Journal of Intelligent & Robotic Systems (2021) 102: 58
https://doi.org/10.1007/s10846-021-01421-2

REGULAR PAPER

Self-Selection Salient Region-Based Scene Recognition Using Slight-Weight Convolutional Neural Network

Zhenyu Li[1] · **Aiguo Zhou**[1]

Received: 31 August 2020 / Accepted: 13 May 2021 / Published online: 3 June 2021

Abstract

Visual scene recognition is an indispensable part of automatic localization and navigation. In the same scene, the appearance and viewpoint may be changed greatly, which is the largest challenge for some advanced unmanned systems,e.g. robot,vehicle and UAV,etc., to identify scenes where they have visited. Traditional methods have been subjected to hand-made feature-based paradigms for a long time, mainly relying on the prior knowledge of the designer, and are not sufficiently robust to extreme changing scenes. In this paper, we cope with scene recognition with automatically learning the representation of features from big image samples. Firstly, we propose a novel approach for scene recognition via training a slight-weight convolutional neural network (CNN) that overall has less complex and more efficient network architecture, and is trainable in the manner of end-to-end. The proposed approach uses the deep-leaning features of self-selection combining with light CNN process to perform high semantic understanding of visual scenes. Secondly, we propose to employ a salient region-based technology to extract the local feature representation of a specific scene region directly from the convolution layer based on self-selection mechanism, and each layer performs a linear operation with end-to-end manner. Furthermore, we also utilize probability statistics to calculate the total similarity of several regions in one scene to other regions, and finally rank the similarity scores to select the correct scene. We have conducted a lot of experiments to evaluate the results of performance by comparing four methods (namely, our proposed and other three well known and advanced methods). Experimental results show that the proposed method is more robust and accurate than other three well-known methods in extremely harsh environments (e. g. weak light and strong blur).

Keywords Scene recognition · Deep learning · Slight-weight CNN · End-to-end training · Salient regions · Self-selection mechanism

1 Introduction

To recognize a previous scene where robot visited is a challenging problem for mobile robot in navigation community. In recent years, many methods are proposed for scene recognition. Recognition methods mainly can be classified into three different categories: manual feature annotation method, sequence-based image matching method and CNN feature extraction method. Some popular methods of handcrafted feature include FAB-MAP [1], SIFT [2] and SURF [3]. FAB-MAP is a probabilistic framework for appearance-based navigation and mapping, which using spatial and visual appearance data. Like the latest work of appearance-based navigation, it utilizes the bag-of-words to distinguish the visited place from the new place by using the positive or negative observation of visual words in the scene. SIFT features detect and describe local features in an image based on interest points in the local appearance of objects. It finds the extremum points in the space scale and extracts its position, scale and rotation invariant. Therefore, it is invariant to rotation, scaling and brightness changes, and robust to viewpoint, affine transformation and noise to a certain extent. However, the process of SIFT feature representation relies heavily on hardware acceleration and retrieval or matching of specialized image processors. Therefore, it is not easy for general computers to detect,extract SIFT features and represent scenes in real time. Speeded-Up Robust Features (SURF) borrowed the thought of simplified approximation from SIFT, and simplified the Gaussian second-order difference template

✉ Zhenyu Li
zhenyu.li@tongji.edu.cn

[1] School of Mechanical Engineering, Tongji University, Shanghai, China

in DoH, thus greatly improving the efficiency of feature extraction. It is a challenging task for mobile robots to be able to recognize the same scene in different seasons and day and night. However, the effect of feature method at that time was not enough to accomplish such a hard task, especially the artificial features, such as SIFT and SURF. Therefore, a new view-based visual localization and navigation approach is proposed, which is the extension of SeqSLAM and called as SeqSLAM++ [4]. This method is not only capable of robustly estimating the position of the robot, but also robustly responding to changes in the heading and speed of the robot. In recent years, the extensive application and rapid development of deep feature technology in images has gradually shown considerable results. Therefore, a large number of scientific research workers related to vision topics have devoted themselves to the discussion of methods combining deep learning and visual detection, e.g. image recognition and object detection. Compared with the traditional hand-crafted feature (e.g. SIFT, SURF), the methods of deep learning can automatically extract features and learn feature representation based on these features from big data. Due to the Non-linear deep convolution operation, the deep features are sufficient discrimination and do not require any complicated aggregation technology that commonly used in embedded and hand-crafted functions, and greatly improves the performance of feature extraction and network training. In the paper, we also adopt deep learning feature for visual scene recognition. The proposed method is a light-weight visual scene representation, and utilizes the light-weight neural network structure to learn these features. This light-weight network architecture is designed specifically for scene recognition tasks, which has the capability of identifying visited scenes although severe changes in appearance and viewpoint. The proposed network is based on VGG-16, but using only some of the modules, and discarding the fully connection layer. At present, CNN and other neural networks are rapidly developing and being applied. In order to pursue high accuracy, the number of deeper and more complex of network models are increasing. Such as some related applications of VGG, GoogleNet and ResNet series in some recognition tasks. However, in some real application scenarios such as mobile or embedded devices, such a large and complex model is difficult to apply. Therefore, it is very important to develop a small and efficient CNN model. At present, the research on this kind of problem mainly focuses on two aspects: to compress the trained model to get a small model or to design a small model directly. In general, considering the limitation of the training model compression, it is a feasible scheme to design a small and precise network structure based on the existing CNN model. In addition, this method is region-based feature representation. As we all known, some regions in an image just contain meaningless parts, and some of the regions contain important parts that can represent scene information. So, in our work, we need to discard meaningless context and save meaningful information. As shown in Fig. 1, the proposed method can identify salient regions, and creates their regional representations directly by the convolutional operation. The major contributions of this paper can be summarized as follows:

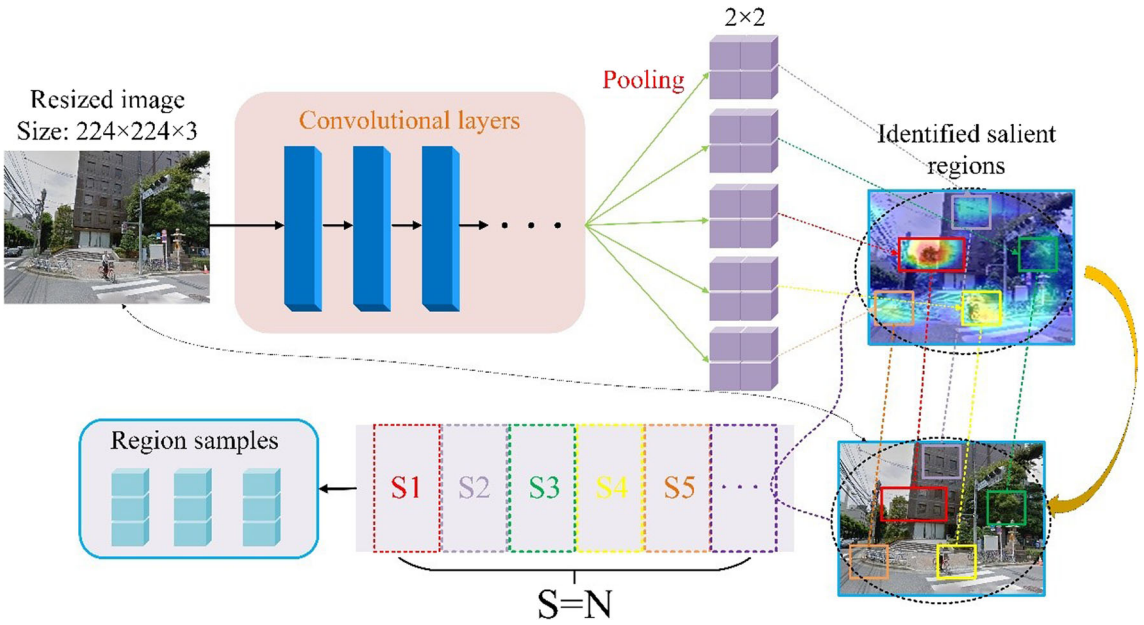

Fig. 1 High salient regions are recognized by convolutional operation, each of pooled vector describes one image region, and all high salient regions make up region samples, which is used to represent the similarity between referenced image and query images

1) We propose a slight-weight convolutional neural network for scene recognition. The network overall has less complex and more efficient network architecture, which has both good performance and high computational efficiency.
2) The proposed network extracts the local representation of a specific image region directly from the convolution layer activation, and each layer performs a linear operation with end-to-end.
3) We also proposed a self-selection mechanism, which is able to collect all the salient features in each convolution module to more efficient represent scenes. And in the back-end of network, utilizing combination of self-selection mechanism and similarity function to offset the performance loss caused by the simplified network.
4) We utilize confidence interval and probability statistics to calculate the total similarity of several regions in one scene to other regions, and finally ranked the similarity scores to select the correct scene. The organizational structure and writing ideas of the paper are as follows. In Section 2, We briefly summarize the related research in the domain of visual scene recognition in recent years based on different feature representations. Section 3 introduces the process of network training and image matching methods. The experimental results and analysis will be presented in Section 4. Finally, this paper is summerized and future work is presented in Section 5.

2 Related Work

Scene recognition is a hot research topic in the robot navigation community [5–7]. The successful application of deep learning technology in the field of computer vision has led to more extensive further research on scene recognition. Many CNN-based methods are proposed for scene recognition, which all using deep features extracted from deep network that are trained for recognition tasks [8–10]. In this section, we briefly overview the previous work and methods of scene recognition by using CNN-based methods.

2.1 Visual Scene Recognition with Global or Local Feature Representation

Global feature representation refers to the overall property of the image, Common global features include colour features, texture features shape features and intensity histogram, etc. As it is a low-level visual feature at the pixel level, the global feature is characterized by good invariability, simple calculation and intuitive representation. Early research on global feature extraction mostly focused on the extraction of low-level features such as colour, edge and shape of RGB images, referring to [11–14], which almost applied low-level features to image classification, image retrieval and object detection. Although the global feature has good scale invariance, its high feature dimension and large amount of calculation are its fatal weakness. In addition, the global feature representation is not suitable for the case of image aliasing and occlusion. Compared with global features such as line feature, texture feature, structure feature, etc., local image features have rich implication amount in the image, low correlation among features, and will not affect the detection and matching of other features due to the disappearance of some features under occlusion. In the literature [15], a graph-based visual place recognition method is proposed, which constructs graph by jointing the deep local visual features (CNN features) and the temporal information of the images in a sequence. Large-scale visual scene recognition is extremely important for robots. Due to the large amount of data collected from scenes, visual classification and recognition is difficult. To tackle this challenge, Vector of Locally Aggregated Descriptors (VLAD) is presented, which uses the distribution of local features in each cluster to construct global features, such as [16–18], all of them use it as a network backend to perform detection tasks, such as scene recognition, image classification, and object recognition. Nowadays, in order to improve the accuracy of recognition, the network tends to learn more deep layers, but this undoubtedly increases the computational burden. To cope with such a challenge, the researchers do the opposite, using simplified convolutional networks to extract local features and other technologies, e.g. effective loss function, to construct modular networks. These methods meet computational efficiency as well as accuracy requirements [19–21].

2.2 Visual Scene Recognition with Region-Based Feature Representation

All the methods mentioned above extract global or local representations from a whole image. However, many of the extracted features are meaningless, but they are still extracted along with the useful features, which results in a serious waste of computing resources. Deal with such problems, in the recently, a method relied on region scene representation is proposed, which was first applied in image retrieval. For example, Kim, J. et al. [22] presents a novel image retrieval method, which utilizes region-based feature aggregation to deal with the problem what the background clutters and varies importance of regions. This method is a simple and effective, salient regional attention network that focuses on an attentive score of a region considering global attentiveness. Lankton, S. et al. [23] just considers

local rather than global image information and proposed a method of re-representing the segmentation energy based on the region in a local manner. Carson, C. et al. [24] presents a region-based method, which is used to automatically extract the colour and content information related to the objects in each image. Image retrieval is based on some regions of interest, rather than a representation of the entire image. Region-based method is also used to object detection, such as [25], which uses region-based fully convolutional networks to obtain accurate and efficient object detection results. This method adopts region-based detectors, which is fully convoluted, to share almost all computations across the entire image. In addition, to cope with the large computing cost of recognition algorithms, the literature [26] presents a lightweight CNN-based scene recognition technique that captures multilayer region-based attentions robust under changing environment and viewpoints. In the literature [27], a holistic visual scene recognition approach is presented, which also using light-weight CNNs for severe viewpoint and appearance changes. With the rapid development of artificial intelligence and image processing technologies, compound multi-column convolution network has been widely used in visual recognition domain. For example, Researchers interpret scene recognition as a region-based image retrieval problem and propose a new method that uses an end-to-end trainable multi-column convolutional neural network (MCNN) structure for scene recognition [28, 29]. These methods have the ability of multi-level perception of the external environment. Similar to the above methods, we also utilize the region-based method for scene recognition. As we all known,the network structure is getting deeper and deeper, and the weight parameters are also getting larger and larger, which undoubtedly causes great computing pressure on mobile microcomputers, and the consequences may be the adverse effects such as reduced transmission response, accelerated computer wear and excessive consumption of electric energy. The difference is that we don't use the fully connectional layer in the process of learning features in the end of the constructed network to reduce so big weight data, and just rely on some convolutional modules to represent special scene. In addition, in order to make up for the lack of network performance caused by the simplification of convolution layer, we propose a feature self-selection mechanism to selectively extract significant feature information as well as discard the useless information in the scene. At the same time, in the back end, we calculate the similarity score according to the correlation of regions of interest between reference image and current image. The proposed method utilizes a slight-weight CNN for scene recognition, which is trainable in an end-to-end manner. To sum up, simply considering the network performance, the presented network of this paper is not as good as that of the current popular networks. However, the weight parameter of our network is much less, so our improvement is obvious in the calculation speed and energy consumption. At the same time, we introduce the feature self-selection mechanism into the network, which helps our network to find out more meaningful scene information. To a certain extent, this effectively improves the accuracy on the premise of ensuring a small increase in computation. All of which has created the basis for deep learning technologies to run on onboard microcomputers.

3 Proposed Method

In this section, we describe the key components of the proposed approach in detail. We first describe the structure of the proposed network and training processing, then we describe how the local features in an image to be extracted from convolutional layers, and we also describe how to extract the local descriptor to represent full image. Further, the same salient region between image A and image B are detected, and the total similarity between two images can be calculated by Summation of similarity scores for all regions. Last, according to the ranking of similar scores, the best matched scene can be selected.

3.1 Training the Proposed Network

The main goal of training a network is to accumulate a large number of weight parameters for learning the representation of some scenes, which is helpful for us to solve the scene recognition of subsequent tasks. The Fig. 2 describes the slight-weight structure of our proposed network, which is re-designed relied on VGG-16. However, the structure in the above figure cannot completely represent our network. In the sake of clearly showing the process of network operation, the pooling process is not described in this picture, but it is real running in the whole network operation. The proposed network structure consists of three parts: input module, convolution module and output module. The image dataset is treated as the input of the network, and all the images of the input network are pre-processed and the size of images is redefined. After convolution operation, the image is output as descriptor with the function of representing scenes. We define the an image can be represented as I $(H \times W \times D)$,in which H, W and D represent high, width and depth of image respectively.All images are first adjusted to the size of $224 \times 224 \times 3$ before imported into convolutional layers. A pre-trained convolution network on ImageNet is used for the total process of feature learning. The used network consists of five modules and 5 pooling layers (each convolution layer is immediately followed by a pooling layer), in which, the first two block all include two convolution layer. The third and

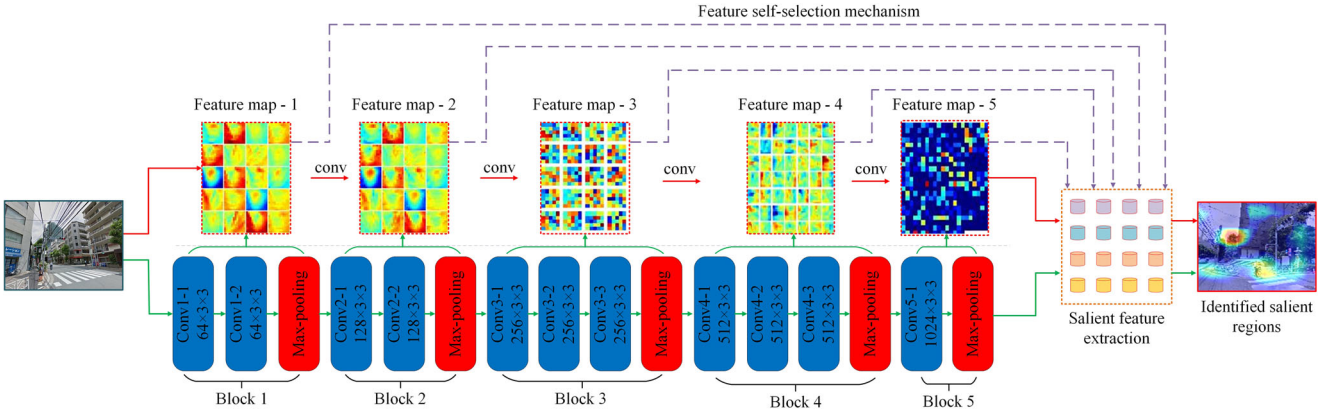

Fig. 2 The architecture of proposed network. The feature map of each convolutional block and the self-selection mechanism of salient features are presented at the top of picture. The size of convolutional block and the number of filter and the size are presented at the bottom of picture. The same size convolution kernel (3 × 3) is used in all convolutional operation. We abandon the fifth block of VGG-16 and choose the first convolution layer of the sixth block as the end of the whole convolution operation

the fourth convolution modules consist of three convolution layers, respectively. The last modules just include one layers. The whole network is fully convolutional operation, and the last layer is not followed fully connected layers. The number of filters with five modules from the first to the five are 64, 128, 256, 512 and 1024, respectively. All of the convolution kernels have the same size of 3 × 3. All these convolution layers use the same padding size of 2 × 2. All of pooling layers are applied with max pooling, which use the kernel size of 2 × 2 and the stride size of 2 × 2. In addition, all of the activation layers in the convolution blocks are applied with Rectified Linear Unit (Rule) function. Before the proposed network is trained, the learning rate is set as 0.001 and the batch size is set as 30. The output of the image trained by the last convolutional layer is 1024 neurons. We also pre-process all images used in the paper, and we adjusted the size of the original image to 224 × 224 pixel. The operation of each convolution module and the image changes after convolution are presented in the Table 1.

In our training network, we use Rectified Linear Unit (Relu) as an activation function. The Relu layer applies (1) to all values of the input, as well as turns all negative activation to 0. In addition, this layer will increase the nonlinearity of the model and even the entire convolutional network, and will not affect the expression effect of the convolutional layer, so as to solve the problem of slow learning convergence of the neural network caused by the disappearance of the gradient [37].

$$f(x) = max(0, x_i) \tag{1}$$

Where $x_i \in \mathbf{R}^{H \times W \times D}$ is the i feature map, which is a 3D metrix, and obtained from the last convolution layer of each convolutional modules, and treated as the input of the next convolution layer. The whole convolutional operation is carried out at the pixel level, and we first parameterize this operation, in which, $x_{i,j}$ describers the row and column elements of the images. To parameterize each weight of the filter, $w_{m,n}$ is used for describing the row and column weight of filters (m is row and n is column), and w_b is used for presenting the bias item of filters. In addition, we also parameterize each feature map after each convolutional operation. $a_{i,j}$ is used to present the row and column element of the each feature map (represent the element of row and column are presented as i and j , respectively). Therefore, in order to enhance the nonlinear processing ability of the network, we introduce the nonlinear factor of Eq. 1 into the whole process of convolutional representation. The convolution output of each layer is shown in Eq. 2.

$$a_{i,j} = f\left(\sum_{d=0}^{D}\sum_{m=0}^{E}\sum_{n=0}^{E} w_{m,n}x_{i+m,j+n} + w_b\right) \tag{2}$$

Where D is image depth, E is the size of filter.

3.2 Extracting Local Descriptors by Convolutional Operation

In our work, a redesigned slight-weight convolutional neural network is used for descriptors extraction, and a pre-training weight is obtained by training big data, e.g. ImageNet dataset. Although CNN was initially trained on public datasets, as well as used for image retrieval and automatically extract deep features, which has been turned out to be highly robust and more effective than traditional hand-made features in other visual recognition or object detection tasks, e.g. place recognition, image segment and face detection. CNN network transforms an input image into feature representations though continuous simple convolutional operations or block's training, and each convolutional block carries out a linear operation with end-to-end manner. The main reason why we redesigned

Table 1 The statistics of network parameters of the slight-weight network architecture

Block	Type	Patch size/stride	Dimensions	Input size (RGB image)
Block 0 (Input)	———	———	———	224 × 224
Block 1	Conv 1-1	3 × 3	64 × 3 × 3	112 × 112
	Conv 1-2	3 × 3		
	Maxpooling	2 × 2/2		
Block 2	Conv 2-1	3 × 3	128 × 3 × 3	56 × 56
	Conv 2-2	3 × 3		
	Maxpooling	2 × 2/2		
Block 3	Conv 3-1	3 × 3	256 × 3 × 3	28 × 28
	Conv 3-2	3 × 3		
	Conv 3-3	3 × 3		
	Maxpooling	2 × 2/2		
Block 4	Conv 4-1	3 × 3	512 × 3 × 3	14 × 14
	Conv 4-2	3 × 3		
	Conv 4-3	3 × 3		
Block 5	Conv 5-1	3 × 3		
	Maxpooling	2 × 2/2	1024 × 2 × 2	

The original size of images is 224 × 224, and the final output image size is 14 × 14. The size of the output feature map of each block is the input size next one

convolutional network is to implement the function that extracts the local representation of salient image regions directly from the last activation layer of convolutional network. Referring to [30], we use I to represents a reference image, and describe its activation that extracted from a convolutional layer as a tensor, who's size can be represented as $H \times W \times D$ that presents the height, width and depth of each feature map is H , W and D respectively. After being activated, the pooled feature map can be represented as a descriptor, which is a feature vector with local features and multi-dimensional space characteristics. This vector can represent a specific region of interest in a scene. The size of this region is consistent with the perceiving region of the filter. So each feature descriptor can represent a special and meaningful region in a scene. Then, a three-dimensional tensor can be responded as a feature channel of two-dimension (2D) though the Eq. 3.

$$X = X_i, i = 1, 2, ..., K \tag{3}$$

In above equation, we describe X_i as a 2D tensor that represents the response of the i^{th} feature channel, and $X_i(S)$ as the response that corresponds to a special position in an scene. Therefore, the feature descriptor is constructed by aggregating the maximum pool of feature map information in the feature space and at each pixel position in a scene [31]. Generally speaking, the features generated by the front-end convolution operation express the low-level image meaning, e.g. point and line, while the features generated by the intermediate operation express the higher level image meaning, e.g. table and chair, while the features generated by the back-end operation express the high-level semantic meaning of the image. To this end, we utilize a self-selection technique to extract the discriminating features from each convolutional module to make the feature representation more robust, and to make the aggregated features have scale invariance. The feature selection mechanism can be described by Eq. 4:

$$X' = \left\{X'_i\right\} = max\left\{X_i\right\}, i = 1, 2, ..., K \tag{4}$$

Where the $max\{\}$ represents the more salient features comparison with nearby features. The difference in the number of features selected in each scene will also affect the accuracy of the total matching, at the same time, the proportion of selected features in each convolution module in the total features will also affect the matching, as shown in Fig. 3 (left). It can be seen that the effect of feature representation is improved with the increase of feature number. However, when the number of features increases to a certain one, the effect of feature representation is not improved obviously, or even no longer improved. Furthermore, the feature self-selection mechanism we use can automatically extract high-level features from each module. In the Fig. 3 (right), we can see that block 3 and block 4 have the highest proportion of selected features, which also shows that the feature self- selection mechanism

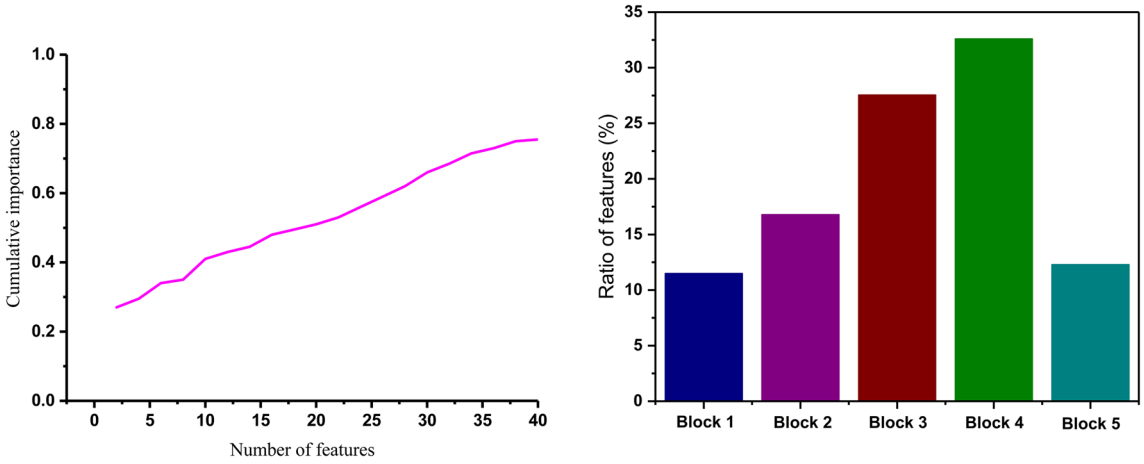

Fig. 3 Cumulative feature importance and the different ratio of feature at different convolutional block.We define that $A(x_i^k)$ and $A(x^k)$ represnet the number of self-selection feature in k^{th} convolutional modules. Therefore, the ratio of selected features on each module $\ell = \frac{A(x_i^k)}{A(x^k)}$ and the cumulative importance $\jmath = \frac{A(x_i)}{A(x)}$

used in light-weight network can extract high-level features from any scene.

In our work, $f_{I,i}$ is used for representing the i special position, and $f_{S,i}$ is used for representing a special salient region in an scene.

$$f_I = [f_{I,1}, f_{I,2}, ..., f_{I,k}] \tag{5}$$

$$f_{S,i} = max\{X_i'(S)\} \tag{6}$$

3.3 Image Retrieval

The purpose of this paper is to utilize the spatial feature vector to describe the local salient regions in the scene. By aggregating all the meaningful and salient features in a scene, we can directly compare the similarity of each region between the two any scene, and then calculate the similarity level between the two scenes. For any two scenes, the similarity is not achieved by comparing the similarities between all the pixels in the two scenes, because this will consume a lot of computing resources, but only by comparing the salient and meaningful things (such as people, trees, tables and chairs, etc.) in the two scenes. Therefore, for two any scenes, there is no need utilize every neuron to perceive the global image to decide whether they are similar, and just to compare the local regions of interest. Then, integrating the local information of each scene to a higher level to obtain the global representation. The key aim of feature vector we extract is to estimate the scene characteristics through vector representation, and then pave the way for further other recognition tasks. In this paper, the statistical method of salient region clustering proposed in this paper can easily detect the local salient regions, as shown in Fig. 4, $f_{I,1}$, $f_{I,2}$, $f_{I,3}$ and $f_{I,4}$ represent the salient regions, which are the most important and meaningful content in the image. The scene of high activation values in this image indicates that filters are searching for visual patterns around them. The characteristics of feature map extracted from the last layer of convolution is very sparse and very relevant to the semantically interesting region. Therefore, when a scene is visited by a visual robot from different viewpoint or appearance, some iconic visual information will be retained as 'landmarks' and re-detected in subsequent images of the same regions using the same convolution filter. To achieve the matching of two scene scenarios, it is necessary to calculate the similarity between the two scenes. We assume that image A is a scene that was taken from a scene where the robot has visited, and image B is a scene that is taken when the robot visits again. Our goal is to achieve the matching between two scenes, and then

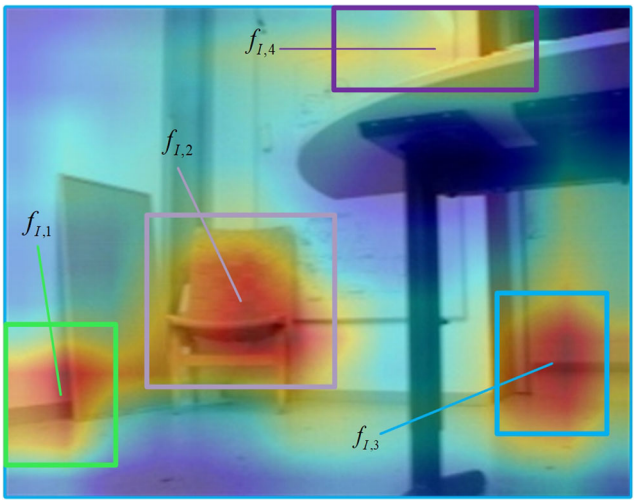

Fig. 4 Illustration of high salient regions. All of regions are extracted from the last convolutional layer activations in an example scene

realizes the scene recognition for mobile robots. In order to describe the similarity between two scenes and parameterize the similarity, we utilize proposed method to extract all the highly salient regions in the two images, and then use the cross-match method to match these vectors [32]. For example, in the Fig. 5, each salient region can be represented as a space vector after convolution. We can judge whether the two regions belong to the same place by comparing the cosine similarity of the region vectors of corresponding positions in the two images. The range of cosine value covers [- 1, 1]. The closer the value is to 1, the closer the distance between the two vectors, indicating that the two regions are more similar. The closer they are to negative 1, the farther the distance between the two vectors, indicating that the two regions are more dissimilar; approaching 0 means that the two vectors are almost orthogonal, indicating that the two regions have no similar relationship.

According to the Fig. 5, cosine similarity of any pair of images can be calculated. For example, we can obtain the similarity scores of region 1, region 2, and region 3 between image A and image B in the Fig. 4. Then , we utilize equation (7) to calculate the similarity between region i and region j, in which region A and region B are part of reference image and query image respectively.

$$S_{i,j} = cos(\theta_{i,j}) = \frac{f_{I,i}^{A^T} \times f_{I,j}^{B^T}}{\left\| f_{I,i}^{A^T} \right\| \left\| f_{I,j}^{B^T} \right\|} \tag{7}$$

According to the Eq. 7, the high salient regions in the reference scene and query scenes can be matched. However, we first need to calculate the similarity scores between all the corresponding salient regions in the two scenes, and then get the final similarity value through addition operation. In our work, the confidence estimation method is applied to the corresponding regions similarity calculating between two scenes, and all of interest of regions in a scene make up a region sample, we assume that the probability of all regions in scene corresponding to the salient regions in the reference image is P_r , in addition, $\{C_1 \le \mu \le C_2\}$ is the overlapping area of any two corresponding similar regions and is significance level, then the score of similarity between image A and image B can be represented by Eq. 9:

$$P_r \{C_1 \le \mu \le C_2\} = (1 - \alpha) \times 100\% \tag{8}$$

$$Q_{A,B} = \frac{\sum_{k=1}^{n} (A_i \times B_j)}{\sqrt{\sum_{k=1}^{n} A_i^2} \times \sqrt{\sum_{k=1}^{n} B_j^2}} \times P_r \tag{9}$$

For the sake of searching a reference scene A that is a best matcher for scene B in some visual scene datasets, we traverse all referenced scenes in a dataset. Then the scene with the highest similarity score is selected as the matching object. The similarity function can be represented as Eq. 10:

$$Q = f(A, B)_{MAX} = argmax \left\{ Q_{A,B}^1, Q_{A,B}^2, ..., Q_{A,B}^N \right\} \tag{10}$$

In order to show our scene recognition method more clearly, we show our algorithm in the way of logical language, as shown Table 2:

4 Experiment Setup

This section objectively evaluates our proposed method of salient region-based scene recognition though training a re-designed slight-weight CNN. For all of experiments, in order to ensure the objectivity and fairness of experimental

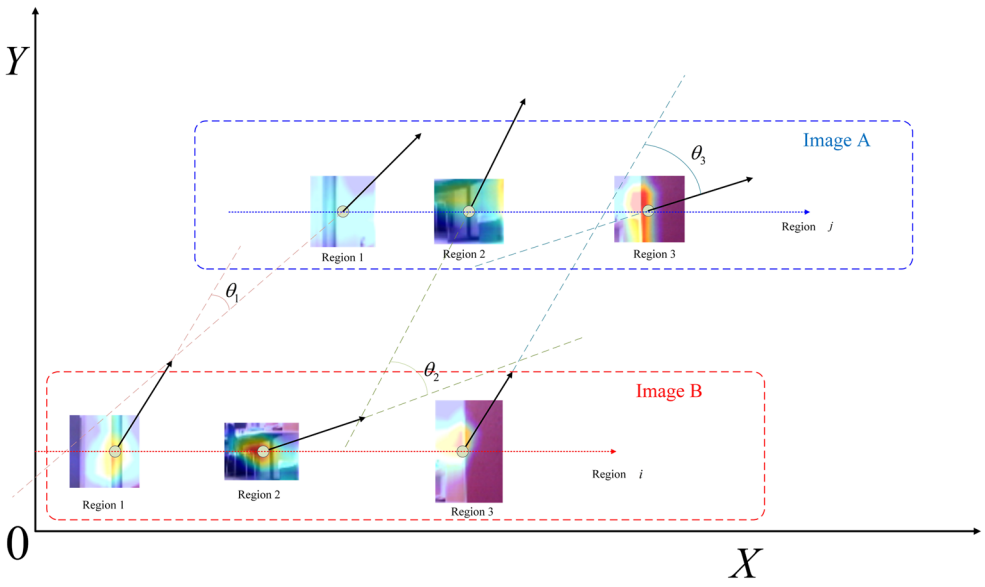

Fig. 5 Schematic diagram of vector similarity calculation of corresponding similar regions in two images

Table 2 Scene recognition algorithm

Input: Datasets images A_i, $A_i \in (A_1, A_2, ..., A_m)$
Query images B_j, $B_i \in (B_1, B_2, ..., B_n)$
Output: Matched image $X_{i,j}$, $X_{i,j} \in (X_{1,1}, X_{1,2}, ..., X_{m,n})$
1.The region $i$is a part of image A_i,and region j is a part of images B_j,we first calculate the similarity $S_{i,j}$ between region i and region j.
2. According to the Eq. 8, calculating the similarity $Q_{A,B}$ score between A_i and B_j.
3.for $i = 1$ to $i = m$, $j = 1$ to $j = n$
4.**if** $Q_{A_1} > Q_{B_1}$
5.$B_1 = B_2$
6.**repeat** step 2
7.**until when** $Q_{A_1} > Q_{B_k}$, $Q_{B_k} \in max\, Q_B$
8.**return** image $X_{1,j} = A_1$
9.$A_2 = A_1$ and **repeat** step 4
10.**else if** $Q_{A_1} < Q_{B_1}$
11.$A_1 = A_2$
12.**repeat** step 2
13.**until when** $Q_{A_k} < Q_{B_1}$
14.**return** image $X_{i,1} = B_1$
15. $B_2 = B_1$ **repeat** step 10
16.**end**
17.**end**
18.**end**
return $X_{i,j}$,image A_i matched image B_j

verification, all our experimental results are obtained on the same hardware platform.

4.1 Datasets

In order to evaluate the wide applicability of our proposed method, we not only selected the indoor public datasets as the experimental data source, but also selected the outdoor scene with different degrees of change in viewpoint and appearance as the experimental data source for methodology verification. In our work, four popular public datasets are adopted, namely KITTI [33], KTH-IDOL2 [34], Tokyo 24/7 [35] and Tokyo Time Machine, which captures different types of environments and display different changes in appearance and viewpoint. Images on the KITTI were taken by camera mounted on the car, which includes 22 stereo sequences. We select the first 11 sequences (00-10) with ground truth trajectories as the training data, and select the other 11 sequences (11-21) as evaluating data, which have not ground truth. Images on the KTH-IDOL2 are taken in indoor scene, which includes 24 sequences. All of the sequences are taken at a fix frame rate, e.g. 5-fps, under extremely harsh illumination conditions. We also select TOKYO 24/7 as the one of experimental datasets that includes 76 thousand reference images and 315 query images. The images on the TOKYO 24/7 are taken by mobile phone camera, which presents great changes in the aspect of viewpoint and appearance. Our experimental verification is very extensive, including both indoor and outdoor road scenes, as well as street scenes, such as Tokyo time machine, which includes enough daily street view images.

4.2 Evaluation Metrics

The proposed method is evaluated and analysed though making comparisons against to some state-of-the-art methods. The visual evaluation results of performance are shown by recall curve. In this paper, the ones that as compared with our methods not only include existing well-known manual feature methods, e.g. SeqSLAM [36], and but also deep learning feature methods, e.g. VGG-16 and NetVLAD [37]. We evaluate the performance results on the same hardware platform and four common datasets.

4.3 Results and Analysis

In this section, we not only use quantitative analysis to evaluate the performance of the proposed methods, but also use qualitative analysis to visually evaluate the regional recognition ability of several methods in different scenes. The Fig. 6 presents the salient regions from query images corresponding to reference images by using our proposed approach. It can be seen that these highlighted regions are almost these positions with significant characteristics in the scene. According to the Fig. 7a, we can see that the results of performance comparison between four methods that includes our proposed and oher existing advanced and well-known methods, namely, VGG-16, VLADNet and SeqSLAM. It is obvious from the figure that our proposed approach is outperform other approaches, and the SeqSLAM also shows similar performance to our proposed. However, the VGG16 based method does not exhibit a good performance when the appearance changes significantly. In the light of the Fig. 7b, we also see the results of performance of four methods on the KTH-IDOL2. The results show that the performance though using the proposed method has reached an advanced level, and exceeds other three methods. In addition to our method, the best performance of the other three methods is VGG-16 based method, which is almost equal to that of our method on performance. In the Fig. 7c, it is also obvious that the results of performance using our proposed method is almost the same to that of VGG-16 based method, but slightly better than the other two methods. On the Tokyo time machine dataset, according to the Fig. 7d, We can see the performance comparison of our method with the

Fig. 6 Some partial images taken from public datasets, from the first row down, the images are taken from KITTI, KTH-IDOL2, TOKYO 24/7 and Tokyo time machine, which presents some salient regions that are detected by using our proposed approach. The darker the color, the stronger the incentive

other three methods. It is obvious that the best performance is use of VGG-16 based method on scene recognition, however, comparatively speaking, our method also reached the advanced level, slightly inferior to VGG-16 based method and superior to the other two methods. The worst performers are obtained by using SeqSLAM method. We also describe the highest recall of four methods when the precision of recognition reaches 80% in Table 3. The best

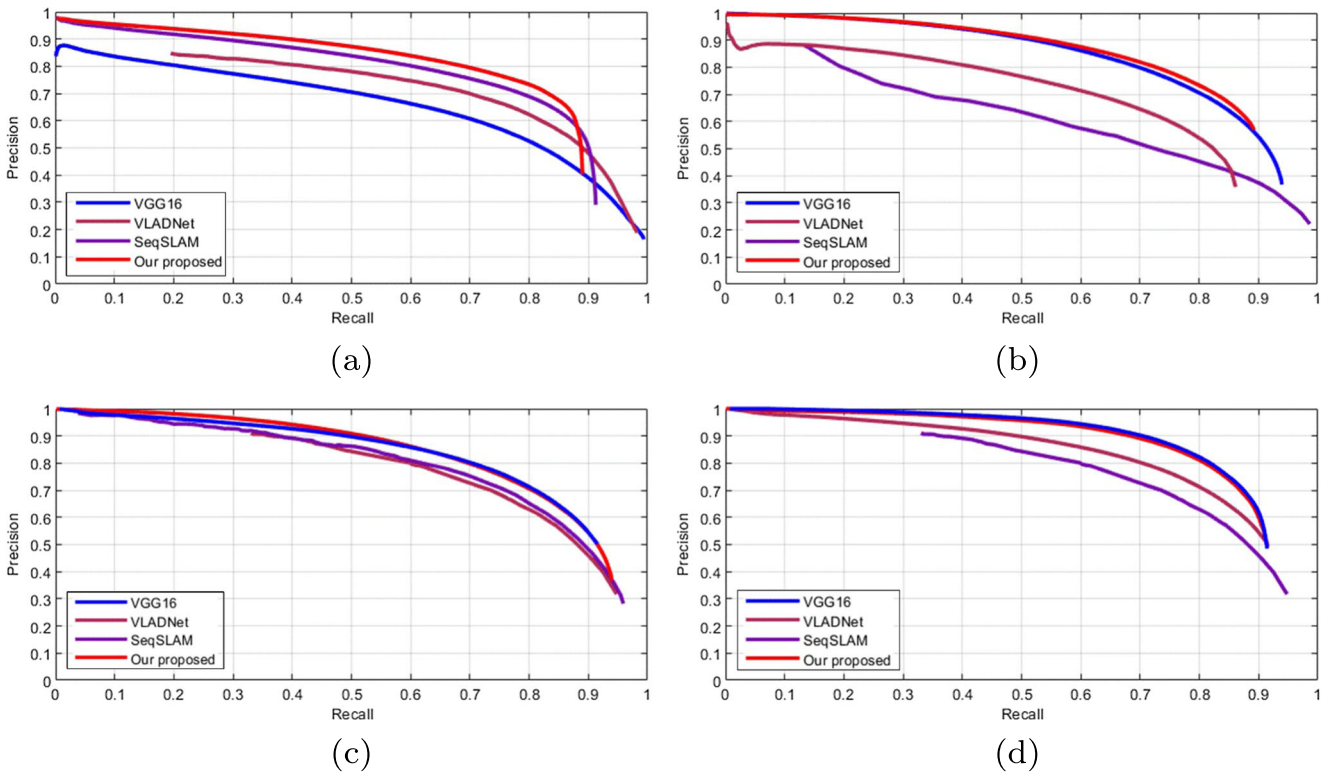

Fig. 7 The performance comparison between our method against state-of-the-art on the **a** KITTI dataset, **b** KTH-IDOL2 dataset, **c** TOKYO 24/7 dataset and **d** Tokyo time machine dataset

Table 3 The accuracy of four different methods with a recall rates of 80%

Dataset	VGG-16	VLADNet	SeqSLAM	Our
KITTI	20.14	42.25	61.87	71.92
KTH-DOL2	69.86	41.92	20.04	73.79
Tokyo 24/7	70.65	59.89	60.96	70.82
Time machine	81.77	70.92	58.67	81.21

results of performance on each dataset are described in red. For KITTI, the recall rate (71.92%) of our proposed approach is obviously well above the method of VGG-16-based (20.14%), VLADNet (42.25%) and SeqSLAM (61.87%). The key is that the appearance and viewpoint of scenes in the KITTI change relatively slight, and the scenes at the same scene has little changes and high similarity, which can give full play to the advantages of handcraft feature methods. In addition, the methods of VGG16 and VLADNet all use local feature to match the image, which has no advantages over SeqSLAM. Although we also adopt the method of local feature, we use end-to-end technology, which strengthens the expressive ability of local features. So our method performs better in the KITTI dataset. For KTH-DOL2, the recall rate (73.79%) of our method is obviously exceed the method of VGG-16-based (69.86%), VLADNet (41.92%) and SeqSLAM (20.04%). The main reason is that SeqSLAM is based on the image sequence for scene searching and matching, the performance is not very stable in the environment where the viewpoint changes greatly. So the method based on CNNs performs better. For Tokyo 24/7, the appearance and viewpoint all changes greatly, which makes all methods confronted with great challenges. However, due to the training flexibility of end-to-end manner and the efficiency of region-based feature representation method, our method performs slightly better than others. For Tokyo Time machine, the recall rate (81.21%) of our method is slightly below that of VGG16 (81.77%), and performs better over VLADNet (70.92%) and SeqSLAM (58.67%). The main reason is that the front-end of our proposed network just takes some modules of VGG16 considering computational cost. However, to greatly improve the performance while reduce computing burden, an end-to-end and region-based salient agglomeration technology are adopted in this paper. In addition, the performance of VLADNet is better than SeqSLAM. The reason is that the former is also adopt end-to-end method, which have the advantages in terms of performance in the environment of changing viewpoint. In order to verify the recognition ability of our proposed method in a wide range of large scenes, referring to [38], we further use the activations from a high-level CNNs layer for performance evaluation on some different scene benchmarks that includes SUN397 [39], MIT Indoor67 [40], Scene15 [41], SUN Attribute [42], Caltech101 [43], Caltech256 [44], Action40 [45] and Event8 [46]. Activations from a higher-level of CNNs, also known as deep features, which has been proven to be an effective universal feature with the most state-of-the-art performance on various image datasets. In our work, we evaluate the recognition performance of the proposed method by comparing with well-known CNNs, namely, MOP-CNN [47], CNNaug-SVN [48], Deep-Optimized [49], Hybrid1369-VGG [50], ImageNet-AlexNet [50], Places-AlexNet [38]. At the same time, in order to verify the superiority of feature self-selection mechanism, we use the deep features extracted by the two modes with self-selection mechanism and without this mechanism to compare with the state-of-the-art CNN features. The difference between the two mechanisms is that the former uses the feature self-selection mechanism proposed by us, which can independently select the scene feature information with salient significance in the convolution process. The latter has not this mechanism and has no feature self-selection ability. Therefore, after the convolution process under this mechanism, there will be a lot of meaningless features in the extracted scene features, which will increase the probability of feature mismatch and have a bad impact on the final scene recognition.In our experiment, the size of training set in each scene is set as 1, 5, 10, 20 and 50, respectively. With the increasing size of training samples, the recognition performance of different CNNs also is improved in different degree, as shown in Fig. 8. Figure 8 plots the recognition accuracy for different visual features on eight different scene datasets. It can be seen that the performance of the proposed feature self-selection mechanism is better than that of method without using this mechanism in the same the skeleton of proposed no matter in which scenes. In addition, by comparing with other state-of-the-art skeletons,we also can see that the performance of our proposed Self-selection method almost better than most others, which fully demonstrates the strong robustness and feature representation of using the self-selection mechanism. In addition, in our work, we also adopt qualitative analysis method to further illustrate and verify the effectiveness of the proposed method in feature detection under challenging scenarios. We adopt the proposed feature self-selection mechanism to detect

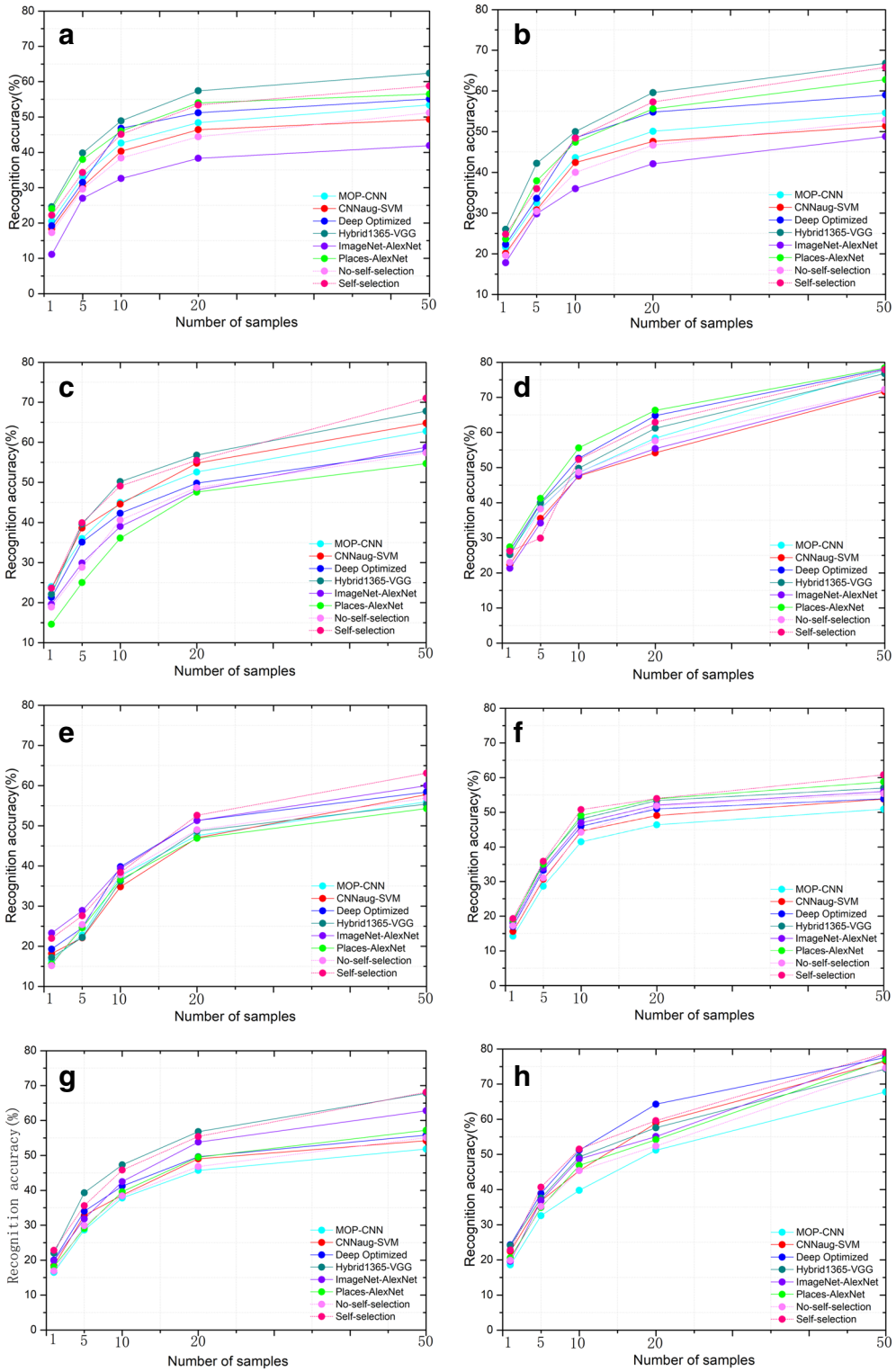

Fig. 8 Comparision of recognition performance with the number of training samples percategory. "No-self-selection" and "self-selection" are all our proposed methods. The different between them is that the former adopts the proposed network architecture without feature self-selection mechanism, while the latter includes it. (a)-(b): SUN397, MIT Indoor67, Scene15, SUN Attribute, Caltech101, Caltech256, Action40 and Event8

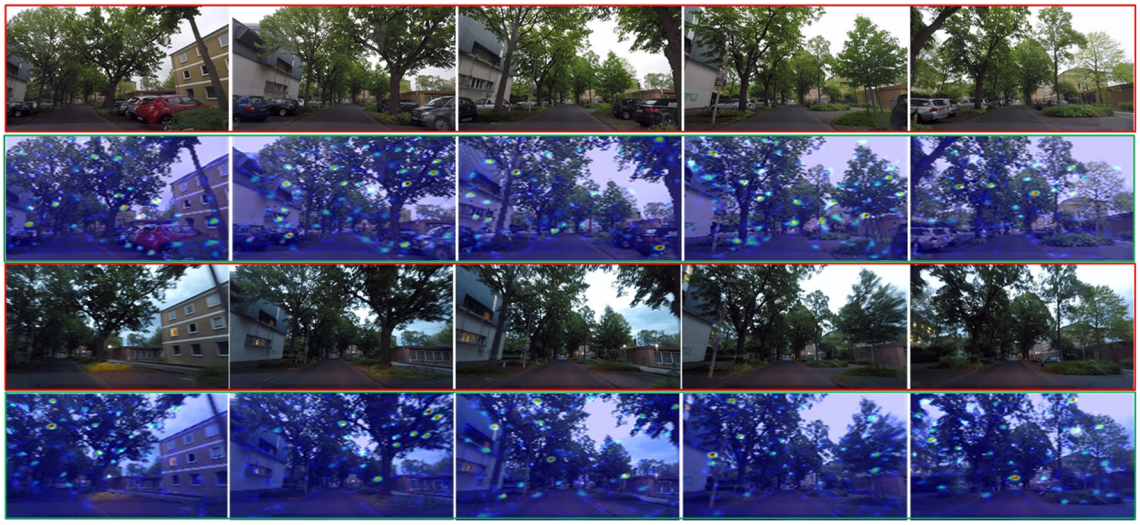

Fig. 9 Qualitative analysis of feature detection.The images are partly taken from VPRICE dataset, From left to right, are continuous road scenes in time. The first and third lines (red rectangle) are scenes collected at different times (day and evening time). The second line and the fourth line (blue rectangle) correspond to the feature detection of day and night scenes respectively

the features of scenes collected in different time periods (day and evening time) at the same location. As we all know, it is very difficult to distinguish some features of certain things with human eyes when the night falls. Therefore, it is quite challenging to detect features of relatively blurry scenes with the proposed method. The detection results are shown in Fig. 9, which shows that the position and quantity of the detected features are basically consistent with the daytime detection effect, even in the evening. We also adopt quantitative analysis to demonstrate the effectiveness of various mainstream feature extraction methods on this challenging dataset. The metric represents the relation between matched rate and precision, which can be represented as "$1 - precision = \frac{\#false-matches}{\#correct-matches+\#false-matches}$", The results as shown in Fig. 10 that shows the results for the seven matching strategies based on seven different feature extraction methods. It can be seen that the feature extraction methods based on CNNs, e.g. our proposed, VGG16-places365 and GoogleNet-places365, is better than that based on manual feature, e.g. SURF, SIFT and ORB. As the same time, comparison with method based on free self-selection mechanism, the proposed method (self-selection mechanism) has a better perfomance on matching rate. To sum up, our method also has certain robustness in relatively extreme environment, which can basically meet the detection requirements of unmanned system in some scenes in the evening.

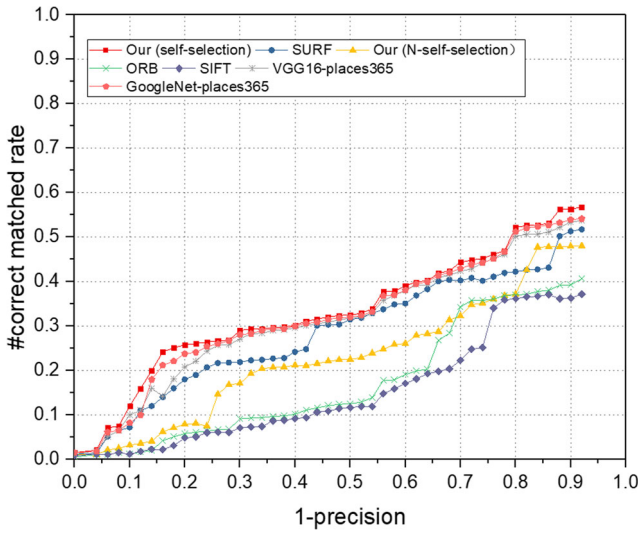

Fig. 10 Comparison of different feature matching strategies on VPRICE dataset. The "Our(self-selection)" represents feature extraction method based proposed self-selection mechanism and the "Our(N-self-selection)" represents feature extraction method that has not adopt self-selection mechanism

5 Conclusion and Future Work

5.1 Conclusions

We propose a novel approach relied on a self-selection slight-weight CNN for scene recognition, as well as evaluate the performance curve based on comparing some different methods (between our proposed method and three other existing advanced and well-known methods) on four popular scene datasets. From the performance on datasets, the method of SeqSALM performs well on datasets with no sever changes both appearance and viewpoint, but not on sever changes in appearance or viewpoint. On the contrary, the method of VGG-16 performs well on datasets

with changing in appearance and viewpoint, but not on datasets with no changes in appearance or viewpoint. The method of VLADNet performs mediocre either on datasets with changing in appearance and viewpoint, and or with no changes in appearance and viewpoint. In addition, we also evaluate the role of adopted feature self-selection on some scene benchmarks, by comparing with state-of-the-art skeletons, we have verified that the feature self-selection mechanism has better feature clustering ability. In all, our method performs well under both severe appearance and viewpoint changes, which shows advanced level. The proposed method is able to effectively detect salient features when dealing with local small scenes. However, it is difficult to deal with large scene object detection. At present, Places-CNNs has better performance for scene classification on large scene datasets, e.g. Places365-Challenge. In addition, most of the existing scene recognition methods, including proposed by us, do not consider the depth information, which will cause scale error to some unmanned systems in the case of localization or navigation to a certain extent. The relevant solution is to obtain the depth scene information by loading LIDAR, but this will cause a serious increase in cost. In order to deal with this problem, some researchers try to combine convolution and recurrent neural network to extract RGB features with depth. To sum up, for scene recognition, the existing methods basically have advantages and disadvantages, researchers need to spend more time and energy to continue to improve their respective methods.

5.2 Future Work

In the future, we will focus on the application of deep learning technology on mobile augmented reality (AR). The AR is able to create a real learning situation and a strong sense of immersion, which is conducive to enhance the learning experience and stimulate learning motivation. However, the existing mobile outdoor augmented reality applications can only reconstruct the 3D scene of a single and simple scene, which can not meet the needs of unmanned systems, e.g. robots, intelligent vehicles and UAVs, running in more complex and real-time scenes. Therefore, the development of a more lightweight mobile outdoor augmented reality method, combined with deep learning and knowledge modeling to perceive the learning scene, in order to improve the learning experience, has become the future development direction of AR application in life.

Acknowledgments The authors would like to thank the anonymous referee for the helpful comments. The research was supported by the National Key Research and Development Program of China (no. 2016YFB0100902).

Author Contributions Conceptualization: Zhenyu Li and Aiguo Zhou; Methodology: Zhenyu Li and Aiguo Zhou; Formal analysis and investigation: Aiguo Zhou; Writing - original draft preparation: Zhenyu Li; Writing - review and editing: Zhenyu Li and Aiguo Zhou; Funding acquisition: Aiguo Zhou;

Funding The research was supported by the National Key Research and Development Program of China (no. 2016YFB0100902).

Data Availability Not applicable.

Competing interests The authors have no conflicts of interest to declare that are relevant to the content of this article.

References

1. Cummins, M., Newman, P.: Appearance-only SLAM at large scale with FAB-MAP 2.0. Int. J. Robot. Res. **30**, 1100–1123 (2011)
2. Ng, P.C., Henikoff, S.: SIFT Predicting amino acid changes that affect protein function. Nucleic Acids Res. **31**, 3812–3814 (2003)
3. Bay, H., Ess, A., Tuytelaars, T., Van Gool, L.: Speeded-up robust features (SURF). Comput. Vis. Image Understand. **110**, 346–359 (2008)
4. Oishi, S., Inoue, Y., Miura, J., et al.: SeqSLAM++: View-based robot localization and navigation. Robot. Auton. Syst. **112**, 13–21 (2019)
5. Gálvez-López, D., Tardos, J.D.: Bags of binary words for fast place recognition in image sequences. IEEE Trans. Robot. **28**, 1188–1197 (2012)
6. Knopp, J., Sivic, J., Pajdla, T.: Avoiding confusing features in place recognition. In: European Conference on Computer Vision, Berlin, Heidelberg, pp. 748–761 (2010)
7. Sünderhauf, N., Dayoub, F., Shirazi, S., et al.: On the performance of convnet features for place recognition. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Barcelona, Spain, pp. 4297–4304 (2015)
8. Chen, Z., Lam, O., Jacobson, A.: Convolutional neural network-based place recognition. Comput. Sci. (2014)
9. Arroyo, R., Alcantarilla, P.F., Bergasa, L.M.: Fusion and Binarization of CNN features for robust topological localization across seasons. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Daejeon, Korea, pp. 4656–4663 (2016)
10. Chen, Z., Jacobson, A., Sunderhauf, N., et al.: Deep learning features at scale for visual place recognition. In: IEEE International Conference on Robotics and Automation (ICRA), Marina Bay Sands, Singapore, pp. 3223–3230 (2017)
11. Fang, Y., Yan, J., Li, L., et al.: No reference quality assessment for screen content images with both local and global feature representation. IEEE Trans. Image Process. **27**, 1600–1610 (2017)
12. Stanchev, P.L., Green, D. Jr., Dimitrov, B.: High level colour similarity retrieval. Int. J. Inf. Theor. Appl. **10**, 363–369 (2003)
13. Islam, M.M., Zhang, D., Lu, G.: A geometric method to compute directionality features for texture images. In: Proc. ICME, pp. 1521–1524 (2008)
14. Zhang, D., Lu, G.: Review of shape representation and description techniques. Pattern Recognit. **37**, 1–19 (2004)

15. Zhang, X., Wang, L., Zhao, Y., et al.: Graph-based place recognition in image sequences with CNN features. J. Intell. Robot. Syst. **95**, 389–403 (2019)
16. Arandjelovic, R., Gronat, P., Torii, A., et al.: NetVLAD: CNN architecture for weakly supervised place recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5297–5307 (2016)
17. Chen, B., Li, J., Wei, G., et al.: M-SAC-VLADNet: a multi-path deep feature coding model for visual classification. Entropy **20**, 341 (2018)
18. Fan, R., Shuai, H., Liu, Q.: PointNet-Based channel attention VLAD network. In: Chinese Conference on Pattern Recognition and Computer Vision (PRCV), pp. 320–331 (2019)
19. Gomez-Ojeda, R., Lopez-Antequera, M., Petkov, N., et al.: Training a convolutional neural network for appearance-invariant place recognition. arXiv:1505.07428 (2015)
20. Quan, Y., Li, Z.: Zhang F.others. DNet-65 R-CNN: Object detection model fusing deep dilated convolutions and light-weight networks. In: Pacific Rim International Conference on Artificial Intelligence., pp. 16–28 (2019)
21. Park, C., Jang, J., Zhang, L., et al.: Light-weight visual place recognition using convolutional neural network for mobile robots. In: 2018 IEEE International Conference on Consumer Electronics (ICCE), Las Vegas, NV, USA, pp. 1–4 (2018)
22. Kim, J., Yoon, S.E.: Regional attention based deep feature for image retrieval. In: Proc. British Machine Vision Conference (BMVC), Newcastle, England (2018)
23. Lankton, S., Tannenbaum, A.: Localizing region-based active contours. IEEE Trans. Image Process. **17**, 2029–2039 (2008)
24. Carson, C., Thomas, M., Belongie, S., et al.: Blobworld: A system for region-based image indexing and retrieval. In: International Conference on Advances in Visual Information Systems, Berlin, Germany, pp. 509–517 (1999)
25. Dai, J., Li, Y., He, K., et al.: R-fcn: Object detection via region-based fully convolutional networks. In: Advances in Neural Information Processing Systems, pp. 379–387 (2016)
26. Khaliq, A., Ehsan, S., Milford, M., et al.: CAMAL: Context-Aware Multi-Scale Attention framework for Lightweight Visual Place Recognition. arXiv:1909.08153 (2019)
27. Khaliq, A., Ehsan, S., Chen, Z., et al.: A Holistic Visual Scene Recognition Approach using Lightweight CNNs for Severe ViewPoint and Appearance Changes. arXiv:1811.03032 (2018)
28. Li, Z., Zhou, A., Wang, M., et al.: Deep fusion of multi-layers salient CNN features and similarity network for robust visual place recognition. In: IEEE International Conference on Robotics and Biomimetics (ROBIO), Dali, China, pp. 22–29 (2019)
29. Li, Z., Zhou, A., Shen, Y.: An end-to-end trainable multi-column CNN for scene recognition in extremely changing environment. Sensors **20**, 1556 (2020)
30. Wan, L., Zeiler, M., Zhang, S., et al.: Regularization of neural networks using dropconnect. In: International Conference on Machine Learning, pp. 1058–1066 (2013)
31. Chen, Z., Maffra, F., Sa, I., et al.: Only look once, mining distinctive landmarks from convnet for visual place recognition. In: 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Vancouver, Canada, pp. 9–16 (2017)
32. Azizpour, H., Sharif Razavian, A., Sullivan, J., et al.: From generic to specific deep representations for visual recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Boston, USA, pp. 36–45 (2015)
33. Geiger, A., Lenz, P., Stiller, C., et al.: Vision meets robotics: The KITTI dataset. Int. J. Robot. Res. **32**, 1231–1237 (2013)
34. Luo, J., Pronobis, A., Caputo, B., Jensfelt, P.: The kth-idol2 database. KTH, CAS/CVAP, Tech Rep. 304 (2006)
35. Torii, A., Arandjelovic, R., Sivic, J., et al.: 24/7 place recognition by view synthesis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, USA, vol. 8-10, pp. 1808–1817 (2015)
36. Milford, M.J., Wyeth, G.F.: SeqSLAM: Visual route-based navigation for sunny summer days and stormy winter nights. In: 2012 IEEE International Conference on Robotics and Automation, St. Paul, MN, USA, vol. 14-18, pp. 1643–1649 (2012)
37. Arandjelovic, R., Gronat, P., Torii, A., et al.: NetVLAD: CNN architecture for weakly supervised place recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, Nevada, pp. 5297–5307 (2016)
38. Zhou, B., Lapedriza, A., Khosla, A.: Places: A 10 million image database for scene recognition. IEEE Trans. Pattern Anal. Mach. Intell. **40**, 1452–1464 (2017)
39. Xiao J., Hays J., Ehinger, KA., et al.: Sun database: Large-scale scene recognition from abbey to zoo. In: Proc. CVPR (2010)
40. Quattoni, A., Torralba, A.: Recognizing indoor scenes. In: Proc. CVPR (2009)
41. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: Proc. CVPR (2006)
42. Patterson, G., Hays, J.: Sun attribute database: Discovering, annotating, and recognizing scene attributes. In: Proc. CVPR (2012)
43. Fei-Fei, L., Fergus, R., Perona, P.: Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. Comput. Vis. Image Understand. (2007)
44. Griffin, G., Holub, A., Perona, P.: Caltec 256 object category dataset (2007)
45. Yao, B., Jiang, X., Khosla, A., et al.: Human action recognition by learning bases of action attributes and parts. In: Proc. ICCV (2011)
46. Li, L.J., Fei-Fei, L.: What, where and who? classifying events by scene and object recognition. In: Proc. ICCV (2007)
47. Gong, Y.C., Wang, L.W., Guo, R.Q.: Multi-scale orderless pooling of deep convolutional activation features (2014)
48. Razavian, A.S., Azizpour, H., Sullivan, J.S., et al.: CNN features off-the-shelf: an astounding baseline for recognition. In: IEEE Conference on Computer Vision and Pattern Recognition Workshop, pp. 806-813. IEEE, Columbus (2014)
49. Azizpour, H., Razavian, A.S., Sullivan, J., et al.: From generic to specific deep representations for visual recognition. In: Conference on Computer Vision and Pattern Recognition Workshop, pp. 36–45. IEEE, Boston (2015)
50. Zhou, B., Garcia, A.L., Xiao, J., et al.: Learning deep features for scene recognition using places database. In: Advances in Neural Information Processing Systems, NIPS, Montréal, Quebec, Canada, pp. 487-495 (2015)

Zhenyu Li is currently a Ph.D. candidate at Tongji University. He obtained the M.S. degree in Mechanical Engineering from Shandong University of Technology, Zibo China in 2018. He has over 20 publications including conference and journal papers. His research topics mainly cover deep learning, scene perception and autonomous localization and navigation for autonomous driving.

Aiguo Zhou received his Ph.D. degree from Shanghai Jiaotong University of Mechatronics Engineering in 2004. He is presently an associate professor of Tongji University, Shanghai China. In 2010, he was a Visiting Scholar at Okland University, Michigan, USA. He has over 30 publications including conference and journal papers. His research interests span the area of smart sensor, intelligent vehicle with a focus on system design and control strategy.