The Research Status of Visual Place Recognition

Bo Wang, Xin-sheng Wu, An Chen*, Chun-yu Chen and Hai-ming Liu
School of Automation Science and Engineering, South China University of Technology, Guangzhou, 510640

*Corresponding author’s Email: chenan@scut.edu.cn

Abstract. Visual place recognition is a challenging problem in the field of Machine Vision and Robotics. Unlike image classification and retrieval, Machine Vision lags far behind humans in place recognition. Generating image descriptors is the basic problem of VPR. They suppose to be insensitive to changes in illumination and angle of view, and can ensure the stability in the long running process. This paper summarizes the solutions of visual place recognition from two directions: traditional methods and deep learning methods. And explains the key technologies and analyzes the advantages and disadvantages of different methods also the implementation difficulties. In particular, the latest research results are introduced: Using image descriptors generated by semantic segmentation algorithms to solve VPR problems can obtain better performance. Finally, the possibility of using semantic segmentation image descriptors combined with landmark topological relations to solve VPR problems is prospected.

1. Introduction

With the rapid development of unmanned driving technology, VSLAM technology gain more attention from academia and industry. Compared with laser SLAM, visual SLAM has the advantages of stronger perception and replace ability, diversified installation methods and lower cost [1]. VSLAM technology obtains the motion path through the front-end visual odometer and the back-end nonlinear optimization step. However, because the visual odometer is an iterative estimation, the error is inevitably accumulated and amplified after a long estimation. The result is unreliable. To solve the error accumulation in the long running process, we need the closed-loop detection algorithm [2]. A true positive closed loop can significantly reduce the cumulative error of the system.

How to judge whether two images represent the same position during the operation of unmanned vehicle? That in itself is an open question. Some foreign scholars try to model the place and propose a general framework for identifying the place, namely Visual place recognition (VPR). Visual place recognition is similar to image classification in principle. Image classification classifies the image to be searched as one of a limited number of known categories [3]. The current excellent solution is the CNN scheme [4]. However, there are problems in applying image classification algorithm to spot recognition directly. The two are quite different: place recognition looks for an image that is consistent with the current view, and the view usually contains many classes of objects [5], and the light and Angle may change (as shown in figure 1), as well as moving objects that interfere with the recognition.
In general, place recognition needs to address the following issues. Firstly, features of places may change, how to abstract stable features over time. For instance, image descriptors should ensure that when judging different weather conditions but the same locations, the differences are small, but there should be large differences when judging different locations but with similar weather environments. Secondly, how to balance the trade-off between viewpoint and appearance invariance. Finally, the map scale of mobile robot in a long time state is huge, the process of comparing high-dimensional features places great emphasis on computing power, so how to perform accurate place recognition for long time running [6]. All these factors make the problem of visual place recognition still be challenging in the field of Machine Vision and robot applications.

2. Traditional visual place recognition problem solutions
The traditional scheme mainly uses the hand-designed image feature descriptor, which is highly subjective. There are two kinds of descriptor based on local feature and global feature.

The method based on local feature description is widely used on mobile robot platform. That is, the description of Visual word Bag model (bag-of-visual Words (BOVW)). The BOVW algorithm firstly extracts the SIFT features from the image frames, and then constructs the Visual word list through k-means clustering on the features. Then, for each image, the extracted features are represented by words, and the frequency of words is counted to obtain the image description vector [7]. The problems existing in the SIFT algorithm is very time-consuming and later developed algorithms such as SURF (speeded up robust feature) algorithm, the ORB (oriented FAST and rotated BRIEF) algorithm are mostly at the expense of the performance to improve efficiency. Since visual vocabulary is pretrained, the ability to generalize in different situations is a very important problem. Nicosevici et al. [8] proposed a method to update vocabulary online based on observation information, which did not require pre-training of visual vocabulary and thus acquired the ability to adapt to the environment. The global feature descriptor is represented by Gist feature extraction. The algorithm divides the image into several grids, and extracts the texture information of the image by Gabor filtering in different directions and scales, finally obtains the global feature descriptor. Krose [9] directly used the principal component analysis method to generate linear image features for the whole image, and directly generated descriptors based on the linear image features.

This chapter focuses on traditional place recognition solutions. Local feature descriptor has certain reliability for perspective change and illumination change, but cannot guarantee good robustness in the face of big changes of external scene and interference of key points. Global feature descriptor is not robust enough for light changes and wide range of view changes.

3. Deep learning-based solutions
In recent years, deep learning methods has been widely used in visual place recognition. It include object recognition schemes and semantic segmentation schemes. For large number of data sets are used for training, both have achieved better results than manually designing image feature descriptors [10].

3.1. Object detection based scheme
The place recognition scheme based on object detection surrounds the scene image with rectangular frames of different sizes. When it is identified that the candidate frame contains an object of known classification, the feature information of the frame is extracted. The CNNs scheme itself can better recognize various objects [11]. Meanwhile, it is not sensitive to changes in shooting angle [12].

Since there are many research results on the application of convolutional neural network (CNN) in
the direction of object recognition. Which is shown in figure 2(a): first they construct an offline database, then perform target recognition, and extract feature vectors through CNN to form a place feature database. In the process of online place recognition, the camera collects image data, and performs target recognition in the same way, extracts features from the identified target candidate frames. Calculate the similarity between them to find the top-rank candidate.

![Flow chart of VPR scheme](image)

(a) based on object recognition (b) based on semantic segmentation

**Figure 2.** Flow chart of VPR scheme

For instance, Hou et al. [13] used AlexNet pre-trained CNN model for object recognition. Obtain the target area of interest (Landmark) in the scene, and then generate feature vectors. It can achieve better results than BoVW in the case of image lighting changes, and the recognition accuracy is higher. However, the disadvantages are the higher feature dimension it obtained. Sünderhauf [14] and others also proposed the idea of landmark proposal. He used the Edge Boxes object proposal algorithm to generate candidate landmarks. For candidate landmarks, a neural network is used to extract feature vectors, and the convolution features are subjected to dimensionality reduction to compare similarities. The method proves that it has better robustness to appearance changes and viewing angle changes. Sünderhauf [15] et al. Studied the value of each layer of the AlexNet network as an image feature representation, seeing that the third convolutional layer of the convolutional network is robust to changes in appearance. The first fully connected layer is robust to changes in viewing angle. Gao [16] uses a stacked autoencoder to build a closed-loop detection model, improves the network structure, removes the decoding layer, and uses the output of the hidden layer as a feature extractor. Construct the similarity matrix to use in recognition. However, the stability is poor in a wide scene changes.

In recent years, with the rapid development of target detection technology, VPR schemes have also been enriched. Lopez-Antequera [6] and other training CNNs extract features and embed images in low-dimensional space. It was suitable for long-running embedded devices. The disadvantage is that the network generalization performance is not high. Nate [17] proposed an unsupervised deep neural network architecture. Bao [18] pointed out that the mid-level features of CNN contain more geometric spatial information. High-level features contain more semantic information and are more resistant to image perspective changes. The author fuses the extracted features and performs dimensionality reduction. It improved recognition accuracy. But, the filtering of interference terms and the extraction of key frame features take a long time. Zhang [19] used YOLOv3 to obtain scene object information. At the time of detection, a similarity matrix is constructed according to the target category and the SURF feature for similarity calculation. The indoor data set verification has obtained better results. The disadvantage is that the calculation amount is large, and poor light adaptation. Hong [20] and others used the TextBoxes ++ algorithm [21] to detect the text of the target scene, and the position was represented by the text recognition result. The proposed method is able to cope with partial lighting changes, perceptual aliasing, dynamic occlusion and viewpoint changes. Luis [22] proposed a visual place recognition scheme that fuses object features and spatial information. The VGG16 network was used to identify the object and verify the consistency of the points around the object's feature points, added the judgment of the positional relationship of objects in space. The recognition accuracy is improved, which is helpful for solving the problem of perceived aliasing.
3.2. Scheme based on semantic segmentation

Semantic segmentation technology based on deep learning has developed rapidly in recent years. The performance has far exceeded the traditional semantic segmentation method.

Semantic segmentation technology explores how to label image information at the semantic level [23]. Among them, Full Convolutional Network (FCN) [24] truly realizes the semantic segmentation mapping from pixel to pixel. DeepLab [25] optimized the segmentation results to get better edges. The model proposed by PSPNet [26] performs better on the recognition of small objects and distant objects. RefineNet [27] solves the problem of reduced spatial resolution in traditional convolutional networks and can generate high-resolution segmentation maps.

Place recognition schemes based on semantic segmentation are closer to human cognitive methods. Humans often do not identify the rectangular frame where the object is located and then make recognition judgments when identifying the place. Instead, the shape features of the object are directly obtained and used for place comparison [28]. Semantic segmentation technology can perform pixel-level segmentation on the image to obtain the classification result of each pixel [29]. This allows the edge information of the image to be preserved. Maximize the true state of the scene target. For example, Xin Z [30] and others designed a significant landmark recognition network (LLN) to obtain a significant map. The higher the degree of activation of a pixel region, the more representative the region is. The recognition effect has been improved. This method emphasizes the important role of landmark outline information. In general, as shown in figure 2(b), existing Semantic segmentation solutions mainly segment the image to obtain different target regions, select targets of interest, and use the combination of their pixel classification regions as place descriptors. When performing recognition, it is also necessary to compare similarities to get a high ranking Candidate results.

Niko Sünderhauf [31] emphasizes the edge information of objects rather than the objects within the rectangular frame, which plays an important role in navigation. This paper performs target detection on key frames and uses the improved depth image recovery algorithm [32] to restore the true boundary of the object. A semantic map with pixel-level information is obtained from the associated object recognition results and object boundary information. Used for loop detection to improve recognition accuracy. But limited by depth information must be known. Ma Lingni [33] used the FuseNet network to perform semantic segmentation of RGB images and depth images, and obtained the pixel-level semantic segmentation results by combining the features of RGB and depth images. This method can improve the effect of semantic segmentation. He expects that this method can be applied to attitude tracking and SLAM closed-loop detection. But the limit is the acquisition of depth information.

Naseer T [34] and others have a relatively new approach to the problem of place recognition. They use Fast-net network [35] to learn the regions of interest in the image. The feature vectors of the region of interest obtained by the semantic segmentation network are accumulated and normalized to obtain landmark descriptors. The similarity is calculated by a cosine operation. This method has stable expression under harsh environmental conditions, shows that the semantic enhancement features of image regions can significantly improve the place recognition scheme based on object recognition. The problem studied by Sourav G[36] and others is the problem of visual place identification including opposite perspectives. The author first constructs full-image semantic feature descriptors and uses a combination of information from different semantic categories (roads, buildings, vegetation, etc.) to form feature vectors. The key points are extracted from the convolutional feature map of the pre-trained network to determine whether the key points correspond to further verification. Realize the matching from coarse to fine. That paper enhances the scoring of local images. Improved perceived aliasing caused by appearance changes.

We analyze the performance information of existing deep learning schemes, such as the time consumption of a single picture, the difficulty of model deployment, debugging, and the legibility of the code. These information have important reference significance for academic research and deployment applications, then compared with the traditional descriptor algorithm based on manual design features. The performance comparison of each method is shown in the table 1:
This chapter introduces solutions to VPR problems based on deep learning algorithms. Among them, the solution based on the object recognition candidate frame is relatively mature and perfect. The VPR scheme combined with the semantic segmentation algorithm is more novel and achieves better results than object recognition.

| Traditional handcraft descriptors | Deep learning descriptors |
|----------------------------------|---------------------------|
| speed                            | ★★                        |
| Perspective invariance           | ★★                        |
| Light invariance                 | ★★                        |
| Recognition rate                 | ★★★                       |
| Model deployment difficulty      | ★★★                       |
| Code legibility                  | ★★★                       |

This chapter introduces solutions to VPR problems based on deep learning algorithms. Among them, the solution based on the object recognition candidate frame is relatively mature and perfect. The VPR scheme combined with the semantic segmentation algorithm is more novel and achieves better results than object recognition.

4. Future direction

In the future, one of the development directions of VPR problem solutions is to adopt faster and better semantic segmentation networks. Existing networks mainly identify and segment close-range pictures, and the problem of place recognition also requires the identification and segmentation of objects at greater distances. If the semantic segmentation network also has a good effect on pixel segmentation of distant objects, and the constructed image descriptor can contain distant buildings or signs, then the performance of the place recognition scheme will be further improved.

Furthermore, existing solutions do not consider the role of relative positional relationships of objects in the scene for recognition. Existing algorithms require that the descriptors have a high spatial dimension when ensuring the recognition accuracy. Therefore, another development direction of visual place recognition is how to express sparse features of the image so that it can run on embedded platforms. We envisage that the future technical solution is to construct a place descriptor combined with the topological relationship to realize identification. The image to be identified is semantically segmented to obtain different target regions, and then the target of interest is selected, and their pixel classification regions are recorded as $x_1, x_2, ..., x_n$, respectively. Calculate the vectors of the first pixel area of each landmark of interest relative to the upper left corner of the image (position of pixel point (0,0)). Then make a difference between $x_i$ and $x_j$ to get the relative position $x_{ij}$. The feature vector of the pixel region and the position reference vector $x_{ij}$ of the region are combined into new place description vectors $v_i$ and $v_j$. This place descriptor contains both the characteristic information of the landmark and the spatial location information of the landmark.

Finally, we analyze the process of human recognition of the scene and find that humans do not rely on a certain frame of scene photos to make judgments. Instead, the place was modeled 360°. So another direction to solve the VPR problem in the future is omnidirectional semantic topological image feature descriptors. That is, pictures of different positions (such as front, back, right, and left directions) of the current position are obtained through the shooting posture of the camera. Semantic segmentation is performed on the pictures from these perspectives respectively. On the premise that the camera pose angle is known, the spatial position relationship of the landmarks in the current scene is fused, and the topological semantic descriptor of the surround view can be obtained through combination. This omnidirectional semantic descriptor can have better fault tolerance.

5. Conclusion

This paper summarizes the research status of visual place recognition. summarizes the descriptors based on traditional image feature and deep learning, and focuses on the latest image feature descriptors based on semantic segmentation. We also look forward to several possible directions for future research. However, place recognition still faces some problems. Further development in related fields such as deep learning, semantic segmentation, and topological positioning is needed. The
development of sensors and multi-source information fusion will also help improve the robustness of place recognition.

This work is supported by the Science and Technology Planning Project of Guangdong Province (Grant No. 2017A040405025).

References
[1] Liu Qiang, Duan Fuhai, Sang Yong, Zhao Jianlong. Overview of closed-loop detection methods for visual SLAM in complex environments [J]. Robot, 2019, 41 (01): 112-123 + 136.
[2] Xiaoyu Li. Research on loop detection algorithm based on image information [D]. Jilin University, 2019.
[3] Xin Z, Cai Y, Lu T, et al. Localizing Discriminative Visual Landmarks for Place Recognition[J]. arXiv preprint arXiv:1904.06635, 2019.
[4] Jiao Pengfei. Research on image classification and retrieval method based on deep learning and sparse representation [D]. Xi'an University of Technology, 2019.
[5] Huang Kaiqi, Ren Weiqiang, Tan Tieniu. Overview of Image Object Classification and Detection Algorithms [J]. Chinese Journal of Computers, 2014, 37 (06): 1225-1240.
[6] Xin Z, Cui X, Zhang J, et al. Real-time visual place recognition based on analyzing distribution of multi-scale cnn landmarks[J]. Journal of Intelligent & Robotic Systems, 2019, 94(3-4): 777-792.
[7] Kato H, Harada T. Image reconstruction from bag-of-visual-words[C]/Proceedings of the IEEE conference on computer vision and pattern recognition. 2014: 955-962.
[8] Nicosevici T, Garcia R. Automatic visual bag-of-words for on-line robot navigation and mapping[J]. IEEE Transactions on Robotics, 2012, 28(4): 886-898.
[9] Kr ose B J A, Vlassis N, Bunschoten R, et al. A probabilistic model for appearance-based robot localization[J]. Image & Vision Computing, 2001, 19(6): 381-391.
[10] Qiu Chenli, Huang Dongzhen, Liu Huawei, Yuan Xiaobing, Li Baqing. Closed-loop detection algorithm combining Gist features and convolutional self-encoding [J / OL]. Progress in Laser and Optoelectronics: 1-15 [2019-11-05].
[11] Zhang Dongdong. Overview of image recognition in artificial neural network [J] .Electronic World, 2018 (09): 97.
[12] Lopez-Antequera M, Gomez-Ojeda R, Petkov N, et al. Appearance-invariant place recognition by discriminatively training a convolutional neural network[J]. Pattern Recognition Letters, 2017, 92: 89-9
[13] Hou Y, Zhang H, Zhou S. Convolutional neural network-based image representation for visual loop closure detection[C]. IEEE International Conference on Information & Automation, 2015: 2238-2245.
[14] Sünderrauf N, Shirazi S, Jacobson A, et al. Place recognition with convnet landmarks: Viewpoint-robust, condition-robust, training-free[J]. Proceedings of Robotics: Science and Systems XII, 2015.
[15] N. Sünderrauf, S. Shirazi, F. Dayoub, B. Upcroft, M. Milford, N. Sünderrauf,F. Dayoub, S. Sareh, U. Ben, M. Michael, On the performance of ConvNet features for place recognition, in: International Conference on Intelligent Robots and Systems (IROS), 2015.
[16] Xiang Gao. Loop Closure Detection for Visual SLAM Systems Using Deep Neural Networks [A]. Professional Committee of Control Theory, Chinese Society of Automation. Proceedings of the 34th China Control Conference (Volume D) [C]. Chinese Society of Automation Control Theory Committee: Control Theory Committee, 2015: 6.
[17] Merril N, Huang G. Lightweight unsupervised deep loop closure[J]. arXiv preprint arXiv:1805.07703, 2018.
[18] Bao Zhenqiang, Li Aihua, Cui Zhigao, Su Yanzhao, Zheng Yong. Closed-loop Detection Algorithm Fusion of Multi-level Convolutional Neural Network Features [J]. Progress in Laser and Optoelectronics, 2018, 55 (11): 375-381.
[19] Zhang Kuojia, Zhang Yunzhou, Lu Guanghao, Gong Yiqun. Vision SLAM Closed-Loop Detection Based on Local Semantic Topology Graph [J]. Robot, 2019, 41 (05): 649-659.

[20] Hong Z, Petillot Y, Lane D, et al. TextPlace: Visual Place Recognition and Topological Localization Through Reading Scene Texts[C]/Proceedings of the IEEE International Conference on Computer Vision. 2019: 2861-2870.

[21] Liao M, Shi B, Bai X. Textboxes++: A single-shot oriented scene text detector[J]. IEEE transactions on image processing, 2018, 27(8): 3676-3690.

[22] Camara L G, Přeučil L. Spatio-Semantic ConvNet-Based Visual Place Recognition[C]//2019 European Conference on Mobile Robots (ECMR). IEEE, 2019: 1-8.

[23] Zhang Xinming, Zhu Xiaobin, Cai Qiang, Liu Xinliang, Shao Wei, Wang Lei. A Survey of Deep Learning Models for Image Semantic Segmentation [J]. High Technology Letters, 2017, 27 (Z1): 808-815.

[24] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 3431-3440.

[25] Chen L C, Papandreou G, Kokkinos I, et al. Deep Lab:semantic image segmentation with deep convolutionalnets, atrous convolution, and fully connected CRFs [ J ]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017, PP( 99) : 1-1.

[26] Zhao H, Shi J, Qi X, et al. Pyramid scene parsing net-work [ J ]. arXiv preprint arXiv , 2016: 1612. 01105.

[27] Lin G, Milan A, Shen C, et al. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 1925-1934.

[28] Sudderhauf N, Newman P, et al. Visual place recognition: A survey[J]. IEEE Transactions on Robotics, 2015, 32(1): 1-19.

[29] Zhang Wei. SLAM Research of Indoor Vision Based on Object Semantic Information [D]. Shandong University, 2019.

[30] Xin Z, Cai Y, Lu T, et al. Localizing Discriminative Visual Landmarks for Place Recognition[J]. arXiv preprint arXiv:1904.06635, 2019.

[31] Sudderhauf N, Pham T T, Latif Y, et al. Meaningful maps with object-oriented semantic mapping [C]// Proc of IEEE/RSJ International Conference on Intelligent Robots and Systems. [S.1]:IEEE Press, 2017: 5079-5085.

[32] Sudderhauf N, Pham T T, Latif Y, et al. Meaningful maps with object-oriented semantic mapping [C]// Proc of IEEE/RSJ International Conference on Intelligent Robots and Systems. [S.1]:IEEE Press, 2017: 5079-5085.

[33] Pham T T, Eich M, Reid I, et al. Geometrically consistent plane extraction for dense indoor 3D maps segmentation[C]//2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2016: 4199-4204.

[34] Ma Lingni, Stü ckerl J, Keri C, et al. Multi-view deep learning for consistent semantic mapping with rgb-d cameras [C]// Proc of IEEE/RSJ International Conference on Intelligent Robots and Systems .[S.1]:IEEE Press, 2017: 598-605.

[35] Naseer T, Oliveira G L, Brox T, et al. Semantics-aware visual localization under challenging perceptual conditions[C]//2017 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2017: 2614-2620.

[36] Garg S, Sudderhauf N, Milford M. Semantic–geometric visual place recognition: a new perspective for reconciling opposing views[J]. The International Journal of Robotics Research, 2019: 0278364918397961.