Copyright ⓒ 2021 KSII

Efficient Visual Place Recognition by Adaptive CNN Landmark Matching

Yutian Chen[1*], Wenyan Gan[1], Yi Zhu[1], Hui Tian[1], Cong Wang[1], Wenfeng Ma[1], Yunbo Li[2], Dong Wang[1] and Jixian He[3]

[1] Institute of Field Engineering, Army Engineering University of PLA
Nanjing 210007, China
[e-mail: 791443626@qq.com; 2372570122@qq.com; zhuyi73@126.com; jaytianhui@foxmail.com; lgd_dolphin@139.com; 19805171502@139.com; dyhkxydfbb@163.com]

[2] Institute of Command and Control Engineering, Army Engineering University of PLA
Nanjing 210007, China
[e-mail: 183155316@qq.com]

[3] Changsha Vocational and Technical College
Changsha 410000, China
[e-mail: hejixian@163.com]

[*]Corresponding author: Yutian Chen

Received March 26, 2021; revised September 7, 2021; revised September 23, 2021; accepted October 27, 2021; published November 30, 2021

Abstract

Visual place recognition (VPR) is a fundamental yet challenging task of mobile robot navigation and localization. The existing VPR methods are usually based on some pairwise similarity of image descriptors, so they are sensitive to visual appearance change and also computationally expensive. This paper proposes a simple yet effective four-step method that achieves adaptive convolutional neural network (CNN) landmark matching for VPR. First, based on the features extracted from existing CNN models, the regions with higher significance scores are selected as landmarks. Then, according to the coordinate positions of potential landmarks, landmark matching is improved by removing mismatched landmark pairs. Finally, considering the significance scores obtained in the first step, robust image retrieval is performed based on adaptive landmark matching, and it gives more weight to the landmark matching pairs with higher significance scores. To verify the efficiency and robustness of the proposed method, evaluations are conducted on standard benchmark datasets. The experimental results indicate that the proposed method reduces the feature representation space of place images by more than 75% with negligible loss in recognition precision. Also, it achieves a fast matching speed in similarity calculation, satisfying the real-time requirement.

Keywords: visual place recognition, CNN, adaptive, landmark, matching

This work was supported in part by National Natural Science Foundation of China (60974086, 62001515, 61771486) and Jiangsu Planned Projects for Postdoctoral Research Foundation (2019K090).

http://doi.org/10.3837/tiis.2021.11.012 **ISSN : 1976-7277**

1. Introduction

Visual place recognition (VPR) is a fundamental but challenging task of mobile robot navigation and localization. In this task, a robot needs to determine whether or not the given image contains a place it has already seen [1]. VPR can be regarded as a special case of image retrieval problems based on some pairwise similarity of image descriptors. However, the solutions to the image retrieval problem are usually sensitive to visual appearance change and also can be computationally expensive. Considering that a place may not always be revisited from the same viewpoint and position and the appearance of a place can change drastically due to environmental changes such as season, illumination, weather, etc., eliminating the influence of the environmental variations and improving the accuracy and efficiency of place recognition has become a critical challenge in VPR.

Traditional VPR methods are mainly based on some hand-crafted image descriptors including both local features and global features, such as Scale-Invariant Feature Transform (SIFT) [2], Speeded-up Robust Features (SURF) [3], and Oriented Fast and Rotated Brief (ORB) [4]. Although these methods achieve promising results, the local features suffer from appearance variations while global features are prone to viewpoint changes. With the tremendous success of deep learning, convolutional neural networks (CNNs) have been widely exploited in VPR recently [5]. The early research focuses on directly selecting appropriate CNN layers to extract features for global image representation. Although this method can achieve high recognition accuracy, it fails to simultaneously handle environment and viewpoint variations. To address viewpoint invariance, a landmark-based VPR framework based on CNN's description of local features is proposed. In this framework, a set of local regions of an image is detected as landmarks and described by a set of CNN feature vectors. Meanwhile, the problem of VPR is reduced to landmark matching by calculating the overall similarity between images from the matched landmarks. Although the VPR methods based on CNN landmarks can achieve good resistance to environment variations and significant robustness to viewpoint changes, the success of such methods heavily depends on the quality of the landmarks detected to represent images, which essentially is a trade-off between the VPR accuracy and computational overhead.

Our research aims to design a more efficient VPR method based on the pre-trained CNN models (**Fig. 1**). The rationality of landmark selection in our method is ensured by a new evaluation index in combination with the removal of location-mismatched landmark pairs. Also, adaptive landmark matching is introduced into the CNN features for accurate similarity calculation. Experimental results demonstrate that our method can obtain a smaller representation space of images and more rapid landmark matching with high precision. The main contributions of this paper are summarized as follows:

(1) A new evaluation index called significance score is proposed for selecting potential landmark areas from the feature maps generated by CNN models.

(2) A more effective landmark matching algorithm is designed by using outlier analysis to remove location-mismatched landmark pairs.

(3) An adaptive similarity calculation algorithm is proposed to assign more weights to landmark matching pairs with higher significance scores.

(4) An efficient visual place recognition method is proposed, which can reduce the feature representation space of place images by more than 75% and can tackle great environmental changes.

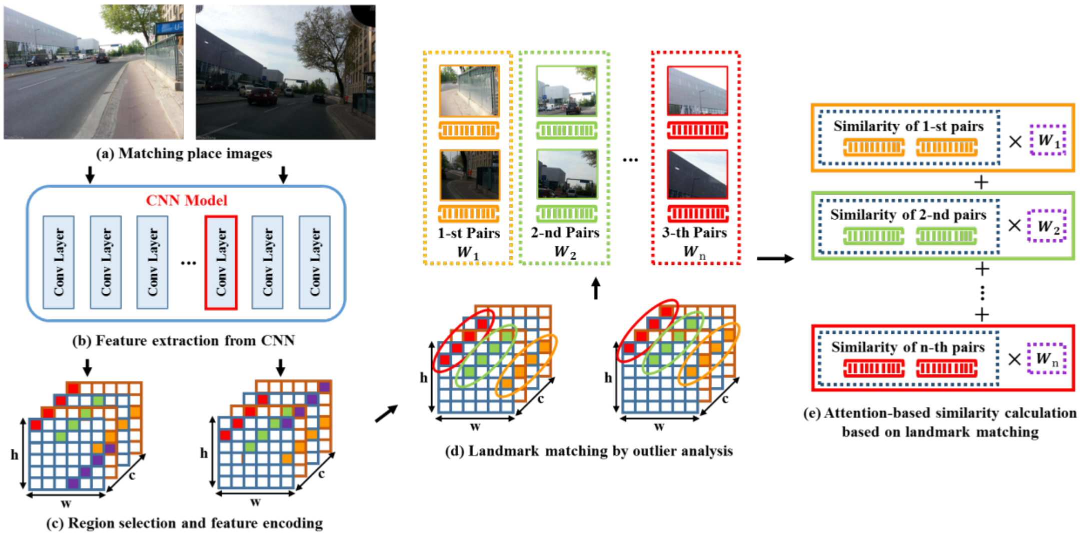

(a) Matching place images
CNN Model
Conv Layer
(b) Feature extraction from CNN
h
w
c
(c) Region selection and feature encoding
1-st Pairs
W_1
2-nd Pairs
W_2
3-th Pairs
W_n
(d) Landmark matching by outlier analysis
Similarity of 1-st pairs
Similarity of 2-nd pairs
Similarity of n-th pairs
(e) Attention-based similarity calculation based on landmark matching

Fig. 1. Illustration of our proposed method

The rest of the paper is organized as follows. Section II provides a brief overview of related VPR studies. Section III describes our method in detail; Sections IV presents the testing datasets and experimental details. In section V, the proposed method is compared to four state-of-the-art VPR methods to verify the robustness of our method. Section VI summarizes our method and the future work.

2. Related Works

Traditional VPR methods [6, 7, 8] applied hand-crafted features to represent place images and aggregated them through pre-trained dictionaries of visual words. The representative hand-crafted features were SIFT [2], SURF [3] and ORB [4]. However, these descriptors failed to deal with complex changing scenes. As a supervised learning method, CNN attracted extensive attention due to its high accuracy and good transferability. In 2012, the AlexNet model [9] achieved great success in large-scale image classification. Since then, CNN had been widely used in computer vision applications, such as image retrieval [10], image classification [11], and object detection [12]. After inputting an image into a pre-trained CNN model, the output features extracted from one layer were regarded as the feature representation of this image. Features extracted from the shallow layers embodies the contour, texture and other shallow information of the image and features extracted from the top layers are more robust with respect to deep semantic information. Fundamental works of CNN-based VPR methods [13, 14] proved that the features generated from the middle layers in combination with shallow texture features and depth semantic features can resist great appearance changes, especially for the conv3 and pool5 layers of the AlexNet. This indicated that the pre-trained CNN models can also be applied to VPR. Thus, CNN-based VPR systems have become a research hotspot. To address environmental changes better, landmark-based VPR frameworks combined with CNN features gradually become the mainstream [15]. These frameworks considered that VPR should only reserve salient landmark regions and it was not necessary to keep the whole CNN features of images. In this case, the problem of VPR can be transformed to landmark matching that calculates the similarity between the images of the landmark matching pairs. Sunderhauf et al. [16] proposed a standard landmark-based process that consisted of four major steps, including landmark detection, CNN feature extraction, dimensionality reduction, and image matching. This method relied on landmark regions selected by landmark detector rather than the whole image to describe one place image. Similarly, the selected landmark area needed to be scaled to adapt to the input size of CNN models. The dimensionality reduction such as Gaussian random projection was used to transfer the original features to a space of much lower dimension. Overall, landmark selection was the most important step in this process. Since landmark-based VPR frameworks depend on the landmark detector, Hou et al. [17] made a comprehensive investigation on the existing landmark extraction methods that were not special for VPR, including Edge-Boxes [18], Binarized normed gradients (BING) [19], YOLOv2 [20], and Selective Search [21]. The results in [17] indicated that the landmarks extracted from BING [19] are the most robust to severe environmental changes. Inspired by the window-scoring mechanism in Edge-Boxes [18], Yang et al. [22] proposed a new landmark generation method for VPR by using multi-scale sliding window (MSW), and the method obtained good results. This method formed a uniform distribution in multiple landmark scales within an appropriate range by a process that sampled an image with a sliding window, which ensured the uniformity of landmark detection distribution. In summary, these landmark-based VPR methods required screenshots for all the selected landmark regions and decomposed an image into multiple images, leading to low efficiency of landmark matching. Meanwhile, because these landmark detectors relied on shallow texture information and lacked semantic understanding of place images, landmarks selected were not accurate, and their levels of attention were not distinguished. This was not consistent with the human visual perception of the scene.

In contrast, some researchers attempted to train CNN models specifically for landmark-based VPR. Chen et al. [23] firstly collected a massive Specific PlacEs Dataset (SPED) for VPR, and the trained CNN model achieved better performance than other pre-trained CNN models. SPED collected nearly 2500000 images by 30000 cameras and all the images captured by these cameras were taken in February 2014 and August 2014. This dataset covered a wide variety of outdoor scenes, ranging from forest landscapes, country roads to urban scenes. Meanwhile, it studied environmental variations, such as lighting changes, day-night circles or season changes. Based on a large number of training datasets, these methods had made great progress. Arandjelovic et al. [24] designed a CNN layer called NetVLAD, which can be trained in an end-to-end pattern for the VPR task. Inspired by the dictionaries of visual words, NetVLAD regarded the output of a convolutional layer as several local descriptors and aggregated them with a specifically designed pooling layer to obtain the effective feature representation. Chen et al. [25, 26] determined salient regions by grouping non-zero CNN feature values into individual clusters and building a training dataset to assign weights to these regions. For each cluster, its energy was calculated by averaging all the activations in this cluster, and only the clusters with large energy could be retained as landmarks. Mao et al. [27] put forward a learning-based attention model from the feature pyramid by weighting the spatial grids on the original CNN features. This attention model built a multi-scale feature pyramid by applying multi-scale pooling at all the spatial locations and concatenating the pooled feature maps, which achieved a feature fusion to improve the robustness of the learned visual representations. Inspired by the brain architectures of fruit flies, Chancan et al. [28] presented a two-layer neural network called FlyNet, which can be combined with a continuous attractor neural network to achieve high performance in VPR. This method designed an insect-based shallow neural network model without resorting to full deep learning architectures. Chen et al. [29] proposed a new multi-constraint loss to optimize the distance constraint relationship in the Euclidean space for efficient CNN model training. In this standard triplet loss method, the tuple composed of three images was used as an input, in which two images were of the same category and the other image was of another different category. The training purpose is to learn a distinguishing image representation, where the distance between images belonging to the same category was minimized and the distance between images belonging to different categories was maximized. In summary, the training-based landmark VPR methods embodied the characteristics of the attention-based model, and the importance of each landmark area for VPR was different. With the support of the training datasets and computing resources, these methods performed well in allocating landmark weights and achieved the highest positioning accuracy. However, they were not efficient in image representation and landmark matching.

Aiming at these problems, a group of scholars attempted to find another way for efficient VPR systems. Based on pre-trained CNN models, they fully analyzed the meaning of the original CNN features and made full use of the activations for landmark selection. Chen et al. [30] evaluated the feature effectiveness of feature maps obtained from the layer of CNN by variance and proposed a novel method that reserved salient feature maps to achieve fast image matching. This method greatly reduced the space of image representation to half of the original CNN features with a tolerable loss in accuracy. Camara et al. [31, 32, 33] proposed semantic and spatial matching VPR (SSM-VPR) that involved global matching-based candidate selection, spatial-constrained candidate reduction, and frame correlation processing. SSM-VPR extracted feature representations from two layers of a pre-trained CNN model by sliding along the layer's horizontal and vertical directions. Later, Principal Component Analysis (PCA) is used to reduce the dimension of image feature representation. The spatial-constrained candidate reduction was based on evaluating location consistency of matching vectors in both images with respect to the anchor points. Frame correlation processing introduced some prior knowledge into the image matching under the premise that there existed some time correlation between frames in a mobile robotics environment. Experimental results showed that SSM-VPR achieved an impressive recognition effect. However, in terms of image representation space and matching efficiency, SSM-VPR was inferior to the method in [30]. Besides, the inability to distinguish the saliency of landmark areas was also a disadvantage of these methods. A summary of various existing representative methods mentioned above is shown in **Table 1**.

Table 1. The summary of various existing representative methods in VPR

Category		Method
Hand-crafted features		Cummins et al. [6] Galvez-Lpez et al. [7] Torii et al. [8]
CNN-based features	Without landmark retraining-free	Sunderhauf et al. [13] Hou et al. [14]
	Landmark-based retraining-free	Sunderhauf et al. [16] Hou et al. [17] Yang et al. [22] Chen et al. [30] Camara et al. [31, 32, 33] Ours
	Landmark-based retraining	Chen et al. [23, 25, 26] Arandjelovic et al. [24] Mao et al. [27] Chancan et al. [28] Chen et al. [29]

3. Method

In this section, our proposed method is illustrated in **Fig. 1**. The sub-graph (a) presents two place images for matching; sub-graph (b) represents the feature maps extracted from CNN; sub-graph (c) describes the step of region selection and feature encoding by calculating significant scores; sub-graph (d) makes landmark matching by using outlier analysis to remove spatial-mismatched pairs; sub-graph (e) illustrates the adaptive image matching process based on different weights of landmark matching pairs.

3.1 Feature Extraction from CNN

The first step in our method is feature extraction from the CNN model, and it is shown in the sub-graph (b) of **Fig. 1**. After an image is input to a pre-trained CNN model, the output feature maps of one layer can be defined as the overall characteristic description of this image. These feature maps could be described as a cube with a size of $c \times w \times h$, i.e., $F \in R^{w \times h \times c}$, where w and h are respectively the width and height of each feature map, and c is the number of feature maps. Formally, the feature maps extracted from the layer of CNN are described in Eq. (1). The c-dimensional feature vector $f_{i,j}$ can be regarded as the feature representation of a certain region in the image. The size of this region is equal to the receptive field of the corresponding CNN layer.

$$F = \left\{ f_{i,j} \in R^c \middle| i \in \{1, \dots, w\}, j \in \{1, \dots, h\} \right\} \tag{1}$$

3.2 Region Selection and Feature Encoding

In the above step, the feature extracted from the CNN layer can be considered as a set of feature representations of $w \times h$ regions in the place image. Obviously, not every region needs to be preserved. Therefore, the second step in our method is to set evaluation indexes for regions and select potential landmark regions from all the $w \times h$ regions.

Typically, a feature map is derived from a continuous operation on the input image, such as convolution, pooling, and activation functions. Among them, convolution is the most important operation. A feature map can be described as the detection scores of the input image after a set of convolution filters. The regions in a place image with high detection scores indicate that there exist different kinds of visual patterns that are searched for by the convolution filters. In fact, the feature maps at a late convolutional layer are generally sparse. The non-zero values in the feature maps are responses to the visual patterns corresponding to some semantically meaningful regions, such as a significant fixed goal. When a place is visited from different viewpoints and conditions, these visual landmark regions are likely to be detected by applying the same series of convolutional filters.

Based on the above observations, a new evaluation index called significance score is set for the c-dimensional feature vector $f_{i,j}$, the feature representation of a certain region in the image. The significance score includes two factors: the sum of the vector $f_{i,j}$ and the number of non-zero values of the vector $f_{i,j}$, which respectively represent the total detection score and the number of responses of the region to various visual patterns. These two factors are positively correlated with the significance score. With the continuous increase of the complexity of CNN models, more and more learning parameters are needed, and the number of extracted feature maps will increase accordingly. In other words, the feature vector $f_{i,j}$ may have a very large dimension. According to the visualization result of the feature maps mentioned in [30], a lot of depth feature values are close to zero. The calculation of the number of non-zero values of the vector $f_{i,j}$ may exaggerate the contribution of the close-to-zero values to the significance of the region. To avoid too many non-zero values in the vector $f_{i,j}$, its square root is used in our method.

The significance scores of all the $w \times h$ regions in image A is stored in matrix E^A, i.e., $E^A \in \mathrm{R}^{w \times h}$. The c-dimensional feature vector $f_{i,j}^A$ is the feature representation of a region with the coordinate (i, j) in image A, and the significance score of vector $f_{i,j}^A$ is stored in the element $e_{i,j}^A$ of matrix E^A. Define the function $S(f)$ as summing the values in vector f and the function $Q(f)$ as counting the number of non-zero values on vector f. The significance score of vector $f_{i,j}^A$ is calculated in Eq. (2).

$$e_{i,j}^A = S\left(f_{i,j}^A\right) \times \sqrt{Q\left(f_{i,j}^A\right)} \tag{2}$$

The threshold t_l is set to determine the proportion of the reserved image regions. The proposed method chooses potential landmarks according to the significance scores sorted from high to low by the threshold t_l. In this way, a limited number of regions are reserved, and the dimension of original CNN features is reduced. Denote the k-th largest value of the matrix E as r_E^k. After region selection and feature encoding, the feature representations $LM\_F \in R^{n \times c}$ of potential landmarks are obtained by the proposed method, and the corresponding horizontal and vertical coordinates are defined as $LM\_P \in R^{n \times 2}$. Note that the number of potential landmarks is $n = \lfloor w \times h \times t_l \rfloor$. The algorithm of region selection and feature encoding is listed in **Table 2**, and its execution process is shown in the sub-graph (c) of **Fig. 1**. The time complexity of this algorithm is analyzed as follows. From the 1-st row to the 7-th row, the time complexity is $O(whc)$; from the 9-th row to the 15-th row, the time complexity is $O(wh)$. Since the value of r_E^k can be determined by the Divide-and-conquer method, the time complexity of the 8-th row is $O(wh)$. Overall, the time complexity of this algorithm is $O(whc)$.

Table 2. The algorithm for region selection and feature encoding

```
Algorithm 1: Region selection and feature encoding
Input: F ∈ $\mathrm{R}^{w \times h \times c}$, $t_l$
Output: $LM\_F$, $LM\_P$, $n$
Procedure:
1   $n = \lfloor w \times h \times t_l \rfloor$
2   Initialize $LM\_F$, $LM\_P$
3   for $i = 0$ to $w$
4       for $j = 0$ to $h$
5           calculate the significance score $e_{i,j}$ in Eq. (2)
6       end for
7   end for
8   Initialize $r_E^n$
9   for $i = 0$ to $w$
10      for $j = 0$ to $h$
11          if $\mathrm{E}_{i,j} > r_E^n$
12              add $f_{i,j}$ to $LM\_F$
13              add $(i, j)$ to $LM\_P$
14      end for
15  end for
16  return $LM\_F$, $LM\_P$, $n$
```

3.3 Landmark matching by outlier analysis

After obtaining the coordinates and feature representations of potential landmarks, our method determines whether the landmarks from different images match. If two images are taken in the same place from different viewpoints and they contain a proportion of the same scene content, a fixed spatial relationship between the landmarks will be preserved. Therefore, landmark matching pairs are likely to have a similar position offset. The coordinate (i, j) mentioned above can be regarded as the spatial position of landmarks. Also, the potential landmarks can be further optimized according to the spatial location relations.

The areas with higher feature similarity are more likely to be the same regions in one place. Denote the feature representations of potential landmarks of the place images A and B as LM_F_A and LM_F_B, respectively. Meanwhile, the corresponding horizontal and vertical coordinates are denoted as LM_P_A and LM_P_B, respectively. Then, the feature representation of the i-th potential landmark in place image A is defined as $LM\_F_A[i] \in R^{1\times c}$, and the feature representation of the j-th potential landmark in place image B is defined as $LM\_F_B[j] \in R^{1\times c}$. The feature similarity between the potential landmarks of images A and B are calculated by cosine distance, and the calculation formula is shown in Eq. (3). The time complexity of calculating the cosine distance between two c-dimensional feature vectors is $O(c)$.

$$Cos(LM\_F_A[i], LM\_F_B[j]) = \frac{\sum_{k=1}^{c}(LM\_F_A[i])_k \times (LM\_F_B[j])_k}{\sqrt{\sum_{k=1}^{c}[(LM\_F_A[i])_k]^2} \times \sqrt{\sum_{k=1}^{c}[(LM\_F_B[j])_k]^2}} \tag{3}$$

The matrix $\mathrm{D} \in \mathrm{R}^{n\times n}$ is a similarity matrix that stores the cosine distance between $LM\_F_A[i]$ $(i \in \{1, \dots, n\})$ and $LM\_F_B[j]$ $(j \in \{1, \dots, n\})$. According to the significance scores ordered from high to low, the potential landmark of A picks the best matching potential landmark of B in a non-returning manner. The formal description of the matching pairs M_p is shown described in Eq. (4), where x_{ai} and y_{ai} are respectively the horizontal and vertical coordinates of the i-th pair in image A; x_{bi} and y_{bi} are respectively the horizontal and vertical coordinates of the i-th pair in image B. Note that $(x_{ai}, y_{ai}) \in LM\_P_A$ and $(x_{bi}, y_{bi}) \in LM\_P_B$.

$$M_p = \left\{\left(f^{A}_{x_{ai},y_{ai}}, f^{B}_{x_{bi},y_{bi}}\right)\middle| i \in \{1, \dots, n\}\right\} \tag{4}$$

The calculation of the position offset is shown in Eq. (5):

$$P_o = \left\{(x_{ai} - x_{bi}, y_{ai} - y_{bi})\middle| i \in \{1, \dots, n\}\right\} \tag{5}$$

Each element of P_o is a two-dimensional point. If the landmark matching above is effective, most of the points in P_o are concentrated in one piece, and the points away from the center areas should be abandoned. This requires data clustering and outlier analysis to ensure the quality of landmark matching. Clustering algorithms have been extensively studied in recent years [34, 35, 36, 37]. Local Outlier Factor (LOF) [38] can be used to find outliers and remove these matching pairs. LOF is a density-based classical algorithm, and its core idea is that the anomaly of a point depends on the local environment. Especially, the fewer neighbors, the more likely a point is an outlier. Our method calculates the number of neighbors for each point based on the k-neighbors-graph. The proportion of the normal point is denoted as t_a. After the potential landmark matching is conducted by outlier analysis on the spatial-location relations, the number of landmark matching pairs $v = \lfloor n \times t_a \rfloor$. The process is shown in the sub-graph

(d) of **Fig. 1**.

Define the function $F(D_i)$ as obtaining the row number of the largest value in the i-th column of the matrix D. The algorithm of landmark matching is listed in **Table 3**. After the execution of this algorithm, our method obtains the feature representations of landmarks of the place images A and B, i.e., $NLM\_F_A \in R^{v\times c}$ and $NLM\_F_B \in R^{v\times c}$. Also, $NLM\_P_A \in R^{v\times 2}$ and $NLM\_P_B \in R^{v\times 2}$, the corresponding horizontal and vertical coordinates of the landmarks in the place image A and B, are also available. Specifically, $(NLM\_F_A[j], NLM\_F_B[j])$ are the feature representations of the j-th matching pair between the images A and B; $(NLM\_F_A[j], NLM\_F_B[j])$ are the corresponding coordinates of the j-th landmark pair between images A and B. The time complexity of this algorithm shown in **Table 2** is analyzed as follows. From the 1-st row to the 6-th row, the time complexity is $O(w^2h^2c)$; from the 7-th row to the 14-th row and from the 18-th row to the 26-th row, the time complexity is $O(w^2h^2)$. The time complexity of the LOF algorithm is proved to be $O(w^2h^2)$ [38]. Overall, the time complexity of this algorithm is $O(w^2h^2c)$.

Table 3. The algorithm for landmark matching by outlier analysis

```
Algorithm 2: The algorithm for landmark matching by outlier analysis
Input: $LM\_F_A, LM\_F_B, LM\_P_A, LM\_P_B, n, t_a$
Output: $NLM\_F_A, NLM\_F_B, NLM\_P_A, NLM\_P_B, v$
Procedure:
1  Initialize $NLM\_F_A, NLM\_F_B, NLM\_P_A, NLM\_P_B, M_P, P_O$
2  for $i = 0$ to $n$
3      for $j = 0$ to $n$
4          $D_{i,j} = Cos(LM\_F_A[i], LM\_F_B[j])$
5      end for
6  end for
7  for $i = 0$ to $n$
8      $k = F(D_i)$
9      set all the values of the k-th row of matrix D to 0
10     add $(LM\_F_A[i], LM\_F_B[k])$ to $M_P$
11     add $(LM\_P_A[i] - LM\_P_B[k])$ to $P_o$
12     add $LM\_F_B[k]$ to $NLM\_F_B$
13     add $LM\_P_B[k]$ to $NLM\_P_B$
14 end for
15 do LOF for $P_o$
16 get K-neighbors-graph $G \in R^{n\times n}$
17 sum each column of $G$ to get a new vector $L \in R^n$
18 $v = \lfloor n \times t_a \rfloor$
19 for $i = 0$ to $n$
20     if $L[i] < r_L^v$
21         remove $LM\_F_A[i]$ from $LM\_F_A$
22         remove $NLM\_F_B[i]$ from $NLM\_F_B$
23         remove $LM\_P_A[i]$ from $LM\_P_A$
24         remove $NLM\_P_B[i]$ from $NLM\_P_B$
25 end for
26 $NLM\_F_A = LM\_F_A, NLM\_P_A = LM\_P_A$
27 return $NLM\_F_A, NLM\_F_B, NLM\_P_A, NLM\_P_B, v$
```

3.4 Adaptive Similarity Calculation based on Landmark Matching

After the correct landmark matching pairs are obtained, the adaptive similarity between two place images is calculated by giving more weights to the landmark matching pairs with higher significance scores. The calculation process is shown in the sub-graph (e) of **Fig. 1**. The distinction of the importance of landmark areas is consistent with the VPR system of humans. In this proposed algorithm, the importance degree of one matching pair is calculated by multiplying the two significant scores of the landmarks in this pair and normalizing the result. Later, the similarity of this pair is obtained by multiplying the cosine distance between the landmarks in this pair by the corresponding importance degree. Finally, the overall similarity $S_{A,B}$ between the images A and B is the sum of the similarity of all the landmark matching pairs, which is shown in Eq. (6). The time complexity of calculating $S_{A,B}$ is $O(whc)$.

$$S_{A,B} = \sum_{j=1}^{v} Cos(NLM\_F_A[j], NLM\_F_B[j]) \times \frac{E^{A}_{NLM\_P_A[j]} \times E^{B}_{NLM\_P_B[j]}}{\sum_{i=1}^{v} E^{A}_{NLM\_P_A[i]} \times E^{B}_{NLM\_P_B[i]}} \tag{6}$$

Searching for the best-matching image with image A needs to go through all the images in the dataset and pick the one with the highest similarity score.

4. Datasets and Experimental Details

To verify the effectiveness of our method, three benchmark datasets for VPR are selected to compare our method with several state-of-the-art methods. The experimental details and datasets are described as follows.

4.1 Benchmark Datasets

Three representative benchmark datasets are used to test the proposed method. The datasets involve complex scenario changes with great condition changes and viewpoint changes, and they are widely used to verify the effectiveness of the VPR algorithm. The key information of these benchmark datasets is summarized in **Table 4**.

Table 4. Descriptions of the testing datasets

Dataset/Sub-dataset	Ref. /Query images	Condition changes	Viewpoint changes
Gardens-Point	200/200	Severe	Severe
Synthesized Nordland	1622/1622	Severe	Moderate
Berlin A100	81/85	Moderate	Severe
Berlin Haleenseetrasse	157/67	Moderate	Severe
Berlin Kudamm	201/222	Moderate	Severe

Specifically, the Gardens-Point dataset [13] recorded a single route through the Gardens-Point Campus in Queensland University of Technology by iPhone 5. The route was traversed three times, twice during the day and once at night. The day route traversing the left side was compared with the night route traversing on the right side to reflect scene changes, including viewpoint, illumination, and walking person. 200 images were recorded in the day-left or night-right conditions. This dataset provided severe conditions and viewpoint changes.

Synthesized Nordland dataset [25] included two video footages that were provided by Norwegian Broadcasting Corporation (NRK). It recorded Norway's northernmost railway linking Trondheim and Bodø across spring and winter seasons. The duration of each video was nearly 10 hours. Since these videos were taken on the train, this dataset involved severe condition changes related to season factors and moderate viewpoint changes. 1622 images were filmed in summer or winter.

Mapillary dataset [39] was built by Google Street View that allowed users to upload sequences of GPS-tagged photos and download these sequences. Since many roads had been mapped by different people, this dataset was suitable for VPR under normal conditions. In this dataset, three sub-datasets (Berlin Haleenseetrasse, Berlin Kudamm, and Berlin A100) were constructed by taking 224, 423, 166 images in the street of Berlin city, presenting severe viewpoint changes with moderate condition changes. Each sub-dataset needed to be tested independently.

4.2 Experimental Details

VGG-16 [40] model is a CNN model for image classification. It is more complex than AlexNet and achieves better performance in VPR. The conv5-2 layer of VGG-16 is a middle layer that fuses deep and shallow feature information, and it is suitable for place feature representation [17, 25, 26, 27]. Therefore, the conv5-2 layer of VGG-16 is used to extract feature representations. After one place image is input to the VGG-16 model, the size of the feature maps extracted from the conv5-2 layer of VGG-16 is 512×14×14, indicating that the image is divided into 196 regions. The receptive field of the conv5-2 layer on the input image is 164×164. Before inputting into the VGG-16 model, the images are resized to 224×244 to fit the input size. For good illustration effect, the most significant three landmark pairs obtained by our method are illustrated. Meanwhile, several sets of thresholds t_l and t_a are selected to test our method. Putting all the feature maps extracted from the conv5-2 layer of VGG-16 into a one-dimensional vector for similarity calculation by cosine distance is used to verify the effectiveness of our method, which is independent of the thresholds t_l and t_a. Besides, to prove the effectiveness of our method, normal cosine distance and adaptive similarity calculation are performed for a comparative experiment.

Several state-of-the-art methods are chosen for performance comparison, including the landmark-based retraining (LBR) method based on the VGG-16 model proposed by [26], the landmark-based retraining-free (LRF) method based on BING and VGG-16 model proposed by [17], the reserving salient feature maps (RSF) method based on AlexNet model proposed by [30], and one SSM-VPR pipeline (SSM) method proposed by [31].

The area under the Curve (AUC) [41] is suitable to evaluate the classifier's performance. The higher the value of AUC, the better the classifier is. In this study, the AUC is chosen as the evaluation metric. Like [26], a match is true positive (TP) if it is within 0±3 for the Gardens-Point dataset or 0±2 frames for the Mapillary dataset and Synthesized Nordland dataset. The precision is calculated as the proportion of TP matches to all the matches selected; the recall is calculated as the proportion of TP matches to the total number of correct matches. Besides, the VPR performance is also evaluated by the overall recognition accuracy, which is the precision at 100% recall.

5. Results

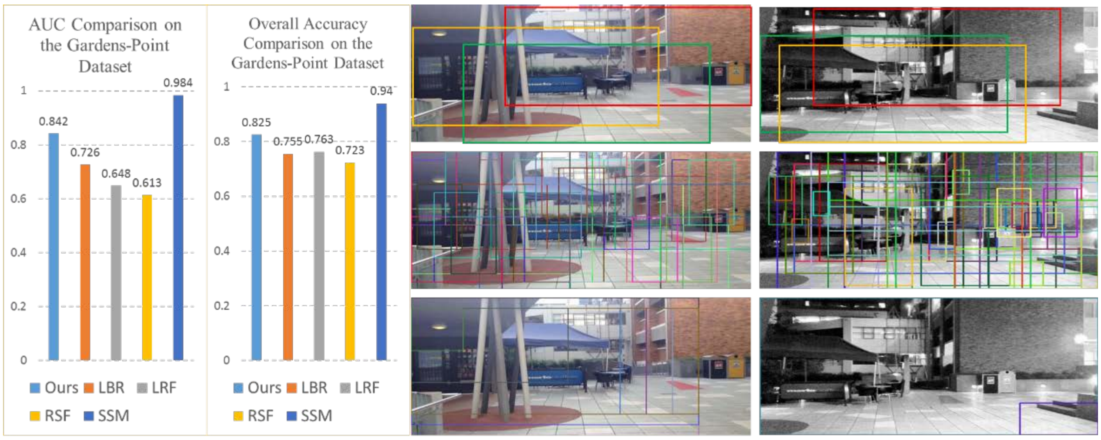
5.1 Results on the Gardens-Point Dataset

Table 5 lists the AUC and overall accuracy of the proposed method under different thresholds on the Gardens-Point dataset based on the VGG-16 model. It can be seen that properly discarding some image areas and mismatched landmark pairs could lead to better recognition results, especially in the case of t_l=0.4 and t_a=0.6. This proves the validity of the main innovations of this paper.

Table 5. The AUC and overall accuracy of the proposed method under different thresholds on the Gardens-Point dataset

Threshold $\mathbf{t_l}$ in step 2	Threshold $\mathbf{t_a}$ in step 3	Similarity Calculation in step 4	AUC	Overall Accuracy
None	None	Cosine distance	0.532	0.491
0.2	1.0	Cosine distance	0.735	0.605
0.4	1.0	Cosine distance	0.762	0.675
0.6	1.0	Cosine distance	0.718	0.650
0.8	1.0	Cosine distance	0.637	0.633
1.0	1.0	Cosine distance	0.608	0.580
0.4	0.2	Cosine distance	0.793	0.695
0.4	0.4	Cosine distance	0.801	0.730
0.4	0.6	Cosine distance	0.813	0.795
0.4	0.8	Cosine distance	0.783	0.770
0.4	0.6	Adaptive Similarity (Ours)	0.842	0.825

Fig. 2 presents the AUC and overall accuracy of different methods for landmark selection on the Gardens-Point dataset. From top to bottom, the sub-graph (c) of **Fig. 2** shows the landmarks selected by the proposed method, BING-based LRF, and LBR methods. The two images in the same row are taken in the same place. There are a variety of visual signs in this dataset. The SSM method achieves the highest recognition accuracy due to the large area of the selected landmarks. Our method achieves the second-best performance, which is mainly attributed to the effectiveness of landmark selection and matching. It can be seen from the sub-graph (c) that our method reserves the main significant regions and maintains the spatial consistency in tentages, buildings, etc. Compared with the BING-based LRF and LBR method, our method is more effective in selecting landmarks.



(a) (b) (c)

Fig. 2. The AUC and overall accuracy of the landmark selection generated by different methods on the Gardens-Point dataset

5.2 Results on the Synthesized Nordland Dataset

Table 6 lists the AUC and overall accuracy of our method under different thresholds on the Synthesized Nordland dataset based on the VGG-16 model. It can be seen that landmarks in this dataset are intensive and limited in the area. Also, properly discarding some image areas and mismatched landmark pairs could lead to better recognition results, especially in the case of t_l=0.6 and t_a=0.4. This proves the validity of the main innovations mentioned in this paper.

Table 6. The AUC and overall accuracy of our method under different thresholds on the Synthesized Nordland dataset

Threshold $\mathbf{t_l}$ in step 2	Threshold $\mathbf{t_a}$ in step 3	Similarity Calculation in step 4	AUC	Overall Accuracy
None	None	Cosine distance	0.518	0.473
0.2	1.0	Cosine distance	0.562	0.508
0.4	1.0	Cosine distance	0.597	0.541
0.6	1.0	Cosine distance	0.613	0.585
0.8	1.0	Cosine distance	0.579	0.527
1.0	1.0	Cosine distance	0.557	0.483
0.6	0.2	Cosine distance	0.675	0.663
0.6	0.4	Cosine distance	0.703	0.757
0.6	0.6	Cosine distance	0.662	0.732
0.6	0.8	Cosine distance	0.646	0.698
0.6	0.4	Adaptive Similarity (Ours)	0.736	0.794

Fig. 3 presents the AUC and overall accuracy of different methods for landmark selection on the Synthesized Nordland dataset. From top to bottom, the sub-graph (c) of Fig. 3 shows the landmarks selected by the proposed method and BING-based LRF and LBR methods. The two images in the same row are taken in the same place. Under strong conditions changes in the natural environment, the rail track information is the key effective landmark information. Because of the feature preservation of rail and its adjacent areas, the SSM method achieves the best performance. Our method obtains the second-best result and completely retains track regions. As for the BING-based LRF method, it retains the critical landmarks, though there is confusion in the landmarks. Due to the lack of landmark information on rails, the accuracy of the LBR method is low.

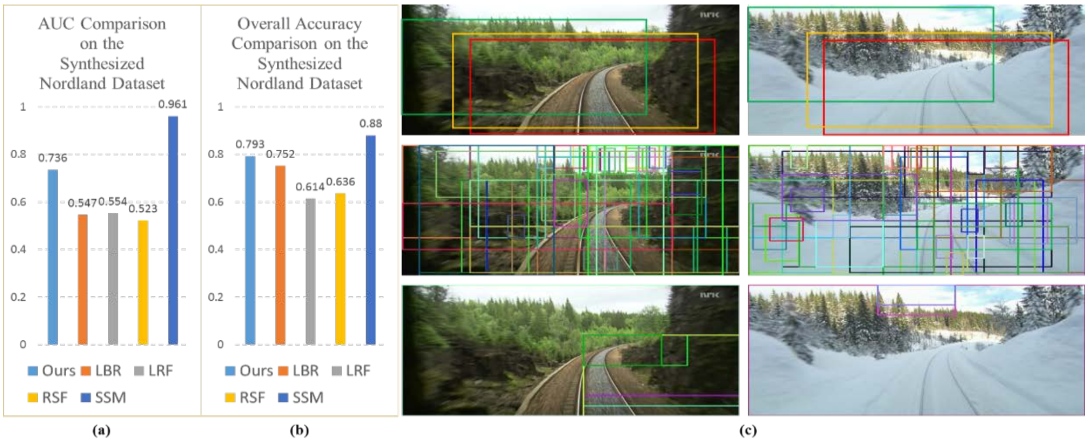


Fig. 3. The AUC and overall accuracy of different methods for landmark selection on the Synthesized Nordland dataset

5.3 Results on the Mapillary Dataset

Table 7 lists the AUC and overall accuracy of our method under different thresholds on the Mapillary dataset based on the VGG-16 model. It can be seen that properly discarding some image areas and mismatched landmark pairs could lead to better recognition results, especially in the case of t_l=0.4 and t_a=0.4. This proves the validity of the main innovations mentioned in this paper.

Table 7. The AUC and overall accuracy of our method under different thresholds on the Mapillary dataset

Threshold $\mathbf{t_l}$ in step 2	Threshold $\mathbf{t_a}$ in step 3	Similarity Calculation in step 4	AUC	Overall Accuracy
Berlin Kudamm sub-dataset				
None	None	Cosine distance	0.558	0.527
0.2	1.0	Cosine distance	0.591	0.584
0.4	1.0	Cosine distance	0.628	0.606
0.6	1.0	Cosine distance	0.622	0.590
0.8	1.0	Cosine distance	0.583	0.562
1.0	1.0	Cosine distance	0.546	0.525
0.4	0.2	Cosine distance	0.689	0.776
0.4	0.4	Cosine distance	0.712	0.818
0.4	0.6	Cosine distance	0.706	0.723
0.4	0.8	Cosine distance	0.658	0.685
0.4	0.4	Adaptive Similarity (Ours)	0.743	0.821
Berlin Haleenseetrasse sub-dataset				
None	None	Cosine distance	0.589	0.543
0.2	1.0	Cosine distance	0.648	0.591
0.4	1.0	Cosine distance	0.651	0.623
0.6	1.0	Cosine distance	0.663	0.634
0.8	1.0	Cosine distance	0.642	0.587
1.0	1.0	Cosine distance	0.617	0.548
0.6	0.2	Cosine distance	0.716	0.652
0.6	0.4	Cosine distance	0.769	0.725
0.6	0.6	Cosine distance	0.712	0.677
0.6	0.8	Cosine distance	0.683	0.629
0.6	0.4	Adaptive Similarity (Ours)	0.795	0.759
Berlin A100 sub-dataset				
None	None	Cosine distance	0.515	0.497
0.2	1.0	Cosine distance	0.566	0.568
0.4	1.0	Cosine distance	0.589	0.592
0.6	1.0	Cosine distance	0.551	0.531
0.8	1.0	Cosine distance	0.534	0.506
1.0	1.0	Cosine distance	0.503	0.488
0.4	0.2	Cosine distance	0.637	0.676
0.4	0.4	Cosine distance	0.659	0.705
0.4	0.6	Cosine distance	0.636	0.648
0.4	0.8	Cosine distance	0.601	0.613
0.4	0.4	Adaptive Similarity (Ours)	0.687	0.724

Fig. 4 presents the AUC and overall accuracy of different methods for landmark selection on the Mapillary dataset. From left to right, the sub-graph (a) of **Fig. 4** shows the landmarks selected by the proposed method and BING-based LRF and LBR methods. The two images in the same column are taken in the same place. There are too many dynamic objects such as cars and pedestrians in the urban environment, which makes the task of VPR challenging. Our method achieves satisfactory results and reflects the benefits of using small landmark regions. It can be seen from the sub-graph (a) that our method reserves the main significant regions with guideposts and maintains spatial consistency. The LBR method is suitable for urban scenery with a lot of moving objects and severe environmental changes. Compared with the BING-based LRF method, our method is more effective in selecting landmarks.

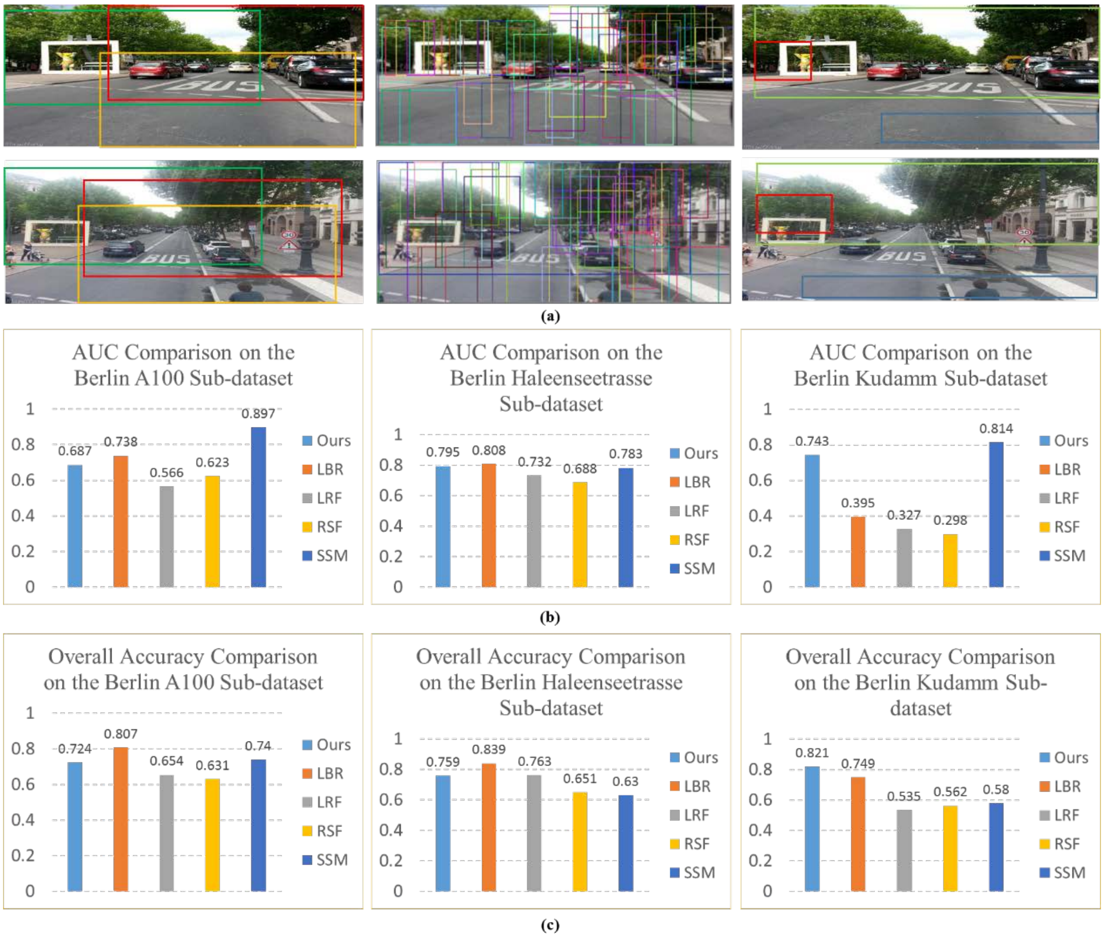


Fig. 4. The AUC and overall accuracy of different methods for landmark selection on the Mapillary dataset

5.4 Image Storage and Runtime

When the thresholds t_l and t_a are respectively set to 0.4 and 0.6, 24% original features are extracted from the CNN model, and our method achieves almost the best results. When the size of features extracted from the conv5-2 layer of VGG-16 is 512×14×14, our method uses nearly 24000 depth feature values to represent one image. Since double-precision floating points values are used in CNN, the storage space of a depth feature value from CNN is 2 bytes, and the feature representation of one image in our method takes a disk space of about 48kB. In comparison, the LRF method selects 100 landmarks, and each landmark is encoded by 1024 depth feature values, so it uses 200kB of disk space to represent one image. The LBR method has 200 regions in an image based on the non-zero values in the conv5-3 layer of VGG-16. Therefore, it uses nearly 786kB of disk space to represent one image [42]. The RSF method is a salient feature selection algorithm that reserves half original features from CNN models, indicating that the representation of each image takes a disk space of 98kB. The SSM method needs 175kB of disk space to represent each image. Compared with these algorithms mentioned above, our method achieves the highest efficiency of feature representation.

Table 8 lists the image storage and runtime comparison between these methods. Generally, after finishing the feature extraction from the CNN layer, the time complexity of our method to calculate the similarity of two images is $O(w^2h^2c)$. As listed in **Table 7**, the process region selection and landmark matching by outlier analysis occupy most of the time. For a single image, one forward pass through the VGG16 model costs about 1.04s, and depth feature encoding takes about 1.28s. The use of our method for image similarity matching within 200 images only takes 6.35s, while the use of the RSF method takes 4.86s less time. The RSF method is faster than the proposed method because of the binarization of CNN features. It is worth noting that the above experiments are conducted under the same experiment configuration as [30], and the Caffe deep learning framework [43] is used for feature extraction from CNN. Besides, it can be seen from the comparison with the RSF method that when the recognition accuracy is improved, the mining and processing time of depth feature information will increase accordingly. Generally, compared with other methods, the running time of our method is satisfactory, especially in the case of weak computing power.

Table 8. Image storage and runtime comparison between different methods

Method	Image storage	Need for retraining	Time-consuming situation	Hardware configuration
Ours	48kB	No	Forward pass by GPU: 1.04s Feature encoding by CPU: 1.28s Image matching by CPU: 31.8ms	GPU: NVIDIA GT940M CPU: Intel i5-5200U
LBR [26]	786kB	Yes	Forward pass by GPU: 0.31s Feature encoding by CPU: 1.33s Image matching by CPU: 50ms	GPU: NVIDIA P100 CPU: Intel Xeon Gold-6134
LRF [17]	200kB	No	Forward pass by GPU: 1.39s Feature encoding by CPU: 0.03s Image matching by CPU: 23ms	GPU: NVIDIA TITAN X CPU: Unknown (4.00GHz)
RSF [30]	98kB	No	Forward pass by GPU: 0.26s Feature encoding by CPU: 0.48s Image matching by CPU: 24.3ms	GPU: NVIDIA GT940M CPU: Intel i5-5200U
SSM [31]	175kB	No	Forward pass by GPU: 0.18s Feature encoding by CPU: 0.22s Image matching by CPU: 26.7ms	GPU: NVIDIA RTX2080Ti CPU: Intel i7-7700

6. Conclusion

This paper proposes a simple yet effective four-step method that can obtain lower -dimensional representations with impressive results by using fewer computing resources. Experimental evaluations demonstrate that our method reduces the feature representation space of place images by more than 75% with negligible loss in recognition precision. Also, it achieves a fast matching speed, and the similarity calculation between two images takes only about 0.03 seconds on an old laptop.

Although our method greatly reduces the image representation space, the process of image feature processing in our method is time-consuming, especially the process of region selection and landmark matching. Therefore, our future work will investigate salient feature map selection to reduce CNN features in two dimensions.

Acknowledgement

The authors gratefully acknowledged the support from the National Natural Science Foundation of China under Grant 60974086, 62001515, 61771486 and Jiangsu Planned Projects for Postdoctoral Research Funds under Grant 2019K090.

References

[1] S. Lowry, N. Sunderhauf, P. Newman, J. J. Leonard, D. Cox, P. Corke, and M. J. Milford, "Visual Place Recognition: A Survey," *IEEE Trans. Robot.*, vol. 32, no. 1, pp. 1–19, Feb. 2016. Article (CrossRef Link)

[2] D. G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91-110, Nov. 2004. Article (CrossRef Link)

[3] H. Bay, A. Ess, T. Tuytelaars, and L. V. Gool, "Speeded-Up Robust Features (SURF)," *Comput. Vis. Image Underst.*, vol. 110, no. 3, pp. 346–359, Jun. 2008. Article (CrossRef Link)

[4] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: An efficient alternative to SIFT or SURF," in *Proc. of IEEE International Conference on Computer Vision*, pp. 2564–2571, 2011. Article (CrossRef Link)

[5] X. Zhang, L. Wang, and Y. Su, "Visual place recognition: A survey from deep learning perspective," *Pattern Recognit.*, vol. 113, pp. 107760, May 2021. Article (CrossRef Link)

[6] M. Cummins, and P. Newman, "Appearance-only SLAM at large scale with FAB-MAP 2.0," *Int. J. Robot. Res.*, vol. 30, no. 9, pp. 1100-1123, Aug. 2011. Article (CrossRef Link)

[7] D. Galvez-Lpez, and J. D. Tardos, "Bags of Binary Words for Fast Place Recognition in Image Sequences," *IEEE Trans. Robot.*, vol. 28, no. 5, pp. 1188-1197, Oct. 2012. Article (CrossRef Link)

[8] A. Torii, R. Arandjelovic, J. Sivic, M. Okutomi, and T. Pajdla, "24/7 place recognition by view synthesis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 2, pp. 257 – 271, Feb. 2018. Article (CrossRef Link)

[9] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84-90, 2017. Article (CrossRef Link)

[10] H. Liu, R. Wang, S. Shan, and X. Chen, "Deep supervised hashing for fast image retrieval," *Int. J. Comput. Vis.*, vol. 127, no. 9, pp. 1217–1234, Sep. 2019. Article (CrossRef Link)

[11] D. Zoran, M. Chrzanowski, P. S. Huang, S. Gowal, and P. Kohli, "Towards Robust Image Classification Using Sequential Attention Models," in *Proc. of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp, 9480-9489, 2020. Article (CrossRef Link)

[12] T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal Loss for Dense Object Detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 318–327, Feb. 2020. Article (CrossRef Link)

[13] N. Sunderhauf, S. Shirazi, F. Dayoub, B. Upcroft, and M. Milford, "On the performance of ConvNet features for place recognition," in *Proc. of IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 4297–4304, 2015. Article (CrossRef Link)

[14] Y. Hou, H. Zhang, and S. Zhou, “Convolutional Neural Network-Based Image Representation for Visual Loop Closure Detection,” in *Proc. of IEEE International Conference on Information and Automation*, pp. 2238–2245. 2015. Article (CrossRef Link)

[15] C. Masone and B. Caputo, "A Survey on Deep Visual Place Recognition," *IEEE Access*, vol. 9, pp. 19516-19547, 2021. Article (CrossRef Link)

[16] N. Sunderhauf, S. Shirazi, A. Jacobson, F. Dayoub, E. Pepperell, B. Upcroft, and M. Milford, “Place Recognition with CNN Landmarks: Viewpoint-Robust, Condition-Robust, Training-Free,” in *Proc. of Robotics: Science and Systems*, vol. 11, 2015. Article (CrossRef Link)

[17] Y. Hou, H. Zhang, and S. Zhou, “Evaluation of Object Proposals and CNN Features for Landmark-Based Visual Place Recognition,” *J. Intell. Robot. Syst.*, vol. 92, no. 3, pp. 505–520, Dec. 2018. Article (CrossRef Link)

[18] Z. C. Lawrence, and P. Dollár, “Edge Boxes: Locating Object Proposals from Edges,” in *Proc. of European Conference on Computer Vision*, pp. 391–405, 2014. Article (CrossRef Link)

[19] M. M. Cheng, Z. Zhang, W. Lin, and P. Torr, “BING: Binarized Normed Gradients for Objectness Estimation at 300fps,” in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3286–3293, 2014. Article (CrossRef Link)

[20] J. Redmon, and A. Farhadi, “YOLO9000: Better, Faster, Stronger,” in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6517–6525, 2017. Article (CrossRef Link)

[21] J. R. Uijlings, K. E. Sande, T. Gevers, and A. W. Smeulders, “Selective Search for Object Recognition,” *Int. J. Comput. Vis.*, vol. 104, no. 2, pp. 154–171, Apr. 2013. Article (CrossRef Link)

[22] B. Yang, X. Xu, J. Li, and H. Zhang, “Landmark Generation in Visual Place Recognition Using Multi-Scale Sliding Window for Robotics,” *Appl. Sci.*, vol. 9, no. 15, pp. 3146-3162, Aug. 2019. Article (CrossRef Link)

[23] Z. Chen, A. Jacobson, N. Sunderhauf, B. Upcroft, L. Liu, C. Shen, I. Reid, and M. Milford, “Deep Learning Features at Scale for Visual Place Recognition,” in *Proc. of IEEE International Conference on Robotics and Automation*, pp. 3223–3230, 2017. Article (CrossRef Link)

[24] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, “NetVLAD: CNN Architecture for Weakly Supervised Place Recognition,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 6, pp. 1437–1451, Jun. 2018. Article (CrossRef Link)

[25] Z. Chen, F. Maffra, I. Sa, and M. Chli, “Only Look Once, Mining Distinctive Landmarks from CNN for Visual Place Recognition,” in *Proc. of IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 9–16, 2017. Article (CrossRef Link)

[26] A. Khaliq, S. Ehsan, Z. Chen, M. Milford, and K. McDonald-Maier, “A Holistic Visual Place Recognition Approach Using Lightweight CNNs for Significant Viewpoint and Appearance Changes,” *IEEE Trans. Robot.*, vol. 36, no. 2, pp. 561−569, Apr. 2020. Article (CrossRef Link)

[27] J. Mao, X. Hu, X. He, L. Zhang, L. Wu, and M. Milford, “Learning to Fuse Multiscale Features for Visual Place Recognition,” *IEEE Access*, vol. 7, pp. 5723–5735, Jan. 2019. Article (CrossRef Link)

[28] M. Chancan, L. Hernandez-Nunez, A. Narendra, A. B. Barron, and M. Milford, “A Hybrid Compact Neural Architecture for Visual Place Recognition,” *IEEE Robot. Auton. Let.*, vol. 5, no. 2, pp. 993–1000, Apr. 2020. Article (CrossRef Link)

[29] L. Chen, S. Jin, and Z. Xia, “Towards a Robust Visual Place Recognition in Large-Scale vSLAM Scenarios Based on a Deep Distance Learning,” *Sensors*, vol. 21, no. 1, Jan. 2021. Article (CrossRef Link)

[30] Y. Chen, W. Gan, S. Jiao, Y. Xu, and Y. Feng, “Salient Feature Selection for CNN-Based Visual Place Recognition,” *IEICE Trans. Inf. Syst.*, vol. 101, no. 12, pp. 3102–3107, Dec. 2018. Article (CrossRef Link)

[31] L. G. Camara, and L. Preucil, “Spatio-Semantic ConvNet-Based Visual Place Recognition,” in *Proc. of European Conference on Mobile Robots*, pp. 1–8, 2019. Article (CrossRef Link)

[32] L. G. Camara, C. Gabert, and L. Preucil, "Highly Robust Visual Place Recognition Through Spatial Matching of CNN Features." in *Proc. of IEEE International Conference on Robotics and Automation*, pp. 3748–3755, 2020. Article (CrossRef Link)

[33] L. G. Camara, and L. Preucil, "Visual Place Recognition by Spatial Matching of High-Level CNN Features," *Robot. Auton. Syst.*, vol. 133, Nov. 2020. Article (CrossRef Link)

[34] B. Diallo, J. Hu, T. Li, G. A. Khan, and Y. Zhao, "Deep embedding clustering based on contractive autoencoder," *Neurocomputing*, vol. 433, no. 3, pp. 96-107, Jan. 2021. Article (CrossRef Link)

[35] B. Diallo, J. Hu, T. Li, G. A. Khan, and A. S. Hussein, "Multi-view document clustering based on geometrical similarity measurement," *Int. J. Mach. Learn. Cybern.*, Mar. 2021. Article (CrossRef Link)

[36] G. A. Khan, J. Hu, T. Li, B. Diallo, and Y. Zhao, "Multi-view low rank sparse representation method for three-way clustering," *Int. J. Mach. Learn. Cybern.*, Aug. 2021. Article (CrossRef Link)

[37] G. A. Khan, J. Hu, T. Li, B. Diallo, and H. Wang, "Multi-view data clustering via non-negative matrix factorization with manifold regularization," *Int. J. Mach. Learn. Cybern*., Mar. 2021. Article (CrossRef Link)

[38] M. M. Breunig, H. P. Kriegel, R. T. Ng, and J. Sander, "LOF: Identifying Density-Based Local Outliers," in *Proc. of ACM SIGMOD International Conference on Management of Data*, vol. 29, pp. 93–104, 2000. Article (CrossRef Link)

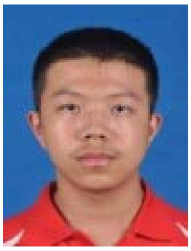
[39] G. Neuhold, T. Ollmann, S. R. Bulo, and P. Kontschieder, "The Mapillary Vistas Dataset for Semantic Understanding of Street Scenes," in *Proc. of IEEE International Conference on Computer Vision*, pp. 5000–5009, 2017. Article (CrossRef Link)

[40] K. Simonyan, and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," in *Proc. of International Conference on Learning Representations*, 2015. Article (CrossRef Link)

[41] J. A. Hanley, and B. J. McNeil, "The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve," *Radiology*, vol. 143, no. 1, pp. 29–36, 1982. Article (CrossRef Link)

[42] M. Zaffar, S. Ehsan, M. Milford, D. Flynn, and K. Mcdonald-Maier, "VPR-bench: an open-source visual place recognition evaluation framework with quantifiable viewpoint and appearance change," *in arXiv*, 2020. Article (CrossRef Link)

[43] 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. of ACM International Conference on Multimedia*, pp. 675–678, 2014. Article (CrossRef Link)



Yutian Chen received the B.S. degree in software engineering from Beijing Institute of Technology, Beijing, China, in 2016, and the M.S. degree in computer science and technology from Army Engineering University of PLA, Jiangsu, China, in 2018. He is currently a teaching assistant with the Institute of Field Engineering, Army Engineering University of PLA, Jiangsu, China. His research interests include visual place recognition and deep learning.

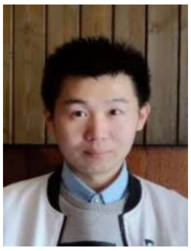
Wenyan Gan received the Ph.D. degree from Tsinghua University, Beijing, China, in 2004. She is currently an associate professor with the Institute of Field Engineering, Army Engineering University of PLA, Jiangsu, China. Her research interests include artificial intelligence and data mining.

Yi Zhu received the Ph.D. degree from Nanjing University of Aeronautics and Astronautics, Jiangsu, China, in 2017. He is currently a lecturer with the Institute of Field Engineering, Army Engineering University of PLA, Jiangsu, China. His research interests include machine vision and aircraft design.

Hui Tian received the Ph.D. degree from PLA University of Science and Technology, Jiangsu, China, in 2016. He is currently a lecturer with the Institute of Field Engineering, Army Engineering University of PLA, Jiangsu, China. His research interests include wireless communication and heterogeneous network.

Cong Wang received the Ph.D. degree from PLA University of Science and Technology, Jiangsu, China, in 2004. He is currently an associate professor with the Institute of Field Engineering, Army Engineering University of PLA, Jiangsu, China. His research interests include computer networks and wireless communications.

Wenfeng Ma received the Ph.D. degree from PLA University of Science and Technology, Jiangsu, China, in 2002. He is currently an associate professor with the Institute of Field Engineering, Army Engineering University of PLA, Jiangsu, China. His research interests include wireless communication networks and broadband wireless communications.



Yunbo Li received the B.S. degree in computer science and technology from Shandong University, Shandong, China, in 2016, and the M.S. degree in computer science and technology from Army Engineering University of PLA, Jiangsu, China, in 2018. He is currently a lecturer with the Institute of Command and Control Engineering, Army Engineering University of PLA, Jiangsu, China. His current research interests include artificial intelligence and machine learning.

Dong Wang received the Ph.D. degree from PLA University of Science and Technology, Jiangsu, China, in 2013. He is currently a lecturer with the Institute of Field Engineering, Army Engineering University of PLA, Jiangsu, China. His research interests include unmanned technology, fault diagnosis and signal processing.

Jixian He received the B.S. degree from Hunan Agricultural University, Hunan, China, in 1998, and the M.S. degree in mechanical engineering from PLA University of Science and Technology, Jiangsu, China, in 2014. He is currently an associate professor with Changsha Vocational and Technical College, Hunan, China. His research interests include electromechanical intelligence and control technology.