweighted jumps by Eq. (23);
9: **end**
10: Recover $\mathbf{x}$ to $\mathbf{X}$;
11: Discretize $\mathbf{X}$ using the Hungarian algorithm;
12: **return** $\mathbf{X}$

In Step 3, we convert $\mathbf{M}, \mathbf{N}, \mathbf{z}$ to stochastic forms in order to normalize the original matrix:

$$\mathbf{M}' = \mathbf{M} / \max_i \sum_{j,k} \mathbf{M}_{i,j,k} \tag{18}$$

$$\mathbf{N}' = \mathbf{N} / \max_i \sum_{j} \mathbf{N}_{i,j} \tag{19}$$

$$\mathbf{z}' = \mathbf{z} / \max_i \mathbf{z}_i \tag{20}$$

In Step 5, $\mathbf{X}$ is updated by:

$$\mathbf{x}^{r+1} = \mathbf{M}' \otimes_2 \mathbf{x}^r \otimes_3 \mathbf{x}^r + {\mathbf{x}^r}^\top \mathbf{N}' + \mathbf{z}' \tag{21}$$

where $\mathbf{x}^{r+1}$ denotes the update from $\mathbf{x}^r$.

In Step 6, in order to jump out local optima, inspired by the Page-Rank algorithm (Haveliwala 2002), we implement a re-weighting jump vector $\mathbf{j} \in \mathbb{R}^{nn'}$ as:

$$\mathbf{j} = \exp(\mathbf{x}^r \circ \mathbf{q} / \max(\mathbf{x}^r \circ \mathbf{q})) \tag{22}$$

where $\mathbf{q}$ is applied to guide the jumps toward a direction to match nodes representing regions with similar scales.

Step 7 employs a bistochastic normalization to normalize each row and column in $\mathbf{j}$ so that to enforce the one-to-one correspondence. Then, in Step 8, to facilitate $\mathbf{x}$ to jump out of local optima, $\mathbf{x}$ is updated by:

$$\mathbf{x}^{r+1} = \alpha(\mathbf{M}' \otimes_2 \mathbf{x}^r \otimes_3 \mathbf{x}^r + {\mathbf{x}^r}^\top \mathbf{N}' + \mathbf{z}') + (1 - \alpha)\mathbf{j} \tag{23}$$

where $\alpha$ is a hyperparameter that controls the update rate.

Since $\mathbf{X}$ uses real-valued numbers in the optimization, we discretize it to obtain the binary correspondence matrix $\mathbf{X} \in \{0, 1\}^{n \times n'}$ using the Hungarian algorithm in Step 11.

**Complexity.** The complexity of the optimization problem in Eq. (11) is $O(n^6)$, dominated by $\mathbf{O} \circ \mathbf{T}$. Since we can apply nearest neighborhood search to compute matches locally (Nguyen, Gautier, and Hein 2015), the complexity reduces

Table 1: The St Lucia (Glover et al. 2010) and CMU-VL (Badino, Huber, and Kanade 2012) datasets are used in the experiments to benchmark long-term loop closure detection methods.

| Dataset | Scenario | Statistics | Challenge |
|---|---|---|---|
| St Lucia | Different times of a day | 10 instances $\times$ 22,000 frames, 640 $\times$ 480 | Lighting changes, shadows, dynamic objects |
| CMU-VL | Different months of a year | 5 instances $\times$ 13,000 frames, 1024 $\times$ 768 | Vegetation, weather, and view changes, dynamic objects |

to $O(n^2k)$, where $k$ is the number of nearest neighborhoods. We set $k = n^2$, and the final complexity becomes $O(n^4)$. In our experiments, given 10-25 detected landmarks per image, the runtime of our matching approach is around 160 Hz.

## Experiments

We employ two large-scale long-term loop closure detection datasets to benchmark our approach, including St. Lucia and CMU-VL datasets. Information of this benchmark dataset is presented in Table 1. The precision-recall curve is used as the evaluation metric, which is a standard metric used in the loop closure detection literature (Lowry et al. 2015).

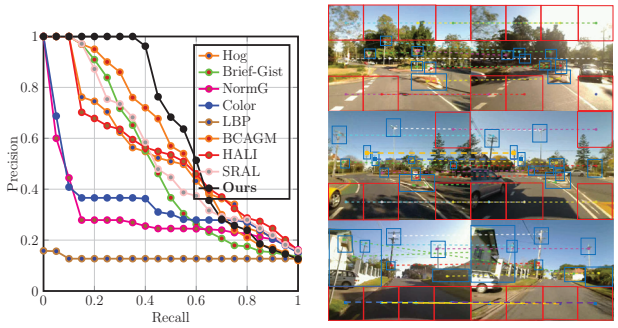
In the experiment, only stable and static landmarks (e.g., houses, traffic signs, and fire hydrants) are used, following recent landmark-based methods (Han et al. 2018; Liu et al. 2019). We use histogram of oriented gradient (Hog) features to describe the visual appearance of each region and landmarks are fully connected to generate graph-based representations. We compare the proposed visual-spatial information preserving multi-order graph matching approach with methods based on visual features, including **SRAL** (Han et al. 2017) that integrates multiple types of visual features by representation learning, and methods based on a single type of features, including **Color** that uses color features, **LBP** that uses local binary pattern features, **Hog** that uses Hog features, **Brief-Gist** that uses Brief-Gist features (Sünderhauf and Protzel 2011), and **NormG** that computes normalized gradients of grayscale images (Milford and Wyeth 2012). We also compare our method with the landmark-based approach **HALI** (Han, Wang, and Zhang 2018), which learns a projection from semantic landmarks to a vector representation, and graph matching-based approach **BCAGM** which only uses spatial relationship of landmarks (Nguyen, Gautier, and Hein 2015) for loop closure detection.

### Results on the St Lucia Dataset

The St Lucia dataset was recorded in the suburban of St Lucia in Australia at different times of a day. GPS information was also collected as the ground truth for vehicle locations.

Quantitative results obtained by our approach are demonstrated in Figure 3(a) based on the precision-recall curve as an evaluation metric. Results from the previous methods are also shown in Figure 3(a) for comparison with our approach. It is seen that the proposed approach outperforms the previous methods compared in this experiment. In order to further study this observation, we assess our approach and compare with other methods using the area below the precision-recall curve as a single-value evaluation metric, which takes values in $[0, 1]$ with a greater value indicating a better performance, and a value 1 indicating the perfect performance. The results are listed in Table 2. It is observed that our approach obtains the score of 0.7207, which significantly outperforms the best visual-appearance based SRAL approach, and the landmark-based HALI method, which uses visual appearances of landmarks but not embeds their spatial relationships. The results show that integrating visual-spatial information in our multi-order graph matching approach can improve performance of long-term loop closure detection.

The qualitative results obtained by our approach on the St Lucia dataset are demonstrated in Figure 3(b), which illustrates three matched places in the query and template images recorded at different times of a day. We observe from these matching results that different numbers of landmarks can be extracted in a same place at different times, and the environment of the same place can look very differently at different times of the day, e.g., due to lighting and shadow variations. Furthermore, we observe that the proposed method can well address these long-term variations, and obtains good matching results for loop closure detection based upon multi-order graph representations with varying numbers of nodes.




(a) Precision-recall curves (b) Matched places

Figure 3: Experimental results on the St Lucia dataset. Figure 3(a) presents quantitative results based on the precision-recall curve as the evaluation metric. Figure 3(b) illustrates qualitative results of place matches in query images recorded at 8:00 AM (left) and template images recorded at 3:00 PM (right). The figures are best viewed in color.

### Results on the CMU-VL Dataset

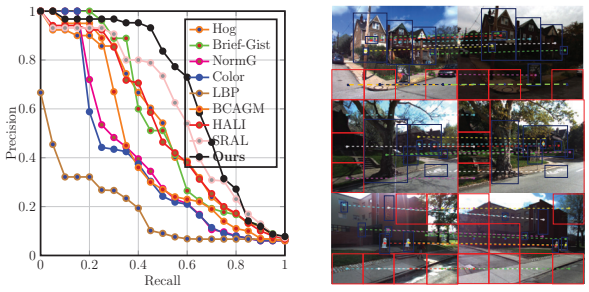
CMU Visual Localization (CMU-VL) benchmark dataset was collected by two cameras installed on a car in different months of a year. GPS data was recorded to provide the ground truth location.

The quantitative results obtained by our method and compared approaches are presented Figure 4(a). We further calculate the area under the precision-recall curve as the single-value performance indicator, as demonstrated in Table 2. We can observe that our approach obtains a score of 0.7452, and

Table 2: Experimental results obtained by our and compared approaches on both datasets. The area under the precision-recall curve is used as a single-value evaluation metric, with a greater value in [0,1] indicating a better performance.

| Approach | St Lucia | CMU-VL |
|---|---|---|
| LBP | 0.1363 | 0.1958 |
| Color | 0.3186 | 0.3970 |
| NormG (Milford and Wyeth 2012) | 0.3672 | 0.4170 |
| Hog (Naseer et al. 2014) | 0.5517 | 0.5514 |
| HALI (Han, Wang, and Zhang 2018) | 0.5206 | 0.5558 |
| Gist-Brief (Sünderhauf and Protzel 2011) | 0.5569 | 0.5612 |
| BCAGM (Nguyen, Gautier, and Hein 2015) | 0.6122 | 0.4878 |
| SRAL (Han et al. 2017) | 0.5630 | 0.6274 |
| **Ours** | **0.7207** | **0.7452** |

it performs better than the compared methods on the CMU-VL dataset that exhibits challenging vegetation and weather variations. The qualitative results obtained by the proposed approach on CMU-VL are demonstrated in Figure 4(b) to illustrate matched places. We observe that our approach well matches places to perform loop closure detection when long-term changes are present (e.g., trees with or without leaves and different lighting conditions).



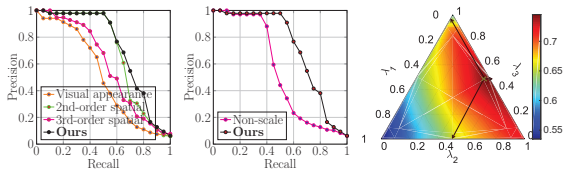

(a) Precision-recall curves (b) Matched locations

Figure 4: Experimental results on the CMU-VL benchmark dataset. Figure 4(a) demonstrates the quantitative results on precision-recall curves. Figure 4(b) shows qualitative results of place matches in query images recorded in March (left) and template images recorded in September (right).

## Discussion

We study the characteristics of our method using the CMU-VL dataset, including importance of visual-spatial cues, robustness to scale changes, and hyperparameter analysis.

**Importance of Visual-Spatial Cues.** The results from approaches that utilize different visual-spatial cues are presented in Figure 5(a). We can see that the spatial relationships are more important than visual appearance cues. Methods utilizing the second and third-order spatial relationships obtain a score of 0.7179 and 0.6178, respectively, based on the area under the precision-recall curve as a metric, which is much better than the score of 0.5350 obtained by the method using visual appearance cues. Combining the second and third-order spatial relationships can lead to a score of 0.7448. Our complete approach that integrates both visual and spatial information further improves performance of loop closure detection and obtains a score of 0.7452.




(a) Info cues (b) Scale (c) Hyperparameters

Figure 5: Analysis of our approach on the CMU-VL dataset. Figure 5(a) compares methods using different visual-spatial cues. Figure 5(b) compares the performance when the scale change is addressed or not. Figure 5(c) depicts performance variations given different hyperparameter values.

**Robustness to Scale Changes.** The results of approaches with and without addressing scale changes are illustrated in Figure 5(b). The area score with addressing scale changes is 0.7452, which is better than the score of 0.5661 when scale changes are not addressed. This generally shows the benefit of addressing scale changes in our proposed approach to improve loop closure detection performance.

**Hyperparameter Analysis.** We present our method's performance changes given different hyperparameter values in Figure 5(c), using the area under the precision-recall curve as the metric. Although our approach has three hyperparameters that control relative weights of information cues in Eq. (12), since they satisfy $\sum_{i=1}^{3} \lambda_i = 1$ and $\lambda_i \geq 0, i = 1, 2, 3$, only two of them are independent. Accordingly, the dependent hyperparameters can be presented and illustrated in the standard simplex topological space (Zhang et al. 2014). For a point in this triangular topological space, the three hyperparameter values can be retrieved along the edge directions. For example, the point marked as a cross in Figure 5(c) represents $\lambda_1 = 0.02$, $\lambda_2 = 0.49$, and $\lambda_3 = 0.49$, which results in the best performance. We can also observe that the performance decreases toward the bottom left corner (i.e., $\lambda_1 = 1$, $\lambda_2 = 0$, and $\lambda_3 = 0$), indicating that only utilizing appearance cues without considering spatial relationships reduces performance. This observation is consistent with the results from the analysis of visual-spatial cues in Figure 5(a).

## Conclusion

We propose the novel visual-spatial information preserving multi-order graph matching method for long-term loop closure detection. It implements a graph representation that fuses both visual appearances and multi-order spatial relationships of image regions representing landmarks and the background environment. It is also based upon a fresh formulation that formulates loop closure detection as a multi-order graph matching problem to compute similarity of query and template images directly from graph representations, instead of performing vector-based place matching. Evaluation on two benchmark datasets have shown our approach outperforms the previous state-of-the-art approaches

for long-term loop closure detection.

## Acknowledgments

This work was partially supported by DOT PHMSA Award 693JK31850005CAAP and NSF grants CNS-1823245, IIS-1849348 and IIS-1849359.

## References

Badino, H.; Huber, D.; and Kanade, T. 2012. Real-time topometric localization. In *ICRA*.

Chen, Z.; Maffra, F.; Sa, I.; and Chli, M. 2017. Only look once, mining distinctive landmarks from ConvNet for visual place recognition. In *IROS*.

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

Duchenne, O.; Bach, F.; Kweon, I.-S.; and Ponce, J. 2011. A tensor-based algorithm for high-order graph matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33(12):2383–2395.

Durrant-Whyte, H., and Bailey, T. 2006. Simultaneous localization and mapping: part i. *IEEE Robotics & Automation Magazine* 13(2):99–110.

Gawel, A.; Del Don, C.; Siegwart, R.; Nieto, J.; and Cadena, C. 2018. X-view: Graph-based semantic multi-view localization. *IEEE Robotics and Automation Letters* 3(3):1687–1694.

Glover, A. J.; Maddern, W. P.; Milford, M. J.; and Wyeth, G. F. 2010. FAB-MAP+ RatSLAM: Appearance-based SLAM for multiple times of day. In *ICRA*.

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.; Wang, H.; Huang, G.; and Zhang, H. 2018. Sequence-based sparse optimization methods for long-term loop closure detection in visual SLAM. *Autonomous Robots* 42(7):1323–1335.

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

Haveliwala, T. H. 2002. Topic-sensitive pagerank. In *WWW*.

Ho, K. L., and Newman, P. 2006. Loop closure detection in SLAM by combining visual and spatial appearance. *Robotics and Autonomous Systems* 54(9):740–749.

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

Lee, Jungmin and Cho, Minsu and Lee, Kyoung Mu. 2011. Hypergraph matching via reweighted random walks. In *CVPR*.

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

Liu, K.; Wang, H.; Han, F.; and Zhang, H. 2019. Visual place recognition via robust l2-norm distance based holism and landmark integration. In *AAAI*.

Lowry, S.; Sünderhauf, N.; Newman, P.; Leonard, J. J.; Cox, D.; Corke, P.; and Milford, M. J. 2015. 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 sinter nights. In *ICRA*.

Mur-Artal, R., and Tardós, J. D. 2014. Fast relocalisation and loop closing in keyframe-based SLAM. In *ICRA*.

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

Naseer, T.; Ruhnke, M.; Stachniss, C.; Spinello, L.; and Burgard, W. 2015. Robust visual SLAM across seasons. In *IROS*.

Newman, P.; Cole, D.; and Ho, K. 2006. Outdoor SLAM using visual appearance and laser ranging. In *ICRA*.

Nguyen, Q.; Gautier, A.; and Hein, M. 2015. A flexible tensor block coordinate ascent scheme for hypergraph matching. In *CVPR*.

Panphattarasap, P., and Calway, A. 2016. Visual place recognition using landmark distribution descriptors. In *ACCV*.

Pronobis, A.; Martinez Mozos, O.; Caputo, B.; and Jensfelt, P. 2010. Multi-modal semantic place classification. *International Journal of Robotics Research* 29(2-3):298–320.

Rabanser, S.; Shchur, O.; and Günnemann, S. 2015. Introduction to tensor decompositions and their applications in machine learning. *Machine Learning* 98(1-2):1–5.

Ramos, F. T.; Fox, D.; and Durrant-Whyte, H. F. 2007. Crf-matching: Conditional random fields for feature-based scan matching. In *Robotics: Science and Systems*.

Ramos, F.; Kadous, M. W.; and Fox, D. 2009. Learning to associate image features with crf-matching. In *Experimental Robotics*, 505–514.

Schönberger, J. L.; Pollefeys, M.; Geiger, A.; and Sattler, T. 2018. Semantic visual localization. In *CVPR*.

Stumm, E.; Mei, C.; Lacroix, S.; and Chli, M. 2015. Location graphs for visual place recognition. In *IEEE International Conference on Robotics and Automation*.

Stumm, E.; Mei, C.; Lacroix, S.; Nieto, J.; Hutter, M.; and Siegwart, R. 2016. Robust visual place recognition with graph kernels. In *CVPR*.

Sünderhauf, N., and Protzel, P. 2011. Brief-Gist-Closing the loop by simple means. In *IROS*.

Sünderhauf, N.; Shirazi, S.; Dayoub, F.; Upcroft, B.; and Milford, M. 2015. On the performance of ConvNet features for place recognition. In *IROS*.

Sünderhauf, N.; Neubert, P.; and Protzel, P. 2013. Are we there yet? Challenging SeqSLAM on a 3000 km journey across all four seasons. In *ICRA Workshop*.

Williams, B.; Cummins, M.; Neira, J.; Newman, P.; Reid, I.; and Tardós, J. 2009. A comparison of loop closing techniques in monocular SLAM. *Robotics and Autonomous Systems* 57(12):1188–1197.

Zhang, H.; Zhou, W.; Reardon, C.; and Parker, L. E. 2014. Simplex-based 3D spatio-temporal feature description for action recognition. In *CVPR*.