PAPER

Salient Feature Selection for CNN-Based Visual Place Recognition

Yutian CHEN[†a)], **Wenyan GAN**[†b)], **Shanshan JIAO**[†c)], **Youwei XU**[†d)], ***Nonmembers***,
and **Yuntian FENG**[†e)], ***Member***

SUMMARY Recent researches on mobile robots show that convolutional neural network (CNN) has achieved impressive performance in visual place recognition especially for large-scale dynamic environment. However, CNN leads to the large space of image representation that cannot meet the real-time demand for robot navigation. Aiming at this problem, we evaluate the feature effectiveness of feature maps obtained from the layer of CNN by variance and propose a novel method that reserve salient feature maps and make adaptive binarization for them. Experimental results demonstrate the effectiveness and efficiency of our method. Compared with state of the art methods for visual place recognition, our method not only has no significant loss in precision, but also greatly reduces the space of image representation.

key words: *visual place recognition, CNN, variance, feature map, binarization*

1. Introduction

Visual place recognition is an active research field in the robotic navigation and localization, especially in simultaneous localization and mapping (SLAM), which means the ability to recognize whether current place was visited before by using vision sensors [1]. Typically, a visual place recognition system is composed of three steps: image acquisition, feature extraction & encoding, and matching search, where feature extraction & encoding is the most important step that can determine the accuracy of place recognition and the efficiency of matching search.

As the most popular feature extraction methods, appearance-based methods mainly focus on hand-crafted feature extraction, such as SIFT (scale-invariant feature transforms) [2], SURF (speeded up robust features) [3] and FAB-MAP [4], [5]. However, hand-crafted feature extraction usually faces serious challenge of great appearance changes, which can be divided into two parts: one is called condition change consisting of season, weather, illumination and the movement of objects, the other is described as viewpoint change caused by different shooting angle in the same place. Motivated by the success of deep learning in computer vision, recently researchers pay attention to automatic feature extraction methods based on a CNN model [6], which have been proved to be high precision in large scale place recognition. However, CNN models will lead to a sharp increase in the space of place image representation, which cannot meet the real-time demand for matching search. How to deal with this problem is vital to visual place recognition in large scale dynamic environment.

In this paper, we make a further investigation in the feature representation extracted from CNN. We judge the validity of feature maps based on variance and propose a novel method that reserve salient feature maps and make adaptive binarization for them. In particular, this paper has the following two main contributions:

(1) A novel CNN-based feature representation method to select salient feature maps which are robustness to condition and viewpoint changes;

(2) An adaptive binarization for feature maps based on multi-thresholds segmentation;

The paper proceeds as follows. Section 2 provides a brief overview of related work in CNN for visual place recognition as well as deep hash learning in feature encoding. Our method is described in detail in Sect. 3. Sections 4 presents several benchmark datasets used in place recognition and Sect. 5 shows experimental results. Finally we conclude the paper and propose the future work in Sect. 6.

2. Related Work

2.1 CNN Methods for Visual Place Recognition

Visual place recognition based on CNN models generally falls into two categories: pre-trained CNN models and special CNN models trained over a domain-related image set.

Owing to high accuracy and good transferability, the method proposed by [6] firstly used a pre-trained CNN model to produce feature representations of place images. In both [7] and [8], the authors provided a thorough investigation of the utility on pre-trained models. On the basis of these works, methods of [9]–[11] pushed the whole place images into pre-trained CNN models associated with SVD (singular value decomposition) and sequential image retrieval, while others [12]–[14] selected saliency regions of place images through object detection and feature descriptor.

However, pre-trained CNN models are not trained for

Manuscript received May 17, 2018.
Manuscript revised August 2, 2018.
Manuscript publicized September 26, 2018.

[†]The authors are with Army Engineering University of PLA, Nanjing, Jiangsu, 210007, China.
a) E-mail: 791443626@qq.com (Corresponding author)
b) E-mail: 2372570122@qq.com (Corresponding author)
c) E-mail: 597241031@qq.com
d) E-mail: 812951570@qq.com
e) E-mail: fengyuntian2009@live.cn

DOI: 10.1587/transinf.2018EDP7175

visual place recognition so that the authors of [15] and [16] trained a new CNN model for visual place recognition respectively and illustrated that fine-tuning has a better result than training from scratch. In [17], the authors proposed a CNN layer for visual place recognition named NetVLAD, which is readily pluggable into any CNN models.

In summary, no matter which CNN model is used, the obtained features are high-dimensional, which not only occupy large storage space, but decrease the speed of matching search. In [7], the authors used AlexNet model and LSH (Locality-Sensitive Hashing) [18] to obtain binary feature representation of place images, which causes obvious drop in precision.

2.2 Deep Hash Learning for Feature Encoding

Owing to the dimensionality curse caused by deep learning especially for CNN, the combination of hash learning and deep learning has been a new trend in the field of visual place recognition.

In [19]–[21], the authors constructed similarity matrix to get the binary feature encoding of place images and then trained CNN models to fit them. In another way, the methods of [22]–[24] fine-tuned pre-trained CNN models and then made the operation of dimensionality reduction and simple binarization.

Motivated by the works above, we skip the process of training or fine-tuning CNN models and make a further investigation in the feature maps extracted from a pre-trained CNN model. Different from the methods of [12]–[14], we evaluate their feature effectiveness by variance and propose a novel method that reserve salient feature maps and make adaptive binarization for them. The detailed idea of our method will be described in next section.

3. Method

3.1 Salient Feature Selection for Place Images

AlexNet model [25] was originally proposed for ILSVRC2012 (ImageNet Large Scale Visual Recognition Challenge 2012), which was trained on the ImageNet dataset including 1.2 million images and 1000 classes. Its structure can be described as five convolutional layers followed by three fully connected layers and a softmax layer, as shown in Fig. 1.

For a pre-trained CNN model, after inputting an image, the output vector of each layer can be regarded as the feature representation of this image. In [10], the authors summarized the experimental results of [7] and [8] that feature representation extracted from the middle layers of AlexNet represented by conv3 and pool5 layers exhibits robustness in great appearance changes induced by the time of day, seasons, or weather conditions. Considering that conv3 and pool5 layers both can perform effective feature representation and the dimension of conv3 (64896) is obviously higher than pool5 (9216), in this paper we extract feature representations from the pool5 layer of AlexNet.

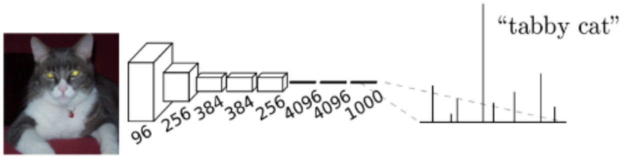


Fig. 1 AlexNet model diagram. It has a landmark significance in image classification.

We put a place image into AlexNet and finally we can obtain 256 feature maps whose size is 6 × 6 from the pool5 layer, as shown in Fig. 2. Each feature map is derived from a continuous operation of the input place image such as convolution, pooling, and activation functions. Owing to the different parameters in the operation, each feature map is not the same and can be described as the expression of a particular pattern for the place image. It is obvious that almost all the pixel values in the feature map with blue box are zero, which means invalid feature representation. In order to reduce the dimension of feature representation, we need to reserve salient feature maps such as the feature map with red box, which contains a variety of pixel values.

We use variance to evaluate the feature effectiveness of feature maps. In mathematics, variance is the measure of dispersion degree of a set of data, as shown in Eq. (1), where σ^2 is the variance of the feature map, x is the pixel value of the feature map, μ is the average of pixel values of the feature map and n is the size of the feature map defined as 36. The larger the variance of the feature map is, the higher the degree of discretization of pixel values is, the better the feature effectiveness is.

$$\sigma^2 = \frac{\sum_{i=1}^{n} (\mathrm{x_i} - \mu)^2}{\mathrm{n}} \tag{1}$$

Given a place image database D and the query image, we obtain the feature representation of each image extracted from pool5 layer and calculate the variance of each feature map of D. Formally, we obtain the variance matrix of the feature maps of D in Eq. (2), where V is the variance matrix of the feature maps of D, m is the image number of D, n is the number of the feature maps in the feature representation of an image defined as 256 and σ_{ij}^2 is the variance of the j-th feature map of the i-th image of D.

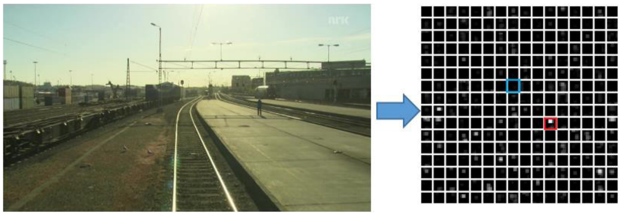


Fig. 2 The feature representation of place image extracted from the pool5 layer of AlexNet. Each small square is a feature map with 6 × 6 size and there are 256 feature maps.

$$V = \begin{bmatrix} \sigma_{11}^2 & \cdots & \sigma_{1n}^2 \\ \vdots & \ddots & \vdots \\ \sigma_{m1}^2 & \cdots & \sigma_{mn}^2 \end{bmatrix} \tag{2}$$

Feature maps represented by each column in V come from the same operation for different place images in D. Therefore, each column of V represents the feature effectiveness of the place image database for a particular feature pattern. The feature effectiveness of this column in V can be expressed as Eq. (3), where d_j is the feature effectiveness of the j-th column.

$$d_j = \sum_{i=1}^{m} \sigma_{ij}^2 \tag{3}$$

The larger d_j is, the more salient the feature maps of the j-th column are. Since n is defined as 256, we get the feature effectiveness of each column from d_1 to d_{256}. We calculate the average of them and reserve the feature maps of place image database whose column's feature effectiveness is higher than the average. Then, we do the same dimension retention for the query image. Owing to setting the average as threshold, our method retains nearly half of the feature maps.

3.2 Adaptive Binary Encoding

In order to further reduce the image representation space and improve the efficiency of similarity calculation, we try to make binarization for place image feature representation obtained in the previous step. Let P be the feature representation of the place image database and Q be the feature representation of the query image, which can be expressed in Eq. (4) and Eq. (5), where m is the number of images in the place image database, n is the feature dimension of an image, p_{ij} is the j-th feature value of the i-th image of the place image database and q_i is the i-th feature value of the query image.

$$P = \begin{bmatrix} p_{11} & \cdots & p_{1n} \\ \vdots & \ddots & \vdots \\ p_{m1} & \cdots & p_{mn} \end{bmatrix} \tag{4}$$

$$Q = [q_1 \ldots q_n] \tag{5}$$

Owing to too many zero-pixel values in the feature maps reserved, we do the binary encoding operation through selecting thresholds for each column of P. Binary operation based on whether the value is zero or not will lose a lot of feature information. Therefore, we make finer divisions of non-zero values. Apart from the zero-feature value, we choose the medians of each column of P as the thresholds expressed in Eq. (6), where T is the matrix of thresholds and t_i is the threshold of the i-th column.

$$T = [t_1 \ \ldots \ t_n] \tag{6}$$

Then we need to do the binarization for the place image database and the query image based on the thresholds. As for the feature value, it has three cases of binary assignment according to Eq. (7), where a is the binary assignment and v_i is the feature value in the i-th column of Q or P represented by q_i or p_{ji}. So far, we have achieved a method of visual place recognition by calculating the hamming distance of images based on linear matching.

$$a = \begin{cases} 00, & \mathrm{v_i} = 0 \\ 01, & 0 < \mathrm{v_i} \le \mathrm{t_i} \\ 11, & \mathrm{v_i} > t_\mathrm{i} \end{cases} \tag{7}$$

4. Evaluation with Datasets

4.1 Dataset

There are three representative large-scale datasets which are used to test our method. These datasets contain a variety of common scenarios with severe appearance changes, which is suitable to test recognition effect. Details are summarized in Table 1.

The Nordland dataset [26] is a video footage recorded in the perspective of the front train across four different seasons for 10 hours. This dataset provides severe condition changes with moderate viewpoint changes, as shown in Fig. 3.

The Gardens Point dataset [7] was taken in QUT (Queensland University of Technology). It includes three traversals of the point garden, two during the day and one at night. One of the day traversals was made on the left side while other traversals have been recorded on the right side that the night datasets were pretreated to grayscale. The dataset has both severe condition and viewpoint changes, as shown in Fig. 4.

The 24/7 Tokyo dataset [27] was captured at different times of day include daytime, sunset and night in Tokyo. It provides both severe condition and viewpoint changes, as shown in Fig. 5.

Table 1 Dataset descriptions.

Dataset	Number	Environment	Condition changes	Viewpoint changes
Nordland	1000	train journey	strong	moderate
Gardens Point	600	campus	strong	strong
24/7 Tokyo	315	urban	strong	strong

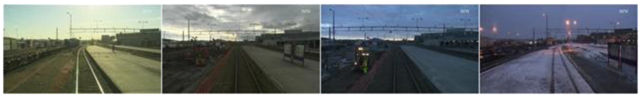


Fig. 3 Images from Nordland in four seasons at same place

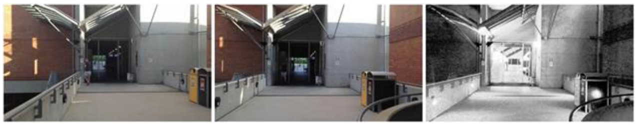


Fig. 4 Images from Gardens Point in day-left, day-right and night-right at same place

4.2 Evaluation Metrics

We analyze the performance of our method in terms of precision-recall (PR) curves and average precision (AP) score. In visual place recognition, correct matches are true positives (TP), incorrect matches are false positives (FP) and the wrongly discarded matches are false negative (FN). Precision is the proportion of selected matches which are true positive matches and recall is the proportion of true positives to the total correct matching number, which can be expressed in Eq. (8) and Eq. (9). PR curves reflects the mutual changes between precision and recall.

$$Precision = \frac{TP}{TP + FP} \tag{8}$$

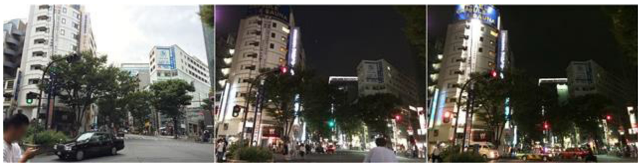


Fig. 5 Images from Japan in daytime, sunset and night at same place

$$Recall = \frac{TP}{TP + FN} \tag{9}$$

AP score summarizes the PR curve as the weighted mean of precisions achieved at each threshold, which can be expressed in Eq. (10), where p_i and r_i are the precision and recall at the i-th threshold.

$$APscore = \sum_{\mathrm{i}} (r_{\mathrm{i}} - r_{\mathrm{i}-1}) p_{\mathrm{i}} \tag{10}$$

We set standard as [7] that a match is TP when it is within ±1 : 5 frames of the ground truth and the threshold for creating the PR curves is the ratio of the distances of the best over the second best match in the nearest neighbor search.

5. Results

5.1 Robustness to Great Appearance Changes

Since our method is based on the AlexNet, we compare it with other similar existing methods for visual place recognition as well as FAB-MAP 2.0 [5]. Namely, the feature extracted by the pool5 layer of AlexNet in [8] is called Pool5,

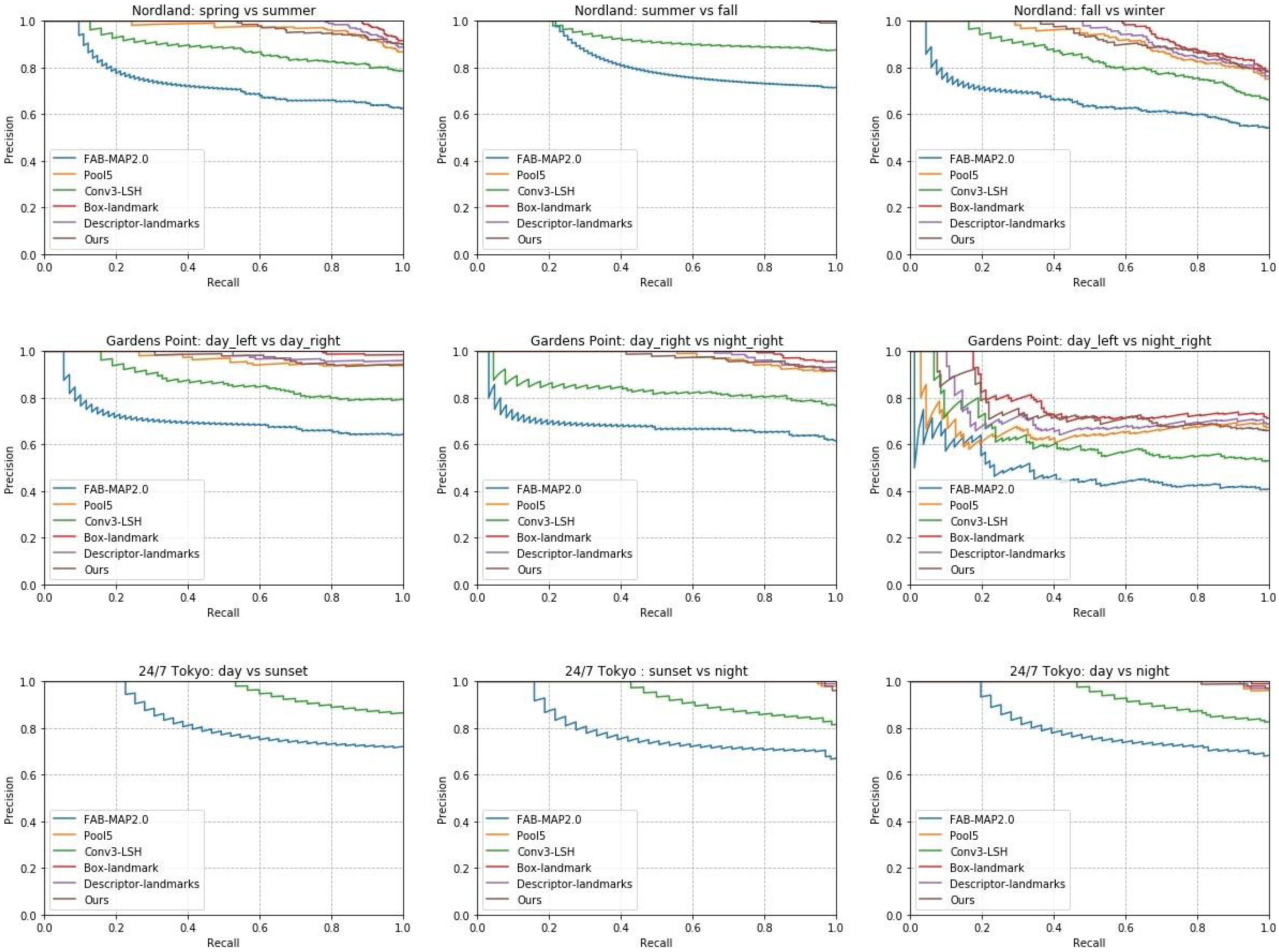


Fig. 6 PR curves for our method against other existing methods

Table 2 Condition for AP score

Test	FAB-MAP 2.0	Pool5	Conv3-LSH	Box-landmarks	Descriptor-landmarks	Ours
Nordland: spring vs summer	0.74	0.97	0.88	0.99	0.99	0.98
Nordland: summer vs fall	0.83	0.99	0.93	0.99	0.99	0.99
Nordland: fall vs winter	0.67	0.92	0.85	0.95	0.94	0.93
Garden point: day_left vs day_right	0.71	0.96	0.87	0.99	0.98	0.98
Garden point: day_right vs night_right	0.69	0.97	0.83	0.99	0.98	0.98
Garden point: day_left vs night_right	0.49	0.67	0.63	0.79	0.72	0.74
24/7 Tokyo: day vs sunset	0.83	0.99	0.93	0.99	0.99	0.99
24/7 Tokyo: sunset vs night	0.79	0.99	0.93	0.99	0.99	0.99
24/7 Tokyo: day vs night	0.81	0.99	0.94	0.99	0.99	0.99

Table 3 Condition for storage space.

Dataset	Pool5	Conv3-LSH	Box-landmarks	Descriptor-landmarks	Ours
Nordland	18432B	1024B	n×2048B	≈129792B	896B
Gardens Point	18432B	1024B	n×2048B	≈129792B	932B
24/7 Tokyo	18432B	1024B	n×2048B	≈129792B	990B

LSH-based depth feature hashing algorithm with 8192 bits used in [7] is called Conv3-LSH. The methods of local salient feature from AlexNet used in [12] and [14] are respectively called Box-landmarks and Descriptor-landmarks.

Figure 6 and Table 2 present the PR curves and AP scores generated by these methods on the datasets presents. It is evident that CNN feature representation is significantly better than hand-crafted feature representation in visual place recognition especially for large-scale dynamic environment. Compared with Box-landmarks and Descriptor-landmarks, experimental results show that our method has no significant loss in precision, which can be explained in this way that the effective global feature representation is no less than the local salient feature representation in terms of precision. Meanwhile, our method outperforms Pool5 and Conv3-LSH. This is probably due to the fact that original CNN features have too much noise information and LSH will lose more CNN feature information in the step of dimension reduction and binarization.

5.2 Condition for Storage Space

Table 3 presents the storage space generated by these CNN-based methods above. Since the double-precision floating points is used in CNN, the storage space of a feature value from CNN is 2 bytes, which is 16 times of a binary coding. Pool5 extracts the feature in the pool5 layer of AlexNet so that it has 9216 feature values. Conv3-LSH turns all the feature representation of the image into 8192 bits. The storage space of Box-landmarks and Descriptor-landmarks depends on the number of distinctive areas, where Box-landmarks encoded each distinctive area as 1024 number of double-precision floating points and Descriptor-landmarks picks 200 regions in an image by grouping all the non-zero feature values in Conv4 layer of AlexNet. Our method makes binarization based on the feature extraction from CNN so that it takes up smaller storage space.

5.3 Runtime Consideration

We evaluate the runtime performance of our method in order to realize its employments for mobile robots. For a single image, one forward pass through AlexNet costs approximately 0.26s using Caffe [28] on an NVIDIA GT940M GPU and encoding the CNN-based features used in our method on the Python platform takes about 0.48s. Thanks to the less bit of binary encoding, searching for the best matching image in the place image database with 200 images takes approximately 4.86s in our method, which is less than 12.37s of Pool5 and 5.28s of Conv3-LSH.

6. Conclusion

The application of CNN in visual place recognition achieves high precision, but it has much invalid feature information. We evaluate the information validity of feature maps obtained from the layer of CNN by variance and propose a novel method that reserve salient feature maps and make adaptive binarization for them. Compared with state of the art methods, our method not only has no significant loss in precision, but also greatly reduces the space of image representation.

At present, the step of matching search in visual place recognition is still based on linear matching, which lacks corresponding index structure for narrowing the matching range. Therefore, we will develop the further work focus on the index structure for the system of visual place recognition.

References

[1] S. Lowry, N. Sunderhauf, P. Newman, J.J. Leonard, D. Cox, P. Corke, and M.J. Milford, “Visual Place Recognition: A Survey,” IEEE Trans. Robot., vol.32, no.1, pp.1–19, 2016.

[2] D.G. Lowe, “Object Recognition from Local Scale-Invariant Features,” iccv IEEE Computer Society, p.1150, 1999.

[3] H. Bay, T. Tuytelaars, and L.V. Gool, “SURF: speeded up robust features,” European Conference on Computer Vision, Springer-Verlag, pp.404–417, 2006.

[4] M.J. Cummins and P.M. Newman, “FAB-MAP: Probabilistic Localization and Mapping in the Space of Appearance,” International Journal of Robotics Research, vol.27, no.6, pp.647–665, 2008.

[5] M. Cummins and P.M. Newman, “Appearance-only SLAM at large scale with FAB-MAP 2.0,” International Journal of Robotics Research, vol.30, no.9, pp.1100–1123, 2011.

[6] Z. Chen, et al., “Convolutional neural network-based place recognition,” Computer Science, 2014.

[7] N. Sunderhauf, et al., “On the performance of CNN features for place recognition,” 2015 IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS), pp.4297–4304, 2015.

[8] Y. Hou, H. Zhang, and S. Zhou, “Convolutional neural network-based image representation for visual loop closure detection,” 2015

IEEE Int. Conf. Inf. Autom., pp.2238–2245, 2015.

[9] M. Milford, S. Lowry, N. Sunderhauf, S. Shirazi, E. Pepperell, B. Upcroft, C. Shen, G. Lin, F. Liu, C. Cadena, and I. Reid, "Sequence searching with deep-learnt depth for condition- and viewpoint-invariant route-based place recognition," Computer Vision and Pattern Recognition Workshops IEEE, pp.18–25, 2015.

[10] D. Bai, et al., "CNN Feature boosted SeqSLAM for Real-Time Loop Closure Detection," Chin. J. Electron., vol.27, no.3, pp.488–499, 2018.

[11] S. Lowry and M. Milford, "Change removal: Robust online learning for changing appearance and changing viewpoint," IEEE International Conference on Robotics and Automation, 2015.

[12] N. Sünderhauf, et al., "Place recognition with CNN landmarks: Viewpoint-robust, condition-robust, training-free," Proc. 2010 Academy of Marketing Science (AMS) Annual Conference, pp.296–296, Springer International Publishing, 2015.

[13] P. Neubert and P. Protzel, "Local region detector + CNN based landmarks for practical place recognition in changing environments," European Conference on Mobile Robots IEEE, pp.1–6, 2015.

[14] Z. Chen, et al., "Only look once, mining distinctive landmarks from CNN for visual place recognition," IEEE/RSJ International Conference on Intelligent Robots and Systems, 2017.

[15] R. Gomez-Ojeda, M. Lopez-Antequera, N. Petkov, and J. Gonzalez-Jimenez, "Training a convolutional neural network for appearance-invariant place recognition," Computer Science, 2015.

[16] Z. Chen, A. Jacobson, N. Sünderhauf, B. Upcroft, L. Liu, C. Shen, I. Reidand, and M. Milford, "Deep learning features at scale for visual place recognition," 2017 IEEE Int. Conf. Robot. Autom. (ICRA), 2017.

[17] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "NetVLAD: CNN Architecture for Weakly Supervised Place Recognition," IEEE International Conference on Computer Vision and Pattern Recognition, pp.5297–5307, 2016, doi: 10.1109/CVPR.2016.572.

[18] A. Gionis, P. Indyk, and R. Motwani, "Similarity search in high dimensions via hashing," 8.2, pp.518–529, 1999.

[19] R. Xia, et al., "Supervised hashing for image retrieval via image representation learning," AAAI Conference on Artificial Intelligence 2012.

[20] H. Lai, et al., "Simultaneous feature learning and hash coding with deep neural networks," Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp.3270–3278, 2015.

[21] F. Zhao, Y. Huang, L. Wang, and T. Tan, "Deep semantic ranking based hashing for multi-label image retrieval," Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp.1556–1564, 2015.

[22] K. Lin, H.-F. Yang, J.-H. Hsiao, and C.-S. Chen, "Deep learning of binary hash codes for fast image retrieval," Computer Vision and Pattern Recognition Workshops IEEE, pp.27–35, 2015.

[23] J. Guo and J. Li, "CNN based hashing for image retrieval," arXiv preprint arXive: 1509.01354, 2015.

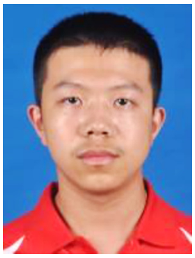
[24] K. Zhou, Y. Liu, J. Song, L. Yan, F. Zou, and F. Shen, "Deep Self-taught Hashing for Image Retrieval," ACM International Conference on Multimedia ACM, pp.1215–1218, 2015.

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

[26] P. Neubert, N. Sünderhauf, and P. Protzel, "Superpixel-based appearance change prediction for long-term navigation across seasons," Robotics & Autonomous Systems, vol.69, no.1, pp.15–27, 2015.

[27] A. Torii, R. Arandjelovic, J. Sivic, M. Okutomi, and T. Pajdla, "24/7 place recognition by view synthesis," IEEE Conference on Computer Vision and Pattern Recognition IEEE Computer Society, pp.1808–1817, 2015.

[28] Y. Jia, et al., "Caffe: Convolutional Architecture for Fast Feature Embedding," Proc. 22nd ACM Int. Conf. Multimedia, pp.675–678, 2014.



Yutian Chen is master candidate in Army Engineering University of PLA. His main research fields are deep learning and image processing.

Wenyan Gan is associate professor in Army Engineering University of PLA. Her main research fields are artificial intelligence and data mining.

Shanshan Jiao is doctor candidate in Army Engineering University of PLA. Her main research fields are machine learning and cloud model.

Youwei Xu is master candidate in Army Engineering University of PLA. His main research fields are data mining and effectiveness evaluation.

Yuntian Feng is doctor candidate in Army Engineering University of PLA. His main research fields are deep learning and natural language processing.