2018 IEEE International Conference on Robotics and Automation (ICRA)
May 21-25, 2018, Brisbane, Australia

Omnidirectional Multisensory Perception Fusion for Long-Term Place Recognition

Sriram Siva and Hao Zhang

Abstract— **Over the recent years, long-term place recognition has attracted an increasing attention to detect loops for large-scale Simultaneous Localization and Mapping (SLAM) in loopy environments during long-term autonomy. Almost all existing methods are designed to work with traditional cameras with a limited field of view. Recent advances in omnidirectional sensors offer a robot an opportunity to perceive the entire surrounding environment. However, no work has existed thus far to research how omnidirectional sensors can help long-term place recognition, especially when multiple types of omnidirectional sensory data are available. In this paper, we propose a novel approach to integrate observations obtained from multiple sensors from different viewing angles in the omnidirectional observation in order to perform multi-directional place recognition in long-term autonomy. Our approach also answers two new questions when omnidirectional multisensory data is available for place recognition, including whether it is possible to recognize a place with long-term appearance variations when robots approach it from various directions, and whether observations from various viewing angles are the same informative. To evaluate our approach and hypothesis, we have collected the first large-scale dataset that consists of omnidirectional multisensory (intensity and depth) data collected in urban and suburban environments across a year. Experimental results have shown that our approach is able to achieve multi-directional long-term place recognition, and identifies the most discriminative viewing angles from the omnidirectional observation.**

I. Introduction

Place recognition (or loop closure detection) is an essential component of Simultaneous Localization and Mapping (SLAM), which has been actively studied to achieve SLAM in a loopy environment over the past decades. Recently, place recognition in long-term autonomy has attracted significant attention. Beyond traditional challenges including perceptual aliasing and vision-related issues, long-term place recognition introduces a new, significant challenge – long-term appearance changes [1], [2]. For example, the same outdoor place on a sunny summer noon and during snowy winter evening can look very different. It is recognized [3] that the ability to address long-term appearance variations is essential for robots to perform SLAM during lifelong operations.

Given the importance of long-term place recognition, several representations and matching techniques were proposed mainly to deal with the long-term appearance variation. Both global [4], [5], [6] and local [7], [8] features were studied by existing approaches to represent the scene of a place, with an observation that representations based upon global features often perform better [9]. Place matching techniques based on individual frames [10], [11] or frame sequences [1], [12], [13] were also implemented. However, previous long-term visual place recognition methods assumed that observations are acquired from traditional cameras with a limited field of view. Also, all existing methods perform uni-directional place recognition, assuming that a robot goes back to a previously visited place facing the same direction.

In this work, we investigate the problem of long-term place recognition based on omnidirectional perception that allows a robot to perceive its whole surrounding environment. In particular, we are interested in answering two new technical questions, which have not been addressed in the existing research yet. The questions include: (1) when an omnidirectional observation is used, is it possible to perform multi-directional long-term place recognition in situations that an autonomous system approaches the same place from opposite directions; and (2) whether all angles of view the same informative, or are certain angles of view more representative than others in the omnidirectional perception?

Moreover, we introduce a novel principled approach under the mathematical framework of sparse optimization, which is capable of automatically learning the importance of the viewing angles and integrating omnidirectional perception data to perform multi-directional long-term place recognition. Furthermore, we introduce a multisensory data fusion paradigm under the same framework to integrate heterogeneous visual features that are computed from different types of sensors. Due to our approach's ability to incorporate omnidirectional observations from different sensors, we name our proposed unified method *Fusion of Omnidirectional Multisensory Perception* (FOMP).

The contributions of this work are threefold:

- We introduce a new research problem, that is long-term place recognition based upon omnidirectional multisensory perception, and introduce two technical questions that are critical in omnidirectional perception but have not yet been studied in existing long-term place recognition literature.
- We propose the novel FOMP approach, which estimates the importance of viewing angles and learns discriminative features, as well as integrates all the information to construct a discriminative representation for long-term place recognition in situations when robots approach the same location from different directions. We also implement a new optimization solver to solve the formulated problem, which is guaranteed to converge to the optimal solution theoretically.
- We collect and make available a new large-scale dataset

Sriram Siva and Hao Zhang are with the Human-Centered Robotics Lab in the Department of Computer Science, Colorado School of Mines, Golden, CO 80401, USA {sivasriram, hzhang}@mines.edu

to benchmark methods for multisensory omnidirectional long-term place recognition. Extensive experiments are performed using this new dataset to evaluate FOMP, and to answer the proposed technical questions.

II. Related Work

A. Scene Matching for Place Recognition

Many techniques have been implemented to match a query observation and the scene template of a previously visited location. Based on the approach they take towards matching locations, these methods could be broadly classified into two categories. That is, either they use a sequence of images to assert match between two scenes or they follow a image-to-image matching paradigm. Sequence-based matching depends on a sequence of images to find the best matches between query and template image sequences, such as used by SeqSLAM [1] and RAT-SLAM [12]. SeqSLAM computes a summation of similarity scores of query images and template sequences to find the best match location. Typically, methods based on image-to-image matching calculates the distance metric between a query image and existing templates, with the maximum score indicating a scene match [14]. Several techniques also use nearest neighbours search for finding the best match. For example, FAB-MAP [11] uses a Chow-Liu tree to get the best match and RTAB-MAP uses a K-d tree to perform the nearest neighbour search.

The previous methods are designed to work with traditional cameras with limited views, and cannot integrate omnidirectional observations, or perform multi-directional long-term place recognition, which is the focus of our research.

B. Representation in Long-Term Place Recognition

Most existing place-recognition methods rely on features to construct a representation with the hope to capture long-term scene variations. When environments show changes in illumination conditions, global features outperform local features [1], [3], [9]. Global features extract features from the whole image, and often create a representation using histograms. For example, HOG [5] uses unsigned gradient changes within each pixels of a grid and stores it as a histogram. GIST features [4], [15] employ Gabor filters at different orientations and frequencies to extract information from the images. Convolutional neural networks (CNN) [16], [17], [18] are employed to create a representation for matching image sequences. Local Binary Patterns (LBP) are used to encode scenes by labelling pixels of an image by thresholding the neighbourhood of each pixel, which constructs a representation denoted as a binary vector [6]. Depth information from Kinect-like sensors [19] is also been used for object-based SLAM.

Recent methods such as [20] show advantages to identify important features and fuse them together to achieve better performance on long-term place recognition. In this research, we follow the same insight and integrate feature learning as a part of the proposed approach under the unified optimization framework to improve place recognition during long-term autonomy, through fusing heterogeneous features from various omnidirectional sensors.

III. The FOMP Approach

We aim at addressing the new problem of utilizing omnidirectional observations to perform long-term place recognition when robots revisit the same place from different directions. To address this challenge, we propose the FOMP approach to estimate the importance of viewing angles and sensor modalities, and to integrate all multisensory omnidirectional data to perform multi-directional long-term place recognition.

Notation. Matrices are denoted using boldface-capital letters, and vectors are denoted by boldface lower-case letters. Given a matrix $\mathbf{U} = \{u_{ij}\} \in \Re^{n\times m}$, we denote the i-th row and j-th column as $\mathbf{u}^i$ and $\mathbf{u}_j$ respectively. The ℓ_1-norm of a vector $\mathbf{u} \in \Re^n$ is defined as $\|\mathbf{u}\|_1 = \sum_{i=1}^{n} |u_i|$. The ℓ_2-norm of a vector $\mathbf{u}$ is defined as $\|\mathbf{u}\|_2 = \sqrt{\mathbf{u}^\top \mathbf{u}}$. The Frobenius norm of a matrix $\mathbf{U}$ is defined as $\|\mathbf{U}\|_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} u_{ij}^2}$.

A. Problem Formulation

Given a collection of omnidirectional images acquired in different scenarios, each image is equally divided into a set of views. We also assume that multimodal features are extracted from each view, where a modality of features is defined as the features computed using a specific descriptor from images acquired by a specific visual sensor (e.g., intensity or depth sensor). Then, the set of multisensory omnidirectional images can be expressed as $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n] \in \Re^{p\times n}$, where $\mathbf{x}_i \in \Re^p$ denotes the feature vector extracted from all views of the i-th image, which is a concatenation of features from m modalities, such that $p = \sum_{i=1}^{m}\sum_{j=1}^{a} d_{ij}$, where d_{ij} is the dimensionality of the i-th modality in the j-th view, and a is the total number of views. The label vector of scenarios (e.g., different seasons) associated with $\mathbf{X}$ is represented by $\mathbf{Y} = [\mathbf{y}_1; \mathbf{y}_2; \ldots; \mathbf{y}_n] \in \Re^{n\times c}$, where c is the number of scenarios, and $\mathbf{y}_i$ is the scenario indicating vector, with elements $y_{ij} \in \{0, 1\}$ representing that the i-th image is collected from the j-th scenario.

Then, the problem of omnidirectional multisensory place recognition is formulated as a regularized sparse optimization problem:

$$\min_{\mathbf{W}} \mathcal{L}(\mathbf{X}, \mathbf{Y}; \mathbf{W}) + \lambda \mathcal{R}(\mathbf{W}) \tag{1}$$

where $\mathcal{L}(\cdot)$ is a loss function and $\mathcal{R}(\cdot)$ is a sparsity-inducing regularizer with $\lambda \geq 0$ as a trade-off hyperparameter. $\mathbf{W}$ denotes the weight matrix, which represents the importance of the features $\mathbf{X}$ to represent the scenarios $\mathbf{Y}$ in general.

We define a new loss function to address multi-directional place recognition (i.e., identification of the same place from different directions) as follows:

$$\min_{\mathbf{W}} \|(\mathbf{R}\mathbf{X})^T\mathbf{W} - \mathbf{Y}\|_F^2 \tag{2}$$

where $\mathbf{R}$ is the rotation matrix for aligning omnidirectional images with the same origin. For example, when two omnidirectional images are take by a car driving on the two sides

of a road respectively, then one image must be rotated $180°$ to align with the other image, which is performed using $\mathbf{R}$.

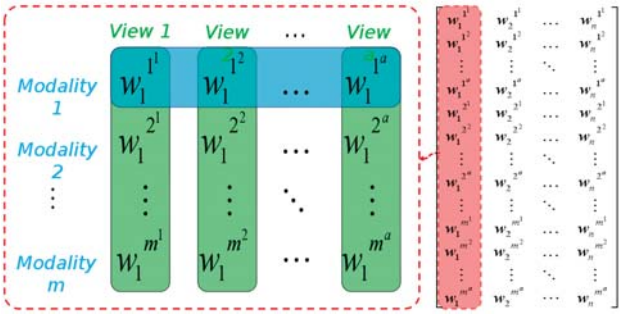


Fig. 1. Illustration of the weight matrix $\mathbf{W}$.

The solution to this optimization task is the weight matrix $\mathbf{W} = [\mathbf{w_1}, \mathbf{w_2}, \ldots, \mathbf{w_c}] \in \Re^{p \times c}$, which contains the weights $\mathbf{w}_i \in \Re^p$ of all modalities and views with respect to the i-th scenario. Each $\mathbf{w}_i$ contains weights of m-modalities from all views, which can be expanded as $\mathbf{w}_i = [\mathbf{w}_i^1, \mathbf{w}_i^2, \ldots, \mathbf{w}_i^m]^T$. In addition, the weight vector of each modality $\mathbf{w}_i^m$ can be further divided into a segments, each from a particular view, as follows $\mathbf{w}_i^m = [\mathbf{w}_i^{m^1}; \mathbf{w}_i^{m^2}; \ldots; \mathbf{w}_i^{m^a}] \in \Re^{d_{ij}}$, representing the weights of the features extracted in different views from m-th modality and i-th scenario. The weight matrix is graphically represented in Fig. 1.

B. Learning Discriminative Views

An omnidirectional image provides $360°$ field of view, that allows robots to observe the entire surrounding environment. We hypothesize that for long-term place recognition, specific views in the omnidirectional image are more discriminative than others. In order to automatically identify these views, we introduce a novel cone-structured sparsity-inducing norm to learn the discriminative features under our unified regularized optimization framework.

Formally, each leaf node of the introduced cone structure contains the features extracted from an individual view, and each internal node of the cone contains the features from its respective child nodes, which represents a combination of the views represented by the child nodes. We represent the set of nodes as $\mathcal{V} = (\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{root})$, each $\mathbf{v}$ containing features collected from a certain view or multiple adjacent views, and we denote the weights of the features from respective views as $(\mathbf{w}_{\mathbf{v}_1}, \mathbf{w}_{\mathbf{v}_2}, \ldots, \mathbf{w}_{\mathbf{v}_{root}})$. For example, the cone structure with four leaf nodes, each including the features obtained from a $90°$ view, is demonstrated in Fig. 2, with the 3D cone structure illustrated in Fig. 2(a) and the unwrapped structure shown in Fig. 2(b).

Then, we propose a method to compute the *weight of each node* $w(\mathbf{v})$ in the cone structure, which is a scalar that indicates the importance of the node, as follows:

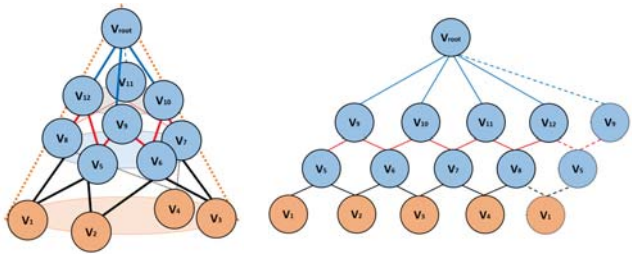
$$w(\mathbf{v}) = \begin{cases} \widetilde{h}_{\mathbf{v}} \sum_{C(\mathbf{v})} \|\mathbf{w}_{C(\mathbf{v})}\|_1 + h_{\mathbf{v}} \sum_{C(\mathbf{v})} \|\mathbf{w}_{C(\mathbf{v})}\|_2 \\ \qquad\qquad \text{if } \mathbf{v} \text{ is an internal node} \\ \sum_{\mathbf{v}} \|\mathbf{w}_{\mathbf{v}}\|_1 \qquad\qquad \text{if } \mathbf{v} \text{ is a leaf node} \end{cases}$$

where $\widetilde{h}_{\mathbf{v}} = 1 - h_{\mathbf{v}}$, and $h_{\mathbf{v}}$ is the normalized height of a node $\mathbf{v}$ with respect to the height of the cone structure. At lower levels of the cone structure, $h_{\mathbf{v}}$ takes smaller values; thus ℓ_2-norm is more significant. Moving toward the upper level of the cone, $h_{\mathbf{v}}$ increases, so ℓ_1-norm becomes more dominant. The variable $h_{\mathbf{v}}$ is designed to incorporate the principle that at lower levels, the grouping effect of different views should be promoted and at higher levels sparsity among the grouped views should be emphasized.

Based upon the weights of all nodes, the cone-structured sparsity-inducing norm is defined as $\|\mathbf{W}\|_C = \sum_{\mathbf{v} \in \mathcal{V}} w(\mathbf{v})$. which allows for discovering discriminative views by assigning a greater weight to features from discriminative views. When using the cone-structure sparsity-inducing norm as a regularization term, we obtain the new objective function:

$$\min_{\mathbf{W}} \|(\mathbf{RX})^T \mathbf{W} - \mathbf{Y}\|_F^2 + \lambda \|\mathbf{W}\|_C \tag{3}$$

where λ is a hyperparameter used to balance the loss function and the regularization term.



(a) 3D cone structure (b) Unwrapped structure

Fig. 2. Illustration of the proposed cone-structured sparsity-inducing norm for learning discriminative views.

C. Learning Discriminative Modalities

Modern robots are usually equipped with different types of sensors (e.g., intensity and depth sensors), and we can extract different type of features from observations obtained by each sensor. We employ the term *modality* to refer to a set of features extracted by a type of feature extraction method from the data obtained by a specific sensor. In this case, some feature modalities are typically more descriptive than others. For example, in a dark environment, observations from depth sensors are often more useful than data from color cameras. In this research, we also propose to identify discriminative features under the unified optimization framework to improve place recognition accuracy.

Inspired by [20], we incorporate a modality norm, named M-norm, as a regularization term to enforce sparsity among different modalities thus identifying discriminative modalities. The M-norm applies the ℓ_2-norm within each modality and the ℓ_1-norm across different modalities. The M-norm is mathematically expressed as $\|\mathbf{W}\|_M = \sum_{i=1}^{c} \sum_{j=1}^{m} \|\mathbf{w}_i^j\|_2$.

Incorporating both sparsity-inducing norms to model the relationship among various views and modalities, we observe the final objective function:

$$\min_{\mathbf{W}} \|(\mathbf{RX})^T \mathbf{W} - \mathbf{Y}\|_F^2 + \lambda_1 \|\mathbf{W}\|_C + \lambda_2 \|\mathbf{W}\|_M \tag{4}$$

where λ_1 and λ_2 are the trade-off hyper-parameters to balance the loss function and the sparsity-inducing norms.

D. Omnidirectional Multisensory Place Recognition

After solving the regularized optimization problem in Eq. 4 (using Algorithm 1, described in the next subsection), we obtain the optimal weight matrix $\mathbf{W}^* \in \Re^{p\times c}$.

Given a feature vector $\mathbf{x}_i \in \Re^p$ of a query omnidirectional multisensory observation, we can compute a similarity score between this query observation and the template as follows:

$$s = \sum_{i=1}^{a}\sum_{j=1}^{m} w_A(i) * w_M(j) * s^{j^i} \tag{5}$$

where s^{j^i} denotes the similarity score between the observation and the template in i-th view of j-th modality, $w_M(j)$ is the optimal weight of the j-th modality, and $w_A(i)$ is the optimal weight of the i-th view. If the similarity score thus calculated is above a user-defined threshold, the query image is decided as matching to the template. The weight of the i-th view is computed as $w_A(i) = \sum_{j=1}^{m} \|\mathbf{w}^{j^i}\|_2, i = 1, 2, \ldots, a$, and the weight of the j-th modality is computed as $w_M(j) = \|\mathbf{w}^j\|_2, j = 1, 2, \ldots, m$.

Algorithm 1: An iterative algorithm to solve the formulated optimization problem in Eq. (4)

Input : feature matrix $\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_n] \in \Re^{p\times n}$ and ground truth matrix $\mathbf{Y} = [\mathbf{y}_1, \cdots, \mathbf{y}_n]^T \in \Re^{n\times c}$ from the training set

Let $t = 1$. Initialize $\mathbf{W}(t)$ by solving $\min_{\mathbf{W}} \|(\mathbf{RX})^T\mathbf{W} - \mathbf{Y}\|_F^2$.

while *not converge* **do**

  Calculate the block diagonal matrix $\mathbf{D}(t+1)$, where the k-th diagonal block of $\mathbf{D}(t+1)$ is $\frac{1}{2\|\mathbf{w}^k(t)\|_2}$.

  Calculate the block diagonal matrix $\widetilde{\mathbf{D}}(t+1)$, where the diagonal block of $\widetilde{\mathbf{D}}(t+1)$ is $\frac{1}{2\|\mathbf{W}(t)\|_C}\mathbf{I}_{id}$.

  For each $\mathbf{w}_i (1 \le i \le c)$, $\mathbf{w}_i(t+1) = ((\mathbf{RX})(\mathbf{RX})^T + \gamma_1\mathbf{D}(t+1) + \gamma_2\widetilde{\mathbf{D}}(t+1))^{-1}(\mathbf{RX})\mathbf{y}_i$.

  $t = t + 1$.

Output: $\mathbf{W} = \mathbf{W}(t) \in \Re^{p\times c}$

E. Optimization Algorithm

The objective function in Eq. (4) comprises of non-smooth regularization terms, which is challenging to solve in general. Thus, we implement a new iterative algorithm to solve this formulated optimization problem.

Taking the derivative of the objective function with respect to the columns of $\mathbf{W}$ (i.e., $\mathbf{w}_i, i = 1, \ldots, c$) and setting the whole equation to a zero vector gives us:

$$(\mathbf{RX})(\mathbf{X}^T\mathbf{R}^T)\mathbf{w}_i - (\mathbf{RX})\mathbf{y}_i + \gamma_1\mathbf{D}\mathbf{w}_i + \gamma_2\widetilde{\mathbf{D}}\mathbf{w}_i = 0 \tag{6}$$

where $\widetilde{\mathbf{D}}$ is a diagonal matrix with the i^{th} diagonal element as $\frac{1}{2\|\mathbf{w}^i\|_2}$, and $\mathbf{D}$ is defined as the diagonal matrix with the diagonal block as $\frac{1}{2\|\mathbf{W}\|_C}\mathbf{I}_{id}$, where $\mathbf{I}_{id}$ is an identity matrix of size p. Since the matrices $\mathbf{D}$ and $\widetilde{\mathbf{D}}$ are dependent on the vectors of $\mathbf{W}$, we develop an iterative algorithm to solve the optimization problem with these unknown variables, as described in Algorithm 1, which holds a theoretical convergence guarantee as described by the following theorem.

***Theorem** 1:* Algorithm 1 converges to the optimal solution to the optimization problem in Eq. 4.

Proof: See supplementary materials[1]. ■

IV. Omnidirectional Multisensory Dataset

One of the contributions of this research is the collection of large-scale omnidirectional multisensory datasets. Although various sensors and omnidirectional cameras are increasingly widely deployed on robots and autonomous cars, before this research, no dataset containing omnidirectional multisensory information is publicly available for benchmarking long-term place recognition. Motivated by this need, we collected the new dataset called MOLP, which stands for *Multimodal Omnidirectional Long-term Place-recognition*. The dataset was collected using a omnidirectional camera installed on a SUV to collect omnidirectional intensity and depth information. The MOLP dataset includes two sub-datasets obtained from two different routes:

- Route-A: *Mines-Downtown Golden*. This route contains scenes from the Colorado School of Mines and downtown of Golden CO. This route is 4.3 miles long and the dataset consists of 3000-7000 images in each of the 16 instances from different long-term scenarios across a year. The dataset also captures the short-term dynamics such as traffics, construction work, and pedestrians.
- Route-B: *Historic Suburban Golden*. This route contains scenes of the trip where the gold-rush era started 150 year ago. This driving route is 7.6 miles long, from the circuitous suburban Golden and to the Rocky Mountains. The dataset consists of 2500-5000 images in each of the 16 instances from different long-term scenarios across a year. Beyond long-term appearance changes, this route has the severe challenge of perceptual aliasing because of the similar winding roads while driving.

The MOLP dataset is publicly available and more details are discussed on the dataset website at `http://hcr.mines.edu/code/MOLP.html`.

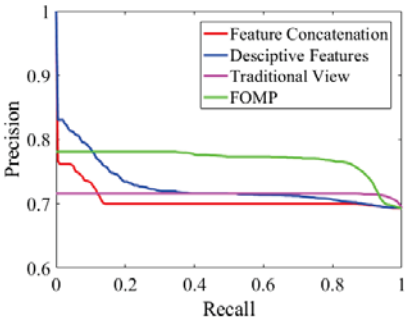
V. Experiments

A. Experimental Setup

The Summer and Fall scenarios from the dataset are used in the experiments. Each omnidirectional image is vertically split into 18 sections, which corresponds to 18 views, each including 20° field of view. Then, each split image is downsampled to a resolution of 210*240, which consists of both

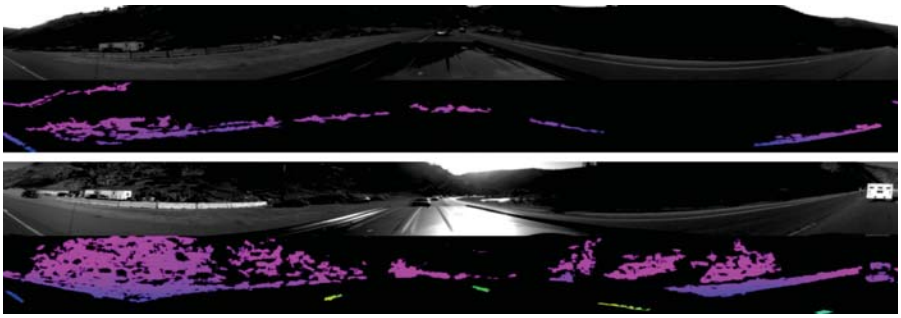
[1]The proof is available at: `hcr.mines.edu/publication/FOMP_Supp.pdf`

(a) Matched omnidirectional multisensory observations

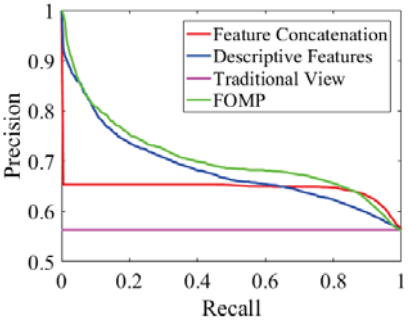


(b) Precision-recall curves

Fig. 3. Results over Route-A across different times of a day (i.e., morning versus evening).



(a) Matched omnidirectional multisensory observations



(b) Precision-recall curves

Fig. 4. Results over the Route-B dataset across different times of a day (i.e., morning versus evening).

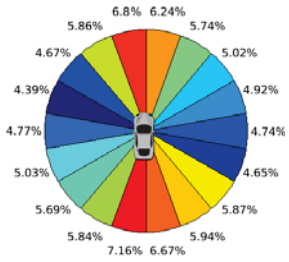
intensity and depth information. No image processing is further performed on these images. For ground truth we use the GPS data recorded at the time of collecting the dataset.

Four different types of visual features are extracted from each of the intensity and depth images in our experiments, including GIST [4], HOG [5], LBP [6], and CNN-based deep features [17], [16]. All features are separately extracted from both intensity and depth images. Moreover, we implement several techniques from the literature to compare with our FOMP approach, including representations by concatenated features and discriminative features. Furthermore, we implement a method using the traditional front 80° field of view as a baseline. For all experiments, we set 0.1 to the hyper-parameter λ_1, and 0.05 to λ_2.

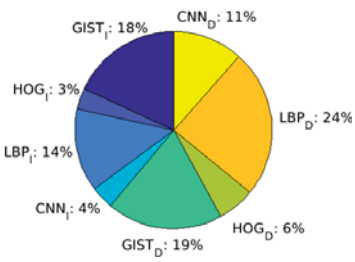
B. Results at Different Times of a Day

To evaluate our approach on place recognition with long-term appearance changes across different times of a day and with environment dynamics, we perform experiments using morning and evening scenarios for both Route-A and Route-B in the MOLP dataset.

The qualitative result on the Route-A dataset is illustrated in Fig. 3(a), which includes a detected match of omnidirectional intensity-depth images in the query and template. It can be observed from the intensity image that the same place exhibits very different illumination conditions in the morning versus evening, and also contains different dynamics because of varying traffic and pedestrians. In this challenging scenario, the match shows our FOMP approach can well perform place recognition with long-term appearance variations. In the experiments over the Route-B dataset, we observe similar qualitative results, with an exemplary match depicted in Fig. 4(a), which shows our FOMP approach is able to match the places with the presence of long-term illumination changes.



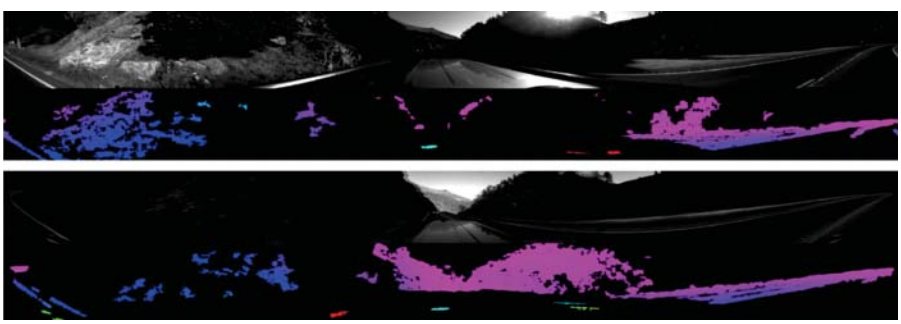
(a) Descriptive views



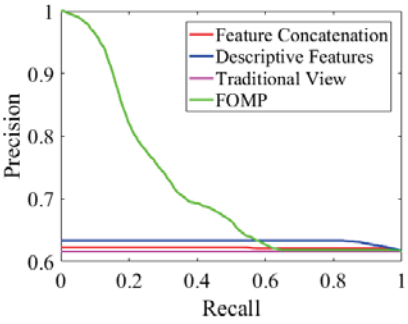
(b) Descriptive modalities

Fig. 5. Experimental results on view and modality importance using the Route-A dataset across various times of a day.

To provide a quantitative evaluation, the standard metric of precision-recall curves is used. Fig. 3(b) shows the precision-recall curves obtained over the morning and evening scenes of Route-A. The results for Route-B is illustrated in Fig. 4(b). Comparisons with baseline techniques are also presented in the respective figures. We observe that integrating omnidirec-



(a) Matched omnidirectional multisensory observations



(b) Precision-recall curves

Fig. 6. Results over the Route-B dataset across different seasons (i.e., Summer versus Fall).

tional information can improve the performance (as shown in Fig. 3(b)). In addition, our FOMP method outperforms the baseline techniques, due to the ability to identify and fuse discriminative views and feature modalities.

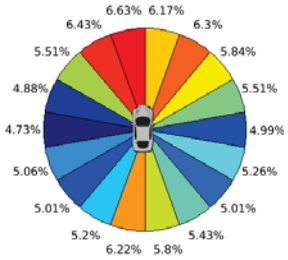
In addition, we perform experiments to evaluate the importance of different viewing angles, with the results shown in Fig. 5(a) for Route-A. The relative importance of each view is presented as a heat chart, with a warmer color denoting greater importance, and the vehicle is facing up as the front. It can be observed that the front and back views in the omnidirectional observation are the most descriptive for long-term place recognition across different times of the day, and the sideways of the observation are less descriptive. Finally, we perform experiments to assess the importance of different sensing modalities, i.e., different types of features acquired from various sensors. The results are shown in Fig. 5(b) for Route-A, with the numbers denoting the importance of the modality. We observe consistent results on modality importance: the four most important modalities are depth-LBP, depth-GIST, intensity-GIST, and depth-CNN.

C. Results across Different Seasons

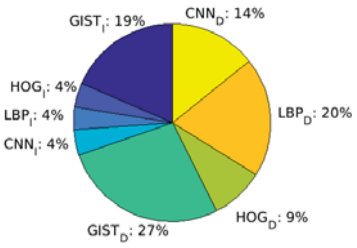
To evaluate FOMP over a longer span of time, we perform experiments using omnidirectional multisensory observations across different seasons, in which the places show significant appearance changes caused by weather and vegetation.

As the qualitative experimental result, examples of the detected omnidirectional multisensory matches between query and template observations are illustrated in Fig. 6(a) for Route-B. The results show FOMP is able to well perform place recognition under different vegetation and illumination conditions across different seasons. The precision-recall curves are illustrated in Fig. 6(b). We can observe that the proposed approach outperforms the baseline techniques on long-term place recognition across different seasons.

The importance of viewing angles for Route-B across different seasons is illustrated in Fig. 7(a). Results in Route-B are a bit different from previous observations: the front view is relatively more discriminative than the back view, although both front and back views are still more important than side views for omnidirectional long-term place recognition. Finally, the quantitative results over modality weights are illustrated in 7(b) for Route-B across different seasons, respectively. Similar to the results obtained from different times of a day, the modality importance is also consistent with the same top four most descriptive modalities. When we sum up the weights of features acquired by either the intensity or depth sensor, we observe depth data weights more over intensity data, which indicates that depth observations are more important. A possible explanation is that the intensity cues are more sensitive to the long-term appearance variation such as various illumination and weather, while the depth information determined by the environment topology is less affected by appearance changes.



(a) Descriptive Views



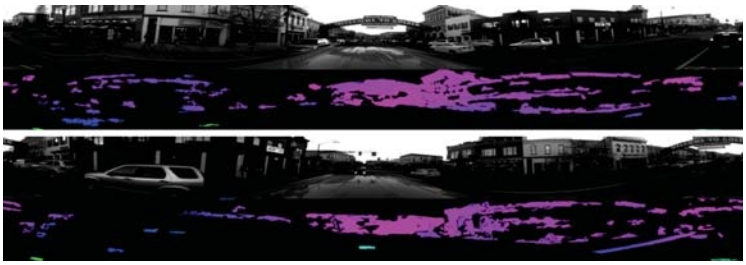
(b) Descriptive Modalities

Fig. 7. Experimental results on view and modality importance using the Route-B dataset across different seasons.

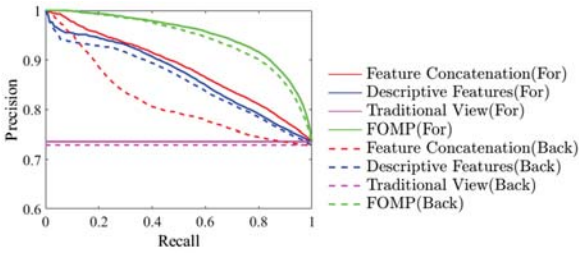
D. Results on Multi-directional Place Recognition

One unique capability of the FOMP approach is to achieve multi-directional place recognition, which is enabled by introducing the rotation matrix in our problem formulation to address different robot's orientations. This new advantage is validated and evaluated in this set of experiments. We train our approach using the data from one direction only (either forward or backward), then we evaluate its performance on long-term place recognition using data from both directions.

An exemplary location match detected by FOMP is illustrated in Fig. 8(a). Because the vehicle drives through the same place from different directions, it is observed that the query observation is rotated around 180° comparing to the template, which is clearly demonstrated by the arched sign (which reads "WELCOME TO GOLDEN"). Since the rotation matrix $\mathbf{R}$ in our formulation is able to model this rotation,



(a) Detected matches when the car drives from opposite directions



(b) Precision-recall curves

Fig. 8. Results of bidirectional place recognition when a car drives through the same place from different directions in different seasons. In the legend of precision-recall curves, "*For*" indicates the FOMP approach is trained using forward-direction data only, and "*Back*" indicates the approach is trained using backward-direction data only. Testing is performed using data collected from both directions.

FOMP can successfully recognize the same place when the car approaches it from different directions.

Precision-recall curves obtained from bidirectional place recognition are illustrated in Fig. 8(b). We can observe three key phenomena. First, for bidirectional place recognition, the methods using traditional camera do not work, as expected. Second, our FOMP approach significantly outperforms other baseline techniques, mainly due to its capability of modeling rotations. Third, because of the same reason, FOMP methods trained on forward-direction data or backward-direction data obtain consistent good performance. This consistent capability highlights our approach for bidirectional place recognition in long-term autonomy.

VI. Conclusion

In this paper, we propose a new problem of place recognition from omnidirectional multisensory observations in long-term autonomy. To address this challenge, we introduce the novel FOMP approach that is able to identify and integrate discriminative multimodal data obtained from heterogeneous sensors in different views. Our approach shows that different viewing angles in the omnidirectional observation have different description powers. Our research also demonstrates that multi-directional long-term place recognition is achievable. To validate our FOMP approach and hypothesis, we collect a large-scale dataset containing omnidirectional multisensory observations. Experimental results on this dataset have demonstrated that FOMP obtains promising long-term place recognition performance.

VII. Acknowledgment

This work was funded in part by the ARO grant W911NF-17-1-0447.

References

[1] M. J. Milford and G. F. Wyeth, "Seqslam: Visual route-based navigation for sunny summer days and stormy winter nights," in *International conference on Robotics and Automation*, 2012.

[2] F. Han, H. Wang, and H. Zhang, "Learning of integrated holism-landmark representations for long-term loop closure detection," in *Association for the Advancement of Artificial Intelligence*, 2018.

[3] T. Naseer, L. Spinello, W. Burgard, and C. Stachniss, "Robust visual robot localization across seasons using network flows," in *Conference on Artificial Intelligence*, 2014.

[4] A. Oliva and A. Torralba, "Building the gist of a scene: The role of global image features in recognition," *Progress in Brain Research*, vol. 155, pp. 23–36, 2006.

[5] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Conference on Computer Vision and Pattern Recognition*, 2005.

[6] R. Arroyo, P. F. Alcantarilla, L. M. Bergasa, and E. Romera, "Towards life-long visual localization using an efficient matching of binary sequences from images," in *International Conference on Robotics and Automation*, 2015.

[7] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features-SURF," *Computer Vision and Image Understanding*, vol. 110, no. 3, pp. 346–359, 2008.

[8] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardos, "Orb-SLAM: a versatile and accurate monocular slam system," *IEEE Transactions on Robotics*, vol. 31, no. 5, pp. 1147–1163, 2015.

[9] S. Lowry, N. Sünderhauf, P. Newman, J. J. Leonard, D. Cox, P. Corke, and M. J. Milford, "Visual place recognition: A survey," *IEEE Transactions on Robotics*, vol. 32, no. 1, pp. 1–19, 2016.

[10] M. Labbe and F. Michaud, "Appearance-based loop closure detection for online large-scale and long-term operation," *IEEE Transactions on Robotics*, vol. 29, no. 3, pp. 734–745, 2013.

[11] M. Cummins and P. Newman, "FAB-MAP: Probabilistic localization and mapping in the space of appearance," *The International Journal of Robotics Research*, vol. 27, no. 6, pp. 647–665, 2008.

[12] M. J. Milford, G. F. Wyeth, and D. Prasser, "Rat-SLAM: a hippocampal model for simultaneous localization and mapping," in *International conference on Robotics and Automation*, 2004.

[13] H. Zhang, F. Han, and H. Wang, "Robust multimodal sequence-based loop closure detection via structured sparsity.," in *Robotics: Science and Systems*, 2016.

[14] C. Chen and H. Wang, "Appearance-based topological bayesian inference for loop-closing detection in a cross-country environment," *The International Journal of Robotics Research*, vol. 25, no. 10, pp. 953–983, 2006.

[15] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *International Journal of Computer Vision*, vol. 42, no. 3, pp. 145–175, 2001.

[16] N. Sunderhauf, S. Shirazi, A. Jacobson, F. Dayoub, E. Pepperell, B. Upcroft, and M. Milford, "Place recognition with convnet landmarks: Viewpoint-robust, condition-robust, training-free," *Proceedings of Robotics: Science and Systems*, 2015.

[17] N. Sünderhauf, S. Shirazi, F. Dayoub, B. Upcroft, and M. Milford, "On the performance of convnet features for place recognition," in *Intelligent Robots and Systems*, 2015.

[18] Z. Chen, O. Lam, A. Jacobson, and M. Milford, "Convolutional neural network-based place recognition," *Australian Conference of Robotics and Automation*, 2014.

[19] R. F. Salas-Moreno, R. A. Newcombe, H. Strasdat, P. H. Kelly, and A. J. Davison, "Slam++: Simultaneous localisation and mapping at the level of objects," in *Conference on Computer Vision and Pattern Recognition*, 2013.

[20] F. Han, X. Yang, Y. Deng, M. Rentschler, D. Yang, and H. Zhang, "SRAL: Shared representative appearance learning for long-term visual place recognition," *IEEE Robotics and Automation Letters*, vol. 2, no. 2, pp. 1172–1179, 2017.