Visual Loop Closure Detection Based on Stacked Convolutional and Autoencoder Neural Networks

Fei Wang\textsuperscript{1,2}, Xiaogang Ruan\textsuperscript{1,2}, Jing Huang\textsuperscript{1,2}

\textsuperscript{1}Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China;
\textsuperscript{2}Beijing Key Laboratory of Computational Intelligence and Intelligent System, Beijing 100124, China;
wf123580@163.com

Abstract. Simultaneous localization and mapping is the basis for solving the problem of robotic autonomous movement. Loop closure detection is vital for visual simultaneous localization and mapping. Correct detection of closed loops can effectively reduce the accumulation error of the robot poses, which plays an important role in building a globally consistent environment map. Traditional loop closure detection adopts the method of extracting handcrafted image features, which are sensitive to dynamic environments and are poor in robustness. In this paper, a method called stacked convolutional and autoencoder neural networks is proposed to automatically extract image features and perform dimensionality reduction processing. These features have multiple invariances in image transformation. Therefore, this method is robust to environmental changes. Experiments on public datasets show that the proposed method is superior to traditional methods in terms of accuracy, recall, and average accuracy, thereby validating the effectiveness of the proposed method.

1. Introduction
In recent years, autonomous mobile robots have become an important research direction in the field of robotics. Visual SLAM is currently the main research direction, that is, the robot acquires environmental information through a visual sensor during the moving process and performs autonomous positioning and map construction according to the collected image information [1]. There are two main problems in loop closure detection, perceptual aliasing and perceptual variability [2]. Perceptual aliasing occurs when two locations that are not closed loops are treated as closed loops. This results in incorrect information being provided to the optimization backend. Perceptual variability means that the real closed-loop position is considered not to be a closed loop. The loop closure detection problem is essentially an image matching problem. Image matching is usually divided into the following two steps, generating features and measuring the similarity.

There are many methods for generating image features, and some of them, such as Bag-of-Words (BoW)[3], have been successfully applied in loop closure detection. In recent years, the deep learning method has shown excellent performance in computer vision field. Considering that the loop closure detection task is essentially image matching [4], the loop closure detection problem has similarities with tasks such as image recognition. Therefore, we can try to use deep learning to solve the problem.
2. Related Work

In traditional methods, BoW is a widely used method that has achieved good results in loop closure detection. BoW was originally designed to classify documents that are considered to be unordered word sets. In computer vision, an original image is represented by extracting a series of visual features, also referred to as “visual words”. FV[5] uses a Gaussian mixture model to build a visual dictionary that contains more image information than BoW, VLAD[6] can be seen as a simplified version of FV, which is a better choice when the tradeoff between performance and computational efficiency is important. The global image descriptor GIST [7] has been applied to loop closure detection. However, given that the GIST descriptor is calculated for the entire image, its limited robustness to changes such as camera motion and illumination may affect its performance in image matching.

Recently, owing to the great success of deep learning in various computer vision tasks [8], researchers began to explore the use of deep neural networks to solve loop closure detection problems. The denoising autoencoder [9] and a neural network model pretrained on the ImageNet data set such as GoogLeNet [2] are typical representatives.

3. Stacked Convolutional and Autoencoder Neural Networks

3.1. SCANN Model

In the loop closure detection problem of visual SLAM, convolutional neural networks can be used to learn to extract image features. We can use the scene classification data to train the network model so that the model learns the ability to extract scene data features.

A stacked convolutional and autoencoder neural network (SCANN) is proposed considering generating the representative image features and computational efficiency. Figure 1(c) shows the basic building module of this network, the Base-Unit. The 3×3 convolution kernels are employed to ensure the size invariance of the features. A 1×1 convolution kernel is used to reduce the number of network parameters, realize feature combination across channels, and increase nonlinearity.

The batch normalization [10] method is used to regularize the data of each layer. Residual learning [11] is used to speed up the training process and enrich feature combinations. Usually, the feature extracted by convolutional neural networks is redundant [12]. Therefore, an autoencoder is proposed to reduce the dimension of the feature vector as shown in Figure 1(a).

Figure 4 shows the Base-Block of the Stacked Convolutional and Autoencoder Neural Network, which is a cascade of n Base-Units as shown in Figure 1(b). The width and height represent the size of the feature map of the module, and depth represents the channels of the feature maps of the module. Figure 2 shows the overall structure of the SCANN network model. The first part is a convolutional...
neural network based on Base-Block. “Maxpool_1, / 2” represents the first max-pooling layer, the pooling step is 2, and so on. The second part is the autoencoder network with the decoding part removed.

3.2. Training Model
In the training process, the convolutional neural network and the autoencoder network are separately trained. First, we remove the autoencoder part of the SCANN model and add a global pooling layer [13] and softmax output layer after the convolutional neural network part. Training is performed on the Places 205 [14] scene classification data sets to enable the classification network to learn to extract scene features. The output of the last four max-pooling layers and the output of the last Base-Block will be saved after the training images pass through the network to train the autoencoder.

![Figure 2. Overall structure of SCANN model](image)

We remove the decoding portion after the autoencoder training is completed. Then, the encoded portion is cascaded to the end of the previously trained convolutional neural network to obtain the final SCANN model.

3.3. Loop Closure Detection
Loop closure detection determines if the current location is a previously visited location. We need to compare the current image with all previous images to determine if the current location has been visited. There is no closed loop at the adjacent time. Incorrect results will be obtained when comparing the current picture with the picture at a nearer time. Therefore, it is necessary to set a time threshold. Those results acquired at a time whose distance from the current time is within the threshold are not considered. The procedure of loop closure detection is shown as Figure 3.

![Figure 3. Loop closure detection process](image)

This similarity between images is not only reflected in the size of the vector but also in the direction of the vector. In this paper, the cosine distance of the vector is used to measure the similarity between vectors, and the two properties of vector size and direction can be compared at the same time [4]. It is determined that the loop is detected when the cosine distance between the two vectors is
greater than the threshold. The accuracy of the detection and the recall rate can be improved by setting
the threshold reasonably in practice.

4. Experimental Results

4.1. Comparison of different features from SCANN model

Experiments were performed on the New College and City Center datasets [15] to verify the feasibility
and effectiveness of the proposed method. We compared the performance of features from different
layers of the SCANN model by using the outputs of Maxpool_2, Maxpool_3, Maxpool_4, Maxpool_5,
and the last Base-Block layer. The last Base-Block layer is labeled as the Final-Block.

A recall-precision curves graph was drawn based on the experimental results, and the curve is
shown in Figure 4. It can be seen from the figure that the performance of the descriptors of the deep
layer is better than the performance of the features from the shallow layer. It can be concluded that as
the layer is deepened, the ability of features to describe images is gradually enhanced. At the same
time, it can be seen that although Final-Block is deeper than Maxpool_5, its performance is slightly
inferior to Maxpool_5. Final-Block is directly connected to the global pooling layer during training
process, and therefore it may be more suitable for solving problems such as image classification.

Table 1 shows the average accuracy of different features from the SCANN model. It can be seen
that, on the whole, the average accuracy gradually increases as layer deepens. Features from deep
layers perform better than those from shallow layers. The best results for our approach on the New
College dataset were 81.41% and 81.76%, and the best results on the City Center data sets were 85.31%
and 84.89%, respectively.

Machine learning theory states that the larger the area enclosed by the precision-recall curve and
the coordinate axis, the better the performance of the algorithm. It is also obvious from the figure that
the area enclosed by the precision-recall curve and the coordinate axis for City Center is larger than
that of New College. On the one hand, this means that the robustness of the method to different data
sets is insufficient. On the other hand, this shows that the uniqueness of different data sets can have an
impact on the experimental results.

Table 1. Average accuracy of different features from SCANN model (%)

| Features       | Maxpool_2 | Maxpool_3 | Maxpool_4 | Maxpool_5 | Final-Block |
|----------------|-----------|-----------|-----------|-----------|-------------|
| New College Odd| 78.9      | 79.82     | 80.11     | 81.41     | 81.12       |
| New College Even| 80.52    | 80.81     | 81.03     | 81.76     | 81.32       |
| City Center Odd| 84.45     | 83.34     | 84.99     | 85.31     | 85.13       |
| City Center Even| 83.24    | 84.14     | 83.96     | 84.89     | 84.08       |

Figure 4. Precision-recall curves of different features from SCANN model
4.2. Comparison of features from deep learning and handcrafted methods

In addition to comparisons between various features of the SCANN model, this paper also compares the features extracted from the SCANN model with those of traditional methods such as BoW, FV, GIST, and the features extracted by the classical CNN [11] model. The two best features of the SCANN model, Maxpool_5 and Final-Block, were chosen to compare with the other methods. The experimental results are shown in Table 4 and Figure 8.

It can be seen from Table 2 and Figure 5 that the performance of the deep learning method, namely the SCANN model and the classical CNN, are generally superior to the handcrafted-based method. The average accuracies of all features used in the experiment are plotted in Figure 6 in order to observe the overall performance of the deep learning methods and artificial methods on different data sets more intuitively. The highest average accuracy of the deep learning method on the two data sets is 81.76% and 85.31%. The highest average accuracy of the traditional method is 80.80% and 83.84%. SCANN performs better than all of the other methods on all data sets.

Traditional methods use handcrafted features, which rely on image knowledge and prior knowledge of tasks. Therefore, traditional methods may find it difficult to cope with complex and changing environments. The deep learning method is trained from the original image to learn to extract image features that are invariant to a variety of image transformations. Therefore, these features are more robust to environmental changes than other methods when applied to loop closure detection.

Table 2. Average accuracy of features from different methods (%)

| Features        | Maxpool_5 | Final-Block | CNN   | BoW  | FV   | GIST |
|-----------------|-----------|-------------|-------|------|------|------|
| New College     |           |             |       |      |      |      |
| Odd             | 81.41     | 81.12       | 80.95 | 80.12| 80.51| 80.43|
| New College     |           |             |       |      |      |      |
| Even            | 81.76     | 81.32       | 81.23 | 79.31| 79.62| 80.80|
| City Center     |           |             |       |      |      |      |
| Odd             | 85.31     | 85.13       | 84.27 | 83.13| 83.94| 83.45|
| City Center     |           |             |       |      |      |      |
| Even            | 84.89     | 84.08       | 84.27 | 82.56| 83.34| 83.28|

(a) New College Odd      (b) New College Even         (c) City Center Odd          (d) City Center Even

Figure 5. Precision-recall curves of different methods

5. Conclusion and Discussion

In this paper, the SCANN model was proposed for loop closure detection in visual SLAM based on the basic structure of convolutional neural networks. Image features extracted using the SCANN model have multiple invariances in image transformation. These features are more suitable for complex and varied real-world environments.

Two experiments were performed to verify the performance of the proposed method. In the first experiment, the features of the different layers of the SCANN model were compared. The experimental results show that features from deeper layers perform better on both datasets than features from shallow layers. We also compared the performance of the SCANN model with classic CNN and traditional methods. From the experimental results, it can be concluded that the SCANN
model has a better precision rate, recall rate, and average accuracy on different data sets than the traditional method and classic CNN.

Because the loop closure detection task is faced with a complex and dynamic environment, in future research, we should focus on a deep neural network that is a more suitable and dynamic environment in order to improve the robustness of the algorithm to environmental changes and further improve the detection accuracy rate and recall rate.

![Figure 6. Average precision of different features](image)

**References**

[1] Fuentes-Pacheco J, Ruiz-Ascencio J and Rendón-Mancha J M 2015 Visual simultaneous localization and mapping: a survey *Artificial Intelligence Review* 43 55–81

[2] Xia Y, Li J, Qi L, Yu H and Dong J 2018 An Evaluation of Deep Learning in Loop Closure Detection for Visual SLAM *IEEE International Conference on Internet of Things* 85-91

[3] Csurka G, Dance C R, Fan L and Willamowski J 2004 Visual categorization with bags of keypoints *ECCV* 44 1–22

[4] Bai D, Wang C, Zhang B, Yi X and Tang Y 2016 Matching-range-constrained real-time loop closure detection with CNNs features *Robotics & Biomimetics* 3 15-21

[5] Perronnin F and Dance C 2008 Fisher kernels on visual vocabularies for image categorization *Computer Vision and Pattern Recognition* IEEE Computer Society 1–8

[6] Jégou H, Douze M, Schmid C and Pérez P 2010 Aggregating local descriptors into a compact image representation *Computer Vision and Pattern Recognition* IEEE 3304–11

[7] Oliva A and Torralba A 2001 Modeling the shape of the scene: a holistic representation of the spatial envelope *International Journal of Computer Vision* 42 145–75

[8] Schmidhuber J 2015 Deep learning in neural networks: an overview *Neural Networks* 61 85–117

[9] Gao X and Zhang T 2017 Unsupervised learning to detect loops using deep neural networks for visual SLAM system *Autonomous Robots* 41 1–18

[10] Ioffe S and Szegedy C 2015 Batch normalization: accelerating deep network training by reducing internal covariate shift *ICML* 448–456

[11] He K, Zhang X, Ren S and Sun J 2016 Deep residual learning for image recognition *Computer Vision and Pattern Recognition* IEEE Computer Society 770–78

[12] Krizhevsky A, Sutskever I and Hinton G E 2012 ImageNet classification with deep convolutional neural networks International Conference on *Neural Information Processing Systems* Curran Associates Inc. 1097–1105

[13] Lin M, Chen Q, Yan S 2013 Network In Network *CoRR* abs/1312.4400

[14] Zhou B, Lapedriza A, Xiao J, Torralba A, and Oliva A 2014 Learning deep features for scene recognition using places database *International Conference on Neural Information Processing Systems* MIT Press 487–95
[15] Cummins M J and Newman P 2008 FAB-MAP: probabilistic localization and mapping in the space of appearance *Int. J. Robot Res* **27** 647–65