Visual Place Recognition Using an Unsupervised CNN Approach

Yuan Miaolong, Liu Ning, Zhang Jie and Wang Zhenbiao
Advanced Remanufacturing & Technology Centre (ARTC)
Agency for Science, Technology and Research (A*STAR)
3 Cleantech Loop, #01/01 CleanTech Two, Singapore 637143
Republic of Singapore
{myuan, liun, zhang_jie and wangzb}@artc.a-star.edu.sg

***Abstract*—Visual place recognition (VPR) plays a key role in many applications, such as mobile robot navigation and localization, location intelligence in smart warehouse etc. Convolutional neural network (CNN) based deep learning (DL) solutions have been demonstrated to outperform traditional bag-of-words (BoWs) solutions for VPR. However, CNN-based DL solutions often require a large amount of high-quality annotated data, where manual annotation is time-consuming and even difficult in some scenarios. In this paper, we propose an unsupervised CNN approach for VPR. First, we use Oriented FAST and Rotated BRIEF (ORB) feature detector, bag-of-words (BoWs) and Hough space verification methods to automatically generate image clusters. Each image cluster is assigned with a numbered keyframe representing one place. Images relating to its keyframes will be stored as a visual experience with relevant available information, such as pose of current keyframe, in the corresponding cluster. Next, the image clusters are used as training datasets to train various CNN models for VPR. Experiments have been conducted on a public dataset and competitive results have been achieved using the proposed unsupervised CNN method.**

***Index Terms*—Visual place recognition, ORB, bag of words, Hough space verification, convolutional neural network (CNN), unsupervised learning**

I. Introduction

Visual place recognition (VPR) plays a key role in many applications, such as augmented reality, visual simultaneous localization and mapping (SLAM) [1]. However, it is an open and challenging task, as the appearance of the environments can change drastically, such as various weather and seasons, lighting conditions or moving objects in dynamic environments [2]–[6]. Another issue is that multiple places in an environment may look similar, which is a well-known perceptual aliasing problem [7]. The key component of visual place recognition is how to describe a specific place against various perceptual changes.

There are two major types of image descriptors commonly investigated for VPR tasks. Conventional compact image descriptors are often constructed by aggregating a set of local features into a global vector to represent an image for place recognition [8]–[10]. In the past decade, convolutional neural network (CNN) solutions have been demonstrated to significantly outperform traditional solutions for place recognition and more robust to changing lighting conditions [15], [17], [24]. However, CNN-based solutions often require a large amount of annotated data, where a high quality of annotation is required to train a CNN model for place recognition. Manual annotation is time-consuming and even difficult in some environments. For example, when an autonomous vehicle is traveling along a road, it is hard to choose the places to be discriminative in some scenarios. They still look similar even the distance between two places maybe far away. For another example, it is prone to labelling errors as human may be confused to label a large-scale warehouse for a long time, leading to unreliable training results. The key issue thus becomes how to identify and describe a keyframe representing a place in a fully automatic way.

In this paper, we propose an unsupervised deep learning approach for visual place recognition by combing traditional BoWs solution with deep CNN techniques. The proposed solution is to automatically generate data clusters for deep CNN model training. First, we use local features, BoW and Hough verification (referred to as BoW+HV) techniques to generate image clusters (i.e., each cluster has a keyframe) while keeping the other frames associated with the corresponding keyframe for training data. Next, the image clusters can be used as labels, together with the training samples of each keyframe to train CNN models for place recognition.

There are three major advantages of the proposed solutions. First, with no human annotations or no metric information (i.e., GPS or IMU), we can generate training datasets with automatically generated cluster labels and use the advanced CNN techniques to achieve high performance of place recognition. Second, by adjusting related parameters, we can choose different sizes of training data samples with the most representative places for training CNN models. Third, if users want to specify each place by a name, only the cluster labels are required to be named instead of labeling each image in the whole training dataset. This is especially efficient when there are a large number of places with a huge size of training dataset.

The rest of the paper is organized as follows. Section II reviews works relevant to the proposed approach. The details of the proposed method are provided in Section III. Section IV

[§]Equal contribution

describes the implementation details and experimental results. Conclusion is given in Section V.

II. Literature Review

While there are a lot of advanced techniques proposed in VPR, all relevant methods can be generally classified into two categories: (1) Traditional computer vision approaches using BoWs model on handcrafted local descriptors to produce image-level features [8]–[10], [14], [20], [22] and (2) CNN-based image abstract feature extraction using deep CNN model for visual place recognition [15], [17], [24].

In the traditional VRP approaches, compact image descriptors are often constructed by aggregating a set of local features into a vector to represent an image using the well-known bag-of-words (BoWs) technique. Over a decade, BoW based solutions are widely used for VPR [8], [9]. The BoW method represents an image as a set of visual words, which are obtained by quantizing local feature spaces [8]. It usually considers occurrences of visual features of each image. Lopez and Tards [10] propose a fast place recognition using bag of binary visual words obtained from FAST+BRIEF [18], [19]. Based on the work of [10], Artal et al. propose an impressive complete vision-based SLAM system using the ORB detector and a hierarchical bag of a visual binary words [20], providing a fast place recognition solution for VRP and making it advantageous to be employed in real robotics applications that require real-time performance.

Khan and Wollherr propose an incremental bag of binary words approach for appearance-based place recognition [21]. It provides an online, incremental formulation of binary vocabulary generation without a prior vocabulary learning phase for loop closure detection. Fidalgo and Ortiz propose a novel visual place recognition approach based on a hierarchical strategy [22], where images with similar visual properties were first grouped using the PHOG descriptor [13] and in each group, binary image features were employed to search for the most likely images inside these groups. This two-level hierarchical scheme can reduce the search space and maintain high accuracy. However, BoW solutions ignore their spatial relationships during word assignments, thus leading to a large percentage of outliers. Therefore, a subsequent spatial verification is required to filter out the unreliable matches which are not geometrically consistent [32].

In the past decade, CNN solutions have been demonstrated to significantly outperform traditional Bag-of-words solutions for VPR, especially under changing lighting conditions. The pioneer works to explore CNN descriptors in the VPR problem can be found in [15], [17], [23], [24], where the authors employed pre-trained networks and explored the capabilities of various layers for VPR tasks. They have demonstrated that the last few layers have more specific information and therefore were more robust under conditions of large viewpoint variances, while middle layers are less affected by appearance.

Hou et al. [23] also investigate the performance of CNN middle layers as image descriptors and have validated that they are faster to compute and have similar or better performance compared to handcrafted image descriptors. Some authors also have trained their networks instead of using pre-trained models, for the VPR task. [24] generated a large dataset of places and trained a network for VPR problem and significantly outperformed other approaches. Arandjelovic et al. [25] design an impressive CNN architecture, NetVLAD, which is a new generalized VLAD layer. It is designed to be effectively plugged into the CNN architecture and amenable to training via back-propagation for VPR tasks. Their results on some very challenging datasets were remarkable, significantly outperforming state-of-the-art works based on pre-trained CNNs. Chen et al. employ a late convolutional layer as a landmark detector and an earlier one to create local descriptors to match the detected landmarks for VPR and improved place recognition performance have been achieved under strong viewpoint and condition variations [15]. Kaliq et al. [26] develop a lightweight visual place recognition approach, capable of achieving high performance with low computational cost, and feasible for mobile robotics under significant viewpoint and appearance changes. This is achieved by detecting CNN-based regional features and combining them with the VLAD encoding methodology, adapted specifically for computation-efficient and environment invariant-VPR problem. Xin et al. investigate how to use CNN features to localize discriminative visual landmarks that positively contribute to the similarity measurement with a Landmark Localization Network (LLN). The proposed approach achieves superior performance against state-of-the-art methods [27]. Recently, Wang et al. propose a novel holistic place recognition model based on vision Transformers with multi-level self- attention aggregation mechanism [6]. This model allows end-to-end training achieves competitive results compared with state-of-the-arts while still maintaining low computational time.

While CNN techniques have been demonstrated significant performance in VPR, a large amount of high quality annotated data are required to train a CNN model for place recognition. Manual annotation is time-consuming and even difficult in certain scenarios. Unsupervised CNN learning has attracted researchers in computer vision and robotics communities. Some impressive works have been reported in literature. Sudeep and Leonard [28] develop a self-supervised approach to VPR for mobile robots in which the task of VPR is cast as a metric learning problem to generate positive and negative training datasets using GPS information. However, the proposed solution is not applicable in GPS-denied environments. Mathilde et al. [29] propose Deep Clustering, which is the first attempt to unsupervised training of CNN on large dataset like ImageNet. It jointly learns CNN feature clustering and models. However, the parameters in the classifier cannot be inherited from the previous epoch and need to be initialized randomly, resulting in instability and representation corruption in training. Radenovic et al. [30] propose a fine-tune CNN model for image retrieval on a large collection of unordered images in a fully automated manner by reconstructing 3D models using structure-from-motion (SfM) methods in order to guide the selection of the hard-positive and hard-negative

training data. However, this method is more applicable in image retrieval rather than in VPR, as the 3D reconstruction of a large navigating environment with SfM technique is far from applicability, especially for an large-scale environment.

Inspired by the existing observation, we propose an unsupervised CNN approach for VPR. The proposed solution aims to design a strategy to automatically generate image clusters for training datasets without manual annotations by combining traditional bag of words solutions in computer vision with advanced deep CNN models in order to achieve better performance of place recognition. It is worth noting that the proposed method can also be generalized well to various pre-trained models, such as VGG, YOLO and ResNet etc.

we use BoW+HV solutions to automatically gener- ate image clusters as training datasets with automatically assigned labels, as shown in Figure 1(a)

III. Methodology

A. Overall Architecture

Figure 1 shows the architecture of the proposed unsupervised CNN approach for visual place recognition. It includes two major tasks:

1) First, we use BoW+HV solutions to automatically generate image clusters as training datasets with automatically assigned labels, as shown in Figure 1(a).
2) Second, we use the outputs of the image clusters from the BoW-HV solutions as labeled training datasets to train a deep CNN model and output the CNN feature for VPR, as shown in Figure 1(b).

To automatically generate image clusters, feature matching is a core component for place recognition. However, the bag of words technique usually considers occurrences of visual features of each image while ignoring their spatial relationships during word assignments, thus leading to a large percentage of outliers. Therefore, we use Hough verification in order to filter out unreliable images in each image cluster. By exploiting local feature shapes, a correspondence between two images can be related by a similarity or an affine transformation [8]. The transformations of each group of correspondences in a Hough transformation space exhibit similarity property. By using Hough voting, most of them likely fall into the same bins and the most promising transformations can thus be identified directly in the Hough space [31]. Furthermore, we employed a hypothesize-and-verify scheme [32] for second-step spatial verification to filter out more unreliable images in each image cluster.

B. Hough Transformation

An image is represented by a set P of local ORB features [33]. Given an ORB feature $p \in P$ with its position (x_p, y_p), scale s_p, orientation θ_p and its visual word $u(p)$, its local structure is formatted as:

$$F(p) = \begin{bmatrix} M(p) & \mathbf{t}(p) \\ \mathbf{0}^T & 1 \end{bmatrix} \tag{1}$$

where $M(p) = s_p R(p)$. $R(p) = \begin{bmatrix} \cos(\theta_p) & -sin(\theta_p) \\ sin(\theta_p) & cos(\theta_p) \end{bmatrix}$ is a 2×2 rotation matrix and $\mathbf{t}(p) = (x_p, y_p)$ stands for positions, respectively. Given two images P and Q, a pair of ORB features $p \in P$ and $q \in Q$ are considered as a correspondence $c = (p, q)$ if $u(p) = u(q)$. An affine transformation mapping p to q can be described as:

$$F(c) = F(p)F(q)^{-1} = \begin{bmatrix} M(c) & \mathbf{t}(c) \\ \mathbf{0}^T & 1 \end{bmatrix} \tag{2}$$

where $M(c) = s(c)R(c)$ and $\mathbf{t}(c) = \mathbf{t}(p) - M(c)\mathbf{t}(q)$ and $s_c = s_q/s_p$, $R(c) = R(q)R(p)^{-1}$ denote relative scaling and rotation from p to q, respectively. $F(c)$ can be written as a 4D transformation vector by

$$F(c) = (x(c), y(c), s(c), \theta(c)) \tag{3}$$

where $\theta(c) = \theta_q - \theta_p$ and $((x(c), y(c))^T = \mathbf{t}(c)$. Eq. [3] denotes that $c = (p, q)$ can be transformed into a Hough space.

We follow [31] and [32] to quantize $F(c)$ for $c = (p, q)$, independently and partition it at L resolutions $l_0, l_1, ..., l_{L-1}$. At each level, the four parameters in 3 are quantized into n_x, n_y, n_s and n_θ bins. We consider the score of each bin at the finest level, similar to [31] and [32]:

$$score(b_{(0i)}) = \sum_{l=0}^{L-1} 2^{-\alpha l} \|b_{li}\| \tag{4}$$

$\|b_{li}\|$ denotes the count of the i^{th} bin at level l. $2^{-\alpha l}$ denotes the factor to the voting score at level l.

Hough voting can identify potential groups of correspondences Eq. [1-4]. Next, we use the vote-and-verify strategy [32] to check geometric consistency in order to identify whether two images are matched according to a pre-defined matching threshold.

C. Image Clustering for Training

For visual place recognition, the training images often exist in the format of image sequences and image streams. Thus, the difference between the current input image and the subsequent image might be very similar. Hence, we normally define an interval value δ_{ep} to sample images. For example, for an image stream with 15FPS, we can set the value of δ_{ep} as 15, meaning that only one image is sampled for each second. For each place, there is a keyframe relating the current incoming image. Each place corresponds an image cluster and we denote each image in the cluster as a visual experience. Each visual experience will store relevant image information, such as image index, keypoint information, pose information if available etc. The image clustering approach is described in **Algorithm 1**.

It is worth noting that based on various value of the interval time δ_{ep}, the clusters can be sparse or dense. The smaller the value of the interval, the denser the cluster. In real scenarios, for example, for robot localization/re-localization, this depends on the navigating speed of the robot. If the cluster is dense, this means it will have more places (clusters) to be trained,

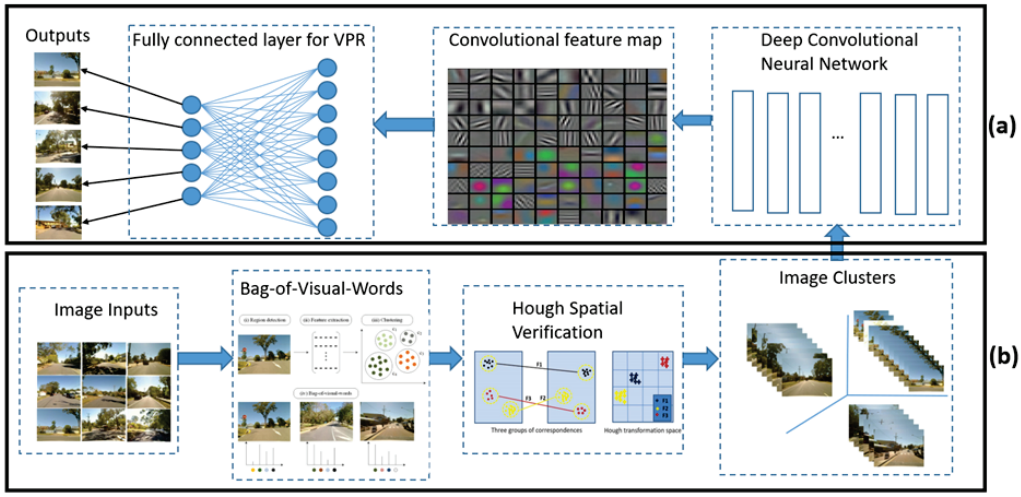


Fig. 1. The Architecture of the proposed unsupervised CNN approach for VPR. It includes two major components. First, Bow+HV is used to process the input sequences to generate labeled training clusters, as shown in Figure 1(a). Second, the automatically labeled image clusters are used to train a deep CNN model with the corresponding generated labels, as shown in Figure 1(b).

Algorithm 1 Automatic Image Clustering

1. Initialization: initialize a Boolean variable $bLoop = false$ and sampling interval δ_{ep}.
2. For each image in an image sequence, use the method in Section III(B) to check if current incoming image is matched with the existing keyframes.
   - If $bLoop = false$, it is considered as a new place and a new image cluster is created and this incoming frame is assigned as a keyframe and its subsequent sampled images within δ_{ep} are stored in this cluster as a new visual experience.
   - Otherwise, it is considered as a re-visited place. We save the current image and relevant information as a visual experience in the existing cluster
5. For other image sequences, we repeat Step 2.

thus resulting more accurate localization. However, it will also increase the computation requirements.

Furthermore, it should be better if each image cluster includes images captured from different viewpoints and different lighting conditions, thus the trained CNN models would be more reliable and have better ability to VPR. This can be further realized in several ways:

1) Generate different vocabularies using images from different lighting conditions and different viewpoints. Then we can use switched vocabularies technique and above VPR technique to recognize the same places under changing lighting conditions for each image cluster.
2) If GPS information is available for outdoor scenarios, or AMCL pose information (e.g., laser) is available for a mobile robot for indoor scenarios, we can also use this pose information to include more training images with different lighting conditions into each image cluster.
3) Use generative adversarial networks (GANs) [34] and image augmentation techniques to generate more synthetic training images for each image cluster.

D. Deep CNN for VPR

Once we have automatically generated the training datasets using the methods discussed in Section III(B) and III(C), we can therefore feed automatically generated clusters as labelled training datasets into a deep CNN model.

A huge amount of significant achievements using DL techniques for image recognition have been reported [35]. Roughly they can be divided into two major categories. One can train a deep CNN model for place recognition from scratch using customer datasets to meet their specific applications. Training a deep CNN model is a non-trivial task. One needs to carefully design proper network architectures including the number of layers, optimizer selection, baseline setup, model validation etc. An easier way is to use the well-known pre-trained models [35], such as VGG, AlexNet, YOLO, SSD, ResNet, Inception, Faster-RCNN etc. The user can fine-tune the pre-trained models by training the last few layers of these models while freezing the remaining layers, making it much easier and achieving promising performance.

The proposed approach is applicable to any pre-trained models and what we need to do is to re-train (i.e., fine-tune) a few layers for image classification (after labeling, the visual place recognition is actually a task of image classification). We can also define our own loss function to re-train the CNN model for VPR.

IV. Experiments and Analysis

The proposed unsupervised CNN place recognition solution has been conducted on a 2.4GHz Intel CPU with GPU NVIDIA Quadro T2000. We validated the effectiveness of the proposed solution on a public dataset which is a single route

TABLE I
NUMBERS OF EXTRACTED IMAGE FOR QUT DATASET

Datasets (Date & Time)	Number of extracted images
180809_0845	4227
190809_1545	4285
190809_1410	4363
210809_1000	4092
210809_1210	3614

TABLE II
RECOGNITION RATE COMPARISON WITH TRADITIONAL BOW+HV

Datasets tested	Unsupervised CNN	BoW+HV
180809_1545	**18.97**	1.07
190809_0845	19.62	**24.91**
190809_1410	**21.20**	0.12
210809_1000	**24.56**	22.48
210809_1210	**28.11**	11.26
Average	**22.49**	11.97

through the suburb of St Lucia, Queensland, Australia [32]. The St Lucia Multiple Times of Day datasets were collected with a forward facing webcam, attached to the roof of a car, through the suburb. The route was traversed ten times during multiple days to capture the difference in appearance between early morning and late afternoon. 640480 RBG images were captured at 15 FPS.

A recording ratio of 5 was set, such that every 1 in every 5 frames would be extracted for the video of 15 fps (3 frames were extracted every second in our experiments). Assuming an average driving speed of 50km/h or 13.8889m/s, each frame extracted would have a difference of roughly 4.2696m between them. The quantities of the extracted images are shown in Table 1. The first set 100909_0845 was selected to generate image clusters using BoW and Hough spatial verification (BoW+HV) based loop detector algorithm. Initially, an interval of 12 images were set to give an estimated coverage of 55.6m of images within a cluster.

After running the algorithm, and using post-processing to remove empty folders, it was found that there were a total of 1333 image matches (out of 4227) in 158 image clusters, with an average of 8.4367 number of images per cluster, as shown in Fig. 2(a) above. The interval value was thus increased from 12 to 50, with attempts to increase the average number of images in each cluster. With the same matching and post-processing algorithms, there were 1312 image matches (out of 4227), but this time in 48 image clusters, with the average increased to 27 images per cluster.

Further removing folders with image quantities less than 20, we have the distribution in Fig. 2(b), with 1107 image matches (out of 4227) in 29 clusters, with an even higher average number of images per cluster of 38.172. These 29 image clusters were used to train a CNN model, and the other datasets were used in testing to obtain prediction results.

In this experiment, we designed a simple CNN architecture including three 2D convolutional layers with an ReLU operation sequentially, followed by a fully connected network (FCN) to output the place recognition. The size of filters were set to 16, 32, 64 for the three CNN layers, respectively. The size of kernel was set to 2. Images were resized to 224×224. It is worth noting that complicated CNN networks and parameter tuning are not discussed in this work. We focus on validating the effectiveness of the unsupervised strategy even with a simple CNN network.

It is observed from Table 2 that the CNN method does consistently better than the BoW-HV method, for most of the given dataset used for testing except for one instance. The average match rate is 10.52% higher for CNN than the BoW method. This validates that the CNN method has better performance on place recognition that the traditional BoW+HV method.

CNN has another advantage that it is more robust to changing lighting conditions than traditional BoW+HV. Fig. 3(a)-Fig. 3(d) show four examples that CNN is able to recognize places under different lighting conditions while BoW+HV is not able to recognize the related places.

V. CONCLUSION

In this paper, we proposed an unsupervised deep CNN approach for visual place recognition. We used an efficient traditional BoWs with a vote-and-verify strategy to automatically generate image clusters, proving an incredibly efficient solution for unsupervised image clustering and labelling. The generated clusters can be used as labelled training datasets for fine-tuning different deep CNN models or training own CNN model for competitive performance of place recognition. We had conducted experiments on a public datasets to validate the effectiveness of the proposed solution. The proposed solution can be used in robot navigation localization, such as lost recovery and location intelligence in smart warehouse.

REFERENCES

[1] V. Shim, B. Tian, M. Yuan, H. Tang and H. Li, Direction-driven navigation using cognitive map for mobile robots, IEEE/RSJ IROS, 2639-2646, 2014.

[2] T. Barros, R. Pereira, L. Garrote, C. Premebida and U. J. Nunes, Place recognition survey: An update on deep learning approaches, arXiv:2106.10458v3, 2021

[3] P. Yin et al., General Place Recognition Survey: Towards the Real-world Autonomy Age, rXiv:2209.04497v1, 2022.

[4] C. Linegar, W. Churchill and P. Newman, Work Smart, Not Hard: Recalling Relevant Experiences for Vast-Scale but Time-Constrained Localisation, IEEE ICRA, 90-97, 2015.

[5] M. Wulfmeier, A. Bewley and I. Posner, Addressing Appearance Change in Outdoor Robotics with Adversarial Domain Adaptation, IEEE/RSJ IROS, 1551-1558, 2017.

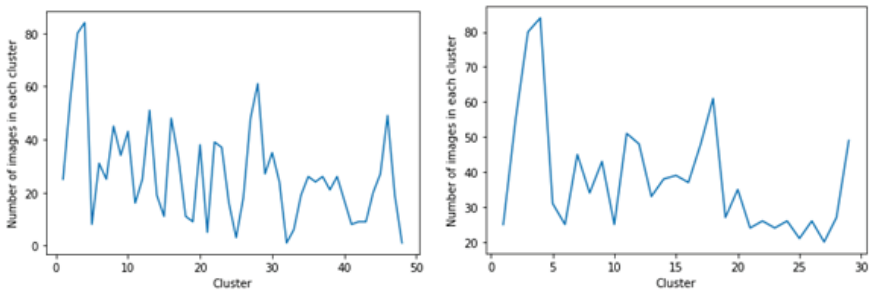
[6] R. Wang, Y. Shen, W. Zuo, S. Zhou and N. Zheng, TransVPR: Transformer-Based Place Recognition with Multi-Level Attention Aggregation, arXiv:2201.02001, 2022.

[7] S. Lowry, N. Sunderhauf, P. Newman, J. J. Leonard, D. Cox, P. Corke and M. J. Milford. Visual place recognition: a survey, IEEE Transactions on Robotics, 32(1), 1-19, 2016.

[8] J. Philbin, O. Chum, M. Isard, J. Sivic and A. Zisserman, Object retrieval with large vocabularies and fast spatial matching, IEEE CVPR, 1-8, 2007.

[9] L. Fei-Fei and P. Perona, A Bayesian hierarchical model for learning natural scene categories, IEEE CVPR, 524-531, 2005.

[10] D. Galvez-Lopez and J. D. Tards, Bags of binary words for Fast place recognition in image sequences, IEEE Transactions on Robotics, 28(5), 1188-1197, 2012.



(a) (b)

Fig. 2. Distributions of image quantities in each image cluster.

(a) (b)

(c) (d)

Fig. 3. Examples which can be recognized with a CNN model but can not be recognized using traditional BoW+HV solution due to varied lighting conditions.

[11] R. Mur-Artal, J. M. M. Montiel and Juan D. Tards, ORB-SLAM: A versatile and accurate monocular SLAM system, IEEE Transactions on Robotics, 31(5), 1147-1163, 2015.

[12] D. Schlegel and G. Grisetti, Visual localization and loop closing using decison trees and binary features, IEEE/RSJ IROS, 4616-4623, 2016.

[13] N. Dalal and B. Triggs, Histograms of oriented gradients for human detection, IEEE CVPR, 886-893, 2005.

[14] E. G. Fidalgo and A. Ortiz, Hierarchical place pecognition for topological mapping, IEEE Transactions on Robotics, 33(5), 1061-1074, 2017.

[15] Z. Chen et al., Convolutional neural network based place recognition, arXiv:1411.1509, 2014.

[16] N. Snderhauf et al. On the performance of convent features for place recognition, arXiv:1501.04158, 2015.

[17] S. Hausler, S., M. Xu, M. Milford and and T. Fischer, Patch-netvlad: Multi-scale fusion of locally-global descriptors for place recognition, IEEE CVPR, 1414114152, 2021.

[18] E. Rosten, R. Porter and T. Drummond, Faster and better: a machine learning approach to corner detection, IEEE TPAMI, 32(1): 105119, 2006.

[19] M. Calonder, V. Lepetit, M. Ozuysal, T. Trzinski, C. Strecha and P. Fua, BRIEF: computing a local binary descriptor very fast, IEEE TPAMI, 34(7): 12811298, 2011.

[20] R. Mur-Artal, J. M. M. Montiel and Juan D. Tards, ORB-SLAM: A versatile and accurate monocular SLAM system, IEEE Transactions on Robotics, 31(5), 1147-1163, 2015.

[21] S. Khan and D. Wollherr, IBuILD: Incremental bag of Binary words for appearance based loop closure detection, IEEE ICRA, 5441-5447, 2015.

[22] D. Schlegel and G. Grisetti, Visual localization and loop closing using decison trees and binary features, IEEE/RSJ IROS, 4616-462, 2016.

[23] Y. Hou, H. Zhang and S. Zhou, Convolutional neural network-based image representation for visual loop closure detection, IEEE ICIA, 2238-2245, 2015.

[24] N. Sunderhauf et al., Place recognition with convent landmarks: Viewpoint-robust, condition-robust, training-free, Robotics: Science and Systems, 2015.

[25] R. Arandjelovic et al., NetVLAD: CNN architecture for weakly supervised place recognition, IEEE CVPR, 5297-5307, 2016.

[26] A. Khaliq et al., A Holistic Visual Place Recognition Approach using Lightweight CNNs for Severe ViewPoint and Appearance Changes, arXiv:1811.03032 2018.

[27] Z. Xin et al., Localizing Discriminative Visual Landmarks for Place Recognition, arXiv: 1904.06635, 2019.

[28] S. Pillai and J. J. Leonard, Self-Supervised Visual Place Recognition Learning in Mobile Robots, arXiv:1905.04453, 2019.

[29] M. Caron, P. Bojanowski, A. Joulin and M. Douze, Deep Clustering for Unsupervised Learning of Visual Features, arXiv:1807.05520, 2018.

[30] F. Radenovic, G. Tolias and O. Chum, Fine-Tuning CNN Image Retrieval with No Human Annotation, IEEE TPAMI, 41(7), 1655-1668, 2019.

[31] Y. Avrithis and G. Tolias, Hough pyramid matching: speeded-up geometry re-ranking for large scale image retrieval, Internation Journal of Computer Vision, 107(1), 1-19, 2014.

[32] J. L. Schonberger, T. Price, T. Sattler, J. M. Frahm M. Pollefeys, A vote-and-verify strategy for fast spatial verification in image retrieval, Asian Conference on Computer Vision, 321-337, 2016.

[33] E. Rublee, V. Rabaud, K. Konolige and G. Bradski, ORB: An efficient alternative to SIFT or SURF, IEEE CVPR, 2564-2571, 2011.

[34] I. J. Goodfellow et al., Generative adversarial networks, arXiv:1406.2661, 2014.

[35] https://github.com/rasbt/deeplearning-models, accessed on 12 Jan 2023.