A Robust Image-Sequence-Based Framework for Visual Place Recognition in Changing Environments

Yong Wang, *Senior Member, IEEE*, Taolue Xue, and Qin Li

***Abstract*—This paper proposes a robust image-sequence-based framework to deal with two challenges of visual place recognition in changing environments: viewpoint variations and environmental condition variations. Our framework includes two main parts. The first part is to calculate the distance between two images from a reference image sequence and a query image sequence. In this part, we remove the deep features of non-overlap contents in these two images and utilize the remaining deep features to calculate the distance. As the deep features of non-overlap contents are caused by viewpoint variations, removing these deep features can improve the robustness to viewpoint variations. Based on the first part, in the second part, we first calculate the distances of all pairs of images from a reference image sequence and a query image sequence, and obtain a distance matrix. Afterward, we design two convolutional operators to retrieve the distance submatrix with the minimum diagonal distribution. The minimum diagonal distribution contains more environmental information, which is insensitive to environmental condition variations. The experimental results suggest that our framework exhibits better performance than several state-of-the-art methods. Moreover, the analysis of runtime shows that our framework has the potential to satisfy real-time demands.**

***Index Terms*—Visual place recognition, changing environments, deep feature, image sequence, distance matrix retrieval**

I. Introduction

Visual place recognition, which enables robots to recognize previously visited places by vision [1], is a hot issue in the robotics field. It can be considered as an image retrieval task, meaning that the robot finds an image in a reference dataset that is most similar to the query image of a given place. This function is important for a robot's autonomous navigation and location [1]–[3]. However, visual place recognition suffers from the influences of changing environments due to two factors: viewpoint variations and environmental condition variations [1]. Viewpoint variations are caused by the shift or rotation of a robot, which lead to non-overlap contents in images from the same place. Environmental condition variations include the changes of illumination, weathers, or seasons, which change the appearances of images. These two variations reduce the similarity of images from the same place, thus increasing the difficulty of image retrieval. Therefore, it is necessary to deal with these two variations. For this purpose, many attempts have been made.

Some researchers have extracted traditional features of images for single image retrieval, such as SIFT [4], [5], SURF [4]–[6], color histograms [6]–[9], and textures [7]. These traditional features are just robust to slight viewpoint variations, and sensitive to environmental condition variations. To solve this problem, other researchers have extracted deep features of images based on convolutional neural networks (CNNs) for single image retrieval because CNNs have shown advantages in extracting desired deep features for a variety of visual tasks [10]–[13]. In this kind of research, the place datasets, which include both viewpoint variations and environmental condition variations, are needed to train CNN models. Thus, CNN models can learn robust deep features for these two variations [14], [15]. However, such robustness is dependent on the diversity of the place datasets as well as the structures of CNN models. To overcome this limitation, several algorithms have been proposed to further handle the deep features extracted by CNN models [16]–[19]. However, the performance of these methods is still unsatisfactory when facing extreme viewpoint variations and environmental condition variations. Since 2012, a few researchers have made use of image sequence retrieval instead of single image retrieval for visual place recognition. An image sequence contains images from multiple adjacent places. Unlike single image retrieval that calculates the distance between two images, image sequence retrieval captures a distance distribution of two image sequences. As this distance distribution is insensitive to environmental condition variations, image sequence retrieval-based methods are robust to extreme environmental condition variations [20]–[26]. However, they cannot deal with extreme viewpoint variations [23].

Based on these considerations, in this paper, a robust image-sequence-based framework is proposed, with the aim of minimizing the influences of viewpoint variations and environmental condition variations. The proposed framework includes two main parts. The first part is to calculate the distance between two images from a reference image sequence and a query image sequence. The second part is to retrieve the reference image sub-sequence which comes from the same places as the query image sequence. Based on these two parts, the places of the query image sequence can be recognized.

The main contributions of this paper are summarized as follows:

- In the first part, we first extract the deep features of two

This work was supported in part by the Innovation-Driven Plan in Central South University under Grant 2018CX010, in part by the National Natural Science Foundation of China under Grants 61673397 and 61976225, and in part by the Beijing Advanced Innovation Center for Intelligent Robots and Systems under Grant 2018IRS06. (*Corresponding author: Qin Li*)

Y. Wang is with the School of Automation, Central South University, Changsha 410083, China, and also with the Mobile Health Ministry of Education-China Mobile Joint Laboratory, Changsha 410008, China. (Email: ywang@csu.edu.cn)

X. Tao and Q. Li are with the School of Automation, Central South University, Changsha 410083, China. (Email: xuetaolue@gmail.com; liqin6@csu.edu.cn)

images from a reference image sequence and a query image sequence using a CNN model. Then, a method named ReNOF is designed to detect and remove the deep features of non-overlap contents in these two images. Finally, the distance between these two images is calculated based on the remaining deep features. Since the deep features of non-overlap contents caused by viewpoint variations are removed, the first part can improve the robustness to extreme viewpoint variations.

- In the second part, the image sequence retrieval is transformed into the distance matrix retrieval. Specifically, we first calculate the distances of all pairs of images from a reference image sequence and a query image sequence to construct a distance matrix. Then, based on this distance matrix, we design two convolutional operators—the sum operator and the difference operator—to retrieve the distance sub-matrix with the minimum diagonal distribution. This part can capture sufficient information about the distance distribution between a reference image subsequence and a query image sequence, which is insensitive to environmental condition variations. Therefore, this part can deal with extreme environmental condition variations.
- In this paper, we use a humanoid robot NAO to collect images during walking in real environments and build a new dataset named CSU-NAO Dataset. This dataset contains images from different places with both viewpoint variations and environmental condition variations. To the best of our knowledge, it is the first time that a humanoid robot has been utilized for visual place recognition in changing environments.
- A series of experiments shows that our framework outperforms several state-of-the-art methods in dealing with viewpoint variations and environmental condition variations. Moreover, additional experiments verify that our framework can satisfy real-time demands.

The rest of this paper is organized as follows. Section II introduces the related work. Section III gives the details of our framework. Section IV describes the experimental settings, including the datasets, evaluation, image features, and parameter settings. The experimental results are presented in Section V. Finally, Section VI concludes this paper.

II. Related Work

At present, many methods for visual place recognition in changing environments have been proposed, which can be divided into two kinds: single-image-based methods and image-sequence-based methods. Next, we briefly introduce them.

A. Single-Image-Based Methods

According to the feature types, we classify the single-image-based methods into two categories.

The first category uses traditional features for single image retrieval [4]–[7], [9], [25], [27]–[29]. Cummins *et al.* [4] proposed FAB-MAP, which is a representative in this category. They converted images into bag-of-words representations and calculated the distance between the word vectors of two images. The bag-of-words representations are built by traditional features such as SIFT [30] or SURF [31], which are robust to slight viewpoint variations. Ulrich and Nourbakhsh [7] presented a new appearance-based place recognition system, in which an image is retrieved based on image histogram matching. Lowry and Andreasson [28] combined the Histogram of Oriented Gradients (HOG) with an efficient and lightweight image description mechanism to perform visual place recognition.

In the second category, deep features are extracted by CNN models for single image retrieval [14]–[16], [18], [19], [26], [32]. Sünderhauf *et al.* [18] utilized deep features extracted by the pre-trained AlexNet [33] as holistic image descriptors and analyzed the robustness of deep features extracted by different CNN layers for viewpoint variations and environmental condition variations. This work provides a reference for deep feature selection. Gomez-Ojeda *et al.* [14] modified CaffeNet [34] to map an image to low-dimensional deep features. The modified CaffeNet is trained with triplets of images, in which two images are from the same place and the third one is from a different place. Arandjelovic *et al.* [15] built a novel CNN model by a new layer named NetVLAD and trained the CNN model with a large-scale dataset based on a weakly supervised ranking loss for deep feature extraction. Qin *et al.* [26] divided images into patches and extracted deep features of these patches based on the pre-trained AlexNet to calculate the similarity between two images. Snderhauf *et al.* [19] combined the pre-trained AlexNet with a landmark proposal technique named Edge Boxes [35] to extract deep features.

B. Image-Sequence-Based Methods

Since 2012, many image-sequence-based methods have been presented [17], [20]–[25], [36], [37]. SeqSLAM [20] is a typical example, which calculates the best candidate matching place within each local reference image sequence. Thus, the places of a query image sequence are achieved by recognizing coherent sequences of these local best candidates. SeqSLAM has shown a performance improvement over single-image-based methods in dealing with extreme environmental condition variations. Pepperell *et al.* [24] added a new image matching technique named SMART to handle great perceptual changes in viewpoint variations on the basis of SeqSLAM. They also removed the sky region in an image to reduce the influence of illumination variations. Milford *et al.* [25] used state-of-the-art CNN models to generate a synthetic viewpoint of a place, and combined the synthetic viewpoint with the image sequence matching technology in SeqSLAM to recognize places. Chen *et al.* [16] extracted deep features of images by the pre-trained Overfeat network [38] to generate a distance matrix between a reference image sequence and a query image sequence. Then they used a spatial and sequential filter to retrieve the most similar reference image sub-sequence on the distance matrix. Naseer *et al.* [39] proposed a novel data association approach for matching streams of query images to an image sequence stored in a reference dataset. This method

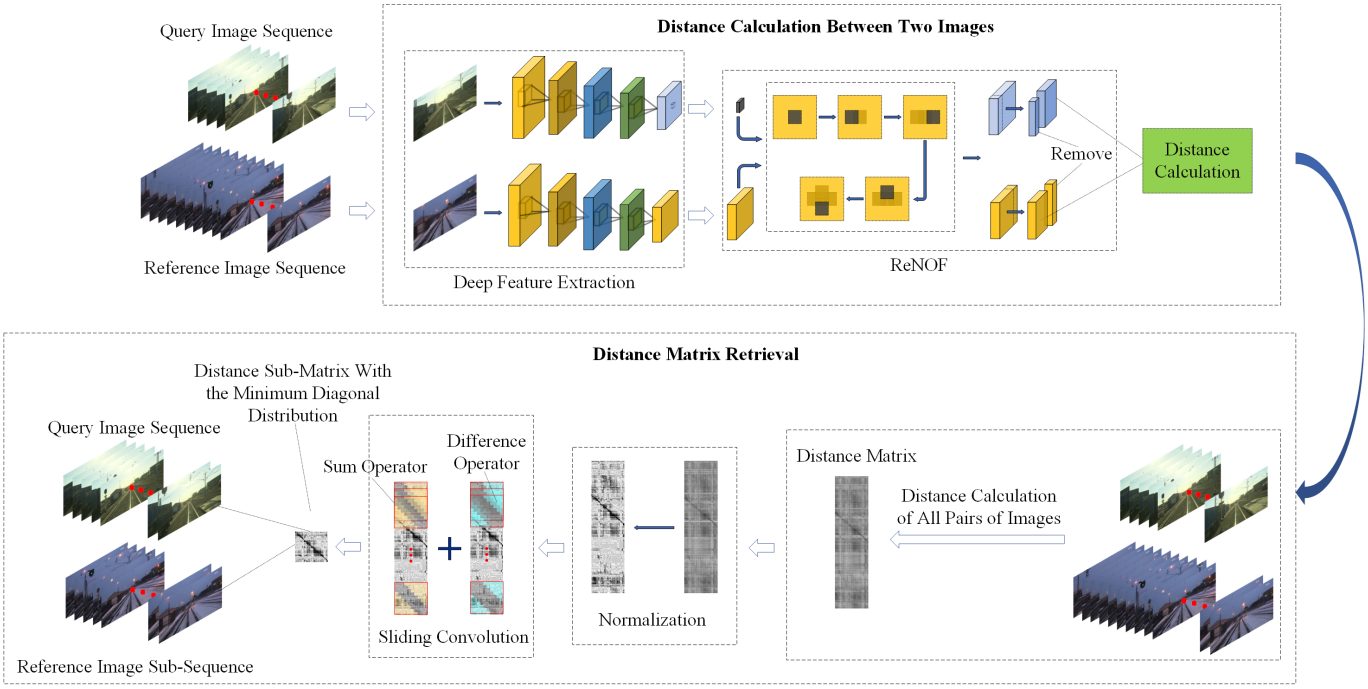


Fig. 1. Illustration of our framework.

exploits network flows to leverage sequential information, with the aim of improving the performance of visual place recognition.

III. Proposed Framework

As introduced in Section I, there exist some weaknesses in current single-image-based methods and image-sequence-based methods for visual place recognition in changing environments. Specifically, single-image-based methods fail when facing extreme viewpoint variations or environmental condition variations. In addition, image-sequence-based methods cannot deal with extreme viewpoint variations. Motivated by these problems, we propose a robust image-sequence-based framework for visual place recognition in changing environments. Fig. 1 illustrates the proposed framework, which contains two main parts: distance calculation between two images and distance matrix retrieval. Next, we introduce these two parts.

A. Distance Calculation Between Two Images

In this paper, the image sequence to be recognized is called the query image sequence (denoted as S_q) and the image sequence that marks places is called the reference image sequence (denoted as S_r). This part is to calculate the distance between two images I_q and I_r from S_q and S_r. We first extract deep features of I_q and I_r, that is, $\boldsymbol{F}_q$ and $\boldsymbol{F}_r$, by using a pre-trained CNN model. Then, we propose a method named ReNOF to detect and remove the deep features of non-overlap contents in I_q and I_r. The process of ReNOF is given in Fig. 2, and each step is introduced below:

- Step 1: Find the center local deep features $\boldsymbol{f}^q_{center}$ and $\boldsymbol{f}^r_{center}$ in $\boldsymbol{F}_q$ and $\boldsymbol{F}_r$. The length and width of $\boldsymbol{f}^q_{center}$ and $\boldsymbol{f}^r_{center}$ are one-third of the length and width of $\boldsymbol{F}_q$ and $\boldsymbol{F}_r$, respectively.
- Step 2: Take the location of $\boldsymbol{f}^r_{center}$ as a starting point and slide $\boldsymbol{f}^q_{center}$ one step to four different positions (i.e., left, right, top, and bottom) on $\boldsymbol{F}_r$. By doing this, five local deep features are overlapped in $\boldsymbol{F}_r$—$\boldsymbol{f}^r_{center}$, $\boldsymbol{f}^r_{left}$, $\boldsymbol{f}^r_{right}$, $\boldsymbol{f}^r_{top}$, and $\boldsymbol{f}^r_{bottom}$.
- Step 3: Calculate the distance between $\boldsymbol{f}^q_{center}$ and the five local deep features obtained in Step 2, respectively:

$$d_i = 1 - \cos\langle \boldsymbol{f}^q_{center}, \boldsymbol{f}_i{}^r \rangle, \quad i \in \{center, left, right, top, bottom\} \tag{1}$$

  Then, the local deep features with the minimum distance (denoted as $\boldsymbol{f}_v{}^r$) are obtained, in which

$$v = \arg\min(d_i), \quad i \in \{center, left, right, top, bottom\} \tag{2}$$

- Step 4: Overlay $\boldsymbol{F}_q$ and $\boldsymbol{F}_r$ by aligning the positions of $\boldsymbol{f}^q_{center}$ and $\boldsymbol{f}_v{}^r$ to detect the non-overlap local deep features which are regarded as the deep features of non-overlap contents in I_q and I_r.
- Step 5: Remove these non-overlap local deep features and obtain the remaining local deep features $\boldsymbol{f}^q_{ol}$ and $\boldsymbol{f}^r_{ol}$.

The reasons why ReNOF can detect the deep features of non-overlap contents are explained in the following. On the one hand, the non-overlap contents between two images can be detected by aligning the positions of overlap contents between these two images. On the other hand, due to the special structure of CNNs, the local deep features in a region correspond to the local contents in the same region of an image. For example, in Fig. 2, $\boldsymbol{f}^q_{center}$ corresponds to the center contents of I_q. Based on these two points, the deep features of

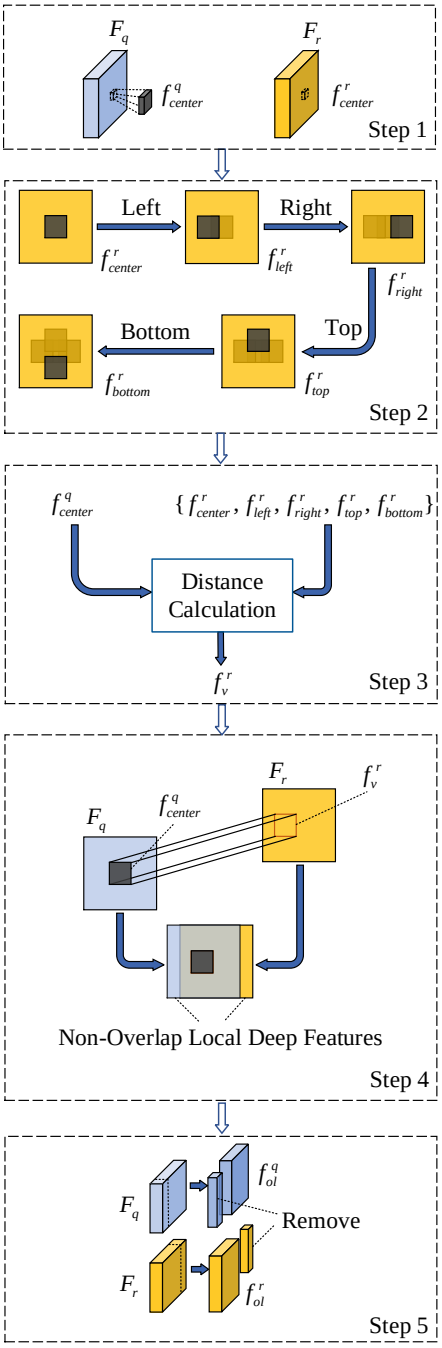


Fig. 2. Process of ReNOF.

non-overlap contents between two images can be detected by aligning the deep features of overlap contents. In ReNOF, we consider two local deep features with the minimum distance (i.e., $\boldsymbol{f}^{q}_{center}$ and $\boldsymbol{f}_{v}{}^{r}$) as the deep features of overlap contents between I_q and I_r, since the minimum distance suggests the most similar image contents. By aligning $\boldsymbol{f}^{q}_{center}$ and $\boldsymbol{f}_{v}{}^{r}$, the deep features of non-overlap contents in I_q and I_r can be detected and removed.

Finally, $\boldsymbol{f}^{q}_{ol}$ and $\boldsymbol{f}^{r}_{ol}$ are used to calculate the distance between I_q and I_r:

$$d = 1 - \cos\langle \boldsymbol{f}^{q}_{ol},\ \boldsymbol{f}^{r}_{ol} \rangle \tag{3}$$

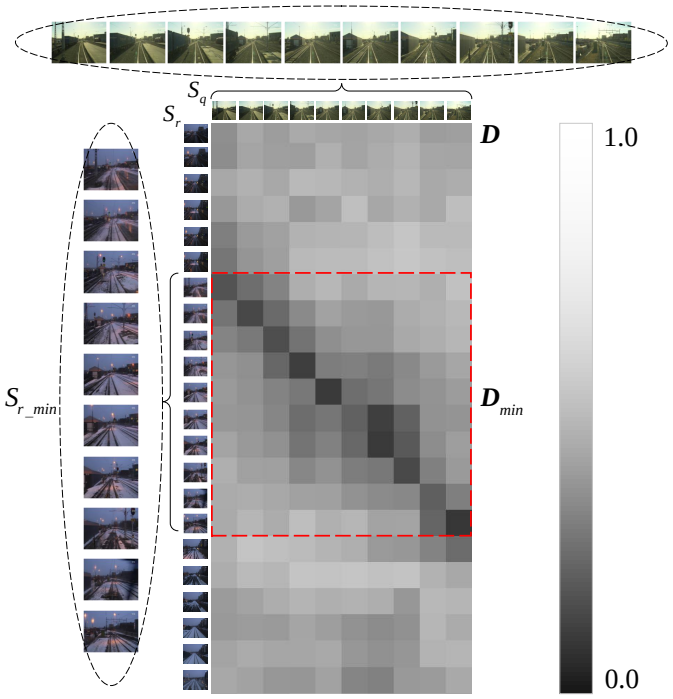


Fig. 3. An example of $\boldsymbol{D}$ and $\boldsymbol{D}_{min}$. $\boldsymbol{D}_{min}$ is the distance sub-matrix with the minimum diagonal distribution.

Remark 1: In current methods, the deep features of two images are directly used to calculate the distance between them. However, there may exist non-overlap contents between two images coming from the same place due to extreme viewpoint variations. Under this condition, the deep features extracted by current methods may belong to the non-overlap contents between these two images. Based on such deep features, the calculated distance between these two images increases, thus leading to the difficulty of visual place recognition in changing environments. Different from current methods, ReNOF detects and removes the deep features of non-overlap contents before calculating the distance between these two images, which makes our framework robust to extreme viewpoint variations.

B. Distance Matrix Retrieval

Suppose that S_q and S_r contain M and N images, respectively: $S_q = \{I_q^1, I_q^2, \ldots, I_q^M\}$ and $S_r = \{I_r^1, I_r^2, \ldots, I_r^N\}$. Based on the first part, we calculate the distances of all pairs of images from S_r and S_q, and obtain distance matrix $\boldsymbol{D}$:

$$\boldsymbol{D} = \begin{bmatrix} d_{1,1} & d_{1,2} & \cdots & d_{1,M} \\ d_{2,1} & d_{2,2} & \cdots & d_{2,M} \\ d_{3,1} & d_{3,2} & \cdots & d_{3,M} \\ \vdots & \vdots & \ddots & \vdots \\ d_{N,1} & d_{N,2} & \cdots & d_{N,M} \end{bmatrix} \tag{4}$$

where $d_{n,m}$ ($n \in \{1, 2, \ldots, N\}$ and $m \in \{1, 2, \ldots, M\}$) indicates the distance between the nth image in S_r and the mth image in S_q. Let $\{I_r^k, I_r^{k+1}, \ldots, I_r^{k+M-1}\}$ be the kth reference image sub-sequence of S_r. The distance sub-matrix corresponding to the kth reference image sub-sequence is

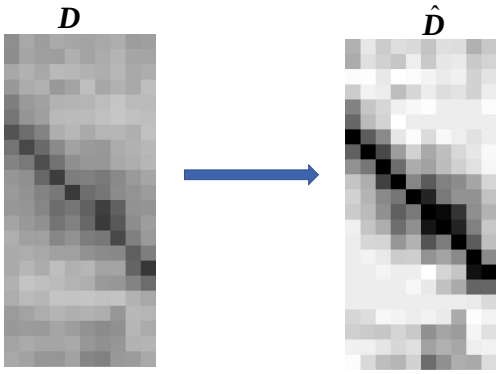


Fig. 4. Normalization of $\boldsymbol{D}$ obtained in Fig. 3.

called the kth distance sub-matrix, denoted as $\boldsymbol{D}_k$. Obviously, there are $(N - M + 1)$ distance sub-matrixes.

In general, if each pair of images at the same position in S_q and a reference image sub-sequence comes from the same place, the corresponding distance sub-matrix has the following two properties: (i) the values of the elements on the main and adjacent diagonals in this distance sub-matrix are smaller than the values of the elements at the same positions in other distance sub-matrixes; and (ii) the values of the elements on the main and adjacent diagonals in this distance sub-matrix are smaller than the values of other elements in the same distance sub-matrix. In this paper, the element distribution that satisfies these two properties is considered as the minimum diagonal distribution, and the distance sub-matrix with the minimum diagonal distribution is denoted as $\boldsymbol{D}_{min}$. If $\boldsymbol{D}_{min}$ can be found, then the reference image sub-sequence matched to S_q is retrieved, denoted as S_{r_min}. Therefore, we transform the image sequence retrieval into the distance matrix retrieval. Fig. 3 gives an example of this process. In Fig. 3, S_{r_min} and S_q come from the same places; therefore, the corresponding distance sub-matrix satisfies the minimum diagonal distribution.

Next, we attempt to retrieve $\boldsymbol{D}_{min}$ from $\boldsymbol{D}$, which includes two steps. In the first step, a normalization method is proposed to increase the difference of the values of all elements in $\boldsymbol{D}$, which can highlight the minimum diagonal distribution. In the second step, two convolutional operators are designed, the sum operator $\boldsymbol{O}_{sum}$ and the difference operator $\boldsymbol{O}_{dif}$, to calculate the degree to which a distance sub-matrix satisfies the minimum diagonal distribution.

The proposed normalization method selects the lth largest element from each column in $\boldsymbol{D}$. If the value of an element in a column is smaller than that of the lth largest element in this column, then this element is normalized based on min-max normalization; otherwise, the value of this element is set to 1. Note that l is a hyper-parameter to control the normalization range, which can reduce the influence of abnormal elements with extremely large values in $\boldsymbol{D}$. This process can be described by Eq. (5) and Eq. (6):

$$d_{j\_min} = min(\boldsymbol{d}_j), \tag{5}$$

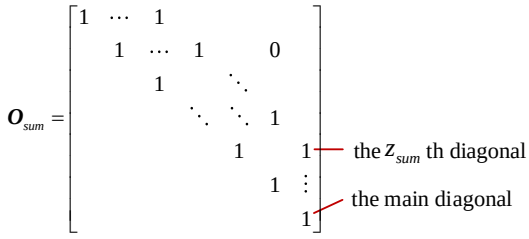


Fig. 5. Structure of $\boldsymbol{O}_{sum}$.

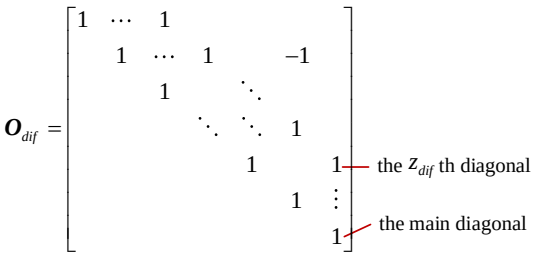


Fig. 6. Structure of $\boldsymbol{O}_{dif}$.

$$\hat{d}_{i,j} = \begin{cases} \dfrac{d_{i,j} - d_{j\_min}}{d_{j\_lmax} - d_{j\_min}}, & d_{i,j} < d_{j\_lmax} \\ 1, & otherwise \end{cases}, \tag{6}$$

where $\boldsymbol{d}_j$ is the jth column of $\boldsymbol{D}$, $d_{i,j}$ is the ith element of $\boldsymbol{d}_j$, d_{j_min} is the minimum element of $\boldsymbol{d}_j$, d_{j_lmax} is the lth largest element of $\boldsymbol{d}_j$, and $\hat{d}_{i,j}$ is the normalized element. After normalizing all elements in $\boldsymbol{D}$, the normalized distance matrix (denoted as $\hat{\boldsymbol{D}}$) is obtained. Fig. 4 shows the normalization of $\boldsymbol{D}$ obtained in Fig. 3. As can be seen, the difference of the values of all elements in $\hat{\boldsymbol{D}}$ is significantly highlighted, which can effectively improve the retrieval performance.

We design sum operator $\boldsymbol{O}_{sum}$ with the size of $M \times M$, which is convoluted with a normalized distance sub-matrix to calculate the sum of elements on the main and adjacent diagonals in this normalized distance sub-matrix. The calculated sum indicates the degree to which a normalized distance sub-matrix satisfies the first property of the minimum diagonal distribution. The smaller the calculated sum, the higher the degree. Fig. 5 shows the structure of $\boldsymbol{O}_{sum}$. As shown in Fig. 5, the values of elements on the main and adjacent diagonals are set to 1, while the values of the remaining elements are set to 0. In this paper, the number of the selected adjacent diagonals is z_{sum}.

In addition, we design another convolution operator $\boldsymbol{O}_{dif}$ with the size of $M \times M$, which is convoluted with a normalized distance sub-matrix to calculate the difference between the elements on the main and adjacent diagonals and the elements in other positions in this normalized distance sub-matrix. The calculated difference indicates the degree to which a normalized distance sub-matrix satisfies the second property of the minimum diagonal distribution. The smaller the calculated

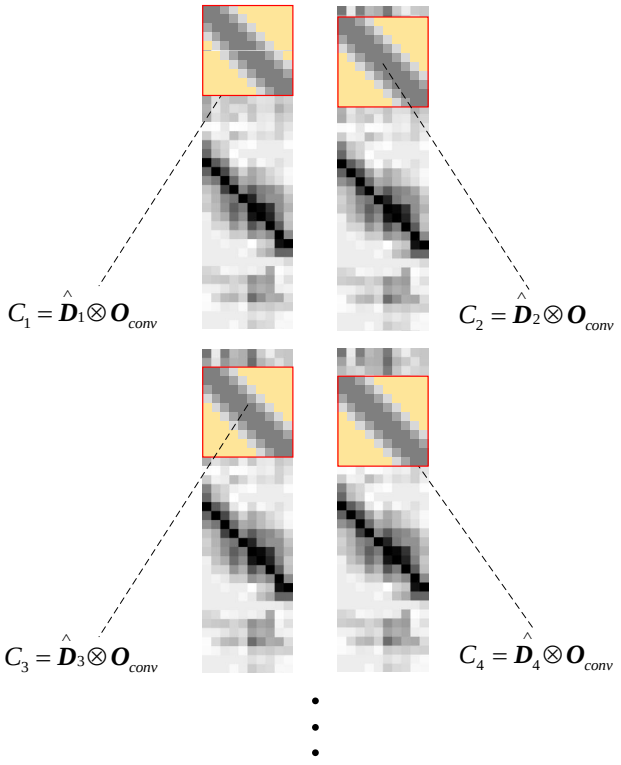


Fig. 7. An example of sliding convolution.

difference, the higher the degree. Fig. 6 shows the structure of $\boldsymbol{O}_{dif}$. As shown in Fig. 6, the values of elements on the main diagonal and the adjacent diagonals are set to 1, while the values of the remaining elements are set to -1. In this paper, the number of the selected adjacent diagonals is z_{dif}.

Afterward, we use $\boldsymbol{O}_{sum}$ and $\boldsymbol{O}_{dif}$ to implement the sliding convolution on $\hat{\boldsymbol{D}}$ to obtain two convolution results of each normalized distance sub-matrix. Then, we take a weighted sum of these two convolution results to indicate the degree to which each normalized distance sub-matrix satisfies the minimum diagonal distribution:

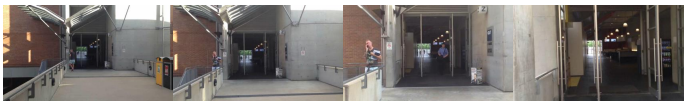
$$\begin{aligned} C_k &= \theta_1 \hat{\boldsymbol{D}}_k \otimes \boldsymbol{O}_{sum} + \theta_2 \hat{\boldsymbol{D}}_k \otimes \boldsymbol{O}_{dif}, \\ &= \hat{\boldsymbol{D}}_k \otimes (\theta_1 \boldsymbol{O}_{sum} + \theta_2 \boldsymbol{O}_{dif}), \\ k &\in \{1, 2, \ldots, N - M + 1\} \end{aligned} \tag{7}$$

where $\theta_1 \in [0, 1]$ and $\theta_2 \in [0, 1]$ are two weights, and $\hat{\boldsymbol{D}}_k$ is the normalized $\boldsymbol{D}_k$. Let $\boldsymbol{O}_{conv} = \theta_1 \boldsymbol{O}_{sum} + \theta_2 \boldsymbol{O}_{dif}$, then Eq. (7) can be simplified as:

$$\begin{aligned} C_k &= \hat{\boldsymbol{D}}_k \otimes \boldsymbol{O}_{conv} \\ k &\in \{1, 2, \ldots, N - M + 1\} \end{aligned} \tag{8}$$

Fig. 7 gives an example of sliding convolution of $\boldsymbol{O}_{conv}$ on $\hat{\boldsymbol{D}}$. Based on the previous analysis, the smaller the value of C_k, the higher the degree that $\hat{\boldsymbol{D}}_k$ satisfies the minimum diagonal distribution. Therefore, we find the normalized distance sub-matrix with the smallest weighted sum, denoted as $\hat{\boldsymbol{D}}_w$, where

$$w = \arg\min(C_k), k \in \{1, 2, \ldots, N - M + 1\} \tag{9}$$



(a) Garden-Day-Left

(b) Garden-Day-Right

(c) Garden-Night-Right

Fig. 8. Image samples in Garden Point Dataset.

Subsequently, a threshold T is set to determine whether $\hat{\boldsymbol{D}}_w$ is the desired normalized distance sub-matrix with the minimum diagonal distribution or not. If $C_w < T$, then we consider that $\hat{\boldsymbol{D}}_w$ is the normalized distance sub-matrix with the minimum diagonal distribution, and the corresponding reference image sub-sequence S_{r_w} and S_q really come from the same places. Under this condition, the visual place recognition is successful. On the other hand, if $C_w \geq T$, then there is no reference image sub-sequence in S_r that comes from the same places as S_q. It means that S_q comes from new places that the robot has not visited. In this paper, T is swept over a range of values to generate the precision-recall curves [1].

Remark 2: As mentioned previously, if a query image sequence and a reference image sub-sequence come from the same places, their corresponding distance sub-matrix satisfies the two properties of the minimum diagonal distribution. The minimum diagonal distribution contains adequate environmental information, which is insensitive to extreme environmental condition variations. Current image-sequence-based methods only utilize the first property to retrieve the reference image sub-sequence. Different from these methods, our framework uses both properties to retrieve the reference image sub-sequence, which is more robust to extreme environmental condition variations.

IV. Experimental Settings

A. Datasets

We used four datasets to validate the effectiveness of our framework, including three public datasets (i.e., Garden Point Dataset [40], North Campus Dataset [41], and Nordland Dataset [23]) and a new dataset named CSU-NAO Dataset. These four datasets were built using collection tools which collect images by using cameras when moving in specific environments. In these four datasets, viewpoint variations were caused by shift or rotation of the collection tools. In addition, environmental condition variations included the changes of illumination, seasons, or moving objects.

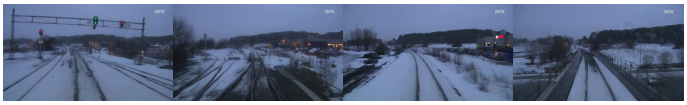
Figs. 8-10 show image samples in the three public datasets. In Garden Point Dataset, there were three image sequences,

(a) North-Summer-Left

(b) North-Autumn-Right

Fig. 9. Image samples in North Campus Dataset.

(a) Nordland-Spring



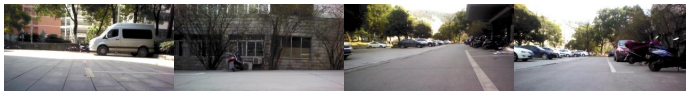
(b) Nordland-Winter

Fig. 10. Image samples in Nordland Dataset.

i.e., Garden-Day-Left, Garden-Day-Right, and Garden-Night-Right (Fig. 8). Garden-Day-Left and Garden-Day-Right were collected on the left and right sides of a path during the day, respectively, and Garden-Night-Right was collected on the right side of the same path at night. Each image sequence contained 200 images. North Campus Dataset had 27 image sequences collected on a path over 15 months. We selected two image sequences in Summer and Autumn: North-Summer-left and North-Autumn-right (Fig. 9). Each image sequence contained 500 images. Nordland Dataset had four image sequences collected in four seasons. Unlike two other datasets, Nordland Dataset recorded images by a camera fixed on a train. Since there was almost no shift or rotation of the fixed camera, this dataset had no viewpoint variation. We selected two image sequences in Spring and Winter: Nordland-Spring and Nordland-Winter (Fig. 10). Each image sequence contained 3600 images. More properties of these three datasets are shown in Table I. Apart from these three public datasets, we also used a humanoid robot NAO to collect a new dataset named CSU-NAO Dataset. We let NAO walk on two paths at the Central South University, Changsha, China, and collected images at a fixed distance interval (i.e., 5m) on the left and right sides of each path. Therefore, we obtained four image sequences: CSU-I-Day-Left, CSU-I-Day-Right, CSU-II-Day-Left, and CSU-II-Night-Right (Fig. 11). CSU-I-Day-Left and CSU-I-Day-Right were collected on the left and right sides of path I during the day, respectively. Both of these two image sequences contained 111 images. CSU-II-Day-Left and CSU-II-Night-Right were collected on the left side of path II during the day and on the right side of path II at night, respectively. Both of these image sequences contain 100 images. To the best of our knowledge, this is the first time that a humanoid robot was utilized to build the dataset for visual place recognition in changing environments.

(a) CSU-I-Day-Left

(b) CSU-I-Day-Right



(c) CSU-II-Day-Left

(d) CSU-II-Night-Right

Fig. 11. Image samples in CSU-NAO Databset.

B. Evaluation

For each query image sequence, if the retrieved reference image sub-sequence was sufficiently close to the correct reference image sub-sequence within a tolerance of one frame for each dataset, it was considered as a true positive match. For example, if the correct reference image sub-sequence is the kth reference image sub-sequence, then the $(k-1)$th, kth, and $(k+1)$th reference image sub-sequences are considered as true positive matches to the query image sequence.

C. Image Features and Parameters

For extracting the deep features, a pre-trained CNN framework Place205-Alexnet [42] was utilized. According to [18], we took the features from the 5th pooling layer of Place205-Alexnet as the deep features, the size of which is $6 \times 6 \times 256$. The size of the deep features of each layer in Place205-Alexnet is shown in Table II. In addition, the parameter settings of our framework are summarized in Table III.

V. Experimental Results

In this section, we conducted extensive experiments to test the performance of our framework for visual place recognition in changing environments. The precision-recall (PR) curve was adopted as the performance indicator.

TABLE I
PROPERTIES OF THE THREE PUBLIC DATASETS

Property \ Name	Garden Point [40]	North Campus [41]	Nordland [23]
Environment	Queensland University of Technology campus	University of Michigan's North Campus	Train ride
Collection tool	Human	Segway robot	Train
Total distance	380m	2.5km	72km
Number of image sequences	3	2	2
Number of images in each image sequence	200	500	3600
Distance between adjacent images	1.9m	5m	20m
Viewpoint variation	Yes	Yes	No
Illumination variation	Yes	Yes	Yes
Seasonal variation	No	Yes	Yes
Moving objects	Yes	Yes	No

TABLE II
THE SIZE OF THE DEEP FEATURES OF EACH LAYER IN PLACE205-ALEXNET

Layer name	conv1	pool1	conv2
Size of features	$55 \times 55 \times 96$	$27 \times 27 \times 96$	$27 \times 27 \times 256$
Layer name	pool2	conv3	conv4
Size of features	$13 \times 13 \times 256$	$13 \times 13 \times 384$	$13 \times 13 \times 384$
Layer name	conv5	pool5	fc6
Size of features	$13 \times 13 \times 256$	$6 \times 6 \times 256$	4096
Layer name	fc7	fc8	
Size of features	4096	205	

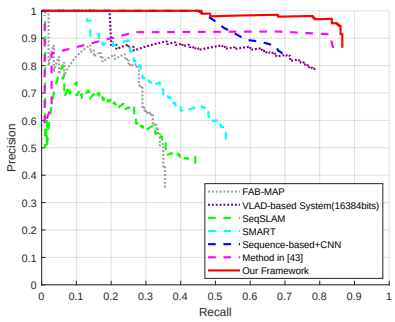
TABLE III
PARAMETER SETTINGS

Parameter	Value	Description
M	10	the number of images in S_q
l	$0.2\times N$	the hyper-parameter in the normalization of $\boldsymbol{D}$
z_{sum}	2	the number of adjacent diagonals in $\boldsymbol{O}_{sum}$
z_{dif}	3	the number of adjacent diagonals in $\boldsymbol{O}_{dif}$
θ_1	0.3	the weight of $\boldsymbol{O}_{sum}$
θ_2	0.7	the weight of $\boldsymbol{O}_{dif}$

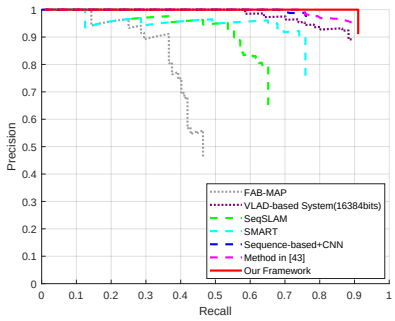
A. Comparison with State-of-the-Art Methods

We compared our framework with five state-of-the-art methods on all datasets. These methods included two single-image-based methods (i.e., FAB-MAP [4] and VLAD-based System(16384bits) [28]) and three image-sequence-based methods (i.e., SeqSLAM [20], SMART [24], and Sequence-based+CNN [16]). Note that the deep features in Sequence-based+CNN were also taken from the 5th pooling layer of Place205-Alexnet [42].

In order to fully validate the performance of our framework in dealing with viewpoint variations and environmental condition variations, we divided comparison experiments into three groups: 1) visual place recognition with viewpoint variations; 2) visual place recognition with environmental condition variations; and 3) visual place recognition with both viewpoint variations and environmental condition variations. For each group, we used image sequence pairs with the corresponding variations in the datasets to complete comparison experiments. Specifically, in Garden Point Dataset, there are viewpoint variations between Garden-Day-Left and Garden-Day-Right, so they were used in the first group. There are environmental condition variations between Garden-Day-Right and Garden-Night-Right, so they were used in the second group. There are both variations between Garden-Day-Left and Garden-Night-Right, so they were used in the third group. In North Campus Dataset, there are both variations between North-Summer-Left and North-Winter-Right, so they were used in the third group. In Nordland Dataset, there are environmental condition variations between Nordland-Spring and Nordland-Winter, so they were used in the second group. In CSU-NAO Dataset, there are viewpoint variations between CSU-1-Day-Left and CSU-1-Day-Right, so they were used in the first group. There are both variations between CSU-2-Day-Left and CSU-2-Night-Right, so they were used in the third group. The comparison results are summarized in the following.



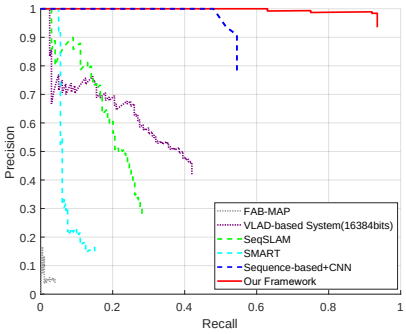
(a) Garden-Day-Left & Garden-Day-Right



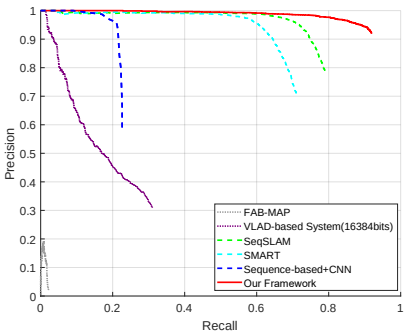
(b) CSU-1-Day-Left & CSU-1-Day-Right

Fig. 12. PR curves of the compared methods on image sequence pairs with viewpoint variations.

1) Visual Place Recognition with Viewpoint Variations: In this group, Garden-Day-Left and CSU-1-Day-Left were for reference, while Garden-Day-Right and CSU-1-Day-Right were for query. Fig. 12 shows the PR curves of the compared methods. From Fig. 12, the performance of our framework is better than that of the five competitors on both image sequence pairs. This is because these competitors directly use the extracted features to calculate the distance between two



(a) Garden-Day-Right & Garden-Night-Right

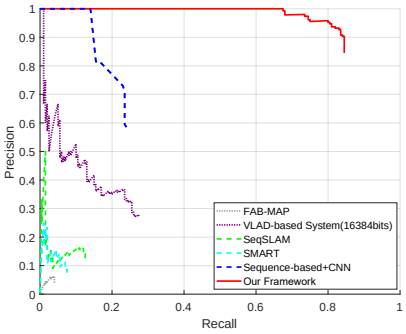


(b) Nordland-Spring & Nordland-Winter

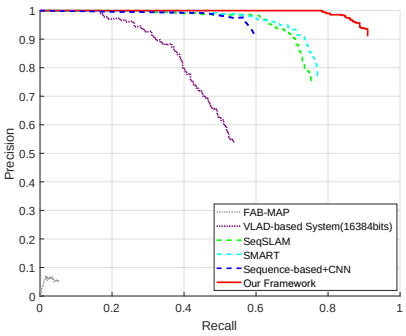
Fig. 13. PR curves of the compared methods on image sequence pairs with environmental condition variations.

images, which are only robust to slight viewpoint variations. Unlike them, our framework detects and removes the features of non-overlap contents in two images, and uses the remaining features to calculate the distance of these two images. In this way, the influence of extreme view variations can be effectively ameliorated. We also compared our framework with the method in [43] which is robust to extreme viewpoint variations. As can be seen in Fig. 12, when using CSU-1-Day-Left and CSU-1-Day-Right, the performance of this method is comparable to that of our framework. However, when using Garden-Day-Left and Garden-Day-Right, our framework performs better. The reason is the following. Images in Garden Point Dataset were collected by a human who is more sensitive to various interference factors in environments than the humanoid robot NAO used in CSU-NAO Dataset. Such human sensitivity will cause obvious camera shift. Therefore, there are extreme viewpoint variations caused by camera shift between Garden-Day-Left and Garden-Day-Right. Our framework can reduce the influence of such extreme viewpoint variations by removing the features of non-overlap contents in two images. However, the method in [43] focuses on dealing with extreme viewpoint variations mainly caused by camera rotation rather than camera shift, thus resulting in performance degradation.

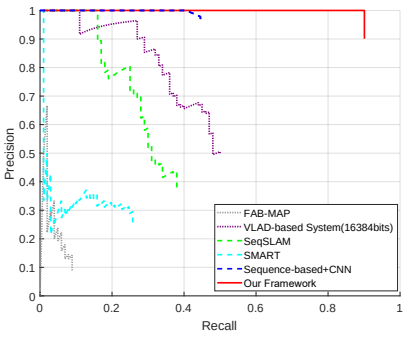
2) Visual Place Recognition with Environmental Condition Variations: In this group, Garden-Day-Right and Nordland-Spring were for reference, while Garden-Night-Right and Nordland-Winter were for query. Fig. 13 shows the PR curves of the compared methods. As can be seen, our framework performs the best on both image sequence pairs. The reasons are the following. Single-image-based competitors (i.e., FAB-MAP and VLAD-based System(16384bits)) cannot deal with environmental condition variations. In addition, three image-sequence-based competitors (i.e., SeqSLAM, SMART, and Sequence-based+CNN) only utilize the first property of the minimum diagonal distribution to retrieve the matched reference image sub-sequence. Unlike these three image-sequence-based competitors, our framework uses both properties of the minimum diagonal distribution to retrieve the matched reference image sub-sequence, which is more robust to extreme environmental condition variations. On the other hand, it can be seen that different types of extreme environmental condition variations lead to different performance of the image-sequence-based methods. For example, there are illumination changes between Garden-Day-Right and Garden-Night-Right. On this image sequence pair, SMART and SeqSLAM exhibit poor performance. There are seasonal changes



(a) Garden-Day-Left & Garden-Night-Right



(b) North-Summer-Left & North-Winter-Right



(c) CSU-1-Day-Left & CSU-2-Night-Right

Fig. 14. PR curves of the compared methods on image sequence pairs with both viewpoint variations and environmental condition variations.

TABLE IV
COMPARISON OF THE MAXIMUM RECALL AT 100% PRECISION

Experiment	Image Sequence Pair	Method	Precision (%)	Recall (%)
Visual Place Recognition with Viewpoint Variations	Garden-Day-Left & Garden-Day-Right	FAB-MAP	100	2
		VLAD-based System(16384bits)	100	19.5
		SeqSLAM	100	1
		SMART	100	13
		Sequence-based+CNN	100	45
		Method in [43]	100	1
		Our Framework	100	**46**
	CSU-1-Day-Left & CSU-1-Day-Right	FAB-MAP	100	14.3
		VLAD-based System(16384bits)	100	59
		SeqSLAM	100	25.9
		SMART	100	12.5
		Sequence-based+CNN	100	67.9
		Method in [43]	100	75
		Our Framework	100	**91**
Visual Place Recognition with Environmental Condition Variations	Garden-Day-Right & Garden-Night-Right	FAB-MAP	100	N/A
		VLAD-based System(16384bits)	100	2.5
		SeqSLAM	100	3
		SMART	100	5
		Sequence-Based+CNN	100	48
		Our Framework	100	**63**
	Nordland-Spring & Nordland-Winter	FAB-MAP	100	N/A
		VLAD-based System(16384bits)	100	2
		SeqSLAM	100	4.6
		SMART	100	4.4
		Sequence-based+CNN	100	9
		Our Framework	100	**22.9**
Visual Place Recognition with Both Viewpoints Variations and Environmental Condition Variations	Garden-Day-Left & Garden-Night-Right	FAB-MAP	100	N/A
		VLAD-based System(16384bits)	100	N/A
		SeqSLAM	100	N/A
		SMART	100	N/A
		Sequence-based+CNN	100	14
		Our Framework	100	**67.5**
	North-Summer-Left & North-Winter-Right	FAB-MAP	100	N/A
		VLAD-based System(16384bits)	100	17
		SeqSLAM	100	31.9
		SMART	100	47.1
		Sequence-based+CNN	100	10
		Our Framework	100	**78.4**
	CSU-2-Day-Left & CSU-2-Night-Right	FAB-MAP	100	N/A
		VLAD-based System(16384bits)	100	11
		SeqSLAM	100	16
		SMART	100	1
		Sequence-based+CNN	100	41
		Our Framework	100	**90**

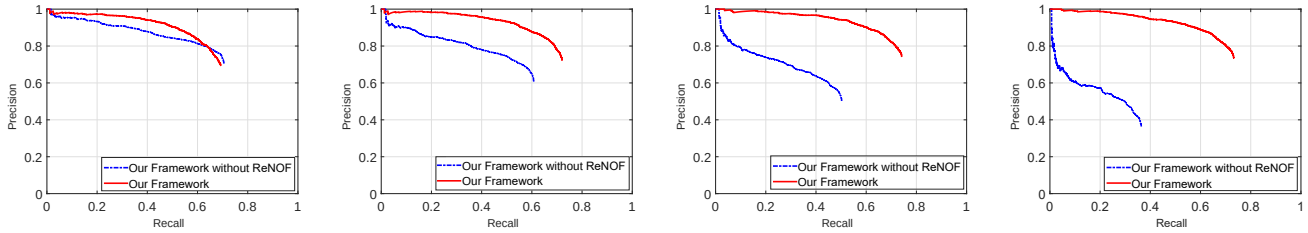
between Nordland-Spring and Nordland-Winter. On this image sequence pair, the performance of Sequence-based+CNN is poor. However, there is little performance degradation of our framework on these two image sequence pairs, which validates the robustness of our framework to different types of extreme environmental condition variations.

3) Visual Place Recognition with Both Viewpoint Variations and Environmental Condition Variations: In this group, Garden-Day-Left, North-Summer-Left, and CSU-2-Day-Left were for reference, while Garden-Night-Right, North-Winter-Right, and CSU-2-Night-Right were for query. Fig. 14 gives the PR curves of the compared methods. As shown in Fig. 14, our framework also obtains the best results. The reason may be that, as mentioned previously, other compared methods are either robust to only one type of variation or have a limited ability to deal with both variations.

In recent visual place recognition systems, the maximum recall at 100% precision has been frequently used for performance evaluation, as introducing false recognition into systems may cause catastrophic failure in practical tasks [4]. Therefore, we also compared the maximum recall at 100% precision of the compared methods in the three groups of experiments, which is shown in Table IV. In Table IV, "N/A" represents that the corresponding method cannot achieve 100% precision. From Table IV, our framework consistently maintains the highest recall at 100% precision in all experiments, which demonstrates that our framework has the potential to provide high-accuracy visual place recognition.

B. Effectiveness of ReNOF

In order to test the effectiveness of ReNOF, we compared the performance of our framework with and without ReNOF for visual place recognition by using Nordland Dataset. In this dataset, Nordland-Spring was for reference and Nordland-Winter was for query. We simulated viewpoint variations between Nordland-Spring and Nordland-Winter by moving all the images in Nordland-Winter to the right by the same distance. Such moving will generate non-overlap contents in these two image sequences. Then, we determined the degree of viewpoint variations based on the ratio of non-overlap contents



(a) 7.5% non-overlap contents (b) 10% non-overlap contents (c) 12.5% non-overlap contents (d) 16% non-overlap contents

Fig. 15. PR curves of our framework with and without ReNOF on Nordland Dataset. The ratios of non-overlap contents are 7.5%, 10%, 12.5%, and 16%, respectively, which reflect different degree of viewpoint variations in Nordland Dataset.

TABLE V
RUNTIME OF DEEP FEATURE EXTRACTION AND DISTANCE CALCULATION BETWEEN TWO IMAGES IN OUR FRAMEWORK

Operation	Runtime
Extracting the deep features of one image	30ms
Calculating the distance between two images	0.08ms

to an entire image. The four ratios were 7.5%, 10%, 12.5%, and 16%, respectively. Fig. 15 shows the performance comparison. From Fig. 15, in the case of 7.5% non-overlap contents, the effect of ReNOF is not significant. With the increase of the ratio, our framework with ReNOF performs better than our framework without ReNOF, which demonstrates that ReNOF is able to effectively remove the deep features of non-overlap contents.

C. Analysis of Runtime

We also carried out the analysis of runtime by using the Matlab implementation on a desktop with a Nvidia1080 GPU. In general, deep feature extraction of one image and distance calculation between two images are the most time-consuming operations; thus, we only provided the runtime of them. Table V summarizes their runtime. Based on them, for a reference image sequence with 1000 images, it takes about 120ms for our framework to retrieve a query image sequence. As a consequence, our framework can run at 8.3 Hz for the dataset with 1000 images, which has the potential to satisfy real-time demands.

VI. CONCLUSION

In this paper, we proposed a robust image-sequence-based framework for visual place recognition in changing environments. Our framework involved two main parts. The first part was to calculate the distance between two images from a reference image sequence and a query image sequence by removing the deep features of non-overlap contents. The second part was to find the matched reference image sub-sequence to the query image sequence by retrieving the distance sub-matrix with the minimum diagonal distribution. Our framework has the following advantages: 1) it is robust to extreme viewpoint variations and extreme environmental condition variations; 2) it has the potential to provide high-accuracy visual place recognition in practical tasks; and 3) it satisfies real-time requirements. On the other hand, it should be noted that our framework relies on image sequence matching. However, there may be some situations where image sequences cannot be acquired for some reasons (e.g., limited experimental conditions or special scenes). In these situations, our framework is not applicable.

There are still some problems to be solved in visual place recognition. For example, in the process of robot movement, the number of images in the reference image dataset will increase, which may lead to an increase in the time cost of image retrieval in visual place recognition. Therefore, it is meaningful to design methods that can ensure the time cost of image retrieval is not significantly affected by the number of images in the reference image dataset. Additionally, in practical tasks, visual place recognition may be affected by dynamic objects such as moving cars and pedestrians. Therefore, it is useful to include object recognition in visual place recognition. By doing this, the recognized static objects (e.g., buildings and trees) are used for visual place recognition, while the recognized dynamic objects (e.g., cars and pedestrians) are ignored. In the future, we will study how to further refine the positions of non-overlap contents to the pixel level, which can make the images from the same places more similar.

REFERENCES

[1] S. Lowry, N. Sünderhauf, P. Newman, J. J. Leonard, D. Cox, P. Corke, and M. J. Milford, "Visual place recognition: A survey," *IEEE Trans. Rob.*, vol. 32, no. 1, pp. 1–19, Nov. 2016.
[2] A. M. Martínez and J. Vitria, "Clustering in image space for place recognition and visual annotations for human-robot interaction," *IEEE Trans. Cybern.*, vol. 31, no. 5, pp. 669–682, Oct. 2001.
[3] N. Piasco, D. Sidibé, C. Demonceaux, and V. Gouet-Brunet, "A survey on visual-based localization: On the benefit of heterogeneous data," *Pattern Recognit.*, vol. 74, no. 2018, pp. 90–109, Feb. 2018.
[4] M. Cummins and P. Newman, "FAB-MAP: Probabilistic localization and mapping in the space of appearance," *Int. J. Rob. Res.*, vol. 27, no. 6, pp. 647–665, June 2008.
[5] C. Valgren and A. J. Lilienthal, "SIFT, SURF & seasons: Appearance-based long-term localization in outdoor environments," *Rob. Auton. Syst.*, vol. 58, no. 2, pp. 149–156, Feb. 2010.
[6] P. Furgale and T. D. Barfoot, "Visual teach and repeat for long-range rover autonomy," *J. Field Rob.*, vol. 27, no. 5, pp. 534–560, Aug. 2010.
[7] I. Ulrich and I. Nourbakhsh, "Appearance-based place recognition for topological localization," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2000, pp. 1023–1029.
[8] P. Neubert, N. Sunderhauf, and P. Protzel, "Appearance change prediction for long-term navigation across seasons," in *Proc. IEEE Eur. Conf. Mobile Robots*, 2015, pp. 198–203.
[9] A. Ranganathan, S. Matsumoto, and D. Ilstrup, "Towards illumination invariance for visual localization," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2013, pp. 3791–3798.

[10] H. Chen, L. Wu, Q. Dou, J. Qin, S. Li, J.-Z. Cheng, D. Ni, and P.-A. Heng, "Ultrasound standard plane detection using a composite neural network framework," *IEEE Trans. Cybern.*, vol. 47, no. 6, pp. 1576–1586, Mar. 2017.

[11] Y. Zhang, M. J. Er, R. Zhao, and M. Pratama, "Multiview convolutional neural networks for multidocument extractive summarization," *IEEE Trans. Cybern.*, vol. 47, no. 10, pp. 3230–3242, Nov. 2016.

[12] J. Han, H. Chen, N. Liu, C. Yan, and X. Li, "CNNs-based RGB-D saliency detection via cross-view transfer and multiview fusion," *IEEE Trans. Cybern.*, vol. 48, no. 11, pp. 3171–3183, Oct. 2017.

[13] Y. Wei, Y. Zhao, C. Lu, S. Wei, L. Liu, Z. Zhu, and S. Yan, "Cross-modal retrieval with cnn visual features: A new baseline," *IEEE Trans. Cybern.*, vol. 47, no. 2, pp. 449–460, Mar. 2016.

[14] R. Gomez-Ojeda, M. Lopez-Antequera, N. Petkov, and J. Gonzalez-Jimenez, "Training a convolutional neural network for appearance-invariant place recognition," *arXiv preprint arXiv:1505.07428*, 2015.

[15] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "NetVLAD: CNN architecture for weakly supervised place recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 5297–5307.

[16] Z. Chen, O. Lam, A. Jacobson, and M. Milford, "Convolutional neural network-based place recognition," in *Proc. Au. Conf. Robot. Autom.*, 2014, pp. 4–12.

[17] D. Bai, C. Wang, B. Zhang, Y. I. Xiaodong, and X. Yang, "CNN feature boosted SeqSLAM for real-time loop closure detection," *Chinese Journal of Electronics*, vol. 27, no. 3, pp. 488–499, May 2018.

[18] N. Sünderhauf, F. Dayoub, S. Shirazi, B. Upcroft, and M. Milford, "On the performance of convnet features for place recognition," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2015, pp. 4297–4304.

[19] N. Sünderhauf, S. Shirazi, A. Jacobson, F. Dayoub, E. Pepperell, B. Upcroft, and M. Milford, "Place recognition with convnet landmarks: Viewpoint-robust, condition-robust, training-free," in *Proc. Robot. Sci. Syst. XII*, 2015, pp. 1–10.

[20] M. J. Milford and G. F. Wyeth, "SeqSLAM: Visual route-based navigation for sunny summer days and stormy winter nights," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2012, pp. 1643–1649.

[21] M. Milford, "Vision-based place recognition: how low can you go?" *Int. J. Rob. Res.*, vol. 32, no. 7, pp. 766–789, July 2013.

[22] M. J. Milford, I. Turner, and P. Corke, "Long exposure localization in darkness using consumer cameras," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2013, pp. 3755–3761.

[23] N. Sünderhauf, P. Neubert, and P. Protzel, "Are we there yet? challenging seqslam on a 3000 km journey across all four seasons," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2013, pp. 1–3.

[24] E. Pepperell, P. I. Corke, and M. J. Milford, "All-environment visual place recognition with SMART," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2014, pp. 1612–1618.

[25] M. Milford, C. Shen, S. Lowry, N. Suenderhauf, S. Shirazi, G. Lin, F. Liu, E. Pepperell, C. Lerma, B. Upcroft *et al.*, "Sequence searching with deep-learnt depth for condition-and viewpoint-invariant route-based place recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 18–25.

[26] Q. Li, K. Li, X. You, S. Bu, and Z. Liu, "Place recognition based on deep feature and adaptive weighting of similarity matrix," *Neurocomputing*, vol. 199, pp. 114–127, July 2016.

[27] C. McManus, B. Upcroft, and P. Newmann, "Scene signatures: Localised and point-less features for localisation," in *Proc. Robot. Sci. Syst.*, 2014, pp. 1–9.

[28] S. Lowry and H. Andreasson, "Lightweight, viewpoint-invariant visual place recognition in changing environments," *IEEE Rob. Autom. Lett.*, vol. 3, no. 2, pp. 957–964, Jan. 2018.

[29] E. Garcia-Fidalgo and A. Ortiz, "Hierarchical place recognition for topological mapping," *IEEE Trans. Rob.*, vol. 33, no. 5, pp. 1061–1074, June 2017.

[30] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vision*, vol. 60, no. 2, pp. 91–110, Nov. 2004.

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

[32] Z. Chen, A. Jacobson, N. Sünderhauf, B. Upcroft, L. Liu, C. Shen, I. Reid, and M. Milford, "Deep learning features at scale for visual place recognition," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2017, pp. 3223–3230.

[33] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[34] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *Proc. ACM Int. Eur. Conf. Multimedia*, 2014, pp. 675–678.

[35] C. L. Zitnick and P. Dollr, "Edge boxes: Locating object proposals from edges," in *Proc. IEEE Eur. Conf. Comput. Vis.*, 2014, pp. 391–405.

[36] D. Gálvez-López and J. D. Tardos, "Bags of binary words for fast place recognition in image sequences," *IEEE Trans. Rob.*, vol. 28, no. 5, pp. 1188–1197, May 2012.

[37] C. Cadena, D. Gálvez-López, J. D. Tardós, and J. Neira, "Robust place recognition with stereo sequences," *IEEE Trans. Rob.*, vol. 28, no. 4, pp. 871–885, Apr. 2012.

[38] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "Overfeat: Integrated recognition, localization and detection using convolutional networks," *arXiv preprint arXiv:1312.6229*, 2013.

[39] T. Naseer, W. Burgard, and C. Stachniss, "Robust visual localization across seasons," *IEEE Trans. Rob.*, vol. 34, no. 2, pp. 289–302, Jan. 2018.

[40] https://wiki.qut.edu.au/display/cyphy/Day+and+Night+with+Lateral+Pose+Change+Datasets.

[41] N. Carlevaris-Bianco, A. K. Ushani, and R. M. Eustice, "University of michigan north campus long-term vision and lidar dataset," *Int. J. Rob. Res.*, vol. 35, no. 9, pp. 1023–1035, Dec. 2016.

[42] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning deep features for scene recognition using places database," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 487–495.

[43] M. Chen, Z. Shao, D. Li, and J. Liu, "Invariant matching method for different viewpoint angle images," *Appl. Opt.*, vol. 52, no. 1, pp. 96–104, Dec. 2013.

Yong Wang (M'08–SM'17) received the Ph.D. degree in control science and engineering from the Central South University, Changsha, China, in 2011.

He is a Professor with the School of Automation, Central South University, Changsha, China. His current research interests include the theory, algorithm design, and interdisciplinary applications of computational intelligence.

Dr. Wang is an Associate Editor for the *IEEE Transactions on Evolutionary Computation* and the *Swarm and Evolutionary Computation*. He was a recipient of Cheung Kong Young Scholar in 2018 and a Web of Science highly cited researcher in Computer Science in 2017 and 2018.

Taolue Xue received the B.S. degree in automation from Hunan University of Science and Technology, Xiangtan, China, in 2016, and received the M.S. degree in control engineering from Central South University, Changsha, China, in 2019. He is researching on computer vision algorithm in DiDi, Beijing, China. His current research interests include deep learning, image classification, image detection, and image segmentation.

Qin Li received the B.S. degree in intelligent science and technology and M.S. degree in automation both from Central South University, Changsha, China, in 2015 and 2018, respectively. She is currently pursuing the Ph.D degree at Central South University. Her current research interests include computer vision, machine learning, and human-robot interaction.