https://doi.org/10.2352/ISSN.2470-1173.2018.17.AVM-284

Loop Closure Detection in Simultaneous Localization and Mapping Using Learning Based Local Patch Descriptor

Dong-Won Shin and Yo-Sung Ho; Gwangju Institute of Science and Technology, 123 Cheomdangwagi-ro, Buk-gu, Gwangju, 61005, South Korea

Abstract

Simultaneous localization and mapping (SLAM) is a computational problem reconstructing the 3D environment map and estimating the camera trajectories simultaneously. This research topic has been studied for several decades in computer vision field. Among many other components in SLAM, we are concentrating on a loop closure detection determining whether a current view is visited by a camera agent before or not. It is an essential procedure for obtaining a consistent 3D environment map. In this paper, we propose a learning based local patch descriptor using a generative adversarial network to solve a problem of the loop closure detection. We trained the generative adversarial network on local patches of a place-oriented dataset and used the network model to extract the local patch descriptor. By using the descriptor on a general bag-of-visual-word method, we achieved better results than the conventional methods in terms of the precision and recall measure.

Introduction

For a human being, understanding the environment map is a simple and straightforward task and it is a necessary procedure to interact with other objects. We can easily achieve a lot of operations such as avoiding obstacles, putting objects somewhere and opening the doors by knowing the appearance of the surrounding map. However, it is a very difficult task for computer machines. For several decades, many computer vision researchers have tried to implement systems to understand the environment map via cameras. More specifically, the machines should know what environment looks like and where the camera is at the same time. These extensive research topics have emerged as a simultaneous localization and mapping (SLAM). Figure 1 shows the main components of the modern SLAM system [1].

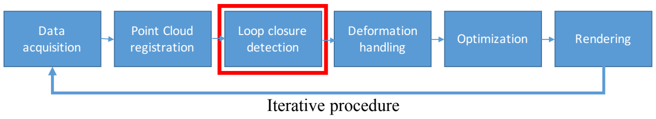

Figure 1 The main components of the modern SLAM system

First of all, the input data from color, depth, global positioning system (GPS) or inertial measurement unit (IMU) sensors is acquired on the data acquisition part with the appropriate noise handling procedure. Next, a 3D point cloud data (PCD) is generated by mainly using images and camera intrinsic parameters, and then the PCD from each time frame is registered in a single point cloud model on the point cloud registration step. A loop closure detection is a step determining whether the current position is visited before or not. If a loop closure is detected, this can be a constraint to make a consistent 3D environment map later on the model optimization step. The deformations are invoked by the moving objects like people, non-rigid, flexible objects like clothes and the SLAM system should properly handle the deformations to construct the accurate map. The optimization step optimizes the camera trajectories and 3D environment map with the constraints such as camera positions, 3D model, the loop closures and deformations from the front end. Finally, the rendering part produces the complete 3D volumetric model from the 3D point cloud data by using various computer graphics techniques. Among the main components of the modern SLAM system, we primarily study on the loop closure detection problem and proposed a method using a learning based local patch descriptor to solve the problem.

If the robot equips the 3D perception ability, there are many potential applications in the robotics field. Currently, the most promising research area is the autonomous vehicle. In order to safely transport the passengers and luggage to the destination, understanding what is around the vehicle and constructing the environment map are important. Any forms of a robot such as a picking robot in a warehouse, a service robot in a store, and drone require the perceiving the 3D environment to complete the assigned missions.

Aside from the robotics field, The SLAM technology is required in the augmented reality (AR). The augmented reality is to seamlessly synthesize the virtual object into the real 3D scene where the viewer is gazing. In order to achieve this goal, knowing the 3D environment map and the position of the viewer are necessary. We can naturally see the virtual character with a real scene and give the user an impressive feeling based on the 3D knowledge from SLAM.

Lastly, the virtual reality (VR) becomes a potential application of SLAM. The present VR technique uses the external position tracking devices which are cumbersome to install to find the location of the head-mounted display (HMD). By using the SLAM technology on the VR, we don't need any external devices causing the additional expenses and it will deliver the compact impression to the user.

In this paper, we propose a loop closure detection method using the learning based local patch descriptors. We combine the descriptors with bag-of-visual word (BoVW) method to get a descriptive vector for an image. By this vector representation, we can measure how much two images are similar so that we can detect the loop closure. For the learning based local patch descriptor, we will exploit a deep convolutional generative adversarial network (DCGAN) which is one of the successful unsupervised learning method. At this point, the training data is a critical issue for the learning based methods. Since we concentrate on the loop closure problem, we will use the place-oriented dataset including millions of scene images from indoor and outdoor environments. In order to evaluate the proposed method for the loop closure detection in SLAM application, we created the ground-truth label for a place recognition by an online survey and a manual classification.

Related Works

There are two main categories for solving the loop closure detection problem: a local descriptor based method and whole image descriptor based method. First, the local descriptor based method is divided into the training and estimation stages. For the training stage, we extract the local image descriptors like SIFT, SURF, and ORB from training images. Next, a hierarchical k-means clustering is applied on the extracted local image descriptors to construct the bag-of-visual-word (BoVW) model in a tree form. At this point, the descriptors having a similar appearance will be clustered and the centroid of the cluster represents the visual-word. For the estimation stage, after extracting the local image descriptors as same as the beginning of the training stage, the histogram of the visual-word can be calculated by stacking the descriptor on the most similar visual-word bin. It is called bag-of-visual-word vector and used for similarity computation between images [2]. Figure 2 shows the overall procedure for the bag-of-visual-word method.

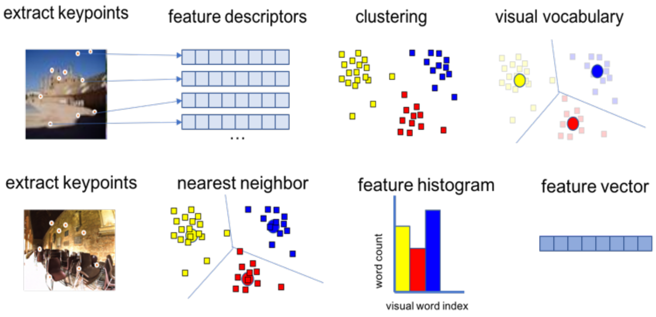

Figure 2 Overall Procedure for the bag-of-visual-word method

As an extension of the bag-of-visual-word model, the Fisher vector (FV) model has been introduced [3]. Although both BoVW and FV models are based on the local feature descriptor, the FV model finds the parameters of Gaussian mixture model while the BoVW model employs the K-means clustering algorithm. Next, the gradient of the log likelihood with respect to the parameters of the model is calculated, then the concatenation of these partial derivatives is defined as the Fisher vector. This model has the advantage to give similar or even better classification performance than BoVW model obtained with supervised visual vocabularies.

As another extension of the bag-of-visual-word model, we can find the vector of locally aggregated descriptors (VLAD) [4]. In the case of VLAD, it accumulates the residual of each descriptor with respect to its assigned cluster while the BoVW model only determines whether the descriptor is assigned to the cluster center or not. Therefore, it adds the more discriminative property in the vector representing an image, which is beneficial for the place recognition.

Second, the whole-image descriptor based method extracts the image descriptor by analyzing the image itself. The GIST descriptor, one of the whole-image descriptor based methods, is computed by measuring the responses of an image to a Gabor filter bank [5]. It has a compact representation with less than 1000 dimensions in its standard implementation. Recently, the whole-image descriptor using the convolutional neural network (CNN) has been introduced [6]. It uses the pre-trained AlexNet model to extract the image descriptor for the loop closure detection. They feed an image to the model and obtain the feature maps at each layer and used the feature maps as the image descriptor. They found out the image descriptor from the fifth convolutional layer shows the best result of the loop closure detection. Figure 3 shows the structure of AlexNet model through all convolution, pooling, and fully connected layer.

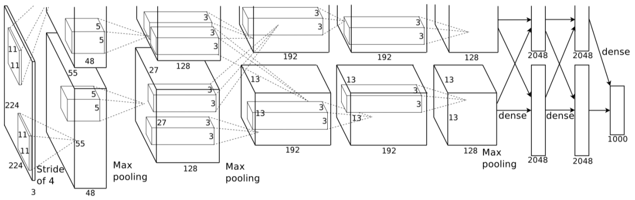

Figure 3 Structure of AlexNet model

Proposed Loop Closure Detection Method

The local image descriptor shows a good loop closure detection result with BoVW method but it is not enough to deal with the harsh variations in the images like an illumination change, view point change, occlusion, and deformation.

Recently, the neural networks approaches have been widely studied and applied to a lot of research fields, especially for the classification, regression, and recognition.

In this paper, we exploit the general BoVW model to detect loop closures on the image sequences. In contrast with the conventional methods, however, we used the descriptors from a deep convolutional generative adversarial network (DCGAN) on the local patches obtained by SURF features to construct the visual-word. In order to obtain the suitable DCGAN model on our problem, we trained the model on the place-oriented dataset. Additionally, we propose the descriptor extraction procedure from the trained DCGAN model for the loop closure detection.

Generative Adversarial Network (GAN)

Among many of neural network models introduced recently, a generative adversarial network (GAN) has become a promising method for the generative model [7]. The generative capability of GAN can handle the variations in images and we focused on that point to solve the loop closure detection problem.

GAN has a competitive relationship between two networks: a discriminator network and a generator network. The generator network (G) generates fake images from a random noise and the discriminator network (D) discriminates whether the input images are real or fake. Figure 4 illustrates the structure of the generative adversarial networks.

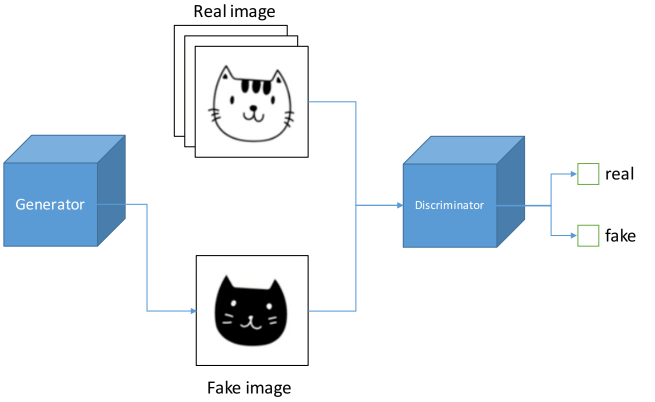

Figure 4 Generative adversarial networks

Deep Convolutional GAN (DCGAN)

The deep convolutional GAN (DCGAN) is one of the variants of GAN containing the convolutional operations over the whole network [8]. More specifically, the discriminator of GAN has a general convolution to abstract the input image and the generator of GAN has a transposed convolution to construct the fake image from the noise vector. Aside from that, DCGAN has a batch normalization procedure improving the regularization ability and making it faster to train the network. Lastly, the Leaky ReLU activation was used in the discriminator for all layers. Combining all the introduced tricks with GAN, DCGAN has achieved astonishing results than the conventional methods. Figure 5 illustrates the overall structure of DCGAN model. In this paper, we exploit the DCGAN to extract the local patch descriptor and tried to combine the descriptor with BoVW model to detect the loop closures.

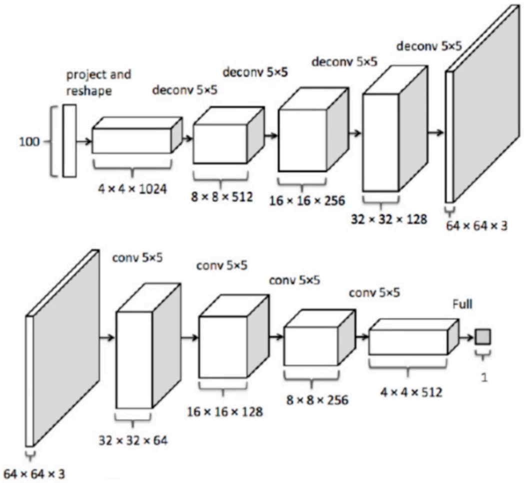

Figure 5 Overall structure of DCGAN model

Local Patch Descriptor using DCGAN

In order to construct the DCGAN model for the loop closure detection, we trained the model on the place365 dataset containg the place-oriented scene images [9]. We first extracted the SURF features from the image and made 64×64 size of local patches having the center as the position of the SURF features. We made approximately 300 local patches per image and applied this process on 36500 images. Therefore, we obtained almost 10,960,000 training patches in total. Figure 6 shows the extracted local patches with SURF features.

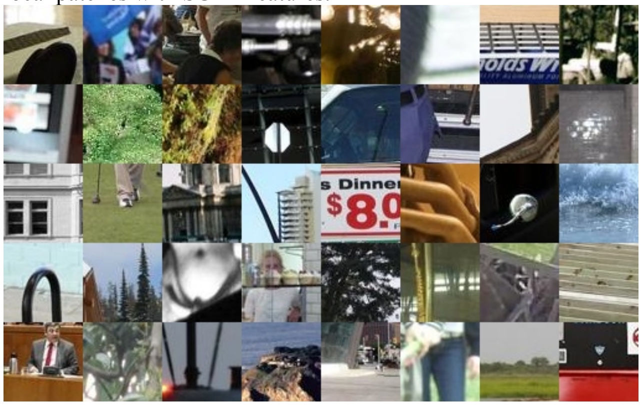

Figure 6 Local patches with SURF features

After training the DCGAN model, we can construct the DCGAN descriptors for local patches from the discriminator part of DCGAN. We applied a 4×4 max-pooling operation on the last layer of the discriminator. Finally, we can get the 512 dimensions vector for a single local patch. Figure 7 indicates the position of the descriptor extraction. There can be several different types of the extraction approaches but we leave the experiments for the future works.

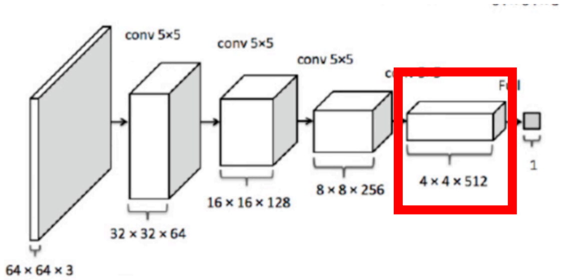

Figure 7 Position of the descriptor extraction

BoVW with the proposed Local Patch Descriptor

The BoVW model exploits the image descriptors like SIFT, SURF, and ORB to construct the hierarchical structure of the visual-word. And then, the histogram of the visual-word can be considered as a representation of a query image.

In this paper, we will combine the DCGAN descriptor with the BoVW model. We first make local patches from an image as the same way to make a training data and extract DCGAN descriptor from the local patches by the trained DCGAN model. Finally, the hierarchical BoVW model is constructed by using the DCGAN descriptor with 10 branches at each level and 5 depth levels. Figure 8 illustrates the construction of BoVW model with DCGAN descriptor.

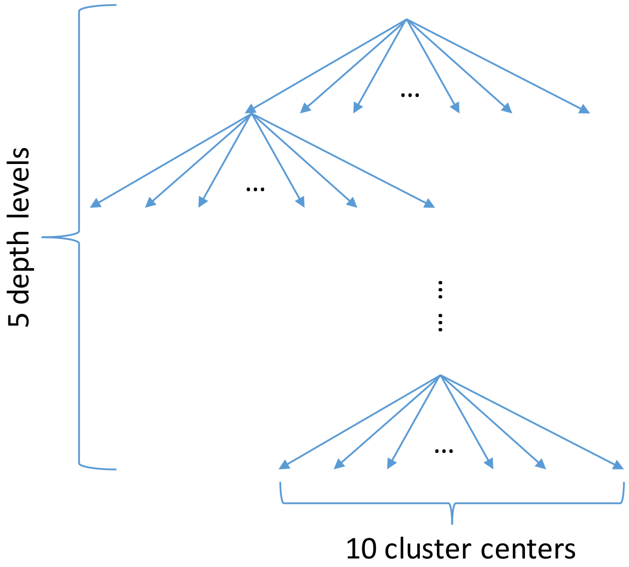

Figure 8 Construction of BoVW model with DCGAN descriptor

Experiment Results

In this section, we will evaluate the loop closure detection compared to many other methods by the metric of the precision and recall curve. We will introduce the dataset we used, training and evaluation details in the following subsections.

Dataset

There are two main categories for datasets: the training and evaluation. For the training dataset, we employed the place365 dataset having the place-oriented scene images by Zhou et. al. [9]. It contains 10 million scene photographs, labeled with scene semantic categories. In our experiments, we used 1000 images among them and extracted local patches as we described.

Figure 10 Example of Places dataset

For the evaluation dataset, we exploited City Centre (CC) and New College (NC) dataset by Mark et. al. [10]. The two datasets are widely used in visual SLAM field and in evaluating loop closure detection since those have the ground-truth GPS data for the robot agent equipped with a stereo camera. At this point, the robot agent passed through a specific route having loop closure.

Figure 11 Example of New College dataset

We made a ground-truth loop closure data by the GPS information and images. We first filtered all image pairs via the proximity of GPS position. If the distance between the image pair is bigger than some threshold (in this experiment, 0.8 meters), we rejected it as the loop closure and if not, we accepted it. And then we selected the definite loop closure pairs by comparing the appearance of images. At this point, although this ground-truth dataset is involved with our subjective opinion, we will make a ground-truth data by surveying over 100 people in the future research.

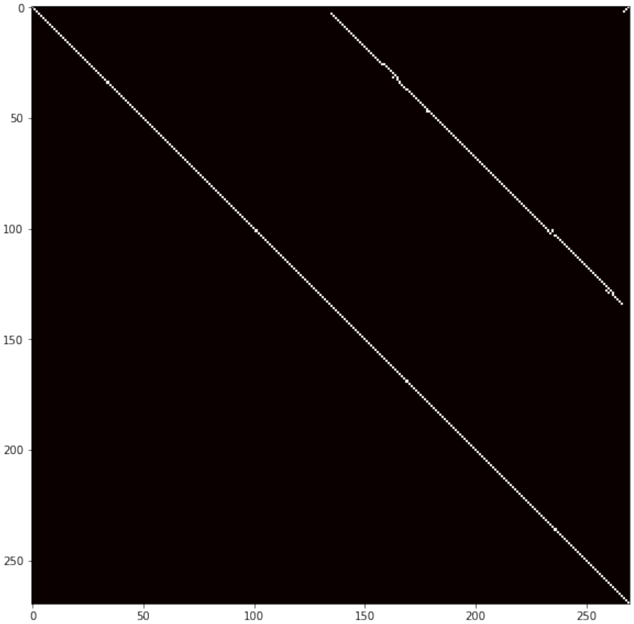

Figure 9 Ground-truth correspondence matrix for City Centre dataset

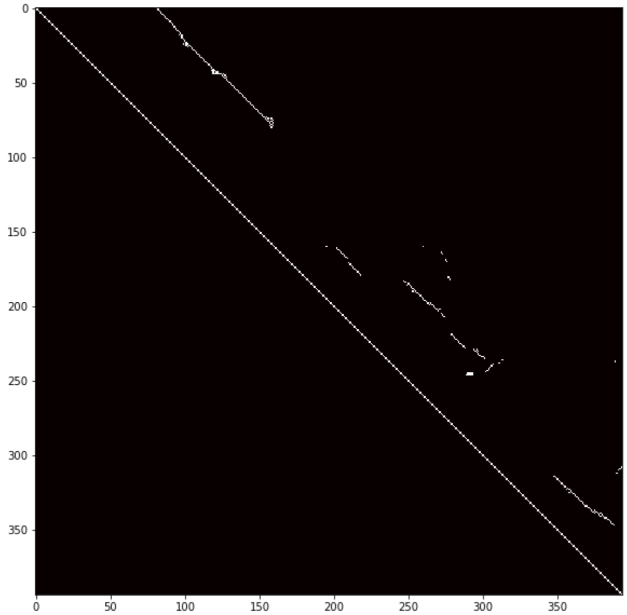

Figure 12 Ground-truth correspondence matrix for New College dataset

Figure 9 and Figure 12 show the ground-truth correspondence matrices for City Centre and New College, respectively. In this case, the element at *(i,j)* position is one if the image *i* and image *j* are considered as the same place. Otherwise, the element at *(i,j)* is zero. The total number of images in the dataset is 274 for City Centre and 394 for New College dataset. Note that the matrices are upper triangular matrices since we eliminated the lower triangular part of the matrices due to the computational redundancy when we construct the estimated correspondence matrices.

Implementation Detail

We used open source machine learning framework, Tensorflow, for the DCGAN model implementation [11]. Tensorflow is a popular and widely used open source framework in various fields ranging from the academia and industry. We implemented DCGAN model over the Tensoflow and there are some hyper-parameters for model training. Table 1 describes the hyper-parameters for model training that we experimented.

Table 1 Hyper-parameters for the model training

Parameter	Value	Parameter	Value
Batch size	128	Input height	64
Epoch	25	Input width	64
Learning rate	0.0002	Output height	64
Momentum	0.5	Output width	64

For our computing machine setting, we exploited Intel Core i7-5960X 3.00GHz CPU and Nvidia Geforce 1070 GTX GPU. Under the environment, the training time took about 8 hours.

Evaluation Measure

In order to evaluate the result of the loop closure detection, we first constructed an image-to-image correspondence matrix generated by several loop closure detection algorithms. The image-to-image correspondence matrix is a matrix containing the element of a similarity score as a probability value (between zero to one) for an image pair.

In order to evaluate the methods, we estimate a correspondence matrix with the evaluation dataset. The estimated correspondence matrix contains the elements calculated by L1 score representing a relevance between image *i* and *j* [12]. The equation for L1 score is shown in Equation 2 where the *desc(i)* represents the descriptor vector for image *i*. The L1 score value is ranging from zero to one. After obtaining the estimated correspondence matrix for each comparison method, we can calculate the precision and recall curve to evaluate the performance [13].

$$L1\_score\ (i,j) = |desc(i)/|desc(i)| - desc(j)/|desc(j)|\,| \qquad ()$$

In the process of comparing the correspondence matrices, we exploited the precision-recall curve which is the generic measure in the recognition field. The precision-recall are described by the following equations.

$$precision = \frac{true\ positives}{true\ positives + false\ positives} \qquad (1)$$

$$recall = \frac{true\ positives}{true\ positives + false\ negatives} \qquad (2)$$

At this point, the value of true positives represents an algorithm determines the image pairs as the loop closures and it is true according to the ground truth data at the same time. Table 2 describes the terms of the recognition result for the precision-recall measure. Typically, the precision-recall are inversely related and we can make the curves by adjusting the thresholds for the estimated correspondence matrix.

Table 2 Description of the recognition result

Loop closure detection		Ground truth data	
		Yes	No
Algorithm	Yes	True positive (tp)	False positive (fp)
	No	False negative (fn)	True negative (tn)

Evaluation Results

In this experiment, we compared the proposed method with several conventional methods such as the BoVW with SURF, ORB and BRISK descriptors in terms of the precision-recall curve. Figure 13 and Figure 14 illustrates the overall result for the precision-recall curves. As you can see in the figures, the proposed method shows the best result than the conventional methods. For the qualitative comparison, we described the average precision value in Table 3 and marked the best result as a bold character.

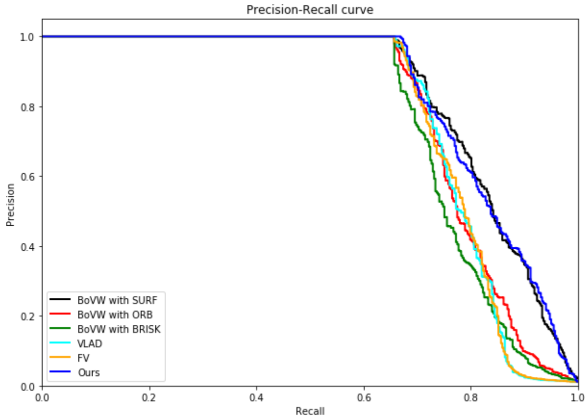

Figure 13 Precision-recall curve for City Centre dataset

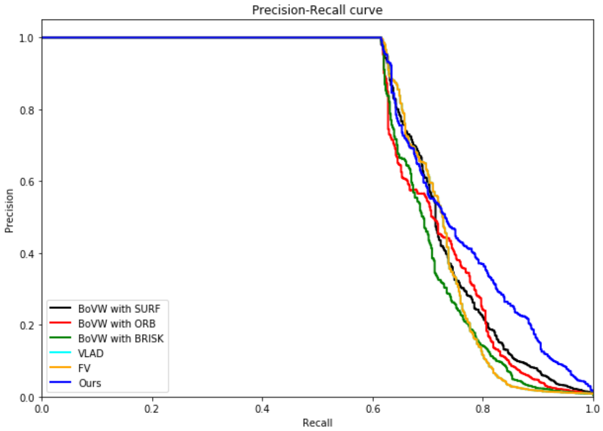

Figure 14 Precision-recall curve for New College dataset

TABLE 3 AVERAGE PRECISION

			BoVW with			
	VLAD	Fisher vector	SURF	ORB	BRISK	Ours
CC	0.78	0.78	**0.84**	0.79	0.77	**0.84**
NC	0.72	0.72	0.73	0.72	0.71	**0.76**

Conclusion

The loop closure detection is an important procedure for SLAM framework. For several decades, many researchers have tried to solve the problem of the loop closure detection via hand-crafted descriptors in images. As the advent of the deep learning approaches in the computer vision field, we proposed the loop closure detection method using the descriptors from the deep learning model. In the experiment result, we verified the proposed method shows the better result than the conventional methods in terms of the precision-recall curve. We, however, believe the method using the deep learning approaches will show the much better result with the large-scale training datasets and we left the task as a further research.

References

[1] C. Cadena *et al.*, “Past, Present, and Future of Simultaneous Localization And Mapping: Towards the Robust-Perception Age,” *IEEE Trans. Robot.*, vol. 32, no. 6, pp. 1309–1332, 2016.

[2] D. Gálvez-López and J. D. Tardós, “Bags of binary words for fast place recognition in image sequences,” *IEEE Trans. Robot.*, vol. 28, no. 5, pp. 1188–1197, 2012.

[3] F. Perronnin, J. Sánchez, and T. Mensink, “Improving the Fisher Kernel for Large-Scale Image Classification,” Springer, Berlin, Heidelberg, 2010, pp. 143–156.

[4] H. Jegou, F. Perronnin, M. Douze, J. Sanchez, P. Perez, and C. Schmid, “Aggregating Local Image Descriptors into Compact Codes,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 9, pp. 1704–1716, Sep. 2012.

[5] A. Oliva and A. Torralba, “Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope,” *Int. J. Comput. Vis.*, vol. 42, no. 3, pp. 145–175, 2001.

[6] 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. 213–221.

[7] I. Goodfellow *et al.*, “Generative Adversarial Nets,” *Adv. Neural Inf. Process. Syst. 27*, pp. 2672–2680, 2014.

[8] A. Radford, L. Metz, and S. Chintala, “Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks,” *arXiv*, 2015. [Online]. Available: http://arxiv.org/abs/1511.06434. [Accessed: 21-May-2017].

[9] B. Zhou, A. Khosla, A. Lapedriza, A. Torralba, and A. Oliva, “Places: An Image Database for Deep Scene Understanding,” *arXiv*, 2016. [Online]. Available: https://arxiv.org/pdf/1610.02055.pdf. [Accessed: 16-Mar-2017].

[10] M. Cummins and P. Newman, “FAB-MAP: Probabilistic Localization and Mapping in the Space of Appearance,” *Int. J. Rob. Res.*, vol. 27, no. 647, pp. 647–665, 2008.

[11] GoogleResearch, “TensorFlow: Large-scale machine learning on heterogeneous systems,” in *Proceedings of the 12th USENIX Conference on Operating Systems Design and Implementation*, 2016, pp. 265–283.

[12] D. Nistér and H. Stewénius, “Scalable recognition with a vocabulary tree,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2006, vol. 2, pp. 2161–2168.

[13] J. Davis and M. Goadrich, “The relationship between Precision-Recall and ROC curves,” in *Proceedings of the 23rd international conference on Machine learning - ICML ’06*, 2006, pp. 233–240.

Acknowledgement

This research was partially supported by the ‘Brain Korea 21 Plus Project’ of the Ministry of Education & Human Resources Development, Republic of Korea (ROK) [F16SN26T2205], and partially supported by the ‘Cross-Ministry Giga Korea Project’ of Ministry of Science, ICT and Future Planning, Republic of Korea (ROK). [GK17C0100, Development of Interactive and Realistic Massive Giga – Content Technology]

Author Biography

Dong-Won Shin received his B.S. in computer engineering from the Kumoh National Institute of Technology, Gumi, Korea (2013) and his M.S. in School of Information and Communications from Gwangju Institute of Science and Technology, Gwangju, Korea (2015). He is currently a Ph. D student. His research interests include 3D computer vision and machine learning.

Yo-Sung Ho received his B.S. in electronic engineering from the Seoul National University, Seoul, Korea (1981) and his Ph.D. in electrical and computer engineering from the University of California, Santa Barbara (1990). He worked in Philips Laboratories from 1990 to 1993. Since 1995, he has been with the Gwangju Institute of Science and Technology, Gwangju, Korea, where he is currently a professor. His research interests include image analysis, 3D television, and digital video broadcasting.