Loop Closure Detection based on Regional Weighted and Siamese Network

Wenju Li*, Qianwen Ma, Liu Cui and Tianzhen Dong
School of Computer Science and Information Engineering, Shanghai Institute of Technology, Shanghai, China
*Corresponding author’s e-mail: wjli@sit.edu.cn

Abstract. Loop closure detection (LCD) is essential for the Simultaneous Localization and Mapping (SLAM) system of an autonomous robot. Aiming at the problem of false positive in traditional loop closure detection methods, in this paper we propose an end-to-end siamese network model. Firstly, the siamese network with two same branches is designed to learn the characteristics of each image pairs. Secondly, the multi-region feature weighting method is introduced to fuse saliency regions of convolutional feature map, which well reflect structured information of the environment. In the end, the geometric consistency verification is used on the candidate convolution feature map to determine the true loop. As a result, experiments on several public datasets have illustrated the superiority of our approach. Compared with traditional methods, under the precision of 100% accuracy, the recall rate is increased by 15%. Our model is more stable in different scenarios, which can achieve robustness loop closure detection.

1. Introduction
Loop closure detection is a key component of long-term SLAM systems. The essence of loop closure detection is to judge whether the mobile intelligent device has returned to its original position. The description and matching between images are the key technology of loop detection, which can reduce the position errors accumulated over time [1]. Bag of Words (BoWs) model is commonly adopted by traditional loop detection methods mostly, which relies on hand-crafted features to construct words lists [2]. In the real running environment, this artificially designed feature is easily disturbed by changes in illumination, season and viewpoint, which makes it difficult to detect the true loop.

Loop closure detection based on image similarity is an image retrieval problem based on image content in essence, we use image retrieval methods and evaluation criteria for reference to design loop detection algorithm and evaluate it. Due to different lighting and weather conditions, changes in viewpoints and dynamic objects cause different conditions in the same place [3]. All these changes increase the difficulty of visual loop closure detection, which limits the accuracy and robustness of loop closure detection. To solve this problem, we propose a multi-region feature weighting method based on the end-to-end siamese network method to fuse the saliency regions on the convolutional feature map. In addition, the geometric consistency verification method combines visual appearance and geometric information, and the experiments have proved that it is effective for reducing false positives in loop closure detection.

2. Related work
With the development of deep learning technology, researchers tend to use deep convolution features to describe an image feature for loop closure detection. Sunderhauf [4] et al. take the AlexNet network as
an example, and evaluated the differences in the appearance and perspective changes of each layer of CNN. The experiment showed that the low-level conv3 of the convolutional network (the third convolutional layer) is more robust to appearance changes, and the high-level fc6 (the first fully connected layer) is more robust to viewing angle changes. Literature [5] uses weakly supervised transfer training method to train ResNet to extract image features, using information entropy improved Vector of Locally Aggregated Descriptors (VLAD) to process image features. Since the training and testing in this article use the Nordlandsbanen dataset, therefore the experimental results are not representative. Literature [6] designed a loop closure detection method based on supervised learning, referred to the siamese residual network. On the one hand, this method draws on the metric learning idea of the twin network, on the other hand, it exerts the good feature expression ability of the residual network. However, the network directly uses global features to construct the feature vector of the picture, and does not make full use of the saliency and distinguishability of local features.

Convolutional neural network depth feature is a data-driven high-level image feature with good image feature expression ability. It has greater advantages than traditional methods in extracting global and local features and contextual information of the image. Commonly used feature vector aggregation methods are divided into encoding and pooling. Literature [7] et al. proposed a convolutional neural network NetVLAD for scene recognition tasks, which can realize end-to-end recognition. The innovation is embedding the traditional VLAD structure into the CNN network structure to obtain a new VLAD layer. NetVLAD has achieved outstanding results in both location recognition and image retrieval data sets. However, this method uses the global convolution feature to describe the image, which requires a large amount of calculation. Literature [8] designed a feature aggregation method based on the weighting of regional saliency and channel sensitivity to enhance the description ability and distinguishability of features. However, the regional significance weight used in the article is fixed. The convolution kernel trained with image classification as the target is more inclined to grab those objects that may become the target of image classification, which is disadvantageous for loop closure detection.

3. Proposed method

3.1. Network architecture

This algorithm uses ResNet50 pre-trained on ImageNet as a benchmark, and fine-tunes the network model parameters on the Matterport3D dataset. At the same time, multi-scale regions are divided, and different local regions are assigned different weights. After the region weighted aggregation, the feature vector is distinguishable for loop closure detection. Figure 1 is the overall architecture of the algorithm. As shown below, assuming that a picture with a resolution of W×H is input to the siamese network, the dimension of the output feature map is W×H×D, it be regarded as D feature matrices of size W×H. The feature map  \( S = \{S_d\} \) \( (d = 1,2,...,D) \),  \( S_d \) defined as the d-th feature map, and is the feature value at the position (x,y) on the d-th feature map. This algorithm introduces a multi-scale area weighting mechanism on the conv4 layer activation map of ResNet50, generate image feature vectors through GeM (Generalized-Mean) pooling. The siamese network compares the spatial distance according to different output vectors to determine whether there is a loop closure.
3.2. Weighted aggregation of salient area features

The more image information contained in the activation response value of a certain spatial location on the feature map, the better its distinguishability. The difference of feature maps and positions on the feature maps have distinctive of importance. Our approach draws on the idea of multi-region feature aggregation proposed in the literature [8]. The difference is that the weights of regional features are not fixed, but are obtained by network parameter learning.

In the multi-region feature aggregation method, divide a number of square regions of different sizes on the feature map \( R = [1, W] \times [1, H] \), then define three scales (s=1, 2, 3). On the largest scale (s=1), R is set to the minimum length and width \( \min(W, H) \) of the feature map. Assuming that the number of regions that can be defined on the scale (s=1) is, the number of square regions that can be evenly divided on other scales is \( f \), the length and width of each area are \( 2 \times \min(W, H)/(s + 1) \). Figure 2 is a diagram of different scale divisions. "\" is the central position of each divided area.

This paper replaces the maximum pooling layer in the multi-region feature aggregation method with the GeM pooling layer, according to above calculation results, each region \( R \) performs GeM-pooling aggregation on all channels.

4. Geometric consistency verification

After that the preliminary similarity measurement between current query frames and previous keyframes. Select the loop closure candidate frames according to threshold, that perform geometric consistency
verification in loop closure candidate frames. Our approach combines visual appearance and geometric information, which can effectively reduce false positives in loop closure detection.

4.1. Key point extraction
Since the feature extraction of convolutional neural network is a down-sampling process, one pixel of the convolutional activation map in the network corresponds to a region of the original image. According to this, we use the extraction of key points from the maximum activation of the conv4 layer on ResNet50, to get a meaningful number of key points, we take the maximum activation area on the \( H/N_w \times W/N_w \) window of each feature map as the key point. Different values of \( N_w \) will produce different numbers of key points. The value of \( N_w \) in our algorithm is set to 4. After the initial key points are extracted, repeated key points are no longer extracted.

Inspired by hand-crafted key point descriptors such as BRIEF, a similar window operation is performed on the convolution feature map where the key points are calculated. We directly compare the activation values around the key points instead of comparing pixel intensity values like BRIEF. Finally, in the 3×3 window, the residuals of each feature vector around the key point relative to the key point feature vector are obtained, and these residuals are connected to obtain a 256-dimensional key point descriptor. Euclidean distance is used in the key point matching stage.

4.2. Determine the true loop closure
In order to determine the true loop closure, performing k-nearest neighbor search with K=7 on the image descriptor database, and filter K candidate frames through the key point matching method described above. The RANSAC algorithm is used to reject candidate frames that do not have enough valid matching items to estimate the fundamental matrix. Finally, if a valid fundamental matrix can be calculated using the matched key points, it is considered that a loop closure has been detected.

5. Experiments

5.1. Datasets
The backbone network of the siamese network in our method uses the ResNet50 model pre-trained on the ImageNet datasets. Furthermore, the transfer learning is used to fine-turning the network weight parameters on the Matterport3D datasets. The Matterport3D dataset is the largest public 3D dataset provided by 3D scanning provider Matterport. It is RGB-D dataset containing 10,800 RGB-D images and include lots of loop closure. The method in the paper has been tested on multiple public datasets. These datasets include indoor and outdoor environment, changes in viewpoint and appearance. The synopsis of testing dataset is shown in Table 1 below.

| Datasets         | Appearance | Viewpoint | Number of frames | Image resolution | Camera position |
|------------------|------------|-----------|------------------|------------------|-----------------|
| Nordland         | Severe     | Moderate  | 4670             | 640 × 360        | Frontal         |
| Campus Loop      | Severe     | Severe    | 200              | 853×840          | Frontal         |
| TUM RGB-D        | Slightly   | Moderate  | 2585             | 680×640          | Frontal         |

The Nordland dataset contains pictures from different seasons. It is one of the most challenging place recognition datasets to date, consists of four time-synchronized videos of train journeys through Norway. Each of the four 9-hour long sequences corresponds to a different season. The Campus Loop dataset consists of two sequences, which sequence contains 100 images. The dataset contains large viewpoint changes and appearance variations from clouds and the presence/absence of snow. The TUM RGB-D dataset is collected by Kinect, including depth images and rgb images, as well as ground truth and other data.
5.2. Experimental setup
Experimental hardware configuration: Inter Core i7-10700U processor, main frequency 2.9GHz, memory 32G, GPU NVIDIA GeForce RTX 2080TI, video memory 11G. Software environment: Windows10, Python3.8, Pytorch deep learning framework, Cuda11.1, PyCharm compiler.

5.3. Model training
Data pre-processing: cut out a square area from the center of the image, and uniformly transform the scale to 250×250, randomly cut out 224×224 images for training, training is divided into 50 epochs, each epoch is 250 pictures. The learning rate was set to 0.0001. The following Figure 3 is the loss change of the processed test dataset during the training process, and Figure 4 is the accuracy change.

5.4. The influence of regional weighting on loop closure detection
In this paper, the convolution feature is presented in the form of visual activation graph, with the purpose of observing the features learned by the network pictures with different appearance and different perspectives. Figure 6 is our method without regionally weighted and Figure 7 is with multi-region feature weighted. It can be seen that the focus of network has shifted from dynamic objects to traffic signs and other static objects that are beneficial to loop closure detection. Finally, the similarity score is about 0.97 as shown in Figure 8.

Figure 3. Training loss of test dataset.

Figure 4. Accuracy of test dataset.

Figure 5. Campus Loop dataset initial image.

Figure 6. Without regionally weighted.

Figure 7. Multi-region feature weighted.

Figure 8. Similarity score after regional weighted.
The PR curve reflects the robustness of the algorithm. Figure 9 illustrates the P-R curves on RGB-D fr3 office dataset. Compared to the DBoW method, our method performs better both on recall rate and precision rate. Moreover, compared with the method without the regional weight mechanism, the precision rate is increased by 15% under the same conditions.

![Figure 9. P-R curve on TUM fr3 office dataset.](image)

6. Conclusion
In this paper, we present a novel loop closure detection algorithm based on siamese networks, which combines visual appearance and geometric information. Based on siamese network, a multi-region weighting method is proposed, which integrates the salient region features of the convolution feature map. Experimental results on three open datasets illustrate that the algorithm is effective for loop closure detection. As an approach used deep learning, the network model can be fine-tuned to adapt to the new environment according to the needs of different scenarios in practical application. Our future works will focus on whether data enhancement can improve the loop closure detection accuracy.

Acknowledgments
This work is supported by the National Natural Science Foundation of China (61806126).

References
[1] Lowry, S., Sniderauf, P., Newman, N., Leonard, J.J., Cox, D., and Milford, M.J. (2016) Visual place recognition: A survey. IEEE Transactions on Robotics, vol. 32, no. 1, pp. 1–19.
[2] Galvez-Lopez, D., Tardos, J.D. (2012) Bags of Binary Words for Fast Place Recognition in Image Sequences. IEEE Transactions on Robotics 28.5, pp. 1188–1197.
[3] Liu Q., Duan F. (2019) Loop closure detection using CNN words[J]. Intelligent Service Robotics, 12(6).
[4] Sunderhauf, N., Shirazi, S., Dayoub, F., Upcroft, B., Milford, M. (2015) On the performance of convnet features for place recognition. In: intelligent robots and systems. Hamburg. pp. 4297–4304.
[5] ANG H J, HAO J L, ZhAO H. (2020) Research on ResNet-based loop detection technology in large-scale and complex scenarios. J. Computer Applications and Software, 37(07): 125-129+135.
[6] QIU K. (2018) Research and implementation of visual SLAM closed-loop detection based on siamese network. D.
[7] Arandjelovic, R., Gronat, P., Torii, A., Pajdla, T., Sivic, J. (2017) Netvlad: Cnn architecture for weakly supervised place recognition. J. IEEE Transactions on Pattern Analysis and Machine Intelligence, 99: 1–1.
[8] Tolias, G., Sicre, R., Hervé Jégou. (2016) Particular object retrieval with integral max-pooling of CNN activations. J. Computer.