Scalable Representation Learning for Long-Term Augmented Reality-Based Information Delivery in Collaborative Human-Robot Perception

Fei Han, Sriram Siva, and Hao Zhang(✉)

Human-Centered Robotics Lab, Colorado School of Mines,
1500 Illinois Street, Golden, CO 80401, USA
{fhan,sriramsiva,hzhang}@mines.edu

Abstract. Augmented reality (AR)-based information delivery has been attracting an increasing attention in the past few years to improve communication in human-robot teaming. In the long-term use of AR systems for collaborative human-robot perception, one of the biggest challenges is to perform place and scene matching under long-term environmental changes, such as dramatic variations in lighting, weather and vegetation across different times of the day, months, and seasons. To address this challenge, we introduce a novel representation learning approach that learns a scalable long-term representation model that can be used for place and scene matching in various long-term conditions. Our approach is formulated as a regularized optimization problem, which selects the most representative scene templates in different scenarios to construct a scalable representation of the same place that can exhibit significant long-term environment changes. Our approach adaptively learns to select a small subset of the templates to construct the representation model, based on a user-defined representativeness threshold, which makes the learned model highly scalable to the long-term variations in real-world applications. To solve the formulated optimization problem, a new algorithmic solver is designed, which is theoretically guaranteed to converge to the global optima. Experiments are conducted using two large-scale benchmark datasets, which have demonstrated the superior performance of our approach for long-term place and scene matching.

Keywords: Collaborative human-robot perception · Representation learning · Augmented Reality · Long-term information delivery

1 Introduction

Augmented Reality (AR) has been attracting an increasing attention in industry and academia, which provides a revolutionary technology to insert virtual objects into the real world through the use of a head-mounted display or a hand-held

J. Y. C. Chen and G. Fragomeni (Eds.): HCII 2019, LNCS 11575, pp. 47–62, 2019.
https://doi.org/10.1007/978-3-030-21565-1_4

mobile device [1,2]. In particular, by overlaying digital information on top of the real scene, AR provides a promising solution to more intuitively and interactively deliver information to humans, which can be applied to improve communications between robots and humans in the critical application of human-robot teaming. For example, such information may include restaurant ratings and descriptions of a building in a city's downtown area, or a tagged damage for further inspection or repair in indoor or underground infrastructure (e.g., power plant boilers, subway tunnels, and pipeline networks), collected by mobile robots and labeled by other teammates. To tether the robot-collected information correctly and stably to reality, the AR system, as a component of communication in human-robot teams, must either estimate the location and orientation of the user, or match the real scene with a database that includes information of the same scene from previous visits. Place and scene matching is especially essential, when human-robot teams work in GPS-denied (e.g., underground infrastructure) and GPS-limited (e.g., a downtown area with tall buildings) areas. With accurate matches of the current scene with previous scenes in the database, labels associated with previous scenes can be displayed over the current scene (Fig. 1).

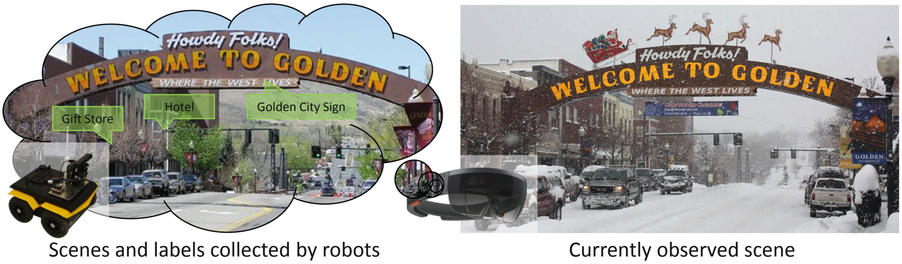


Fig. 1. A motivating example of long-term place and scene matching for long-term AR-based information delivery in collaborative human-robot perception applications. Due to long-term environment changes such as weather, lighting, and vegetation variations, the current scene observed by a camera of AR systems may look significantly different from stored scenes of the same place collected previously by robots.

One biggest challenge of matching a currently observed scene and place with previous scenes is to address long-term changes of the environment during long-term use of an AR system. Occlusion and viewpoint differences are typical problems in conventional scene matching problems. Besides those, the long-term scene and place matching problem is even more challenging since the AR system needs to operate in various scenarios. The appearance of the same place can drastically change in different times of the day, months and seasons. Many factors can cause the appearance changes, for example, lightening changes, weather changes, and vegetation condition changes. In addition, multiple places could have a similar appearance (e.g., two chain stores in Colorado and California may look similarly), which is usually called perceptual aliasing. It is another challenge that makes the long-term scene and place matching problem hard for AR-based information delivery in collaborative human-robot perception applications.

Due to its importance, several approaches on the long-term scene and place matching were investigated, mainly by researchers from robotics and computer vision communities [3–5], for example, to perform camera localization and loop closure detection for simultaneous localization and mapping (SLAM) [6–9]. Many previous techniques formulate long-term scene and place matching into an image-vs-image matching problem using either local features or global features. Typical local features used in long-term scene matching include SIFT [10], SURF [11], ORB [12], while HOG [13], GIST [14], and CNN [15] are widely used global features. However, scene and place matchings based upon single images cannot address perceptual aliasing well. As an improved paradigm, sequence-vs-sequence matching was demonstrated to have better performance to address perceptual aliasing, by introducing additional temporal and spatial information of scenes and places [16,17]. However, image-vs-image and sequence-vs-sequence matchings cannot incorporate the rich information recorded from different scenarios. The methods only compare the currently query scene with *one and only one* existing template acquired from a specific scenario in the database.

In this paper, we propose a novel approach of learning a scalable long-term representation model that adaptively integrates information extracted from multiple environmental conditions to improve encoding power of long-term perceptual variations in order to enable scene matching for long-term use of AR systems. Given its advantage, we refer to the approach as *Learning Of Representation with Scalability* (LORS). Formulated as a regularized optimization problem, LORS learns the representativeness of multiple scene templates (instead of only one template as in conventional methods) recorded in multiple scenarios of the each place. Then LORS adaptively selects the most representative subset of templates to build the representation model for that place, which incorporates representative place information in different scenarios, as shown in Fig. 2. Since LORS is capable of selecting a small subset of the most representative templates, it scales well to large-scale real-world AR applications when identifying a big number of places from observations collected from a big number of long-term scenarios.

The contributions of this research are twofold:

- We propose the novel LORS approach to learn a representation model that adaptively integrates a set of sequence templates extracted under multiple environmental scenarios, which provides a comprehensive representation of long-term perceptual variations for more robust place and scene in long-term use of AR systems. The existing scene matching methods based upon single images and sequences are special cases of the proposed LORS method.
- We introduce a novel formulation to construct the representation model under the general regularized optimization framework, in order to select only a small number of most representative templates, which makes it applicable to large-scale long-term AR-based information delivery in collaborative human-robot

perception applications. A new optimization solver is also implemented to address the formulated problem with a convergence guarantee.

The remainder of the paper is organized as follows. In Sect. 2, we discuss our LORS approach in detail. In Sect. 3, experimental results are presented. Finally, we conclude our paper in Sect. 4.

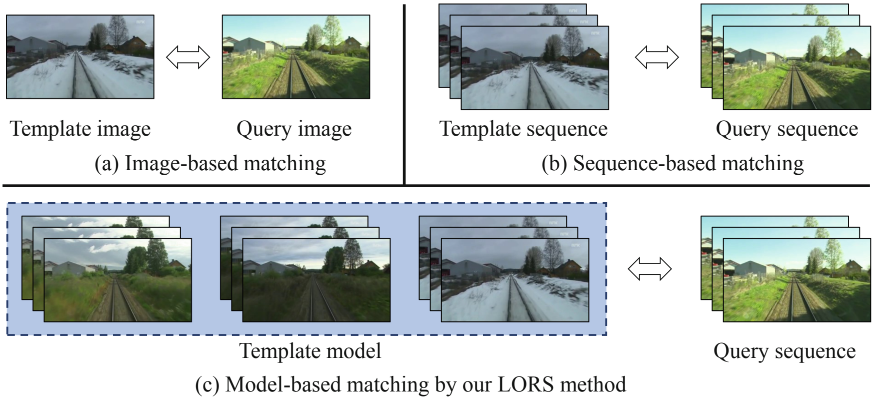


Fig. 2. Overview of the proposed LORS approach to learn an representation model for each place that integrates multiple sequence scene templates extracted in various environmental conditions for long-term AR-based information delivery. Our representation model is constructed by adaptively selecting a small number of the most representative sequence templates. LORS is more general and representative than the previous image-based (Fig. (a)) and sequence-based (Fig. (b)) matching techniques that only use one and only one template to match with the query observation. In addition, LORS is scalable in real-world long-term autonomy applications due to its adaptive, representative sequence selection capability.

2 LORS for Long-Term Scene and Place Matching

In this section, we introduce our novel LORS approach to learn the representation model for each place, which is adaptively constructed by representative templates recorded in different scenarios in a long period of time. We formulate the problem into a novel optimization problem with structured sparsity regularization. In addition, we also developed a new optimization algorithm to solve the formulated non-smooth optimization problem, with the theoretical convergence guarantee.

Notations in this paper follow the following standards: Vectors are denoted as boldface lowercase letters, while matrices use boldface capital letters. For a given matrix $\mathbf{M} = \{m_{ij}\} \in \mathbb{R}^{n \times m}$, its i-th row and j-th column are referred as $\mathbf{m}^i$ and $\mathbf{m}_j$, respectively. The ℓ_1-norm of a vector $\mathbf{v} \in \mathbb{R}^n$ is defined as $\|\mathbf{v}\|_1 = \sum_{i=1}^{n} |v_i|$, and the ℓ_2-norm of $\mathbf{v}$ is defined as $\|\mathbf{v}\|_2 = \sqrt{\mathbf{v}^\top \mathbf{v}}$. The $\ell_{2,1}$-norm of the matrix $\mathbf{M}$ is defined as: $\|\mathbf{M}\|_{2,1} = \sum_{i=1}^{n} \sqrt{\sum_{j=1}^{m} m_{ij}^2} = \sum_{i=1}^{n} \|\mathbf{m}^i\|_2$, and the Frobenius norm is defined as $\|\mathbf{M}\|_F = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} m_{ij}^2}$.

2.1 Problem Formulation

To solve the critical long-term place and scene matching problem, a sequence of frames are collected to represent the place in different scenarios (e.g., different times of the day, months, or seasons). For a specific place p, the feature vectors extracted from the sequential frames in different scenarios are represented as $\mathbf{X}(p) = [\mathbf{x}_1(p), \cdots, \mathbf{x}_s(p)] \in \mathbb{R}^{d \times s}$, where $\mathbf{x}_i(p) = \left[(\mathbf{x}_i^1(p))^\top, \cdots, (\mathbf{x}_i^f(p))^\top\right]^\top \in \mathbb{R}^{d \times 1}$ is a concatenated feature vector of f images in scenario i, and the feature length for each image $\mathbf{x}_i^j(p), j = 1, \cdots, f$ is d^j satisfying $d = \sum_{j=1}^{f} d^j$. s denotes the number of scenarios in the long-term span.

Though sequences of the same place in different s scenarios are recorded for the representation, it is obvious that not all of them are unique and representative. For example, the sequences captured when passing through a tunnel in summer and winter can be largely identical, though it is not true for those on open roads. We are interested in seeking representative sequences that can represent the place in various scenarios in a long period. According to the formulation above, we are trying to select $r (r \leq s)$ template sequences that are most representative in long-term for each place p, respectively, which can be formulated to solve:

$$\min_{\mathbf{W}(p)} \|\mathbf{X}(p)\mathbf{W}(p) - \mathbf{X}(p)\|_F^2 + \lambda \|\mathbf{W}(p)\|_{2,1}, p = 1, \cdots, c \tag{1}$$

where $\mathbf{W}(p) = [\mathbf{w}_1(p), \cdots, \mathbf{w}_s(p)] \in \mathbb{R}^{s \times s}$, and $\mathbf{w}_i(p)$ is the weight of the sequence template candidates to represent the i-th candidate (i-th column) in $\mathbf{X}(p)$. The $\ell_{2,1}$-norm based regularization enforces the sparsity among all sequence template candidates, which means only part of representative sequences are selected to represent all other sequences. c is the total number of places to be distinguished. There is a specific weight matrix $\mathbf{W}(p)$ for place p. For simplicity, we omit p in $\mathbf{X}(p)$ and $\mathbf{W}(p)$ as $\mathbf{X}$ and $\mathbf{W}$ in the following presentation, respectively.

After solving Eq. (1), the rows $\mathbf{w}^i, i = 1, \cdots, s$ are sorted by the value of $\|\mathbf{w}^i\|_1$ in decreasing order, and the resulted row-sorted matrix $\mathbf{W}'$ is obtained. Then, our LORS model enables to adaptively select the most representative sequence templates. This encodes our *insight* that the number of templates in the model should vary according to the degree of the appearance variation of a specific place. For example, a place inside a tunnel requires fewer sequence templates as the appearance does not show significant long-term variations (e.g., it is not affected by snow or sunshine); on the other hand, places on the road in an open area require more templates to represent long-term changes in different times of a day and seasons. Given $\mathbf{W}'$, our model determines the minimum value of r that satisfies $\frac{1}{s}\sum_{i=1}^{r}\|\mathbf{w}'^{i}\|_1 \geq \gamma$. Then, the r sequence template candidates (columns of $\mathbf{X}$) are selected corresponding to the top r rows of $\mathbf{W}'$, where γ is a threshold encoding the expected overall representativeness of the selected sequence template candidates, called the *representativeness threshold*. By this mechanism, not all captured sequences will be treated as the sequence templates, which makes our model highly scalable in real-world applications while still keeps the representativeness among different places and the robustness under different environmental conditions.

Intuitively, when there are no appearance changes during the long-term navigation period, only one sequence template candidate will obtain a high row-sum value (Others have a value close to 0 due to the sparsity effect by the $\ell_{2,1}$-norm regularization), which will be selected as the single template for this place. On the other hand, when the place experiences significant appearance variations, no single sequence template candidate can well represent others. In this case, the rows of $\mathbf{W}$ will become much less sparse and a set of sequence templates can have a high row-sum value, resulting in multiple sequence templates in the top rows of $\mathbf{W}'$ to be selected as templates. Therefore, the proposed LORS model is able to adaptively select a varying number of sequence template candidates based on their different appearance variation degree. Since LORS only requires a subset of templates instead of all, it is highly storage efficient in real-world applications.

Our LORS model is different from the traditional Bag of Words (BoWs) technique. Firstly, the sequence-based representation is applied in our model, which incorporates temporal information while BoWs approaches discard it. Sequence-based scene and place matching has be demonstrated to have better performance than image-based methods [16,18,19]; Secondly, our LORS model enables to select the top representative sequence templates, while BoWs cannot. The LORS mechanism scales well when places are recorded in various scenarios.

2.2 Long-Term Scene and Place Matching

The optimal weight matrix $\mathbf{W}^* = \left[(\mathbf{w}^1)^*; \cdots ; (\mathbf{w}^s)^*\right] \in \mathbb{R}^{s\times s}$ can be obtained after solving the optimization problem in Eq. (1) using Algorithm 1, which is detailed in Sect. 2.3. Then, the representative sequence templates $\mathbf{X}^* \in \mathbb{R}^{d\times r}$ for place p are selected according to the corresponding top r rows satisfying $\frac{1}{s}\sum_{i=1}^{r}\|\mathbf{w}^{i}\|_1 \geq \gamma$.

For long-term scena and place matching in the testing phase, we are given a new query sequence represented by the feature $\mathbf{x}_q \in \mathbb{R}^{d\times 1}$. We then calculate the matching score $\text{score}_{p,i}, p = 1, \cdots, c, i = 1, \cdots, r$ between the query observation sequence and each sequence template i for each place p by feature similarity. Then, the query place q can be identified as

$$q = \underset{p}{\text{argmax}}\, \text{score}_{p,i} \tag{2}$$

Algorithm 1. An iterative algorithm to solve the sparse optimization problem in Eq. (1).

```
Input  : Sequence-based features w.r.t. observations in different scenarios
         $\mathbf{X} \in \mathbb{R}^{d \times s}$
Output: The weight matrix $\mathbf{W} \in \mathbb{R}^{s \times s}$
Let $t = 1$, and initialize $\mathbf{W}(t) \in \mathbb{R}^{s \times s} = \text{argmin}_{\mathbf{W}} \|\mathbf{XW} - \mathbf{X}\|_F^2$;
while not converge do
    Calculate the diagonal matrix $\mathbf{D}(t+1)$ with the $i$-th diagonal element as
    $\frac{1}{2\|\mathbf{w}^i(t)\|_2}$, where $\mathbf{w}^i(t)$ is the $i$-th row of $\mathbf{W}(t)$;
    For each $\mathbf{w}_i$ $(1 \le i \le s)$, calculate $\mathbf{w}_i(t+1) = \left(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{D}(t+1)\right)^{-1}\mathbf{X}^\top\mathbf{x}_i$;
    $t = t + 1$.
end
return $\mathbf{W} \in \mathbb{R}^{s \times s}$.
```

2.3 Optimization Algorithm

The optimization problem in Eq. (1) is convex and can be reformulated and solved as a second-order cone programming (SOCP) or semidefinite programming (SDP) problem. However, solving SOCP or SDP is computationally expensive in general. In this section, we propose an algorithm to solve the formulated optimization in Eq. (1) efficiently with theoretical convergence guarantee.

Taking the derivative of Eq. (1) for each place p with respect to each column of $\mathbf{W}$ and setting it to $\mathbf{0}$, we have

$$\mathbf{X}^\top\mathbf{X}\mathbf{w}_i - \mathbf{X}^\top\mathbf{x}_i + \lambda\mathbf{D}\mathbf{w}_i = \mathbf{0}, \tag{3}$$

where $\mathbf{D}$ is a diagonal matrix with the i-th diagonal element as $\frac{1}{2\|\mathbf{w}^i\|_2}$, $i = 1, \cdots, s$.

Therefore, $\mathbf{w}_i$ can be calculated by

$$\mathbf{w}_i = \left(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{D}\right)^{-1}\mathbf{X}^\top\mathbf{x}_i. \tag{4}$$

It is observed that the matrix $\mathbf{D}$ in Eq. (4) depends on the weight matrix $\mathbf{W}$, which is also unknown. In order to solve this problem, an iterative solver is presented in Algorithm 1. Algorithm 1 can be proved to guarantee the theoretical convergence to the global optima.

3 Experiments

To evaluate the performance of our LORS approach, we conducted experiments on two public benchmark datasets: CMU-VL dataset and Nordland dataset. Our prior work [5] has shown that HOG descriptors can achieve significant performance in comparison to other descriptors (e.g. color, CNN, GIST, etc.) in these two datasets for long-term scene and place matching. Thus, we select the HOG descriptor for every single frame in both experiments. It aims to ensure that the performance increase results from the proposed LORS approach instead of raw feature engineering. Though the HOG descriptor is selected in our experiments, any descriptor can be used in our LORS approach. In addition, multimodal representations by combining multiple descriptors also work in our LORS approach.

3.1 Results over Different Months

The CMU Visual Localization (CMU-VL) dataset [20] is a public benchmark dataset that recorded a 8.8 Km route under a variety of scenarios across different months throughout the entire year. It was recorded by a car with two cameras mounted on the roof of the it and oriented to left and right respectively. GPS data were also measured and recorded to be used as the ground truth of the recorded places. The environmental conditions in the CMU-VL dataset vary a lot across different months of the year (e.g. sunny, snowy, partial cloudy, with green vegetation or reduced colored vegetation, etc.), which makes it very challenging to recognize the same place in such a long period of time. Since multiple recordings of the same route are used to evaluate the proposed LORS method, we have to align them strictly before the experiments. We use the GPS information w.r.t. each frame of different recorded videos to find the same place under different scenarios.

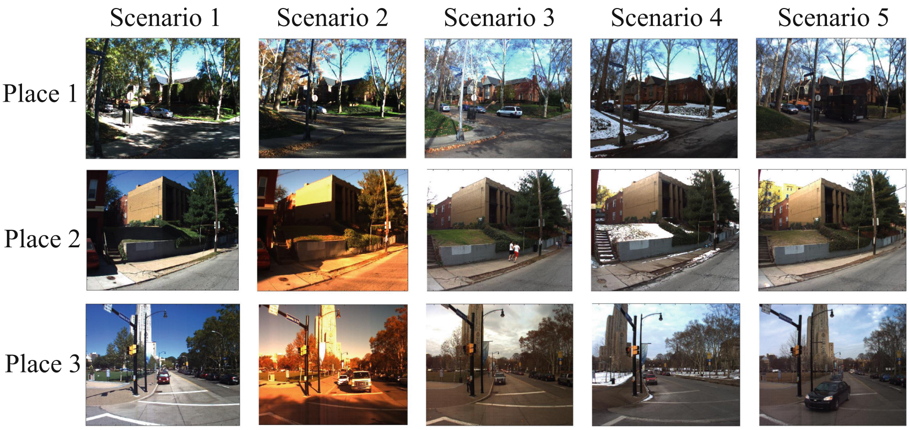


Fig. 3. Three example places and their scenes in five different scenarios in the experiment using the CMU-VL dataset.

The scenarios considered in the experiment via the CMU-VL dataset include:

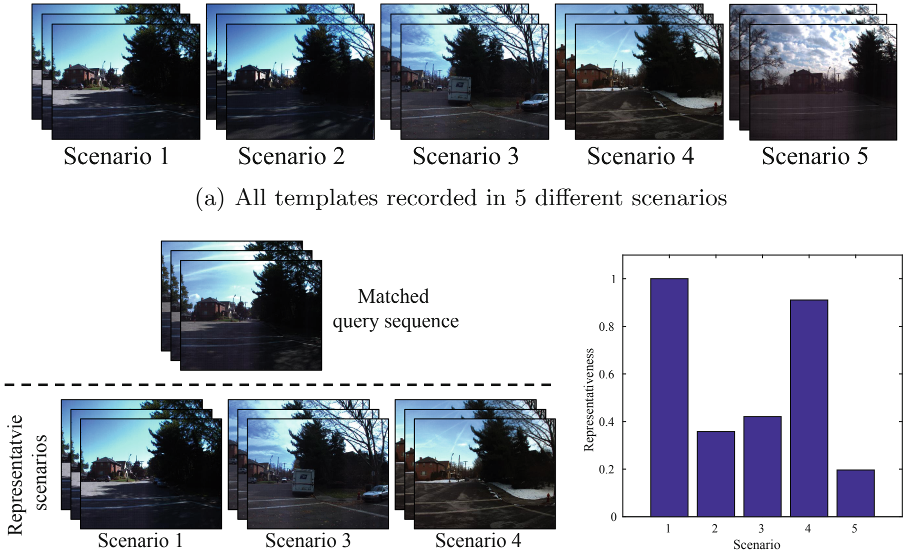
1. Mid September: sunny with abundant green vegetation and vertical shadows
2. Early November: sunny with reduced colored vegetation and fallen leaves
3. Late November: sunny with strong slanted shadows

4. Late December: cloudy with lots of snow on ground
5. Early March: partially cloudy with some shadows

which is also illustrated in Fig. 3. The five videos recorded in these five scenarios respectively are used to train the LORS model. Without loss of generality, a new video recorded in the same first scenario (Mid September) is used as the testing data to evaluate the performance of our proposed LORS method.

The representative templates of each place are obtained during the training phase using the proposed LORS method. After that, the new unseen query observations recorded in Scenario #1 is used to assess the performance. The qualitative evaluation is illustrated in Fig. 4(a), where Fig. 4(a) shows all templates recorded in five different scenarios as illustrated in Fig. 3. Instead of applying all five templates in the testing phase, three representative ones (Scenario #1, 3, and 4, shown in Fig. 4(b)) are identified by our LORS method and are used to represent the place. In the testing phase, the same place has been successfully recognized as shown in Fig. 4(b). The representativeness of each template is quantified in Fig. 4(c), where we can see the templates in Scenario #1, 3 and 4 are three top representative ones for the place when the threshold $\gamma = 40\%$.

We also quantitatively evaluate our LORS method in Fig. 5, where Fig. 5(a) and (b) show the precision-recall curves with respect to different number of templates for place representation and different representativeness threshold in



(a) All templates recorded in 5 different scenarios

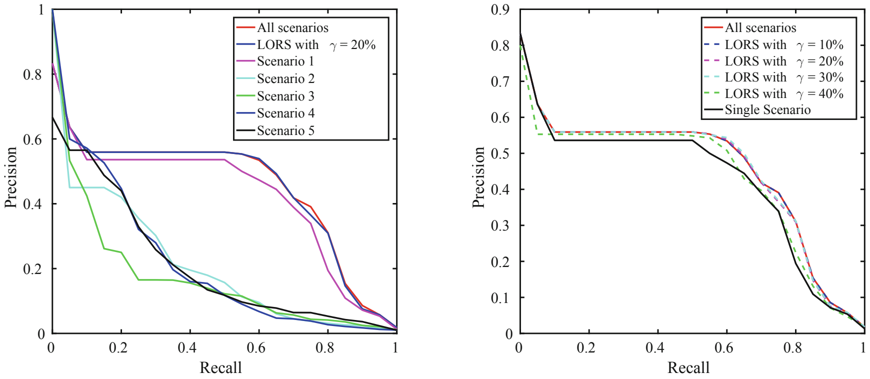
(b) Templates and the matched query sequence

(c) Template representativeness

Fig. 4. Qualitative evaluation of our LORS approach over the CMU-VL dataset across different months.

the LORS method, respectively. From Fig. 5(a) we observe that the best performance is achieved when all 5 templates are used for the place representation, which provides the largest amount of information for each place. Figure 5(a) also shows our LORS method almost has the same performance when the representativeness threshold $\gamma = 20\%$. In addition, our LORS method outperforms other cases when the single template is used for the place representation. Using the single template recorded in Scenario #1 has much better performance than those recorded in Scenario #2, 3, 4 and 5. That is because the query observations in testing are also recorded in Scenario #1, which means it reduces to the 'short-term' place matching problem in this case. On the other hand, the performance is decreased significantly when the training and testing scenarios are inconsistent, showing the poor robustness of the traditional single-scenario-based methods in the long-term scene and place matching problem.

Figure 5(a) demonstrates that incorporating more place information (with more templates) results in better place recognition performance. However, it will require a lot of data storage and suffer from the processing speed. Our LORS method enables to select most representative templates while does not have too much performance decrease. Figure 5(b) and Table 1 further evaluate the LORS performance with respect to different representativeness threshold. Higher representativeness threshold γ indicates more templates (templates with representativeness less than γ) will be discarded for the place representation. From Fig. 5(b) and Table 1, it is observed that the performance (area size below the precision-recall curves) decrease is significantly smaller than the percentage of templates that are discarded, demonstrating the superior place representation capability by our LORS method.



(a) Precision-recall curves computed using different training scenarios

(b) Precision-recall curves computed using varying representativeness threshold

Fig. 5. Quantitative evaluation of our LORS approach on the CMU-VL dataset across different months.

Table 1. Performance decrease with respect to different degrees of information loss (Representativeness threshold γ) over the CMU-VL and Nordland dataset.

Representativeness threshold γ	Performance decrease over CMU-VL	Performance decrease over Nordland
10%	0	0
20%	0.132%	0
30%	0.375%	0
40%	5.05%	2.05%
50%	12.5%	2.22%
60%	15.6%	3.29%

3.2 Results over Different Seasons

We also evaluate the performance of LORS via the Nordland dataset. Nordland dataset [4] is another public benchmark dataset that records the scenes recorded by a self-driving train in a ten-hour long trip traveling around 3000 km in Nordland. Visual data in four seasons were recorded and aligned strictly frame by frame in the dataset. The video has a 1920 × 1080 resolution and 25 frames per second (FPS).

There are significant appearance changes in the Nordland dataset, which are caused by various weather, vegetation and illumination conditions in four seasons. For example, there is almost full snow coverage on the ground in winter while with green vegetation in summer. In addition, the journey passes through many wild places with similar appearances, which means the dataset has strong perceptual aliasing problem. All these difficulties make the Nordland one of the most challenging dataset for long-term place and scene matching. In this experiment, the videos are downsampled with 640 × 360 resolution and 5 FPS.

The previous experiment via the CMU-VL dataset demonstrates significant performance of the proposed LORS approach when the testing scenario is the same as the one of the training scenarios, that is, the testing environmental condition is experienced in the training process. On the other hand, we are also interested in the case when the testing scenario is never experienced during training, since there are numerous combinations of environmental conditions in real-world collaborative human-robot perception applications. In this experiment over the Nordland dataset, the videos recorded in Summer, Autumn and Winter are used in the representation model learning by our LORS method, which are shown in Fig. 6, and the video recorded in Spring is used for testing, which is never experienced before.

The representative templates of each place are obtained during the training phase using the proposed LORS method. After that, the new unseen query observations recorded in Spring is used to assess the performance. The qualitative evaluation is illustrated in Fig. 7, where Fig. 7(a) shows all seasonal templates recorded in Summer, Autumn, and Winter. Instead of applying all three templates in the testing phase, two representative ones (Summer and Autumn shown in Fig. 7(b)) are identified by our LORS method and are used to represent the

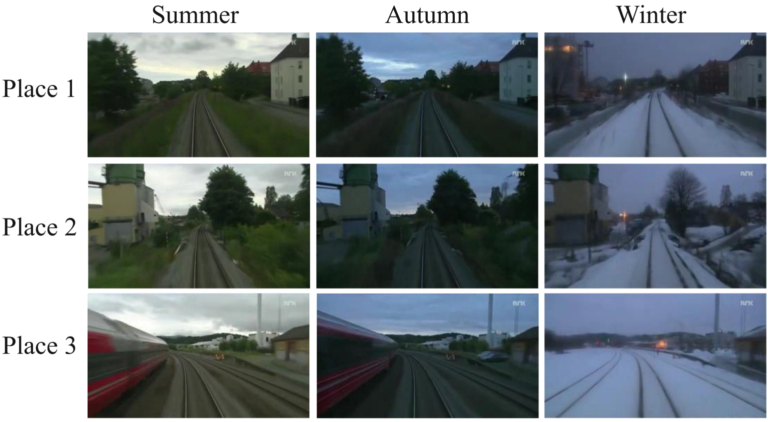
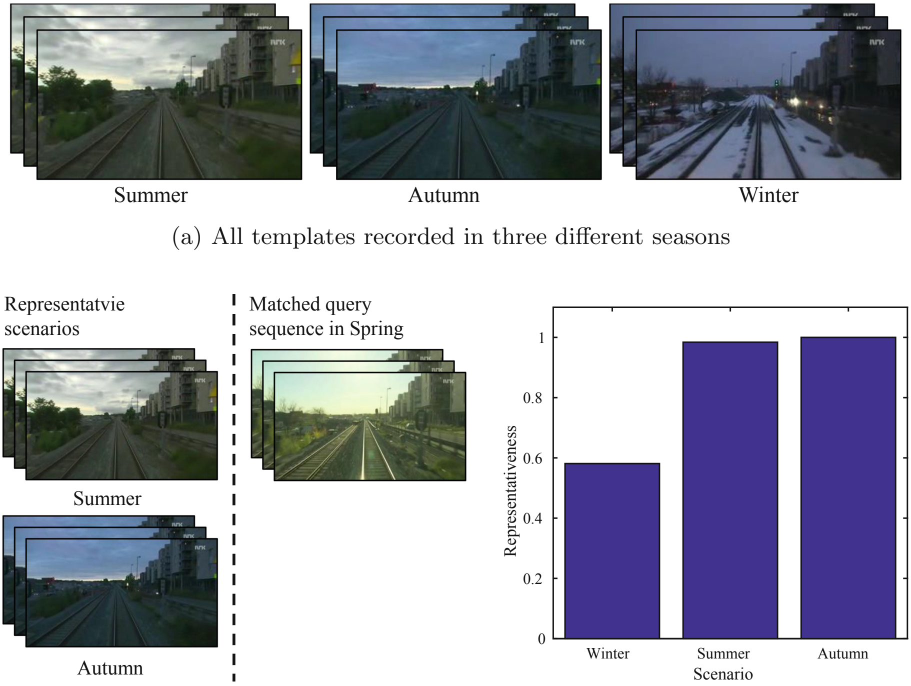


Fig. 6. Three example places and their scenes in three different seasons in the experiment over the Nordland dataset.

place. In the testing phase, the same place has been successfully recognized as shown in Fig. 7(b) though the Spring scenario is never experienced in the training process. The representativeness of each template is quantified in Fig. 7(c), where we can see templates in Summer and Autumn are two top representative ones for the place when the threshold $\gamma = 60\%$.

Similar to the experiment over the CMU-VL dataset, we also quantitatively evaluate our LORS method via the Nordland dataset in Fig. 8, where Fig. 8(a) and (b) show the precision-recall curves with respect to different number of templates for place representation and different representativeness threshold in the LORS method, respectively. From Fig. 8(a) we observe that the best performance is achieved when all 3 seasonal templates are used for the place representation, which provides the largest amount of information for each place. Figure 8(a) also shows our LORS method almost has the same performance when the representativeness threshold $\gamma = 35\%$. In addition, our LORS method outperforms other cases when the single template is used for the place representation, showing the great benefits from multiple template adoption as well as representative template learning even when the environmental condition in testing phase is never experienced before. Different from the previous experiment over the CMU-VL dataset, using the single template recorded in any single season cannot perform well in the long-term scene and place matching (low precision and recall values as shown in Fig. 8(a)). That is because the query observations in testing are recorded in Spring, which is never experienced during training.

Figure 8(b) and Table 1 also evaluate the LORS performance with respect to different representativeness threshold γ, from which it is observed that the performance (area size below the precision-recall curves) decrease is significantly smaller than the percentage of templates that are discarded, demonstrating the superior place representation capability by our LORS method. We are able to balance the long-term place matching performance and scalability degree for



(a) All templates recorded in three different seasons

(b) Templates and the matched query

(c) Template representativeness

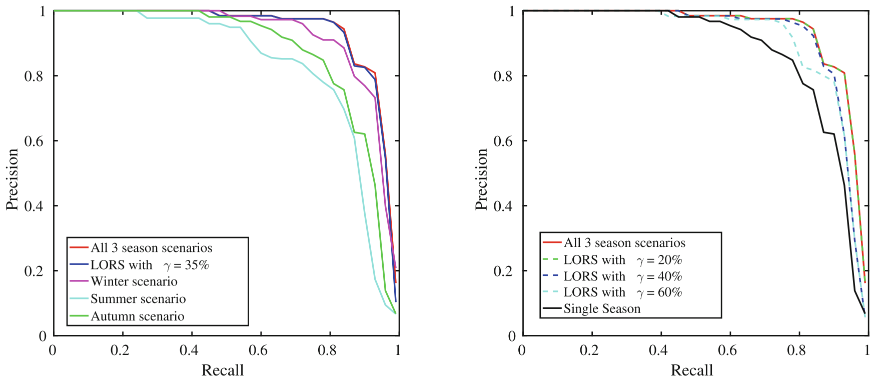
Fig. 7. Qualitative evaluation of our LORS approach over the Nordland dataset across different seasons.

real-world AR-based information delivery in human-robot collaboration applications by the γ parameter of the proposed LORS method.

3.3 Discussion

The main parameters of the LORS approach are discussed and analyzed in this subsection. Without loss of generality, the experimental results via the Nordland dataset is selected to evaluate the effects of the parameter selection in our LORS method, which are illustrated in Fig. 9. We have similar results over the CMU-VL dataset.

The sequence length f is one of the most important parameter in our LORS method. The precision-recall curves in Fig. 9(a) indicate that better long-term place and scene matching accuracy can be achieved when the sequence length f is increased. There is more comprehensive information contained in longer sequences, especially in the Nordland dataset that has strong perceptual aliasing problems. When $f = 1$, the sequence-based place representation reduces to the single image-based representation losing the temporal information, making the performance even worse. Our LORS method is a model-vs-sequence scene



(a) Precision-recall curves computed using different training season scenarios

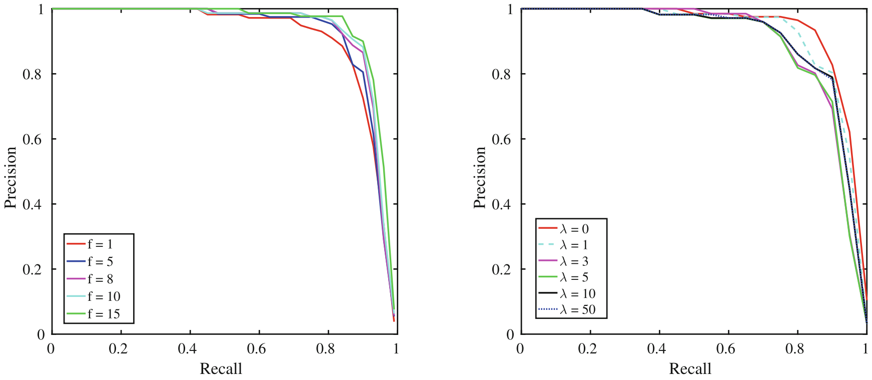
(b) Precision-recall curves computed using varying representativeness threshold

Fig. 8. Quantitative evaluation of our LORS approach over the Nordland dataset across different seasons.

matching method, which has been demonstrated to improve the place matching accuracy in comparison to image-vs-image matching [16], as shown in Fig. 9(a). However, longer sequences include more image frames, indicating the precision of the represented place is low. For example, assuming the speed of the train is 130 km/h, the localization precision when $f = 1$ is 7.2 m, while it is 36.0 m when $f = 5$.

Besides the sequence length f, the LORS's performance will also be affected by the hyperparameter λ in Eq. (1) as all techniques based on optimization with regularization terms [5]. In Fig. 9(b), we compare the LORS approach with different values of λ using the challenging Nordland dataset. In the comparisons in Fig. 9(b), the sequence length $f = 5$ and the representativeness threshold $\gamma = 90\%$ are applied. It is observed from 9(b) that the best performance is achieved when $\lambda = 1$ when the full version LORS is introduced ($\lambda \neq 0$). When $\lambda = 0$, the LORS method reduces to the naive case that all templates in every scenario are used for the place representation. Although it has the highest long-term place matching performance due to the full information utilization, it cannot receive the benefits by the LORS method, including the scalability in real world long-term AR-based information delivery applications.

Our LORS method is a general representation learning framework. The raw feature engineering is not the focus of our LORS method. In our experimental evaluations, the same HOG descriptor is applied based on the prior knowledge that it performs well in both CMU-VL and Nordland datasets [5]. The performance may be further improved if other advanced features (either single feature or multimodal features) are applied in our LORS method.



(a) Precision-recall curves computed using different sequence length f

(b) Precision-recall curves computed using different hyperparameters λ

Fig. 9. Parameter analysis of the proposed LORS approach over the Nordland dataset across different seasons.

4 Conclusion

In this paper, we propose the novel LORS approach that integrates information from multiple environmental scenarios to build a comprehensive representation model to improve long-term place and scene matching, with the ultimate goal to enable long-term AR-based information delivery in collaborative human-robot perception applications. LORS is formulated as a regularized optimization problem, in order to adaptively select only a small subset of most representative scene templates and fuse them into a representation for place representation, which makes LORS highly scalable to long-term changes in real-world AR-based information delivery applications. We further develop an optimization solver that possesses a guarantee to converge to the global optima theoretically. We conduct experiments based upon two public datasets for benchmarking long-term place and scene matching. The promising results have shown performance improvement resulted from the LORS approach.

Acknowledgments. This work was partially supported by NSF CNS-1823245, DOE DE-FE0031650, ARO W911NF-17-1-0447, USAFA FA7000-18-2-0016, and DOT PHMSA 693JK31850005CAAP.

References

1. Billinghurst, M., Clark, A., Lee, G., et al.: A survey of augmented reality. Found. Trends® Hum.-Comput. Interact. **8**(2–3), 73–272 (2015)
2. Chatzopoulos, D., Bermejo, C., Huang, Z., Hui, P.: Mobile augmented reality survey: from where we are to where we go. IEEE Access **5**, 6917–6950 (2017)

3. Lowry, S., et al.: Visual place recognition: a survey. IEEE Trans. Robot. **32**(1), 1–19 (2016)
4. Sünderhauf, N., Neubert, P., Protzel, P.: Are we there yet? Challenging SeqSLAM on a 3000 km journey across all four seasons. In: Workshop of IEEE International Conference on Robotics and Automation (2013)
5. Han, F., Yang, X., Deng, Y., Rentschler, M., Yang, D., Zhang, H.: SRAL: shared representative appearance learning for long-term visual place recognition. IEEE Robot. Autom. Lett. **2**(2), 1172–1179 (2017)
6. Sünderhauf, N., Protzel, P.: BRIEF-Gist - closing the loop by simple means. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (2011)
7. Zhang, G., Lilly, M.J., Vela, P.A.: Learning binary features online from motion dynamics for incremental loop-closure detection and place recognition. In: IEEE International Conference on Robotics and Automation (2016)
8. Han, F., Wang, H., Zhang, H.: Learning of integrated holism-landmark representations for long-term loop closure detection. In: AAAI Conference on Artificial Intelligence (2018)
9. Siva, S., Zhang, H.: Omnidirectional multisensory perception fusion for long-term place recognition. In: IEEE International Conference on Robotics and Automation (ICRA) (2018)
10. Valgren, C., Lilienthal, A.J.: SIFT, SURF and seasons: long-term outdoor localization using local features. In: European Conference on Mobile Robotics (2007)
11. Cummins, M., Newman, P.: FAB-MAP: probabilistic localization and mapping in the space of appearance. Int. J. Robot. Res. **27**(6), 647–665 (2008)
12. Mur-Artal, R., Tardós, J.D.: Fast relocalisation and loop closing in keyframe-based SLAM. In: IEEE International Conference on Robotics and Automation (2014)
13. Naseer, T., Spinello, L., Burgard, W., Stachniss, C.: Robust visual robot localization across seasons using network flows. In: AAAI Conference on Artificial Intelligence (2014)
14. Latif, Y., Huang, G., Leonard, J., Neira, J.: An online sparsity-cognizant loop-closure algorithm for visual navigation. In: Robotics: Science and Systems (2014)
15. Sünderhauf, N., Shirazi, S., Dayoub, F., Upcroft, B., Milford, M.: On the performance of convnet features for place recognition. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (2015)
16. Zhang, H., Han, F., Wang, H.: Robust multimodal sequence-based loop closure detection via structured sparsity. In: Robotics: Science and Systems (2016)
17. Han, F., Yang, X., Zhang, Y., Zhang, H.: Sequence-based multimodal apprenticeship learning for robot perception and decision making. In: IEEE International Conference on Robotics and Automation (ICRA) (2017)
18. Arroyo, R., Alcantarilla, P.F., Bergasa, L.M., Romera, E.: Towards life-long visual localization using an efficient matching of binary sequences from images. In: IEEE International Conference on Robotics and Automation (2015)
19. Johns, E., Yang, G.-Z.: Feature co-occurrence maps: appearance-based localisation throughout the day. In: IEEE International Conference on Robotics and Automation (2013)
20. Badino, H., Huber, D., Kanade, T.: Real-time topometric localization. In: IEEE International Conference on Robotics and Automation (2012)