applied sciences

MDPI

Review

Place Recognition: An Overview of Vision Perspective

Zhiqiang Zeng [1], Jian Zhang [2,*], Xiaodong Wang [1], Yuming Chen [1] and Chaoyang Zhu [3]

[1] College of Computer and Information Engineering, Xiamen University of Technology, Xiamen 361024, Fujian, China; lbxzzq@163.com (Z.Z.); xdwangjsj@xmut.edu.cn (X.W.); ymchen@xmut.edu.cn (Y.C.)

[2] School of Science and Technology, Zhejiang International Studies University, Hangzhou 310023, Zhejiang, China

[3] School of Computer Science, Hangzhou Dianzi University, Hangzhou 310018, Zhejiang, China; zack.zcy@gmail.com

* Correspondence: jzhang@zisu.edu.cn; Tel.: +86-133-0681-2810

Received: 14 October 2018; Accepted: 31 October 2018; Published: 15 November 2018

Abstract: Place recognition is one of the most fundamental topics in the computer-vision and robotics communities, where the task is to accurately and efficiently recognize the location of a given query image. Despite years of knowledge accumulated in this field, place recognition still remains an open problem due to the various ways in which the appearance of real-world places may differ. This paper presents an overview of the place-recognition literature. Since condition-invariant and viewpoint-invariant features are essential factors to long-term robust visual place-recognition systems, we start with traditional image-description methodology developed in the past, which exploits techniques from the image-retrieval field. Recently, the rapid advances of related fields, such as object detection and image classification, have inspired a new technique to improve visual place-recognition systems, that is, convolutional neural networks (CNNs). Thus, we then introduce the recent progress of visual place-recognition systems based on CNNs to automatically learn better image representations for places. Finally, we close with discussions and mention of future work on place recognition.

Keywords: place recognition; Convolutional Neural Network; feature extraction; bag-of-visual words (BoW); vector of locally aggregated descriptors (VLAD)

1. Introduction

Place recognition has attracted a significant amount of attention in the computer-vision and robotics communities, as evidenced by the related citations and a number of workshops dedicated to improving long-term robot navigation and autonomy [1]. It has a number of applications, ranging from autonomous driving and robot navigation to augmented reality and geolocalizing archival imagery.

The process of identifying the location of a given image by querying the locations of images belonging to the same place in a large geotagged database, usually known as place recognition, is still an open problem. One major characteristic that separates place recognition from other visual-recognition tasks is that place recognition has to solve condition-invariant recognition to a degree that many other fields haven't. How can we robustly identify the same real-world place undergoing major changes in appearance (e.g., illumination variation (Figure 1), change of seasons (Figure 2) or weather, structural modifications over time, and viewpoint change)? To be clear, the above changes in appearance are summarized as conditional variations, but exclude viewpoint change. Moreover, how can we distinguish true images from similar-looking images without supervision? Since collecting geotagged datasets is time-consuming and labor-intensive, and locations like indoor places do not necessarily have GPS information. Place-recognition tasks have been traditionally cast as image-retrieval tasks [2] where image representations for places are essential. The fundamental scientific question is what the appropriate representation of a place is that is informative enough to recognize real-world places, yet compact enough to satisfy the real-time processing requirement on a terminal, such as a mobile phone or a robot.

At early stages, place recognition was dominated by sophisticated local-invariant feature extractors, such as Scale-Invariant Feature Transformation (SIFT) [3] and Speed-Up Robust Features (SURF) [4], hand-crafted global image descriptors, such as Generalized Search Trees (GIST) [5,6], and the bag-of-visual-words [7,8] approach. These traditional feature-extraction techniques have gained impressive results.

Recent years have seen a prosperous advancement of visual content recognition using a powerful image representation-extractor—Convolutional Neural Networks (CNNs) [9–12], which offers state-of-the-art performance on many category-level recognition tasks, such as object classification [9,10,13], scene recognition [14–16], and image classification [17]. The principle ideas of CNNs date back to 1980s, and the two major reasons why CNNs are so successful in computer vision are the advances in GPU-based computation power and data volume, respectively. Recent studies show that general features extracted by CNNs can be transferable [18] and generalized well to other visual-recognition tasks. The semantic gap is a well-known problem in place recognition, where different semantics of places may share common low-level features extracted by SIFT, e.g., colors and textures. Convolutional neural networks may bridge this semantic gap by treating an image as a high-level feature vector extracted through deep-stacked layers.

This paper provides an overview of both traditional and deep-learning-based descriptive techniques widely applied to place-recognition tasks, which is by no means exhaustive. The remainder of this paper is organized as follows. Section 2 talks about local and global image descriptors that are widely applied to place recognition. Section 3 presents a brief view of convolutional neural networks and the corresponding techniques used in place recognition. Section 4 discusses future work on place recognition.The primary abbreviations and their extended representations in the paper are listed in Table 1.

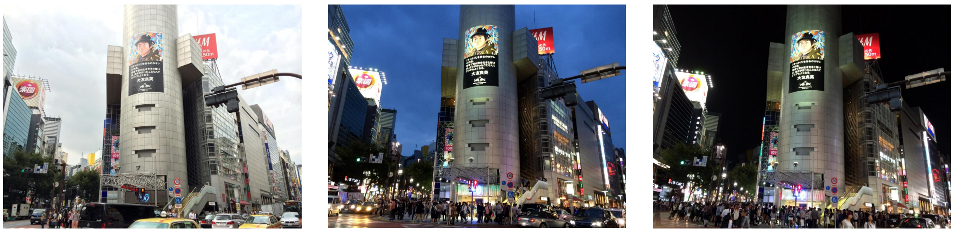

Figure 1. TokyoTimeMachine dataset [19] images from the same place with different illumination conditions: day, sunset, night.

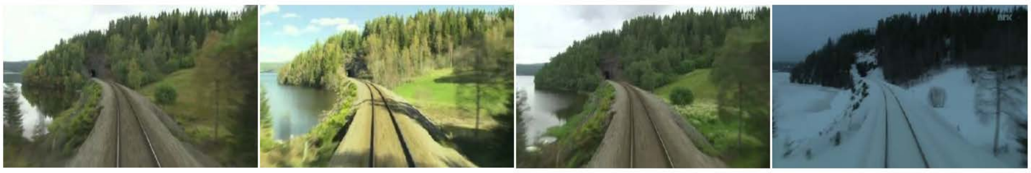

Figure 2. Frames extracted from the Nordland dataset [20] that are from the same place in spring, summer, fall, and winter.

Table 1. Abbreviations used in this paper and their extended representations.

Abbreviation	The Extended Representation
CNN	Convolutional Neural Network
SIFT	Scale-Invariant Feature Transformation
SURF	Speed-Up Robust Features
VLAD	Vector of Locally Aggregated Descriptors
BoW	Bag-of-Words
FAST	Features from Accelerated Segment Test
BRIEF	Binary Robust Independent Elementary Features

2. Traditional Image Descriptors

2.1. Local Image Descriptors

Local feature descriptors, such as SIFT [3] and SURF [4], have been widely applied to visual-localization and place-recognition tasks. They achieve outstanding performance on viewpoint invariance. SIFT and SURF describe the local appearance of individual patches or key points within an image, for an image; $\mathbf{X} = [\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_N}]^T$ represents the set of local invariant features, in which $\mathbf{x}_i \in \mathbf{R}^d$ is the $\mathbf{i}$th local feature. The length of $\mathbf{X}$ depends on the number of key points of the image. These local invariant features are then usually aggregated into a compact single vector for the entire image, using techniques such as bag-of-visual-words [7,8], Vector of Locally Aggregated Descriptors (VLAD) [19], and Fisher kernel [21,22].

Since local-feature extraction consists of two phases, detection and description, a number of variations and extensions of techniques for the two phases have been developed. For example, Reference [23] used the Features from Accelerated Segment Test (FAST) [24] to detect the interesting patches of an image that were then described by SIFT. Reference [25] used FAST as well during the detection phase, whereas they used Binary Robust Independent Elementary Features (BRIEF) [26] to describe the key points instead of SIFT.

State-of-the-art visual simultaneous localization and mapping (SLAM) systems, such as FAB-MAP [27,28] used bag-of-visual-words to construct the final image representation. Each image may contain hundreds of local features, which is impractical in large-scale and real-time processing place-recognition tasks; moreover, they require an enormous amount of memory to store the high dimensional features. The bag-of-visual-words mimics the bag-of-words technique used in efficient text-retrieval fields. It typically needs to form a codebook $\mathbf{C} = [\mathbf{c_1}, \mathbf{c_2}, ..., \mathbf{c_k}]$ with $\mathbf{K}$ visual words, and each visual word $\mathbf{c}_i \in \mathbf{R}^d$ is a centroid of a cluster, usually gained by k-means [29]. Then, each local invariant feature $\mathbf{x}_i$ is assigned to its nearest cluster centroid, i.e., the visual word. A histogram vector of the $\mathbf{k}$ dimension containing the frequency of each visual word being assigned can be formed this way. The bag-of-visual-words model and local-image descriptors generally ignore the geometric structure of the image, that is, different orders of local invariant features in $\mathbf{X}$ do not impact the histogram vector; thus, the resulting image representation is viewpoint-invariant. There are several variations on how to normalize the histogram vector, a common choice being $\mathbf{L}_2$ normalization. Components of the vector are then weighted by inverse document frequency (IDF). However, the codebook is dataset-dependent and needs to be retrained if a robot moves into a new region it has never seen before. The lack of structural information of the image can also weaken the performance of the model.

Fisher Kernel proposed by References [21,22] is a powerful tool in pattern classification, combining the strengths of generative models and discriminative classifiers. It defines a generative probability model $\mathbf{p}$, which is the probability density function with parameter λ. Then, one can characterize the set of local invariant features $\mathbf{X} = [\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_N}]^T$ with the following gradient:

$$\nabla_\lambda \log \mathbf{p}(\mathbf{X}|\lambda) \tag{1}$$

where the gradient of the log likelihood describes the direction in which parameters should be modified to best intuitively fit the observed data. It transforms a variable-length sample $\mathbf{X}$ into a fixed-length vector whose size is only dependent on the number of parameters in the model.

Reference [22] applied the Fisher kernel in the context of image classification with a Gaussian mixture model to model the visual words. In comparison with the bag-of-visual-words representation, they obtained a $(\mathbf{2d}+\mathbf{1}) * \mathbf{k} - \mathbf{1}$ dimensional image representation of a local invariant feature set, while $\mathbf{k}$ dimensional image representation using the bag-of-visual-words. Thus, the Fisher kernel can provide richer information under the condition that their codebook size is equal, or fewer visual words are required by this more sophisticated representation.

Reference [19] proposed a new local-aggregation method, VLAD, which is a state-of-the-art technique compared to the bag-of-visual-words and Fisher vector. The final representation is computed as follows:

$$\mathbf{v}_{i,j} = \sum_{\mathbf{NN(x)=c_i}} \mathbf{x}_j - \mathbf{c}_{i,j} \tag{2}$$

The VLAD vector is represented by $\mathbf{v}_{i,j}$, where the indices $i = 1, ..., k$ and $j = 1, ..., d$ respectively index the $\mathbf{i}$th visual word and the $\mathbf{j}$th component of the local invariant feature. The vector is subsequently $\mathbf{L}_2$-normalized. We can see that the final representation stores the sum of all residuals between local invariant features and its nearest visual word. One excellent property of the VLAD vector is that it is relatively sparse and very structured [19], showing that a principal component analysis is likely to capture this structure for dimensionality reduction without much degradation of representation. They obtain comparable search quality to bag-of-visual-words and the Fisher kernel with at least an order of magnitude less memory.

2.2. Global Image Descriptors

The key difference between local and global place descriptors is the presence of the detection phase. One can easily figure out that local place descriptors turns into global descriptors by predefining the key points as the whole image. WI-SURF [30] used whole-image descriptors based on SURF features, and BRIEF-GIST [31] used BRIEF [26] features in a similar whole-image fashion.

A representative global descriptor is GIST [5]. It has been shown to suitably model semantically meaningful and visually similar scenes in a very compact vector. The amount of perceptual and semantic information that observers comprehend at a glance (around 200 ms) refers to the gist of the scene, termed Spatial Envelope properties, and it encodes the dominant spatial-layout information of the image. GIST uses Gabor filters at different orientations and different frequencies to extract information from the image. The results are averaged to generate a compact vector that represents the "gist" of a scene. Reference [32] applied GIST to large-scale omnidirectional imagery and obtained a good segmentation of the search space into clusters, e.g., tall buildings, streets, open areas, mountains. Reference [33] followed a biological strategy that first computes the gist of a scene to produce a coarse localization hypothesis, then refine it by locating salient landmark points in the scene.

The performance of techniques described in Section 2 mainly depends on the size of codebook $\mathbf{C}$; if too small, the codebook does not characterize the dataset well, while if it is too large, it requires huge computational resources and time. While global image descriptors have their own disadvantages, they usually assume that images are taken from the same viewpoint.

3. Convolutional Neural Networks

Recently, convolutional neural networks have achieved state-of-the-art performance on various classification and recognition tasks, e.g., handwriting-digit recognition [34], object classification [9,10,13], and scene recognition [14,15].The most representative convolutional neural networks include AlexNet [9], VGGNet [10], ResNet [35] and GoogleNet [13]. Features extracted from convolutional neural networks trained on very large datasets significantly outperform SIFT in a variety of vision tasks [9,36]. The core idea behind CNNs is the ability to automatically learn high-level features trained on a significant amount of data through deep-stacked layers in an end-to-end manner. It works as a function that takes some inputs, such as images, and puts out the image representations characterized by a vector. A common CNN model for fine-tuning is VGG-16 [10], and its architecture can be seen from Table 2. For an intuitive understanding of what the CNN model learns in each layer, please see References [19,37] for heatmap graphical explanation.

Table 2. VGG-16 configuration (shown in columns). Configuration depth increases from the left (A) to the right (E) as more layers are added (added layers are shown in bold). Convolutional layer parameters are denoted as "conv(receptive field size)-(number of channels)". ReLU activation function is omitted for simplicity.

ConvNet Configuration					
A	**A-LRN**	**B**	**C**	**D**	**E**
11 weight layers	11 weight layers	13 weight layers	16 weight layers	16 weight layers	19 weight layers
input (224 × 224 RGB image)					
conv3-64	conv3-64 LRN	conv3-64 conv3-64	conv3-64 conv3-64	conv3-64 conv3-64	conv3-64 conv3-64
maxpool					
conv3-128	conv3-128	conv3-128 conv3-128	conv3-128 conv3-128	conv3-128 conv3-128	conv3-128 conv3-128
maxpool					
conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256 conv1-256	conv3-256 conv3-256 conv3-256	conv3-256 conv3-256 conv3-256 conv3-256
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 conv1-512	conv3-512 conv3-512 conv3-512	conv3-512 conv3-512 conv3-512 conv3-512
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 conv1-512	conv3-512 conv3-512 conv3-512	conv3-512 conv3-512 conv3-512 conv3-512
maxpool					
FC-4096					
FC-4096					
FC-1000					
soft-max					

Reference [38] was the first work to exploit CNN in place-recognition systems as a feature extractor. They used a pretrained CNN called Overfeat [39], which was originally proposed for ImageNet Large Scale Visual Recognition Challenge 2013 (ILSVRC2013), and proved that advantages of deep learning can shift to place-recognition tasks. Reference [36] provides an investigation on the performance of CNN and SIFT on a descriptor-matching benchmark. Reference [37] comprehensively evaluated and compared the utility and viewpoint-invariant properties of CNNs. They showed that features extracted from middle layers of CNNs have good robustness against conditional changes, including illumination change, and seasonal and weather changes, while features extracted from top layers are more robust to viewpoint changes. Features learnt by CNNs are proved to be versatile and transferable, i.e., even though they were trained on a specific target task, they can be successfully deployed for other problems, and often outperform traditional hand-engineered features [18]. However, their usage as black-box descriptors in place recognition has so far yielded limited improvements. For example, visual cues that are relevant for object classification may not benefit the place-recognition task. Reference [19] designed a new CNN architecture based on VGG-16 and the VLAD representation; they removed all the fully connected layers and plugged the VLAD layers into it by making it differentiable. The loss function used in this paper is triplet loss, which can be seen in many other recognition tasks. Reference [14] gathered a huge scene-centric dataset called "Places" containing more than 7 million images from 476 scene categories. However, scene recognition is fundamentally different from place recognition. Reference [40] creates, for the first time, a large-scale place-centric dataset called SPED, containing over 2.5 million images.

Though CNNs have the power to extract high-level features, we are far from making full use of them. How to gather a sufficient amount of data for place recognition, and how to train a CNN model in an end-to-end manner to automatically choose the optimal features to represent the image, are still underlying problems to be solved.

4. Discussion and Future Work

Place recognition has made great advances in the last few decades, e.g., on the principles of how animals recognize and remember places, and the relationship between places from a neuroscience perspective [41]; a new way of describing places using convolutional neural networks, and a number of datasets specifically for places were put forward. However, we are still a long way from a robust long-term visual place-recognition system that can be applied well to a variety of scenarios of real-world places. Hence, we highlight several promising avenues of ongoing and future research that are leading us closer to this outcome.

Place recognition is becoming a hot research topic and it is benefiting from related ongoing works in other fields, especially the enormous successes achieved in computer vision through deep-learning techniques, e.g., image classification, object detection, and scene recognition. While features extracted from pretrained CNNs on other vision tasks are shown to be transferable and versatile, they have, so far, yielded unsatisfactory performance on place-recognition tasks. There is a high possibility that we are still not fully exploiting the potential of CNNs. We can improve their performance in two aspects. First, by gathering a sufficient amount of place-centric data covering various environments, including illumination, weather, structure, seasons, and viewpoint changes. An alternative and reliable source is Google Street View Time Machine. If you train a CNN on a small-size dataset, the model usually works awfully on other datasets, since place recognition is dataset-dependent. One also needs to retrain it when a new dataset is fed into it. Since one of the advantages of CNNs is to extract representative features through Big Data, their state-of-the-art performance can be improved. Second, designing optimized CNN architecture for place-recognition tasks. Real-world images from cameras are usually high-resolution, whereas in many cases one needs to downscale the original images. For example, the input size of VGG-16 is 224 × 224. Moreover, an architecture that is well-suited for object detection may not fit well into place-recognition tasks since their visual cues are different. Designing a good loss function is also essential to their features. Developments focused on the above two problems would further improve the robustness and performance of existing deep-learning-based place-recognition techniques.

Place-recognition systems can also benefit from ongoing research on object detection, scene classification [42], and scene recognition [16]. Semantic context from a scene, interpreted as the "gist" of the scene, can help to partition the search space when comparing similarities between image representations, which ensures scalability and real-time processing towards real-world application. Note that different places may share a common semantic concept that needs a further and more precise feature-mapping procedure. Objects such as pedestrians and trees should be avoided, while objects like buildings and landmarks are important for long-term place recognition. Automatically determining and suppressing features that lead to confusion in visual place-recognition systems would also improve place-recognition performance.

Author Contributions: Conceptualization, Z.Z. and J.Z.; methodology, X.W., C.Z., and J.Z.; formal analysis, X.W. and Y.C.; investigation, X.W. and C.Z.; resources, Z.Z. and J.Z.; writing—original draft preparation, C.Z. and J.Z.; writing—review and editing, Y.C.; supervision, Z.Z.; project administration, Z.Z.; funding acquisition, Z.Z. and J.Z.

Funding: This research was funded by the National Natural Science Foundation of China (grant no. 61871464); the National Natural Science Foundation of Fujian Province, China (grant no. 2016J01324); and the Zhejiang Provincial Natural Science Foundation of China (grant number LY17F020009).

Acknowledgments: We want to thank Dongxiaofan Network Technology Co., Ltd. for the technical support when preparing the data to verify some of the papers listed in the bibliography.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

CNN	Convolutional Neural Network
SIFT	Scale Invariant Feature Transformation
SURF	Speed Up Robust Features
VLAD	Vector of Locally Aggregated Descriptors
BoW	Bag-of-Words
FAST	Features from Accelerated Segment Test
BRIEF	Binary Robust Independent Elementary Features

References

1. Yong, N.K.; Dong, W.K.; Suh, I.H. Visual navigation using place recognition with visual line words. In Proceedings of the International Conference on Ubiquitous Robots and Ambient Intelligence, Kuala Lumpur, Malaysia, 12–15 November 2014; p. 676.
2. Yu, J.; Tao, D.; Wang, M.; Rui, Y. Learning to rank using user clicks and visual features for image retrieval. *IEEE Trans. Cybern.* **2015**, *45*, 767–779. [CrossRef] [PubMed]
3. Lowe, D.G. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* **2004**, *60*, 91–110. [CrossRef]
4. Bay, H.; Tuytelaars, T.; Van Gool, L. Surf: Speeded up robust features. In Proceedings of the 9th European Conference on Computer Vision, Graz, Austria, 7–13 May 2006; pp. 404–417.
5. Oliva, A.; Torralba, A. Modeling the shape of the scene: A holistic representation of the spatial envelope. *Int. J. Comput. Vis.* **2001**, *42*, 145–175. [CrossRef]
6. Oliva, A.; Torralba, A. Building the gist of a scene: The role of global image features in recognition. *Prog. Brain Res.* **2006**, *155*, 23–36. [PubMed]
7. Philbin, J.; Chum, O.; Isard, M.; Sivic, J.; Zisserman, A. Object retrieval with large vocabularies and fast spatial matching. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'07), Minneapolis, MN, USA, 17–22 June 2007; pp. 1–8.
8. Sivic, J.; Zisserman, A. Video google: A text retrieval approach to object matching in videos. In Proceedings of the Ninth IEEE International Conference on Computer Vision, Nice, France, 13–16 October 2003; Volume 2, pp. 1470–1477.
9. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*; MIT Press: Lake Tahoe, Nevada, USA, 3–6 December 2012; pp. 1097–1105.
10. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.

11. Yu, J.; Zhang, B.; Kuang, Z.; Lin, D.; Fan, J. iPrivacy: Image privacy protection by identifying sensitive objects via deep multi-task learning. *IEEE Trans. Inf. Forensics Secur.* **2017**, *12*, 1005–1016. [CrossRef]
12. Yu, J.; Yang, X.; Gao, F.; Tao, D. Deep multimodal distance metric learning using click constraints for image ranking. *IEEE Trans. Cybern.* **2017**, *47*, 4014–4024. [CrossRef] [PubMed]
13. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. *arXiv* **2014**, arXiv:1409.4842.
14. Zhou, B.; Lapedriza, A.; Xiao, J.; Torralba, A.; Oliva, A. Learning deep features for scene recognition using places database. In *Advances in Neural Information Processing Systems*; MIT Press: Montréal, QC, Canada, 2014; pp. 487–495.
15. Yuan, Y.; Mou, L.; Lu, X. Scene recognition by manifold regularized deep learning architecture. *IEEE Trans. Neural Netw. Learn. Syst.* **2015**, *26*, 2222–2233. [CrossRef] [PubMed]
16. Yu, J.; Hong, C.; Tao, D.; Wang, M. Semantic embedding for indoor scene recognition by weighted hypergraph learning. *Signal Process.* **2015**, *112*, 129–136. [CrossRef]
17. Yu, J.; Tao, D.; Wang, M. Adaptive hypergraph learning and its application in image classification. *IEEE Trans. Image Process.* **2012**, *21*, 3262–3272. [PubMed]
18. Sharif Razavian, A.; Azizpour, H.; Sullivan, J.; Carlsson, S. CNN features off-the-shelf: An astounding baseline for recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, Ohio, USA, 23–28 June 2014; pp. 512–519.
19. Arandjelović, R.; Gronat, P.; Torii, A.; Pajdla, T.; Sivic, J. NetVLAD: CNN architecture for weakly supervised place recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016.
20. Sünderhauf, N.; Neubert, P.; Protzel, P. Are we there yet? Challenging SeqSLAM on a 3000 km journey across all four seasons. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA 2013), Karlsruhe, Germany, 6–10 May 2013; p. 2013.
21. Jaakkola, T.S.; Haussler, D. Exploiting generative models in discriminative classifiers. In *Advances in Neural Information Processing Systems*; MIT Press: Denver, USA, 1999; pp. 487–493.
22. Perronnin, F.; Dance, C. Fisher kernels on visual vocabularies for image categorization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'07), Minneapolis, MN, USA, 17–22 June 2007; pp. 1–8.
23. Mei, C.; Sibley, G.; Cummins, M.; Newman, P.M.; Reid, I.D. A Constant-Time Efficient Stereo SLAM System. In Proceedings of the 20th British Machine Vision Conference, London, UK, 2009; pp. 1–11.
24. Rosten, E.; Drummond, T. Machine learning for high-speed corner detection. In Proceedings of the 9th European Conference on Computer Vision, Graz, Austria, 7–13 May 2006; pp. 430–443.
25. Churchill, W.; Newman, P. Experience-based navigation for long-term localisation. *Int. J. Robot. Res.* **2013**, *32*, 1645–1661. [CrossRef]
26. Calonder, M.; Lepetit, V.; Ozuysal, M.; Trzcinski, T.; Strecha, C.; Fua, P. BRIEF: Computing a local binary descriptor very fast. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *34*, 1281–1298. [CrossRef] [PubMed]
27. Cummins, M.; Newman, P. FAB-MAP: Probabilistic localization and mapping in the space of appearance. *Int. J. Robot. Res.* **2008**, *27*, 647–665. [CrossRef]
28. Cummins, M.; Newman, P. Appearance-only SLAM at large scale with FAB-MAP 2.0. *Int. J. Robot. Res.* **2011**, *30*, 1100–1123. [CrossRef]
29. Kanungo, T.; Mount, D.M.; Netanyahu, N.S.; Piatko, C.D.; Silverman, R.; Wu, A.Y. An efficient k-means clustering algorithm: Analysis and implementation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2002**, *24*, 881–892. [CrossRef]
30. Badino, H.; Huber, D.; Kanade, T. Real-time topometric localization. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA 2012), Saint Paul, MN, USA, 14–18 May 2012; pp. 1635–1642.
31. Sünderhauf, N.; Protzel, P. BRIEF-Gist-Closing the loop by simple means. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), San Francisco, CA, USA, 25–30 September 2011; pp. 1234–1241.
32. Murillo, A.C.; Kosecka, J. Experiments in place recognition using gist panoramas. In Proceedings of the 12th IEEE Conference on Computer Vision (ICCV Workshops), Kyoto, Japan, 27 September–4 October 2009; pp. 2196–2203.

33. Siagian, C.; Itti, L. Biologically inspired mobile robot vision localization. *IEEE Trans. Robot.* **2009**, *25*, 861–873. [CrossRef]
34. Teow, M.Y.W. A minimal convolutional neural network for handwritten digit recognition. In Proceedings of the IEEE International Conference on System Engineering and Technology, Shah Alam, Malaysia, 2–3 October 2017; pp. 171–176.
35. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'16), LAS VEGAS, NV, USA, 26 June–1 July 2016; pp. 770–778.
36. Fischer, P.; Dosovitskiy, A.; Brox, T. Descriptor matching with convolutional neural networks: A comparison to sift. *arXiv* **2014**, arXiv:1405.5769.
37. Sünderhauf, N.; Shirazi, S.; Dayoub, F.; Upcroft, B.; Milford, M. On the performance of convnet features for place recognition. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Hamburg, Germany, 28 September–2 October 2015; pp. 4297–4304.
38. Chen, Z.; Lam, O.; Jacobson, A.; Milford, M. Convolutional neural network-based place recognition. In Proceedings of the Australasian Conference on Robotics and Automation, Victoria, Australia, 2–4 December 2014.
39. Sermanet, P.; Eigen, D.; Zhang, X.; Mathieu, M.; Fergus, R.; LeCun, Y. Overfeat: Integrated recognition, localization and detection using convolutional networks. *arXiv* **2013**, arXiv:1312.6229.
40. Chen, Z.; Jacobson, A.; Sünderhauf, N.; Upcroft, B.; Liu, L.; Shen, C.; Reid, I.D.; Milford, M. Deep learning features at scale for visual place recognition. *arXiv* **2017**, arXiv:1701.05105.
41. Lowry, S.; Sünderhauf, N.; Newman, P.; Leonard, J.J.; Cox, D.; Corke, P.; Milford, M.J. Visual place recognition: A survey. *IEEE Trans. Robot.* **2016**, *32*, 1–19. [CrossRef]
42. Yu, J.; Tao, D.; Rui, Y.; Cheng, J. Pairwise constraints based multiview features fusion for scene classification. *Pattern Recognit.* **2013**, *46*, 483–496. [CrossRef]