Proceedings of the 40th Chinese Control Conference
July 26-28, 2021, Shanghai, China

Loop Closure Detection Using KPCA and CNN for Visual SLAM

Kun Dai[1], Lan Cheng*[1], Rui Yang[1], Gaowei Yan[1]

1. College of Electrical and Power Engineering Taiyuan University of Technology Taiyuan, China
E-mail: 1791253427@qq.com taolan_1983@126.com 1035446474@qq.com yangaowei@tyut.edu.cn

Abstract: Loop closure detection is often applied to eliminate accumulative track error and mapping error in visual simultaneous localization and mapping (SLAM). Deep convolution neural network (CNN) integrating principal component analysis (PCA) has been recently proposed to implement loop closure detection and to reduce the dimension of features extracted by CNN. However, the combined methods encounter low detection accuracy. To address this problem, Resnet34 pre-trained model is first used to extract features. Then, kernel PCA(KPCA) is applied on the extracted features to reduce the dimension of the features. In the similarity calculation link, restriction range strategy is used to solve the mismatch problem caused by the large similarity of adjacent frames, so as to obtain more accurate recognition results. Finally, the proposed algorithm is analyzed on two open data sets. The experiments show that the proposed algorithm outperforms the traditional CNN and PCA combined method with regard to the accuracy of feature vector matching.

Key Words: Loop closure detection, Visual SLAM, Deep learning, KPCA

1 Introduction

Loop closure detection (LCD) plays an important role in eliminating accumulative track error and map error in visual simultaneous localization and mapping (SLAM) [8]. LCD aims at recognizing the places where a mobile robot has previously visited in an unknown environment. LCD can filter out wrong closed-loops and eliminate the accumulative track error and mapping error.

Calculating the similarity between the current frame and the previous frames is considered as the key problem in LCD. The bag of visual word (BoVW) model is conventionally used to perform LCD [7]. In BoVM-based methods, a place image is usually represented by a vector in the BoVW space, and the image similarity between two places is then computed based on the associated vectors. However, the BoVM methods usually use handcrafted features, such as ORB, to measure the similarity [18], which is vulnerable to occlusion and illumination variation.

In recent years, deep learning and convolution neural network (CNN) have drawn much attention in the applications of LCD because of the rapid development of computer vision [3,10]. It has been found that the introduction of CNN can achieve promising performance regarding in many visual tasks. Since 2015, researchers have tried to use CNN to solve the LCD problem in visual SLAM(VSLAM) [9], and obtained good results. Convolution neural network has a strong feature representation ability, which provides a new method to solve the LCD problem of VSLAM. In practical applications, the high dimension of the feature vectors generated by CNN usually means high computational expense when matching features. Most algorithms use principal component analysis (PCA) to reduce the dimension of features. However, image features have a non-linear relationship which was not emphasized to in previous studies and PCA cannot deal with these non-linear features effectively since PCA is more suitable for reducing the dimension of linear features.

To address this problem, a pre-trained CNN model is first used to generate high-dimensional feature vectors to describe an image. Then, a non-linear feature compression method, named kernel PCA (KPCA), is presented to reduce the dimension of the feature vectors.

The rest of this paper is organized as follows. Section 2 reviews the literature related to artificial design features and CNN-based methods for LCD. Section 3 describes the key components of the proposed method, including the image features representation and the dimension reduction of KPCA. Section 4 introduces the datasets used in this paper, presents the experimental results and conducts discussion on the results. draws the Conclusions are drawn in Section 5.

2 Related Work

Traditional LCD methods can be divided into two categories. One is based on the visual odometer using the geometric relationship. The visual odometer-based methods assume that the camera has returned to its previous position and then perform loop detection. The other is based on the visual image method [11,12], which detects loops according to the similarity between frames, and converts the loops detection problem into a scene recognition problem. The main idea is to get scene image data through a front-end camera, and use the computer vision method to calculate the similarity between images, so as to detect loops. The latter methods can work effectively than the methods based on the visual odometer because the offset error of the odometer will lead to bigger calculation error [2].

The core issue of visual image-based methods is to calculate the similarity between images. At present, the most commonly used method is to calibrate the key points of artificial design features in images, and then to calculate the similarity between feature descriptors. Various feature extraction methods, such as SIFT, SURF, FAST and ORB [13,1,16,17], are used in previous methods. SIFT algorithm is used to extract image features in FAB-MAP [5], and then loop closure detection is carried out by building BoVW. These methods build vocabulary trees based on point

This work was supported by the National Natural Science Foundation China (No.62073232, No. 61973226).

features of different descriptors, typically amenable for fast querying of matches [15]. But the fact is that each of these features has its own characteristics. Some are invariant towards illumination or scale but complex in computing. While others may be efficient but less distinctive. None of these methods is robust towards all scenarios at all times. In addition, those image descriptors describe only the local features of individual patch, which limits their descriptive ability compared with global feature descriptors.

GIST algorithm proposed by Sünderhauf et al. [16] is applied to extract global features of images and improve the detection rate of effective loops. However, both the global features and the local features are designed according to the human experience of a designer. Thus, the performance of GIST in terms of precision rate may decline and not work stably when facing the changes of illumination, weather and seasons in real environments. Consequently, the visual words may not match properly and result in poor performance of the rotation invariance.

Thanks to the rapid development of computer vision and deep learning, CNN has been used to solve visual place recognition (VPR) problems in autonomous robots instead of traditional handcrafted features. Chen et al. [17] has used CNN for location recognition for the first time. In the existing works, a well-trained CNN model is used as a feature extractor. An image is inputted into the feature extractor, the output of one specific layer of the CNN is adopted as the global image feature. The similarity between images is then calculated. The experiments show that the CNN features are more robust to environmental changes such as different viewpoints, scales, lighting conditions than those handcrafted features. Hou et al. [9] used a neural network method to tackle the location recognition problem in the closed-loop detection of VSLAM. After comparing the output characteristics of different layers of VGG16, researchers found that the pool5 of VGG16 and the full-connected layer f are convolution layer conv5 of VGG16. The feature performance of FC6 is good, which proves the feasibility of this method. On this basis, Zhang [19] et al. reduced the dimension of features extracted from VGG16 pre-trained model using PCA and whitening feature processing, which works better than the FAB-MAP method, and has higher average precision than the methods extracting image features directly from a model. Also, the combined method of PCA and the whitening feature processing requires less computation compared with only CNN-based methods.

PCA can reduce the dimension of data while keep the original information in an image as much as possible. As a result, the computational expense caused by the high dimension of data can be reduced. However, PCA works well only when the data is linearly correlated. To address this problem, researchers presented a KPCA method for dimension reduction for nonlinear correlation data [4].

Inspired by the research in [4], we believe that the image feature vectors extracted by CNN are not linearly correlated and after PCA dimensionality reduction the independence between feature vectors changes, which results in reduced accuracy of LCD. In order to solve the problem, a KPCA method is used to reduce the dimension of the extracted feature vectors. Besides, the combined features of a certain length of image sequences are considered. Then, a two-layer matching method of interval search and group matching are employed to perform closed-loop query. More reliable and accurate loop detection results are expected.

3 Method

Fig. 1 illustrates the system that we have designed to implement the proposed LCD method. Two key components that are different from previous work are marked out in box. First, a KPCA step instead of a PCA step is performed on CNN features. Second, a constraint if considered to reduce the matching error caused by adjacent frames. Details of these changes are given in the following subsections.

3.1 Deep CNN-Based Features

In this work we use publicly available pre-trained convolutional neural networks from Torchvision [14] which is an open source machine vision package for Torch. These networks were trained on ImageNet dataset [6], which contains 1.2 million images of 1000 categories.

We use VGG-16 and Resnets34 pre-trained models and modify the network as a feature extractor to obtain specified features. The features of pool5 layer and FC6 layer in VGG-16 network achieves a better result over all layers， which has been validated in [9]. We discard the Resnet34 network

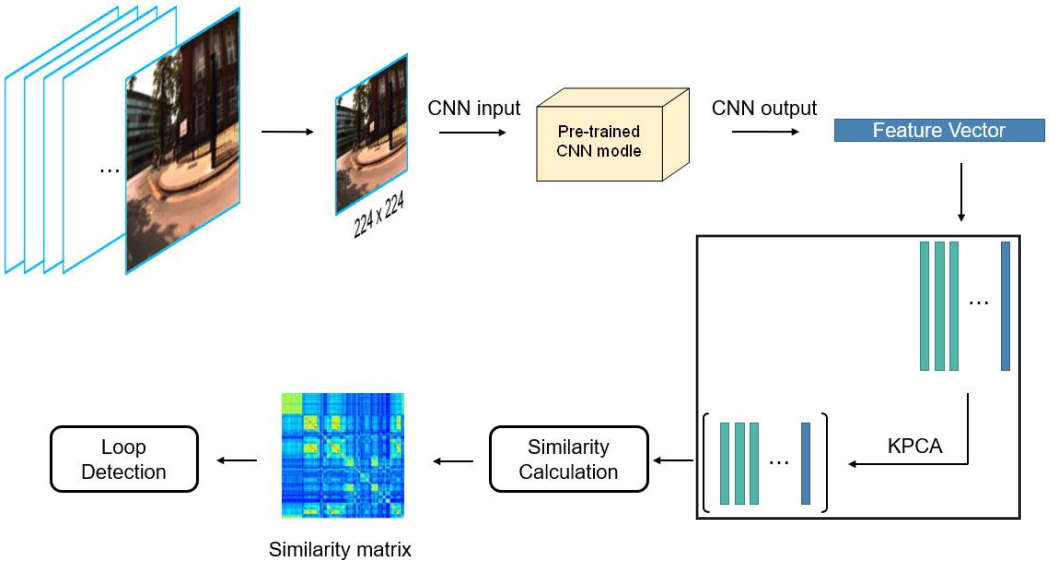


Fig. 1: Overview of the proposed CNN-based loop closure detection system.

softmax layer and only use the part before the softmax layer as an image feature extractor and the feature details are shown in Table 1. Before the image is inputted into the CNN model, the image has to be processed with the size of no less than 224×224. Then these features are outputted as one-dimensional column vector and stored in chronological order.

By putting an image *I* into the pre-trained CNN model, we can get whole-image features, which are basically different high-dimensional vectors at different layers. We use v_n to denote corresponding CNN output of the input image I_n , *n* = 1, …, *N*

$$V = (v_1,\ v_2, \dots, v_N) \tag{1}$$

where *N* denotes the number of images in the dataset, and *V* is a $d \times N$ matrix for storing image vectors. The dimension of the column vector v_n is *d*, and it is the dimension of the layer of the specified CNN output.

Table 1: Different CNN Feature

CNN architecture	Feature Layer	Dimension
VGG-16	FC6	4096
Rsenet-34	FC	512

3.2 Kernel Principal Component Analysis

Previous research usually uses PCA to reduce the dimension of the feature vectors extracted from CNN. KPCA only needs to calculate the kernel function used as the inner product in the original space, and does not need to know the form and parameters of the non-linear mapping function, nor does it need to calculate the non-linear transformation. It overcomes the limitations of determining the structure and parameters of the non-linear function and the dimension of the feature space in general mapping methods.

Defines a non-linear mapping $\Phi(x)$ and map the vector in R^N to the higher-dimensional feature space R^D, $N << D$, we have

$$\Phi(x): R^N \to R^D \tag{2}$$

$$Z = \Phi(V) = [\Phi(v_1)\ \ \Phi(v_2)\ \ \cdots\ \ \Phi(v_N)] \tag{3}$$

$$C_Z = \frac{1}{N}\Phi(V)\Phi^T(V) = \frac{1}{N}\sum_{1}^{N}\Phi(v_i)\Phi^T(v_i) \tag{4}$$

where C_Z denotes the covariance of the mapped data set *Z*. Since the non-linear mapping function $\Phi(x)$ is not easy to obtain, a kernel matrix method is usually used to realize the mapping from the input space to the higher-dimensional feature space [4]. The kernel matrix K can be calculated as

$$K = \Phi(V)\Phi^T(V) = \begin{bmatrix} \Phi^T(v_1)\Phi(v_1) & \Phi^T(v_1)\Phi(v_2) & \cdots & \Phi^T(v_1)\Phi(v_N) \\ \Phi^T(v)\Phi(v_1) & \Phi^T(v_2)\Phi(v_2) & \cdots & \Phi^T(v_2)\Phi(v_N) \\ \vdots & \vdots & \ddots & \vdots \\ \Phi^T(v_N)\Phi(v_N) & \Phi^T(v_N)\Phi(v_2) & \cdots & \Phi^T(v_N)\Phi(v_N) \end{bmatrix} \tag{5}$$

For the sake of description, we define $k(x, y)$ as kernel function,

$$k(x,y) = \Phi^T(x)\Phi(y) \tag{6}$$

The internal product of two vectors in the feature vector space can be expressed as a kernel function of two variables in the input space. Kernel functions are symmetric functions (real positive definite functions) that satisfy Mercer's conditions. Among the commonly used kernel functions, the Gaussian kernel function is chosen to compute K because the Gaussian kernel tends to give good performance under general smoothness assumptions [4].

$$k(x,y) = \exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right) \tag{7}$$

Then, the KPCA dimensionality reduction can be performed following the steps listed below.

1) Use the image vector matrix $V = (v_1\ , v_2\ , \dots, v_N)$ to calculate the kernel matrix *K*.

$$K = \left[k_{ij}\right]_{N\times N}, \qquad k_{ij} = \Phi^T(v_i)\Phi(v_j) \tag{8}$$

2) Perform data centralization and calculate the covariance matrix *K*.

$$\widetilde{K} = K - K\cdot 1_N - 1_N\cdot K + 1_N\cdot K\cdot 1_N\ ,$$

$$1_N = [1, 1, \dots, 1]^T \in \mathbb{R}^{N\times 1} \tag{9}$$

3) Perform singular value decomposition.

$$(U, S, W) = svd(\widetilde{K}) \tag{10}$$

4) Take the feature vectors corresponding to the largest δ feature vector and convert each data into new data V_{new}.

$$V_{new} = V\cdot U\cdot [:, :, \delta] \tag{11}$$

3.3 Similarity Calculation

The adjacent image frames in loop closure detection are highly correlated because of the small difference in features, which may cause a wrong detection of a closed loop and lead to the performance decline of the proposed method.

To cope with this problem, we introduce a constraint to limit the matching range of images for the current position. The constraint can be expressed as: if the current image number is N and the number of excluded images is L, the algorithm only needs to determine whether a loop closure appears from image number 1 to image number N−L. If the range is set properly, the distance between image N−L and image N is long enough so that the difference of descriptors generated by CNN between image N and image N−L is distinguishable for the LCD algorithm. By taking this constraint into consideration, the problem of wrong detection of a closed-loop can be properly addressed.

In addition, the feature vectors are normalized and the similarity of them is calculated by Euclidean distance. Then the distance between image $I_i$and $I_j$can be calculated by

$$D(I_i, I_j) = \left\| \frac{f^i}{\|f^i\|_2} - \frac{f^j}{\|f^j\|_2} \right\|_2 \tag{12}$$

where f^i, f^j represent the feature vectors after dimension reduction of v_i, v_j respectively. $\|\cdot\|_2$ is the L2-norm of the vector.

Then the similarity between image I_i and I_j can be calculated by

$$S(I_i, I_j) = 1 - \frac{D(I_i, I_j)}{\max\{D(I_i, I_j)\}} \tag{13}$$

Where $S(I_i, I_j)$ is the degree of similarity between the two feature vectors. To facilitate the calculation, we use the normalized distance to normalize the resulting similarity score to [0, 1]. If the degree of similarity is greater than or equal to a specific threshold, it will be treated as a closed loop.

4 Experiments

In this section, experiments are conducted to evaluate the proposed method on four commonly used datasets. The datasets and the evaluation method along with the implementation details are first described. The impact of the free parameters in the proposed algorithm on the detection performance is then discussed. All experiments are conducted on a NVIDIA GTX TITAN Xp GPU with 12-GB memory under CUDA 10.0. The CNN model is built under the Pytorch framework in the CentOS system.

4.1 Dataset Description

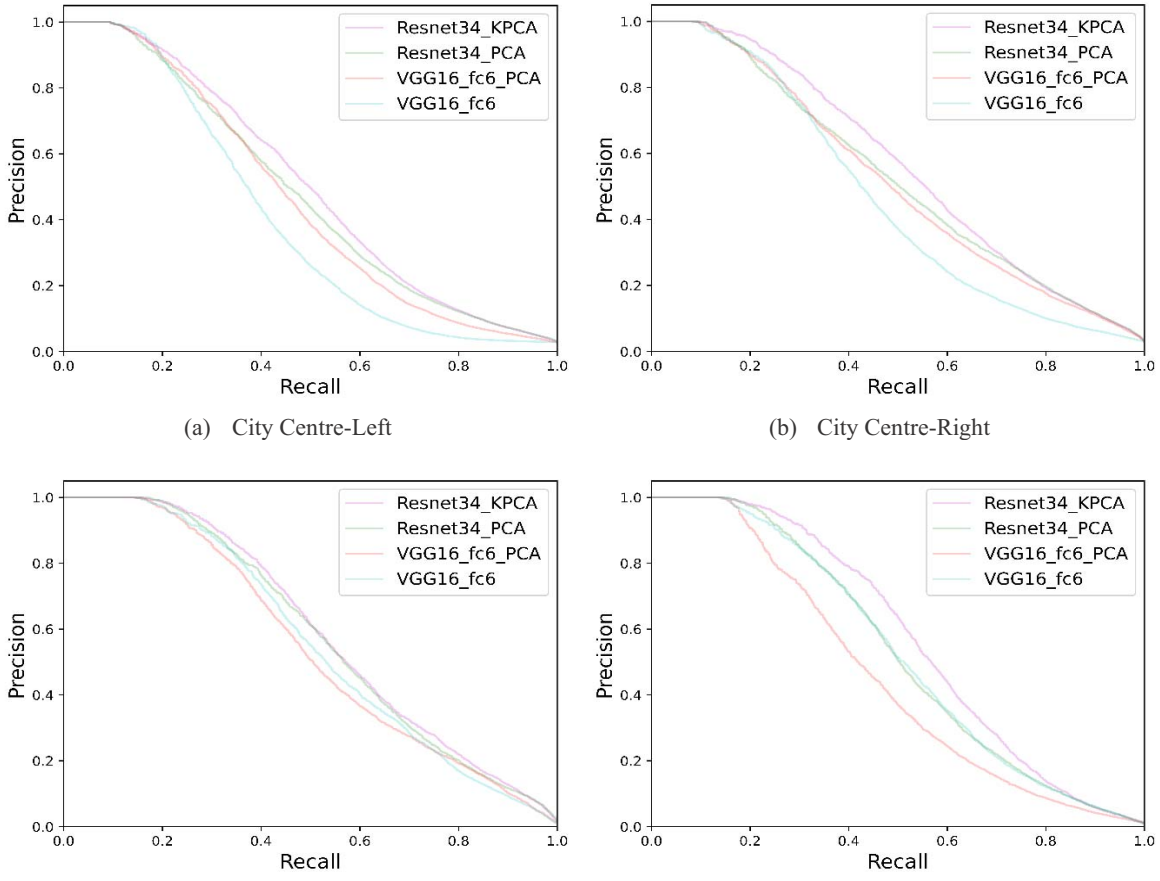
The New College and the City Centre datasets are used to evaluate the proposed method. The two datasets have been commonly used in the field of loop closure detection. The details of each dataset are summarized in Table 2. Both datasets provide images from the left camera and the right camera and the ground-truth loop closures.

Table 2: Datasets Information

Datasets		# Images	Total length(km)	Image size
New College	Left	1073	1.9	640 × 480
	Right	1073		
City Centre	Left	1237	2.0	640 × 480
	Right	1237		

4.2 Precision-Recall Curve Analysis

The proposed algorithm is compared with VGG16- PCA method in [19] and compare KPCA-PCA in [4]. The reduced dimension of image feature δ is 100. In the City Centre dataset, L = 500, and in the New College dataset, L = 70. We chose the best result selected from multiple sets of experimental data to be the value of the parameters. The results are shown in Fig.2. PR curve is plotted according to the following formula.



(a) City Centre-Left (b) City Centre-Right

(c) New College-Right (d) New College-Right

Fig. 2: Comparison of precision-recall curves between PCA-based method and KPCA-based method on (a). City Centre -Left (b). City Centre-Right (a). New College-Left and (b). and New College-Right datasets respectively.

PR curve is plotted according to the following formula.

$$P = \frac{TP}{TP + FP} \tag{14}$$

$$R = \frac{TP}{TP + FN} \tag{15}$$

where P, R represent precision, recall rate respectively. The true positive (TP) is the number of the correct loops detected. The false positive (FP) is the number of the loops that are incorrect by detecting. The false negative (FN) denotes the ground-truth loops which are wrongly detected by the algorithm.

4.3 Average Precision Analysis

Average precision is often used to evaluate a LCD algorithm from a different perspective. The performance of the proposed algorithm and the comparison algorithms regarding average precision is shown in Table 3. Its calculation method is to use integral to calculate the area enclosed by PR curve and coordinate axis.

Table 3: Analysis and Comparison of the Algorithm Average Precision on Different Datasets (%)

Datasets	City Centre-Left	City Centre-Right	New College-Left	New College-Right
Rsenet34_KPCA	50.9	56.3	59.6	57.4
Rsenet-34_PCA	48.3	52.7	58.5	52.9
VGG16_fc6_PCA	46.5	51.6	54.8	45.8
VGG16_fc6	40.6	46.5	56.0	52.8

When only PCA is used for dimensionality reduction in Resnet34 and VGG16_fc6, the Resnet34 model shows higher average precision and a better precision-recall curve. When PCA is adopted in both Resnet34 and VGG16_fc6, we have Resnet34_PCA and VGG16_fc6_PCA and Resnet34_PCA still shows better performance. This is because the Resnet34 model provides a deeper network structure and better performance of feature expression than the VGG16 model. So we can draw the conclusion that Resnet34 works better than VGG16_fc6 with or without PCA.

On the other hand, when looking at the precision-recall curves of VGG16_fc16_PCA on the dataset of New College shown in Fig.2 c) and d), we can see VGG16_fc16_PCA shows even worse performance than VGG16_fc16. This is because PCA can not extract the nonlinear correlation characteristics from the data, which causes that the feature vectors after dimensionality reduction cannot be correctly correlated with the previous feature vectors, leading to the accuracy decline of LCD.

The results of the proposed method (Rsenet34_KPCA) with Rsenet34_PCA) show that Rsenet34_KPCA works better on both datasets, which indicates KPCA is more robust when it comes to reducing the dimension of features. Furthermore, compared with VGG16_fc16_PCA, Rsenet34_KPCA stably shows higher average precision and a better precision-recall curve. So we can conclude that the combination of Resnet34 and KPCA works best among the four the methods for loop closure detection and provides an alternative to increase the accuracy of LCD.

5 Conclusion and Future Research

This paper proposed a loop closure detection method for visual SLAM systems by combing a pre-trained CNN and KPCA. Compared with CNN-PCA based methods, the proposed CNN-KPCA method is more powerful for image representation as KPCA can explore higher order information of the original inputs than PCA by using a nonlinear mapping function.

The results show that the proposed method outperforms CNN-PCA based methods regarding precision and recall rate, which indicates that KPCA is feasible for dimensionality reduction in loop closure detection. However, in the proposed method, the constraint to filter out false detection caused by adjacent frames is designed based on trail and error method. An adaptive constraint chosen method will be studied in our future research.

References

[1] Bay, Herbert , T. Tuytelaars , and L. V. Gool . "SURF: Speeded up robust features." Proceedings of the 9th European conference on Computer Vision - Volume Part I Springer-Verlag, 2006.

[2] Beeson, Patrick , J. Modayil , and B. Kuipers . "Factoring the Mapping Problem: Mobile Robot Map-building in the Hybrid Spatial Semantic Hierarchy." The International journal of robotics research 29.4(2010):p.428-459.

[3] Bengio, et al. "Representation Learning: A Review and New Perspectives." IEEE Transactions on Pattern Analysis & Machine Intelligence 35.8(2013):1798-1828.

[4] Cao L J , Chua K S , Chong W K , et al. A comparison of PCA, KPCA and ICA for dimensionality reduction in support vector machine[J]. Neurocomputing, 2003, 55(1-2):321-336.

[5] Cummins, M., Newman, P.: FAB-MAP: Probabilistic localization and mapping in the space of appearance. Int. J. Robot. Res. 27(6), 647–665 (2008).

[6] Deng, Jia , et al. "ImageNet: A large-scale hierarchical image database." IEEE Conference on Computer Vision & Pattern Recognition IEEE, 2009.

[7] Filliat, David . "A visual bag of words method for interactive qualitative localization and mapping." Proc.ieee Intl Conf.robotics & Automation (2007):3921-3926.

[8] Gao, Xiang , and T. Zhang . "Loop closure detection for visual SLAM systems using deep neural networks." Control Conference IEEE, 2015.

[9] Hou, Yi , H. Zhang , and S. Zhou . "Convolutional Neural Network-Based Image Representation for Visual Loop Closure Detection." (2015).

[10] Krizhevsky, Alex , I. Sutskever , and G. E. Hinton . "Imagenet classification with deep convolutional neural networks." Advances in neural information processing systems 25.2(2012).

[11] Labbe, Mathieu , and F. Michaud . "Appearance-Based Loop Closure Detection for Online Large-Scale and Long-Term Operation." IEEE Transactions on Robotics 29.3(2013):734-745.

[12] Latif, Yasir , C. Cadena , and J. Neira . "Robust loop closing over time for pose graph SLAM." The International Journal of Robotics Research 32.14(2013):1611-1626.

[13] Lowe, D. . "Distinctive Image Features from Scale-Invariant Keypoints." International Journal of Computer Vision 20(2004):91-110.

[14] Marcel, Sébastien, and Y. Rodriguez . "Torchvision the machine-vision package of torch." International Conference on Multimedea DBLP, 2010:1485.

[15] Mur-Artal, Raul , J. M. M. Montiel , and J. D. Tardos . "ORB-SLAM: A Versatile and Accurate Monocular SLAM System." IEEE Transactions on Robotics 31.5(2017):1147-1163.
[16] Niko Sünderhauf, and P. Protzel . "BRIEF-Gist - closing the loop by simple means." 2011 IEEE/RSJ International Conference on Intelligent Robots and Systems IEEE, 2011.
[17] Rublee, Ethan , et al. "ORB: An efficient alternative to SIFT or SURF." International Conference on Computer Vision IEEE, 2012.
[18] Saputra, Muhamad Risqi Utama , A. Markham , and N. Trigoni . "Visual SLAM and Structure from Motion in Dynamic Environments." ACM Computing Surveys (CSUR) (2018).
[19] Zhang X , Su Y , Zhu X . Loop closure detection for visual SLAM systems using convolutional neural network[C]// 2017 23rd International Conference on Automation and Computing (ICAC). IEEE, 2017.