Proceedings of the 23rd International Conference on Automation & Computing, University of Huddersfield, Huddersfield, UK, 7-8 September 2017

Loop Closure Detection for Visual SLAM Systems Using Convolutional Neural Network

Xiwu Zhang*, Yan Su*, Xinhua Zhu*
* School of Mechanical Engineering
Nanjing University of Science and Technology
Nanjing, Jiangsu, 210094, China
Email: {314101002261, suyan}@njust.edu.cn, zhuxinhua@mail.njust.edu.cn

***Abstract*—This paper is concerned of the loop closure detection problem, which is one of the most critical parts for visual Simultaneous Localization and Mapping (SLAM) systems. Most of state-of-the-art methods use hand-crafted features and bag-of-visual-words (BoVW) to tackle this problem. Recent development in deep learning indicates that CNN features significantly outperform hand-crafted features for image representation. This advanced technology has not been fully exploited in robotics, especially in visual SLAM systems. We propose a loop closure detection method based on convolutional neural networks (CNNs). Images are fed into a pre-trained CNN model to extract features. We pre-process CNN features instead of using them directly as most of the presented approaches did before they are used to detect loops. The workflow of extracting CNN features, processing data, computing similarity score and detecting loops is presented. Finally the performance of proposed method is evaluated on several open datasets by comparing it with FabMap using precision-recall metric.**

***Index Terms*—SLAM, Loop Closure Detection, Convolutional Neural Network, Deep Learning**

I. Introduction

The visual simultaneous localization and mapping (SLAM) has generated considerable research interest and has been extensively investigated in the past years both in robotics [1], [2] and computer vision communities [3]. Loop closure detection, which is also called place recognition in the filed of computer vision, is one of the most significant part in visual SLAM systems. It aims at recognizing the places where a mobile robot previously visited. Correct loop closure detection benefits visual SLAM systems a lot because it can significantly reduce the position errors that accumulate over time and it enables system to build a consistent map of the environment. Furthermore, loop closure detection can be used for relocation when robots track lost due to, for example, sudden motions, severe occlusions or motion blur [4], [5].

One class of popular and successful approaches to address loop closure detection problem is based on comparing the current observation, i.e. image captured by a robot, with those in the map that correspond to previously visited places. Once the similarity between them is high enough, then a loop hypothesis is raised. In this case, detecting a loop is essentially an image matching problem. It is the most critical technique that developing the best representation for an image.

Recent development of convolutional neural networks (CNNs) in the area of deep learning have shown their strong power to represent images, which provides a new method to address loop closure detection problem. The CNNs models are trained based on millions of labeled images, for example, ImageNet [6] to obtain large amounts of parameters. Once trained in this way, the models have the ability to learn discriminative and human interpretable feature representations for new images [7]. CNNs have been used to achieve excellent performance on a variety of tasks, such as image classification [8] and image retrieval [9]. Numerous studies have demonstrated that generic descriptors extracted from the CNNs significantly outperform hand-crafted features for visual tasks [8], [10]. Since loop closure detection is essentially similar to image classification and image retrieval – the critical step of these problems are all obtaining good representations for images, it is reasonable to expect that CNN features can be used to detect loops for visual SLAM systems.

In this paper, we present an approach that addresses visual loop closure detection problem using convolutional neural network. A pre-trained CNN model is used to generate whole-image descriptors, which are actually high-dimensional vectors. The vectors are then processed to construct representations of the places. We define similarity score based on the processed data and compute similarity matrix to detect loops. Experiments are demonstrated on open datasets from the *New College* and *City Centre* sequence [11] to test the feasibility of the presented algorithm. Finally, a comparison is provided between our approach and the FabMap [11], which is a widely applied hand-crafted feature-based algorithm in visual SLAM systems.

Authorized licensed use limited to: Skolkovo Institute of Science & Technology. Downloaded on December 23,2021 at 13:49:57 UTC from IEEE Xplore. Restrictions apply.

II. Related Work

Most of state-of-the-art loop closure detection algorithms take advantage of Bag-of-Visual-Words (BoVW) [4], [12], [13] model which is applied to image classification initially. BoVW model clusters the feature descriptors, such as SIFT or SURF using a large number of images, and produces a dictionary that contains many "words", a word can be considered as a representative of several similar features. When a new image comes, the feature descriptors are computed and the image is represented based on whether the features occur in the dictionary. An example for this is FAB-MAP [14] system, which obtained an excellent performance both in accuracy and efficiency perspective, and has become one of the standard algorithms regarding loop closure detection.

However, BoVW-based approaches have several drawbacks because they usually rely on traditional features. These so-called hand-crafted features are manually designed by researchers in computer vision area. There are a various type of features, such as SIFT [15], SURF [16], ORB [17], and BRIEF [18] etc. 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 efficiency but less distinctive. None of them is robust towards all application scenarios at all times. In addition, those image representations describe the local appearance of individual patches, limiting their descriptive power with respect to whole image methods [19].

Due to the great development and success of convolutional neural networks and deep learning in the area of computer vision [20], [21], a recent trend in autonomous robots is to exploit learned features instead of hand-crafted traditional features to tackle visual problems, especially loop closure detection problem for visual SLAM systems. Furthermore, the availability of pre-trained CNN models, such as *Caffe* [22] and *OverFeat* [23] makes it easy to experiment with such approaches for different tasks [24].

To find that how a descriptor should be constructed for loop closure detections purpose, Henning et al. [25] proposed to use a set of building blocks to automatically learn a descriptor which is robust against illumination variations. Zetao et al. [26] presented for the first time to combine powerful features learnt by CNN models with a spatial and sequential filter to address place recognition problem, and the results showed that the presented method outperformed most hand-crafted features-based techniques. Several research groups have tried to use pre-trained CNN models as feature generators to obtain whole-image representations and demonstrated on various of datasets, they concluded that CNN features are more robust against viewpoint, illumination and scale variations of the environment [10], [24], [27]–[29].

In comparison with using CNN models as black-box feature extractors, Relja et al. [30] designed a new CNN architecture to tackle large-scale place recognition problem, the model was trained in an end-to-end manner, and results showed that their representations significantly outperformed off-the-shelf CNN features on two particular datasets. The same idea was also adopted by [31] and [32], they trained a specific model from supervised dataset to perform place recognition under heavy appearance changes. Another interesting work need to be mentioned is conducted by Xiang et al. [33], [34], they proposed a novel method that employs a modified stacked denoising auto-encoder (SDA), a deep neural network trained in an unsupervised way, to solve loop closure detection problem.

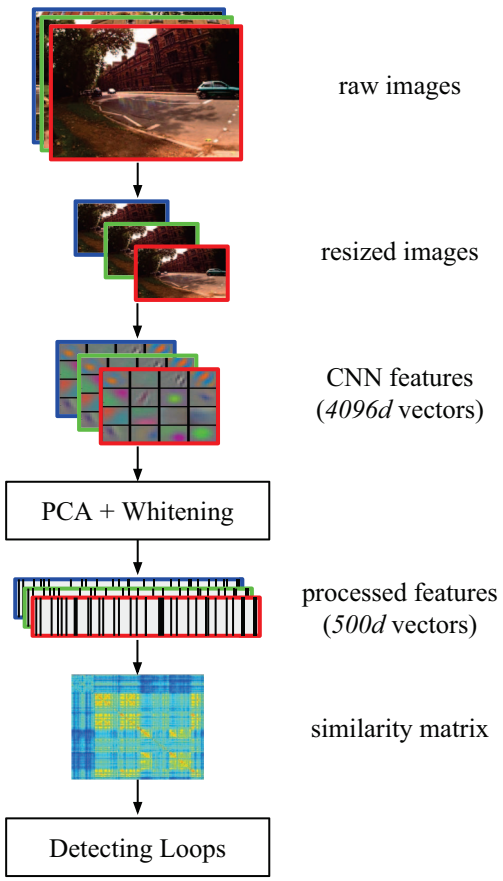


Fig. 1. **Overview of the proposed CNN-based loop closure detection system.** The raw images are resized to the expected size before they are fed into a pre-trained CNN model to obtain CNN features. The features are then dimensionally reduced and are used to compute similarity matrix, based on which we detect loops.

However, CNN or deep learning method has not been fully understood or applied in the area of SLAM. Most of the presented approaches use off-the-shelf features extracted form CNN model directly. In this paper, we pre-process CNN features before they are used to compute the similarity of pairwise images, which is inspired by the successful image retrieval method [35]. We show that the performance of the algorithm is better when doing this pre-processing step.

III. Proposed Method

In this section, we describe the proposed loop closure detection method in details. Fig. 1 illustrates our system, there

Authorized licensed use limited to: Skolkovo Institute of Science & Technology. Downloaded on December 23,2021 at 13:49:57 UTC from IEEE Xplore. Restrictions apply.

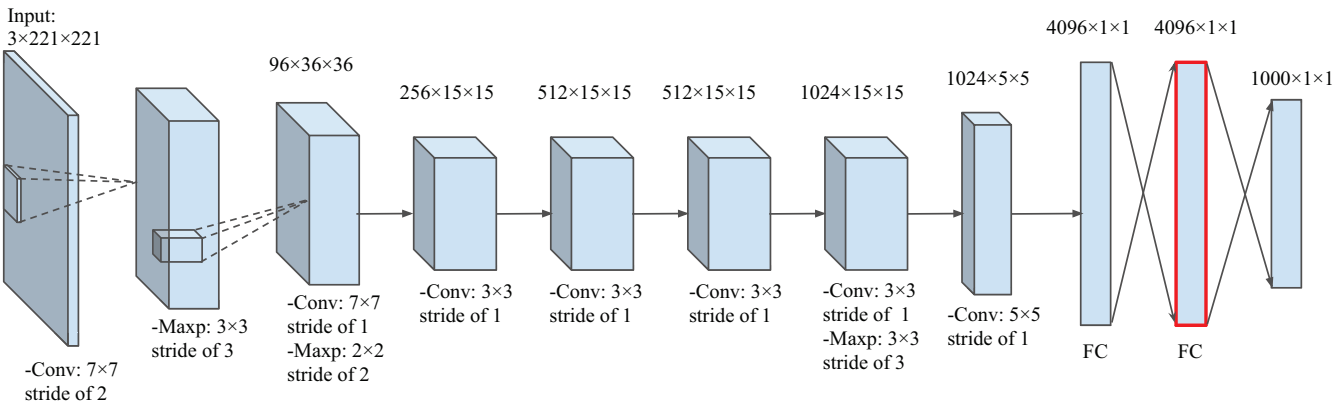


Fig. 2. **Architecture for *Accurate* model of *OverFeat*.** We use "Conv", "Maxp" and "FC" to represent convolution, max-pooling operation and fully connected layer respectively. The network takes color images of size 221×221 as input, and the output is a 1000-element vector corresponding to score for each categories. In this work, we use the output of the first fully connected layer, a 4096-dimension vector as image representation.

are several key components that are different from previous work:

1) In contrast to traditional hand-crafted features, our system uses an off-the-shelf pre-trained convolutional neural network to calculate whole-image level representations to detect loops.
2) We perform a principal component analysis (PCA) and whitening step on CNN features. By doing so, we project those high-dimension features into a lower dimensional space, which makes the detecting process more efficient and more accurate.

A. CNN Architecture

In this work we use a publicly available pre-trained convolutional neural network called *OverFeat* [23] which was proposed for the image classification task of ImageNet Large Scale Visual Recognition Challenge 2013 (ILSVRC2013) and obtained very competitive results for the detection and classifications tasks and won the localization task. This network was trained on ImageNet [6] dataset, which contains 1.2 million images of 1000 categories.

OverFeat has been released as a feature extractor in order to provide powerful features for computer vision research. There are two networks provided: the *fast* and *accurate* one respectively. We used the *accurate* version, which is capable of processing color images of any size equal to or larger than 221×221 pixels. Fig. 2 illustrates the architecture of the *accurate* version with 25 layers totally, which consists of convolutional, max-pooling and fully connected layers. For different convolutional layers, each contain 96 to 1024 kernels of size 3×3 to 7×7 and Rectified Linear Unit (ReLU) is used as nonlinear activation function. Max-pooling layers is used to build robustness to intra-class deformations with 3×3 to 5×5 kernels.

B. CNN-based Image Representation

By feeding an image I into the pre-trained CNN model, we can get whole-image features, which are basically high-dimensional vectors at different layers. These features describe the image at various levels. We use $V_l(I), l = 1, ..., 25$ to denote corresponding output of the l^{th} layer given an input image I:

$$V_l^{(I)} = (v_1^{(I)}, v_2^{(I)}, ..., v_d^{(I)}) \in \mathbb{R}^d$$

where d denote the dimension of the vector. In this study, we use the output of the first fully connected layer as shown in Fig. 2, in this case $d = 4096$.

Most of previous work use the feature vectors extracted from CNN directly to compute distance between images. However, successful image retrieval methods have indicated that performing augmentation steps, for example, principal components analysis (PCA) and whitening, to these raw features can significantly improve their power to represent images and make computation more efficient at the same time [35], [36]. Inspired by these researches, we pre-process the raw CNN features before they are used to detect loops.

1) L_2 Normalization: For each feature vector V extracted from CNN model, we perform L_2 normalization step as follows:

$$(v_1, \cdots, v_d) \leftarrow \left(\frac{v_1}{\sqrt{\sum_{j=1}^{d} v_j^2}}, \cdots, \frac{v_d}{\sqrt{\sum_{j=1}^{d} v_j^2}} \right)$$

2) PCA Dimensionality Reduction: Suppose we have obtained n normalized feature vectors and the corresponding matrix X consists of these vectors is:

$$X = \begin{bmatrix} -V^{(I_1)}- \\ -V^{(I_2)}- \\ \vdots \\ -V^{(I_n)}- \end{bmatrix} \in \mathbb{R}^{n \times d}$$

a principal component analysis procedure is performed as the following algorithm:

1: Let $\overline{V} = \frac{1}{n} \sum_{i=1}^{n} V^{(I_i)}$.
2: **for** $i = 1$ to n **do**

Authorized licensed use limited to: Skolkovo Institute of Science & Technology. Downloaded on December 23,2021 at 13:49:57 UTC from IEEE Xplore. Restrictions apply.

3: Replace $V^{(I_i)}$ in X with $V^{(I_i)} - \overline{V}$
4: **end for**
5: $cov = X^T X$
6: $[U, S, W] = svd(cov)$
7: $V^{(I_i)}_{reduced} = V^{(I_i)} U[:, :500]$

In this algorithm we zero-centre the original features and compute the covariance matrix cov of the data. For $step$ 6, a singular value decomposition (SVD) is performed on cov so that we can get matrix U of which columns are the eigenvectors and matrix S of which diagonal entries λ_j are singular values of cov, i. e. $cov = diag(\lambda_1, \cdots, \lambda_d), \lambda_1 >, ..., > \lambda_d$. Then we project the original vector into a lower dimensional space in $step$ 7, in this work, we reduce the dimensionality of feature to 500.

3) Whitening: We whitened each feature vector according to the form:

$$V^{(I_i)}_{whitened,j} = \frac{V^{(I_i)}_{reduced,j}}{\sqrt{\lambda_j + \epsilon}}$$

where λ_j is the singular value on j^{th} dimensionality obtained in PCA procedure and ϵ is a small constant, typically 10^{-5}, to prevent division by zero.

By performing the above procedures, we have the final CNN-based image representations which formed as 500-dim vectors.

C. Similarity Matrix

The key part of visual loop closure detection problem is the estimation of the similarity between frames (images). For this purpose, we calculate the distance between CNN feature vectors of different frames and define the similarity score for pairwise frames.

We denote the final CNN feature vectors as $V^{(I_i)}_w \in \mathbb{R}^{500}$ of input image i and use the Euclidean distance of vectors to measure the similarity between images. First we normalize the vectors and calculate the Euclidean distance as follows:

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

where $D(i,j)$ is the distance between image I_i and I_j, and $\|\cdot\|_2$ is the L_2-norm of the vector. Then we define the similarity score between images as:

$$S(i,j) = 1 - \frac{D(i,j)}{max\{D(i,j)\}}$$

Note that a normalized distance is used in our case to obtain the score values lies in $[0, 1]$ for measurement purpose. If the similarity score is larger than a specific threshold, we regard it as a loop.

By collecting similarity score of pairwise images in a matrix, we obtain similarity matrix of which each column j stores the score between the j^{th} images and all images in the sequence. Fig. 3 shows an example of the visualized similarity matrix, where main diagonal items $S(i,i)$ are equal to 1, which indicates the same place. Other than these special items, larger values (the bright part in the figure) indicate that the corresponding images are more similar, and they are more likely considered as loops.

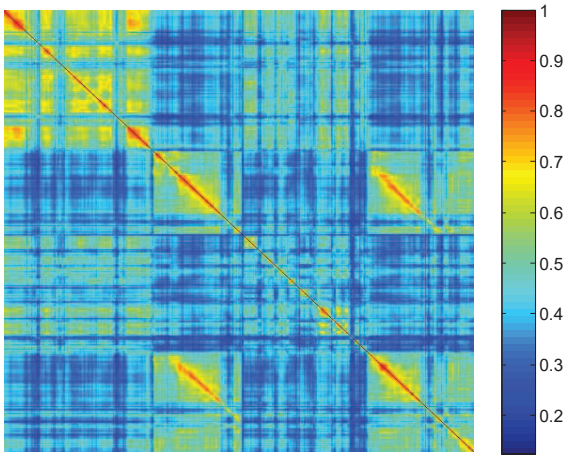


Fig. 3. **An example of similarity matrix.** Larger values indicate higher similarity scores and they are more likely the same places for the corresponding images and vice versa.

TABLE I
DATASETS DETAILS

Dataset	Total Length	# Images	Size	# Ground Truth
New College	1.9km	2146	640×480	14832
City Centre	2.0km	2474	640×480	26976

IV. EXPERIMENTS AND EVALUATION

In this section we describe the details of conducted experiments and evaluate their results. We compare the proposed CNN-based visual loop closure detection method against state-of-the-art methods FAB-MAP [11] (which is based on hand-crafted features: SURF [16]) on several different datasets using precision-recall curve metric.

A. Datasets

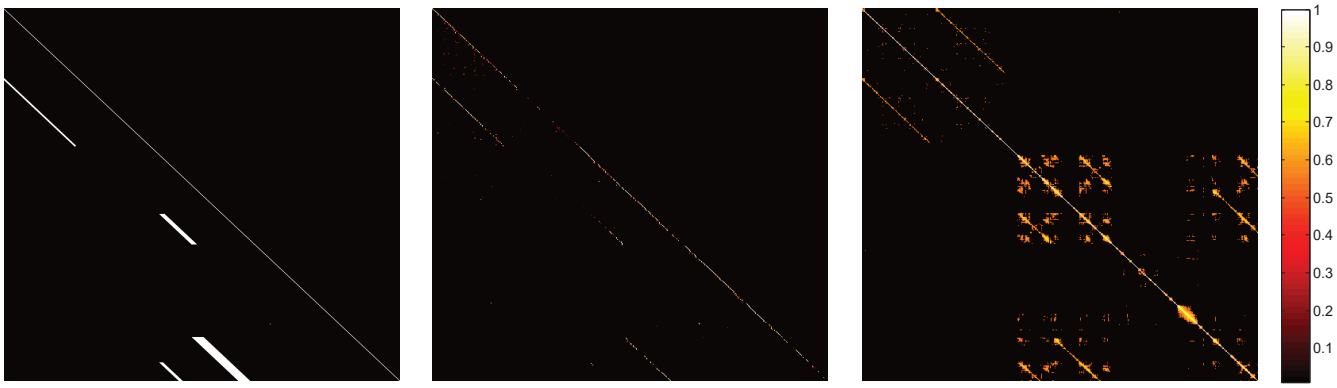
Experiments are conducted on two publicly available [1] datasets called *New College* and *City Centre* [11] which are widely used in visual SLAM research and in evaluating loop closure detection algorithms. The details of these two datasets are show in Table I. The images to the left and right of a robot are collected simultaneously every 1.5m by two cameras mounted on the pan-tilt, so there are 1073 and 1237 image pairs respectively indeed. Both datasets provide ground-truth loop closures, which is convenient for us to measure the correctness of our results.

B. Precision-Recall curve

We use precision-recall curve metric to evaluate the results of proposed loop closure detection method. The correct detections are known as *true positives (TP)*, the incorrect detections are *false positives (FP)*, and the ground-truth loops that the

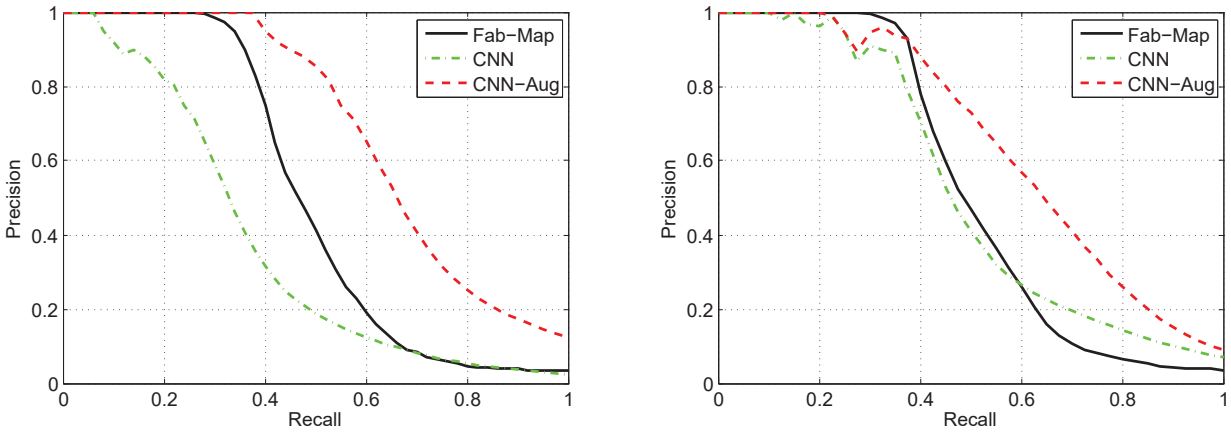
[1] http://www.robots.ox.ac.uk/~mobile/IJRR_2008_Dataset/data.html

Authorized licensed use limited to: Skolkovo Institute of Science & Technology. Downloaded on December 23,2021 at 13:49:57 UTC from IEEE Xplore. Restrictions apply.



(a) Ground-truth (b) Fab-Map (c) CNN-Aug

Fig. 4. Visualization of similarity matrix for *New College* dataset.



(a) New College (b) City Centre

Fig. 5. Comparison of precision-recall curves between CNN-based method and Fab-Map on **(a).** *New College* and **(b).** *City Centre* dataset respectively.

system erroneously discards are *false negatives (FN)*. Precision is defined as the ratio between the number of correct detections and all the detections, while the recall, as the ratio between correct detections and all the loops in the ground-truth. A perfect system would be one that achieves precision of 100% and recall of 100%.

C. Experimental Results

Since the image size of the two datasets is 640×480, we resize the images to the expected input size of 221×221 before they are fed in to CNN model. To make the comparison more reasonable, we use the same resized images for Fab-Map. We denote the proposed CNN-based loop closure detection method as *CNN* without feature augmentation (PCA and whitening) and as *CNN-Aug* when using these procedures.

According to the steps described in section III , we calculate the similarity score for pairwise images using processed CNN features and build similarity matrix. Fig. 4 compares the visualized similarity matrix obtained by Fab-Map and *CNN-Aug* method with the ground-truth loops on *New College* dataset. The figure shows that both Fab-Map and our method can detect most loops. But Fab-Map tends to give a near zero score even if there is a true loop, this may result in discard of true loops in practice. Meanwhile, *CNN-Aug* gives more bright blocks because the nearby images will also be considered as loop closing candidates.

A threshold on the similarity score is then applied to determine if a loop closure has occurred. By scanning this threshold, we obtain precision-recall curve. Fig. 5 shows the results on *New College* and *City Centre* dataset respectively. For *New College* dataset, Fab-Map outperforms CNN-based method when the features are used directly. However, after processing the raw CNN features, *CNN-Aug* method can achieve a higher precision and outperforms the method without augmentation procedures significantly. The similar results can be seen on *City Centre* dataset, what different is that both *CNN* and *CNN-Aug* method achieve higher precision at high recall rate, but in low recall rate the Fab-Map has a better performance. We conclude that our method can obtain more loops when the recall rate is high, which is what we expect to see in practice.

Authorized licensed use limited to: Skolkovo Institute of Science & Technology. Downloaded on December 23,2021 at 13:49:57 UTC from IEEE Xplore. Restrictions apply.

V. Conclusion and Future Research

This paper proposes a loop closure detection method for visual SLAM systems based on convolutional neural network. In comparison with traditional hand-crafted features that most presented approaches use, CNN features are more powerful for image representation as they learn the inner structure of images. The workflow of CNN-based loop closure detection method is described in details. The performance of proposed method is evaluated on open datasets by comparing with Fab-Map. The results show that our method outperforms Fab-Map at a higher recall rate. Which indicates that CNN is feasible for loop closure detection and provide an alternative way for visual SLAM systems.

Considering the new opportunities and challenges that exist in this field. The future work will include:

1) Evaluating features extracted from different layers of CNN model as they represent images at different level. In addition, it enables us to design and train a specific neural network for loop detection purpose.
2) Since we simply detect loops based on similarity scores in this paper, we will use additional constraints, for example spatial consistency to reject false loops and therefor obtain higher precision in the future research.

References

[1] A. J. Davison, I. D. Reid, N. D. Molton, and O. Stasse, "Monoslam: Real-time single camera slam," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 6, pp. 1052–1067, June 2007.

[2] M. J. M. M. Mur-Artal, Raúl and J. D. Tardós, "ORB-SLAM: a versatile and accurate monocular SLAM system," *IEEE Transactions on Robotics*, vol. 31, no. 5, pp. 1147–1163, 2015.

[3] J. Engel, T. Schöps, and D. Cremers, "Lsd-slam: Large-scale direct monocular slam," in *European Conference on Computer Vision*. Springer, 2014, pp. 834–849.

[4] D. Galvez-LoPez and J. D. Tardos, "Bags of binary words for fast place recognition in image sequences," *Robotics IEEE Transactions on*, vol. 28, no. 5, pp. 1188–1197, 2012.

[5] B. Williams, G. Klein, and I. Reid, "Automatic relocalization and loop closing for real-time monocular slam," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 9, pp. 1699–1712, Sept 2011.

[6] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 248–255.

[7] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European conference on computer vision*. Springer, 2014, pp. 818–833.

[8] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *International Conference on Neural Information Processing Systems*, 2012, pp. 1097–1105.

[9] A. Babenko, A. Slesarev, A. Chigorin, and V. Lempitsky, "Neural codes for image retrieval," in *European conference on computer vision*. Springer, 2014, pp. 584–599.

[10] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[11] M. Cummins and P. Newman, "FAB-MAP: Probabilistic Localization and Mapping in the Space of Appearance," *The International Journal of Robotics Research*, vol. 27, no. 6, pp. 647–665, 2008. [Online]. Available: http://ijr.sagepub.com/cgi/content/abstract/27/6/647

[12] J. Sivic, A. Zisserman *et al.*, "Video google: A text retrieval approach to object matching in videos." in *iccv*, vol. 2, no. 1470, 2003, pp. 1470–1477.

[13] D. Filliat, "A visual bag of words method for interactive qualitative localization and mapping," in *Robotics and Automation, 2007 IEEE International Conference on*. IEEE, 2007, pp. 3921–3926.

[14] M. Cummins and P. Newman, "Appearance-only slam at large scale with fab-map 2.0," *The International Journal of Robotics Research*, vol. 30, no. 9, pp. 1100–1123, 2011.

[15] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[16] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," *Computer vision–ECCV 2006*, pp. 404–417, 2006.

[17] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "Orb: An efficient alternative to sift or surf," in *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, 2011, pp. 2564–2571.

[18] M. Calonder, V. Lepetit, C. Strecha, and P. Fua, "Brief: Binary robust independent elementary features," *Computer Vision–ECCV 2010*, pp. 778–792, 2010.

[19] M. J. Milford and G. F. Wyeth, "Seqslam: Visual route-based navigation for sunny summer days and stormy winter nights," in *Robotics and Automation (ICRA), 2012 IEEE International Conference on*. IEEE, 2012, pp. 1643–1649.

[20] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.

[21] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[22] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 2014, pp. 675–678.

[23] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "Overfeat: Integrated recognition, localization and detection using convolutional networks," *arXiv preprint arXiv:1312.6229*, 2013.

[24] N. Sunderhauf, S. Shirazi, A. Jacobson, F. Dayoub, E. Pepperell, B. Upcroft, and M. Milford, *Place recognition with ConvNet landmarks: Viewpoint-robust, condition-robust, training-free*. Springer International Publishing, 2015.

[25] H. Lategahn, J. Beck, B. Kitt, and C. Stiller, "How to learn an illumination robust image feature for place recognition," in *Intelligent Vehicles Symposium (IV), 2013 IEEE*. IEEE, 2013, pp. 285–291.

[26] Z. Chen, O. Lam, A. Jacobson, and M. Milford, "Convolutional neural network-based place recognition," *Computer Science*, 2014.

[27] Y. Hou, H. Zhang, and S. Zhou, "Convolutional neural network-based image representation for visual loop closure detection," in *IEEE International Conference on Information and Automation*, 2015, pp. 2238–2245.

[28] N. Sünderhauf, S. Shirazi, F. Dayoub, B. Upcroft, and M. Milford, "On the performance of convnet features for place recognition," in *Intelligent Robots and Systems (IROS), 2015 IEEE/RSJ International Conference on*. IEEE, 2015, pp. 4297–4304.

[29] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "Cnn features off-the-shelf: An astounding baseline for recognition," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2014, pp. 512–519.

[30] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "Netvlad: Cnn architecture for weakly supervised place recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 5297–5307.

[31] R. Gomez-Ojeda, M. Lopez-Antequera, N. Petkov, and J. Gonzalez-Jimenez, "Training a convolutional neural network for appearance-invariant place recognition," *arXiv preprint arXiv:1505.07428*, 2015.

[32] B. Zhou, A. L. Garcia, J. Xiao, A. Torralba, and A. Oliva, "Learning deep features for scene recognition using places database," *Advances in Neural Information Processing Systems*, vol. 1, pp. 487–495, 2015.

[33] X. Gao and T. Zhang, "Loop closure detection for visual slam systems using deep neural networks," in *Control Conference (CCC), 2015 34th Chinese*. IEEE, 2015, pp. 5851–5856.

[34] ——, "Unsupervised learning to detect loops using deep neural networks for visual slam system," *Autonomous Robots*, vol. 41, no. 1, pp. 1–18, 2017.

[35] H. Jégou and O. Chum, "Negative evidences and co-occurences in image retrieval: The benefit of pca and whitening," *Computer Vision–ECCV 2012*, pp. 774–787, 2012.

[36] Y. Gong, L. Wang, R. Guo, and S. Lazebnik, "Multi-scale orderless pooling of deep convolutional activation features," in *European conference on computer vision*. Springer, 2014, pp. 392–407.

Authorized licensed use limited to: Skolkovo Institute of Science & Technology. Downloaded on December 23,2021 at 13:49:57 UTC from IEEE Xplore. Restrictions apply.