Visual Place Recognition with CNNs: From Global to Partial

Zhe Xin*†, Xiaoguang Cui*, Jixiang Zhang*, Yiping Yang* and Yanqing Wang*
* Institute of Automation, Chinese Academy of Sciences, Beijing, China
† University of Chinese Academy of Sciences, Beijing, China
e-mail: {xinzhe2015, xiaoguang.cui, jixiang.zhang, yiping.yang, yanqing.wang }@ia.ac.cn

***Abstract*—Visual place recognition is one of the most challenging problems in computer vision, due to the large diversities that real-world places can represent. Recently, visual place recognition has become a key part of loop closure detection and topological localization in long-term mobile robot autonomy. In this work, we build up a novel visual place recognition pipeline composed of a first filtering stage followed by a partial reranking process. In the filtering stage, image-wise features are utilized to find a small set of potential places. Afterwards, stable region-wise landmarks are extracted for more accurate matching in the partial reranking process. All global and partial image representations are derived from pre-trained Convolutional Neural Networks (CNNs), and the landmarks are extracted by object proposal techniques. Moreover, a new similarity measurement is provided by considering both spatial and scale distribution of landmarks. Compared with current methods only considering scale distribution, the presented similarity measurement can benefit recognition precision and robustness effectively. Experiments with varied viewpoints and environmental conditions demonstrate that the proposed method achieves superior performance against state-of-the-art methods.**

***Index Terms*—visual place recognition, localization, convolutional neural networks, long-term environment**

I. Introduction

Visual place recognition is the process of identifying images that belong to the same location [18]. With the increasing focus on long-term mobile robot and autonomous driving applications, visual place recognition has become a key part of loop closure detection and topological localization. However, as robots operate in long-term environment, the characteristics of places change gradually with different viewpoints, environmental conditions and dynamic objects, as shown in Fig. 1. All these factors bring great difficulties for visual place recognition and make it still an extremely challenging problem to solve.

Conventional place recognition approaches are mainly based on hand-crafted features, including local keypoints such as SIFT [17], SURF [4] and holistic descriptors such as GIST [20], HOG [7]. These approaches can work effectively in the static scene, but always fail in the challenging environment.

Currently, Convolutional Neural Networks (CNNs) have been used as robust feature generator for place recognition in challenging environments. Sunderhauf [25] showed that features extracted from the middle layers were robust against appearance changes, and features extracted from the top layers were more invariant to viewpoint changes. [26] and [21] extracted region proposals as landmarks and matched CNN features of landmarks to recognize places, achieving satisfactory performance for viewpoint change problems.

978-1-5386-1842-4/17/$31.00 ©2017 IEEE

Fig. 1. The same place during different seasons of a year, (a) summer, (b) fall, (c) winter. The changing characteristics of the environment include illumination, weather, dynamic objects and viewpoint.

In this work, we build up a novel visual place recognition pipeline composed of a first filtering stage followed by a partial reranking process. In the filtering stage, image-wise features are utilized to find a small set of potential places. Afterwards, stable region-wise landmarks are extracted for more accurate matching in the partial reranking process. With the use of filtering stage, more computationally expensive reranking process can just be operated on potential places. Both global and partial features are generated from a CNN model pre-trained on ImageNet [24], and no further environment-specific training is required in the proposed method. Moreover, a new similarity measurement is provided by considering both spatial and scale distribution of landmarks. Compared with the state-of-the-art method [26] only considering scale distribution, the presented similarity measurement can benefit recognition precision and robustness effectively.

The main contributions of this work are:

- A robust place recognition system is provided for challenging appearance and viewpoint changes, requiring no environment-specific training.
- A new pipeline is composed of a first filtering stage followed by a partial reranking process, making use of both image-wise and region-wise information.
- An efficient filtering process is performed to sort out potential places, greatly reducing the search space for computationally expensive reranking.
- A novel similarity measurement encodes not only the scale but also the spatial distribution of landmarks, benefiting recognition precision and robustness effectively.

II. Related Work

Visual place recognition approaches can mainly be classified into two categories, when facing with appearance and view-

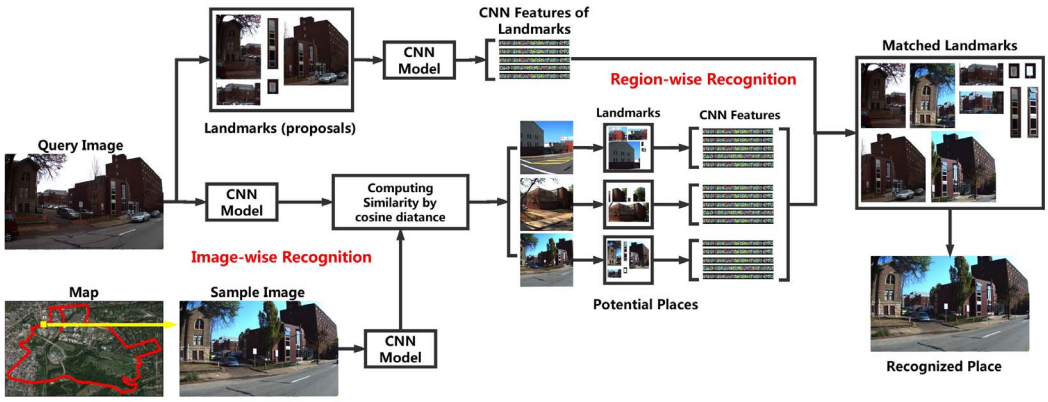


Fig. 2. The architecture of the proposed place recognition method.

point changes: one tries to find invariant features of the place, the other learns how the environment changes over time [18].

A. Feature-based Approaches

Local features such as SIFT [17], SURF [4] and ORB [23] are widely used as visual words in bag-of-words models. The bag-of-words models, such as FAB-MAP [6] and DBoW3 [10], transform the feature space into a visual vocabulary which can be retrieved efficiently. Although local features are pose invariant, they perform poorly in appearance changing environments.

Global descriptors like GIST [20], HOG [7] provide a higher degree of invariance to appearance changes than local descriptors. Instead of matching based on single images, SeqSLAM [19] improved the precision by using sequence search techniques. Also based on sequence matching, ABLE [3] generated the illumination invariant images and used LDB descriptors to extract binary codes. However, global descriptors suffer from the sensitivity of viewpoint changes.

B. Learning-based Approaches

Learning methods try to find the relationship among changed environments. Jacobs [14] observed that the environment changes over time were similar across different scenes. Therefore, the trained model could be generalized to unseen places. Ranganathan [22] built a fine vocabulary and learned a probability distribution over visual words, but the feature correspondences across different illumination had to be matched manually. Han [12] presented a shared representative appearance learning (SRAL) to autonomously learn modality weights of representation features. The inconvenience of learning-based approaches is to obtain the correspondences of training data.

All mentioned approaches above rely on hand-crafted features and perceive on the raw pixel level. Along with the innovative work [16], Conventional Neural Networks (CNNs) have been used as robust feature generators.

Recent studies have certified the possibility of utilizing pre-trained CNNs in visual place recognition. Sunderhauf [25] showed the CNN features derived from the whole image outperformed conventional features. Arandjelovic [2] designed a novel VLAD layer to recognize places in an end-to-end manner. Sunderhauf [26] combined the power of CNNs and region-based proposals, achieving superior performance to viewpoint changes, but with a very large computational cost.

III. Proposed System

The proposed place recognition method is composed of a filtering stage followed by a partial reranking process, as illustrated in Fig. 2. In the filtering stage, image-wise features are utilized to find a small set of potential places. Afterwards, stable region-wise landmarks are extracted for more accurate matching in the partial reranking process. In this section, we firstly describe the generation of image-wise and region-wise features. Then, the filtering stage and patrial reranking process are described in detail.

A. Robust CNN Feature Generation

1) Convolutional Neural Network: In the proposed method, AlexNet [16] is adopted for feature extraction, which consists of five convolutional layers and three fully connected layers. Each kernel of the convolutional layer is convolved across the width and height of the input volume, generating a two-dimensional feature map. The output is formed by the stacking of all feature maps of all kernels.

Sunderhauf [25] discovered that, the features extracted from the third convolutional layer($conv3$) of AlexNet had comparatively favorable invariance against challenging environments. However, fully connected layers were not as effective as convolutional layers because of the loss of spatial information.

Inspired by [25], we use $conv3$ to obtain image-wise and region-wise features. The AlexNet model is pre-trained on ImageNet and provided by Caffe [15], and the output size of $conv3$ layer is $384\times13\times13$. All feature maps are vectorized and concatenated to generate a robust CNN descriptor with 64896 dimensions.

2) Region-based Landmark Extraction: Sunderhauf [26] proposed a method which combined the power of CNNs and object proposal techniques. They applied Edge Boxes [28] to extract reliable landmarks to describe the scene. Edge Boxes mainly relied on the number of contours wholly enclosed in the boxes. All candidate boxes were sorted according to the objectness score. The outperformance of [26] demonstrated that region-based descriptors significantly improved the robustness to viewpoint changes. The proposed method adopts Edge Boxes to extract stable landmarks in every image. The advantage of Edge Boxes is the independence of any objects, so arbitrary objects in the place can be extracted as landmarks.

3) Efficient Feature Reduction: With the dimensions of 64896, it is very inefficient to calculate the cosine distance of $conv3$ features. Efficient feature reduction techniques should be applied to compress features but ensure the accuracy.

Many studies have concluded that a satisfactory precision can be remained by selecting a portion of the whole descriptor. Yang [27] reduced feature dimensions based on a random selection technique, which needed no further training process and calculation. Compared with Local Sensitive Hushing (LSH) [8] and Principal Component Analysis (PCA) [9], this technique is very efficient and effective.

In this work, a fixed set of descriptors are randomly chosen to reduce all generated features. The experiments demonstrate that, when the compression ratio is 93.7%, the precision barely loses and the matching speed is about 20 times faster than the original features.

B. Image-wise Filtering

In the filtering stage, every place is described by the CNN feature extracted from the whole image. The query image is matched with all items in the map, ranking based on the cosine distance of discriptors. The top K items are regarded as potential places. In the following process, more computationally expensive techinques can just be executed on potential places.

If the map size is N, the number of landmarks in each image is P. For [26] that only carries out region-based matching with a mutual crosscheck, the time complexity is $O(2 \times P^2 N)$. Moreover, the matching efficiency can become worse and worse with the increasing of the map size.

With the use of filtering stage, the region-based recognition is just applied to the top K items. Then the time complexity is reduced to $O(N)+O(2\times P^2 K)$. What's remarkable is that the top number K is independent of the map size. The proposed method is demonstrated that the computational efficiency is a dozen even a hundred times faster than [26].

C. Region-wise Reranking

Based on the potential places, the partial reranking process uses region-based features to determine the final match. A new similarity measurement is provided by considering both scale and spatial distribution of landmarks.

To calculate the similarity between two images I_1, I_2, with landmarks $\{l_i^1, l_i^2 ... l_i^P\}$, $i = 1, 2$, the first operation is to find matching landmarks. A crosscheck technique guarantees that only the mutually matched landmarks are preserved.

A weight w is calculated for each mutual match (l_1^m, l_2^n). The weight consists of two parts, the position and shape similarity of landmark proposals. Let (l_m, t_m, w_m, h_m) and (l_n, t_n, w_n, h_n) be the left, top position, the width and the height of the bounding boxes. The horizontal and vertical differences d_w, d_h of two landmarks are calculated as follows.

$$d_w = \lfloor \frac{|l_m - l_n|}{w_{step}} \rfloor, d_h = \lfloor \frac{|t_m - t_n|}{h_{step}} \rfloor \tag{1}$$

Where w_{step} and h_{step} are a quarter of the width and height of the image. If $d_w > 1$ or $d_h > 1$, we regard this mutual match as mismatch. Otherwise, the position similarity is measured as

$$p_{mn} = exp(-\frac{1}{2}(\frac{|l_m - l_n|}{max(l_m, l_n)} + \frac{|t_m - t_n|}{max(t_m, t_n)}) \tag{2}$$

The spatial check can reduce the risk of similar landmarks with different positions to some extent, such as plants and traffic signs. What's more, the w_{step} and h_{step} are the extreme degree of viewpoint changes that we can accept.

The scale distribution adopts the similar technique as [26], to penalize the mutual match which has similar features but different shapes. The shape similarity is measured as

$$s_{mn} = exp(-\frac{1}{2}(\frac{|w_m - w_n|}{max(w_m, w_n)} + \frac{|h_m - h_n|}{max(h_m, h_n)}) \tag{3}$$

Consequently, the weight w of the mutual match and the overall similarity score between I_1, I_2 are calculated as

$$S_{12} = \frac{1}{p_1 \times p_2} \sum_{mn} w_{mn} \times c_{mn} \tag{4}$$

$$w_{mn} = \begin{cases} 0 & d_w > 1 \text{ or } d_h > 1 \\ p_{mn} \times s_{mn} & \text{otherwise} \end{cases} \tag{5}$$

Where p_1, p_2 are the number of landmarks in I_1, I_2, respectively, c_{mn} is the cosine distance of the mutual match. The image with the highest similarity score is chosen as the final match.

IV. Experiments and Evaluation

A. Evaluation Methodology

Experiments with varied viewpoints and environmental conditions are conducted to compare the proposed method with four other state-of-the-art methods, the holistic descriptor HOG, the bag-of-words model DBoW3, the holistic CNN feature (Conv3) [25] and the ConvNet landmarks (Proposals) [26]. For evaluating HOG, we use the implementation provided by OpenCV library [5]. DBoW3 is implemented on the basis of public source code [1]. To evaluate [25] and [26], $conv3$ of AlexNet and Edge Boxes are used to generate features and extract landmarks, respectively. 50 landmarks are extracted in each image for [26] and the proposed method.

The performances are analyzed in terms of the evaluation methodology described in [25], which is based on the precision-recall curve and F1-score.

B. Datasets

Two widespread datasets are used to demonstrate the capability of the proposed global-to-partial method under different appearance and viewpoint changes.

(1) The Gardens Point Dataset: The Gardens Point dataset [11] is recorded on the Gardens Point Campus of QUT by a pedestrian. There are three image sequences of the same route, two during the day and one during the night. The two day sequences are captured on the left and right side of the pathway, respectively. The minor appearance changes in these sequences are mainly caused by pedestrians. The night sequence is recorded on the right side of the pathway and the contrast of all images has been enhanced.

(2) CMU Visual Localization Dataset: CMU Visual Localization dataset [13] consists of several image sequences traveled the same route around Pittsburgh (USA) during different seasons of years. The sequences are recorded by two monocular cameras installed on a car, one is on the left and the other is on the right.

We use the images obtained from the left camera and select three sequences 01/09/2010, 28/10/2010 and 21/12/2010 to carry out the experiments. These sequences belong to summer, fall and winter respectively, including appearance changes generated by illumination, weather, green vegetation and dynamic objects. A set of images are chosen randomly from all sequences based on the GPS information. Besides, two sets of images are selected from 28/10/2010 only to quantify the performance of viewpoint changes.

The selected datasets are partitioned into three groups, based on the characteristics of the changes, as summarized in Table I.

TABLE I
THE SELECTED DATASETS ARE PARTITIONED INTO THREE GROUPS, BASED ON THE CHARACTERISTICS OF THE CHANGES.

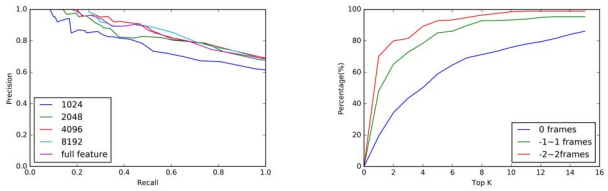
		Variations in	
Datasets	**Sequences**	**Appearance**	**Viewpoint**
GP Walking	day_left vs day_right	minor	medium
CMU-VL	fall_1 vs fall_2	minor	medium
GP Walking	night_right vs day_right	severe	minor
CMU-VL	fall_1 vs summer	severe	minor
CMU-VL	fall_2 vs summer	severe	medium
CMU-VL	fall_2 vs winter	severe	medium

C. Results of Preliminaries

The experiments in this section are performed on the two day sequences of the Gardens Point dataset (day_left, day_right).

The random selection technique is applied to compress the CNN features in this work. The precision-recall curves in the left of Fig. 3 show the performance of CNN features with different dimensions. The satisfactory results can still be maintained even if the features are highly compressed. Table II presents a more detailed analysis, the feature which is reduced to 4096 dimensions has a compression of 93.7%, but barely loses any precision.

The results in the right of Fig. 3 demonstrate the validity of the image-wise filtering stage. The percentage is calculated by the ratio between the number of correctly matched images and the total number of query images. A match is correct if any frame within $\pm x$ frames of the reference image is in the top K items of filtering results. We demonstrate that almost all correct matches can be found when K is about 12, which is much smaller than the number of query images. It is confirmed that the search space for reranking process can be greatly reduced by the filtering stage, on the premise of guaranteeing the precision. We use $x = 2$ and $K = 12$ in the following experiments.



(a) (b)

Fig. 3. (a) Precision-recall curves of CNN features with different dimensions. (b) Results of the image-wise filtering stage.

TABLE II
THE DETAILED RESULTS OF CNN FEATURES WITH DIFFERENT DIMENSIONS.

Dimensions	**Compression ratio**	**Speedup**	**F1-score**
64896	-	-	0.816
8192	87.4%	8x	0.813
4096	93.7%	17x	0.813
2048	96.8%	38x	0.806
1024	98.4%	81x	0.761

D. Results of Viewpoint Changes

In order to demonstrate the viewpoint robustness of the proposed method, we conduct experiments on two datasets: (1) day_left vs. day_right, (2) $fall\_1$ vs. $fall\_2$.

The results in Fig. 4 show that the proposed method outperforms the other state-of-the-art methods. Compared with other CNN based methods, encoding the spatial distribution of landmarks has a notable effect to the improvement of performance.

E. Results of Appearance Changes

Appearance changes are another major challenge for visual place recognition, we conduct experiments on the following datasets: (1) day_right vs. $night\_right$, (2) $summer$ vs. $fall\_1$.

The precision-recall curves in Fig. 5 present the performance of all methods. It is worth noting that in the CMU-VL dataset, all methods based on CNNs have comparative results. But in the Gardens Point dataset, the holistic $conv3$

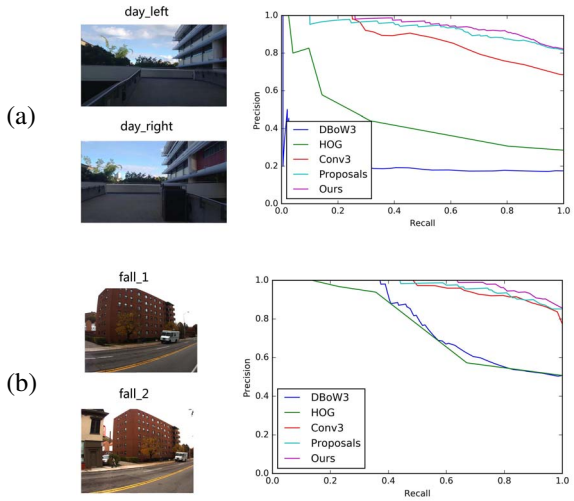


Fig. 4. Precision-recall curves of (a) the two day sequences of the Gardens Point dataset. (b) the $28/10/2010$ sequence of the CMU-VL dataset.

feature [25] significantly outperforms the region-based methods. Methods based on region features are dependent on the detail of images, while too much detailed information has been lost in the severe appearance changes from day to night. As mentioned in [18], the tradeoff between pose invariance and condition invariance is still an unsolved challenge in place recognition field.

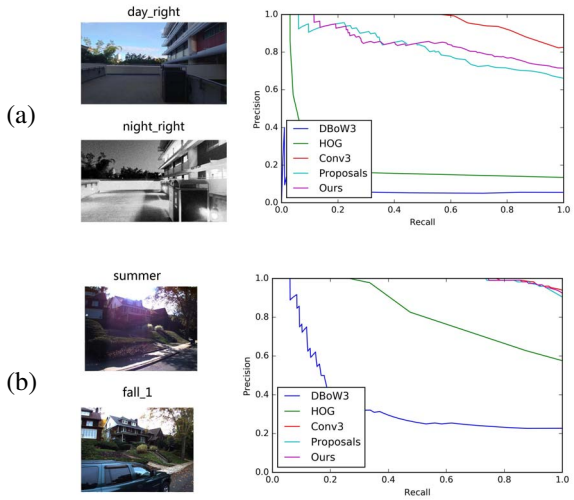


Fig. 5. Precision-recall curves of (a) the day and night sequences of the Gardens Point dataset. (b) the $01/09/2010$ and $28/10/2010$ sequences of the CMU-VL dataset.

F. Results of Appearance and Viewpoint Changes

The sequences used in this section have the viewpoint and appearance changes occurring simultaneously: (1) $summer$ vs. $fall\_2$, (2) $winter$ vs. $fall\_2$.

The performances shown in Fig. 6 demonstrate that considering the spatial distribution of landmarks provides much higher invariance to viewpoint changes. Fig. 7 shows some examples which fail to be matched by [26] but are successfully matched by the proposed method. More correctly matched image pairs are presented in Fig. 8. Only representative matches are colored in each image pair for illustration.

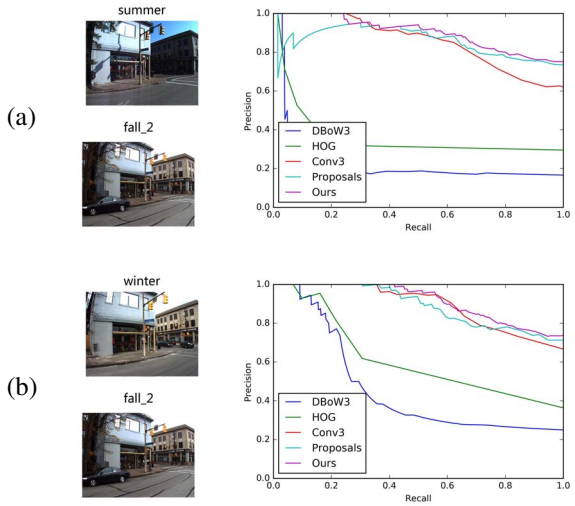


Fig. 6. Precision-recall curves of (a) the $01/09/2010$ and $28/10/2010$ sequences of the CMU-VL dataset. (b) the $21/12/2010$ and $28/10/2010$ sequences of the CMU-VL dataset.

Fig. 7. From left to right in each group: the query image, the result from [26], the result from the proposed method.

V. Conclusions and Discussion

A novel global-to-partial pipeline for visual place recognition is presented in this work, based on CNNs and object proposal techniques. The proposed method is composed of a filtering stage followed by a partial reranking process. In the filtering stage, image-wise features are utilized to find a small set of potential places. Afterwards, stable region-wise landmarks are extracted for more accurate matching in the partial reranking process. In the similarity measurement, both scale and spatial distribution of landmarks are taken into account, which benefits recognition precision and robustness

Fig. 8. More image pairs correctly matched by the proposed method.

effectively. Moreover, we use the CNN model transferred from the image classification task, and no further environment-specific training is required in the proposed method. A set of experiments have demonstrated the capability of the proposed method in challenging environments, where appearance and viewpoint changes occur simultaneously.

For promoting the computational efficiency, the reliable filtering stage greatly reduces the search space for partial reranking process. With the use of random selection technique, CNN features can not only be highly compressed but also maintain the results in the similarity measurement.

There are still some interesting directions for further enhancing the performance of the proposed method.

- For the tradeoff between pose invariance and condition invariance, it may be possible to find a balance between image-wise and region-wise recognition in the similarity measurement. We will investigate whether considering the ranking results of filtering stage can achieve superior performance.
- For each landmark, a forward pass has to be performed through the CNN model to generate the feature. We will analyze the possibility of generating features all of landmarks in just one forward pass.

REFERENCES

[1] Dbow3. https://github.com/rmsalinas/DBow3.

[2] Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Netvlad: Cnn architecture for weakly supervised place recognition. pages 5297–5307, 2016.

[3] R Arroyo, P. F Alcantarilla, L. M Bergasa, and E Romera. Towards life-long visual localization using an efficient matching of binary sequences from images. In *IEEE International Conference on Robotics and Automation*, pages 6328–6335, 2015.

[4] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded up robust features. In *Computer Vision - ECCV 2006, European Conference on Computer Vision, Graz, Austria, May 7-13, 2006, Proceedings*, pages 404–417, 2006.

[5] G. Bradski. The opencv library. *Doctor Dobbs Journal*, 25(11):384–386, 2000.

[6] Mark Cummins and Paul Newman. Fab-map: Probabilistic localization and mapping in the space of appearance. *International Journal of Robotics Research*, 27(6):647–665, 2008.

[7] N Dalal and B Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, pages 886–893 vol. 1, 2005.

[8] Mayur Datar, Nicole Immorlica, Piotr Indyk, and Vahab S Mirrokni. Locality-sensitive hashing scheme based on p-stable distributions. In *Twentieth Symposium on Computational Geometry*, pages 253–262, 2004.

[9] Hans Peter Deutsch. *Principle Component Analysis*. Palgrave Macmillan UK, 2002.

[10] Dorian G'alvez-L'opez and J. D. Tard'os. Bags of binary words for fast place recognition in image sequences. *IEEE Transactions on Robotics*, 28(5):1188–1197, October 2012. ISSN 1552-3098.

[11] Arren Glover. Gardens point walking dataset. 2014. https://wiki.qut.edu.au/display/cyphy/Open+datasets+and+software.

[12] Fei Han, Xue Yang, Yiming Deng, Mark Rentschler, Dejun Yang, and Hao Zhang. Sral: Shared representative appearance learning for long-term visual place recognition. *IEEE Robotics & Automation Letters*, 2 (2):1172–1179, 2017.

[13] Daniel Huber Hernan Badino and Takeo Kanade. The cmu visual localization data set. 2011. http://3dvis.ri.cmu.edu/data-sets/localization.

[14] N Jacobs, N Roman, and R Pless. Consistent temporal variations in many outdoor scenes. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1–6, 2007.

[15] Jia, Yangqing, Shelhamer, Evan, Donahue, Jeff, Karayev, Sergey, Long, and Jonathan. Caffe: Convolutional architecture for fast feature embedding. *Eprint Arxiv*, pages 675–678, 2014.

[16] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *International Conference on Neural Information Processing Systems*, pages 1097–1105, 2012.

[17] David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[18] S. Lowry, N. Sunderhauf, P. Newman, and J. J. Leonard. Visual place recognition: A survey. *IEEE Transactions on Robotics*, 32(1):1–19, 2015.

[19] Michael J. Milford and Gordon. F. Wyeth. Seqslam: Visual route-based navigation for sunny summer days and stormy winter nights. In *IEEE International Conference on Robotics and Automation*, pages 1643–1649, 2012.

[20] Aude Oliva and Antonio Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3):145–175, 2001.

[21] Pilailuck Panphattarasap and Andrew Calway. Visual place recognition using landmark distribution descriptors. 2016.

[22] Ananth Ranganathan, Shohei Matsumoto, and David Ilstrup. Towards illumination invariance for visual localization. In *IEEE International Conference on Robotics and Automation*, pages 3791–3798, 2013.

[23] E Rublee, V Rabaud, K Konolige, and G Bradski. Orb: An efficient alternative to sift or surf. 58(11):2564–2571, 2011.

[24] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, and Michael Bernstein. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[25] N. Sunderhauf, S. Shirazi, F. Dayoub, and B. Upcroft. On the performance of convnet features for place recognition. In *Ieee/rsj International Conference on Intelligent Robots and Systems*, pages 4297–4304, 2015.

[26] Niko Sunderhauf, Sareh Shirazi, Adam Jacobson, Edward Pepperell, Feras Dayoub, Ben Upcroft, and Michael Milford. Place recognition with convnet landmarks: Viewpoint-robust, condition-robust, training-free. In *Robotics: Science and Systems*, pages 296–296, 2015.

[27] Xin Yang and Kwang Ting Tim Cheng. Local difference binary for ultrafast and distinctive feature description. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 36(1):188–94, 2013.

[28] C. Lawrence Zitnick and Piotr Dollar. *Edge Boxes: Locating Object Proposals from Edges*. Springer International Publishing, 2014.