Visual Place Recognition under Severe Viewpoint and Appearance Changes

Ahmad Khaliq

School of Computer Science and Electronic Engineering
University of Essex

The thesis is submitted for the degree of
Master by Dissertation

January 2020

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements.

Ahmad Khaliq
January 2020

Acknowledgements

First and foremost, I would like to thank the God Almighty for giving me the strength, ability and opportunity to undertake this study and complete it successfully. Without His blessings, it would not have been possible to conduct this extensive research work.

My heartiest gratitude to Prof. Klaus McDonald-Maier for offering me to pursue MSc by Dissertation under his supervision. It can be very challenging to take an international student with a different cultural background and way of study. However, undertaking the MSD study under his guidance has been truly a life changing experience. I would like to say thank you to Dr. Shoaib Ehsan for his continuous guidance and support during the whole period of study. His excellent research skills and constructive criticism, coupled with the deep insightful discussions have always been a great source of motivation. During my MSD, he was very generous and put-in a lot of efforts and endless patience to co-supervise me. I have learnt the art to go from idea-to-execution in an efficient manner resulting in novel research contributions.

I extend my gratitude to Prof. Michael Milford, Associate Professor at Queensland University of Technology, Australia, for providing me with his invaluable feedback on my research in all the jointly collaborated works. His survey paper in the field of visual place recognition served as an initial motivation to choose my MSD topic of research in the field of robotic vision. Furthermore, I am very grateful to his student Dr. Zetao Chen, who is currently a postdoctoral research fellow in ETH Zurich for providing me the benchmark datasets and their evaluated results. It helped me in carrying out a fair comparison of my proposed research contributions with state-of-the-art algorithms.

Many thanks to Prof. Dongbing Gu and Dr. Amit Singh, member of my supervisory panel for their encouragement, thoughtful guidance, critical comments and valuable feedbacks. They all were very helpful and provided me with their assistance throughout my research. During my MSD, I had the opportunity to work on a European funded INTERREG V 2 SEAS INCASE project. Under the guidance of Prof. Klaus McDonald-Maier and Prof. Dongbing Gu, I have gained experience in Industry 4.0 which resulted into novel research contributions while collaborating with INCASE project partners including KU Leuven, Technologie campus Gent, Belgium. As a member of the INCASE team, I really enjoyed

working with postdoctoral research fellow Dr. Sangeet Saha. I'm very grateful for his help, advices and really enjoyed the time we have spent together in the lab and attending INCASE meetings held in France and Belgium.

My colleagues in the EIS lab have made my MSc by research a very enjoyable journey. I would like to thank all my office colleagues which include Mubariz Zaffar, Dr. Sangeet Saha, Zeba Khanam, Dr. Bilal Aslam, Somdip Dey, Bina Bhatt, Anca Sticlaru, Grigorios Kalliatakis and many more. I want to thank to Philip George for his administrative support within our EIS lab. I am grateful to all of them - for keeping an atmosphere of friendship and the warmth they extended to me and for always making me feel so welcome. From this experience, I have learnt different culture and will always remember the time spent in enjoying food while having fun talks.

One of my closest friend; research officer and postgraduate research student Mubariz Zaffar, I especially would like to mention him here. The friendship we have built during this whole period, I really appreciate your support. You've always given me your honest opinion and stood by me whenever required. I will never forget the time we have spent staying late at office which resulted into some excellent joint research contributions. I really enjoyed your company and wish you a very bright future ahead.

Finishing this without acknowledging my family support is not conceivable. Coming all over from Pakistan to United Kingdom to pursue postgraduate research was never an easy journey. I want to express my heartfelt appreciation to my lovely mother and father for their great patience, for believing in me and encouraging me to follow my dream. Without the prayer of my parents, I could not have finished this work. I can just say thanks for everything and pray that may God give them long lives and keep them well in body and spirit.

Abstract

Over the last decade, the eagerness of the robotic and computer vision research communities unfolded extensive advancements in long-term robotic vision. Visual localization is the constituent of this active research domain; an ability of an object to correctly localize itself while mapping the environment simultaneously, technically termed as Simultaneous Localization and Mapping (SLAM).

Visual Place Recognition (VPR), a core component of SLAM is a well-known paradigm. In layman terms, at a certain place/location within an environment, a robot needs to decide whether it's the same place experienced before? Visual Place Recognition utilizing Convolutional Neural Networks (CNNs) has made a major contribution in the last few years. However, the image retrieval-based VPR becomes more challenging when the same places experience strong viewpoint and seasonal transitions. This thesis concentrates on improving the retrieval performance of VPR system, generally targeting the place correspondence.

Despite the remarkable performances of state-of-the-art deep CNNs for VPR, the significant computation- and memory-overhead limit their practical deployment for resource constrained mobile robots. This thesis investigates the utility of shallow CNNs for power-efficient VPR applications. The proposed VPR frameworks focus on novel image regions that can contribute in recognizing places under dubious environment and viewpoint variations.

Employing challenging place recognition benchmark datasets, this thesis further illustrates and evaluates the robustness of shallow CNN-based regional features against viewpoint and appearance changes coupled with dynamic instances, such as pedestrians, vehicles etc. Finally, the presented computation-efficient and light-weight VPR methodologies have shown boostup in matching performance in terms of Area under Precision-Recall curves (AUC-PR curves) over state-of-the-art deep neural network based place recognition and SLAM algorithms.

Contents

List of Figures

List of Tables

Abbreviations

AMOS	*Archive of Many Outdoor Scenes*
ASMK	*Aggregtaed Selective Match Kernels*
BRISK	*Binary Robust Invariant Scalable Keypoints*
BoW	*Bag of Words*
CNNs	*Convolutional Neural Networks*
FABMAP	*Fast Appearance Based Mapping*
FAST	*Features from Accelerated Segment Test*
GLD	*Google Landmark Dataset*
GPS	*Global Positioning System*
HCT	*Hull Census Transform*
LBP	*Local Binary Pattern*
LDB	*Local Difference Binary*
PCA	*Principal Component Analysis*
RMAC	*Regional Maximum Activations of Convolutions*
SEQSLAM	*Sequence SLAM*
SIFT	*Scale Invariant Feature Transform*
SLAM	*Simultaneous Localization and Mapping*
SPED	*Specific Places Dataset*

SPP	*Spatial Pyramid Pooling*
SURF	*Speeded – Up Robust Feature*
SVM	*Support Vector Machine*
VGG	*Visual Geometry Group*
VLAD	*Vector of Locally Aggregated Descriptors*
VPR	*Visual Place Recognition*

Chapter 1

Introduction

Over the last few years, significant improvements have been made in autonomous driving and robotic vision [11]. For a safe and continuous operation, a vigorous navigation system is indispensable. To correctly localize within an environment, the object needs to build a map of its surroundings, technically termed as Simultaneous Localization and Mapping. Visual Place Recognition is a prime component of SLAM; a system which can decide whether the place has previously been visited or not, also termed as a loop-closure detection. The aim of this thesis is to improve visual place recognition for battery-operated mobile robots under changing conditions, including appearance and viewpoints variations coupled with dynamic instances. This thesis presents methodologies that increase the robustness of SLAM by improving the performance of visual place recognition at low memory and computation cost. Our proposed frameworks allow mobile robots to globally localize by identifying places which have been previously visited given changed conditions and viewpoints.

This chapter discusses the main topic of this thesis; visual place recognition under severe viewpoint and appearance changes. It starts with the background in Section 1.1, followed up with the introduction of SLAM in Section 1.2. Section 1.3 establishes the link between VPR and visual-SLAM. Section 1.4 highlights the challenges in VPR along with the thesis objectives and contributions are presented in Section 1.5. Section 1.6 outlines the dissertation organization and research contributions made during this research are listed in Section 1.7.

1.1 Background

Over the past few decades, autonomous vehicles and intelligent mobile robots have attracted increased level of attention from the research communities and industrial organizations [12][13]. In 2005, American Department of Defense, DARPA (Defense Advanced Research Projects Agency) organized a vehicle Grand challenge to promote the development of

autonomous vehicles. The challenge winner STANDEY, an autonomous car created by Standford University, USA had employed SLAM as part of its autonomous driving system. It attracted many top level research organizations such as, Google which started its self-driving car project in 2009 and with real-time autonomously driven 1.5 million miles, it is presently steering through the street of Mountain View, California. The first Google self driving car was released in 2014, followed up by many other companies and institutions including Baidu, MIT, BMW and Uber as shown in Fig. 1.1. A number of autonomous mobile systems have been demonstrated including Mars rover car (2014) and Google self driving vehicle (2015).

Fig. 1.1 Examples of human cars with inbuilt self-driven capabilities.

Encouraged by the recent success of autonomous vehicles, some indoor autonomous robotic experiments claim that with minor occasional human interventions, the robot can run autonomously for months using a visual sensor such as, a camera [14][15]. However, real world outdoor scenes are quite challenging and a long-term autonomous navigation system under such extreme environmental changes is still a big question that needs to be answered. Visual information is susceptible to appearance changes because the same place can undergo drastic environmental variations and perceptual changes due to the seasonal, weather and illumination variations. Most localization and navigation systems match places based on the captured RGB information, therefore, such appearance changes cast a severe challenge. Intelligent vehicles or mobile robots comprise the fundamental systems including navigation, localization and perception. Localization is the core component that leads to successful accomplishment of other tasks because by determining the current position/location, path

planning can be performed. Therefore, accuracy of the localization system is directly associated with autonomous vehicles or mobile robots.

Localization can be accomplished through visual place recognition, which tries to retrieve a place from the previously visited places. This dissertation focuses on lightweight visual place recognition with an ultimate goal to localize at low memory and time cost. We expect resource-constraint mobile robots to benefit from this work.

1.2 Simultaneous Localization and Mapping (SLAM)

To recognize a previously visited place, the system needs a map to perform robotic localization. Building such maps of the surrounding environment for localization is termed as Simultaneous Localization and Mapping (SLAM). In the past few decades, SLAM is a very active area of research among the robotic and computer vision communities. It consists of following components:

1.2.1 Mapping

It's the process of internally creating and storing the outside world representation. The generation of world maps can vary from the type of environment and sensory information [16]. For instance, occupancy grid maps [17] subdivides the world into evenly spaced discrete regions or cells, as shown in Fig. 1.2. Each cell is assumed to be occupied or free, named as occupancy grid maps. Few assumptions are made for occupancy grid maps; each cell is either free or occupied and independent upon each other. Given all the past information and states, each region has an obstacle probability. Those regions which are nether occupied nor explored, they have assigned with 0.5 obstacle probability. In dense indoor environments, such occupancy-based grid maps fit perfectly as they can easily be constructed from laser- or sonar-based information but require significant memory resources. These maps have been widely employed in obstacle avoidance algorithms including potential field [18]. In comparison, feature maps are quite compact, sparse and consist of distinctive landmarks with their coordinates. They are suitable in outdoor environments with information about the landmarks obtained from the visual sensors. Pose graphs [19] are another form of maps that graphically defines the robot trajectory. The nodes orientation describe the robot's pose and position. Edges describe the spatial connection between the poses, can be used for loop closure.

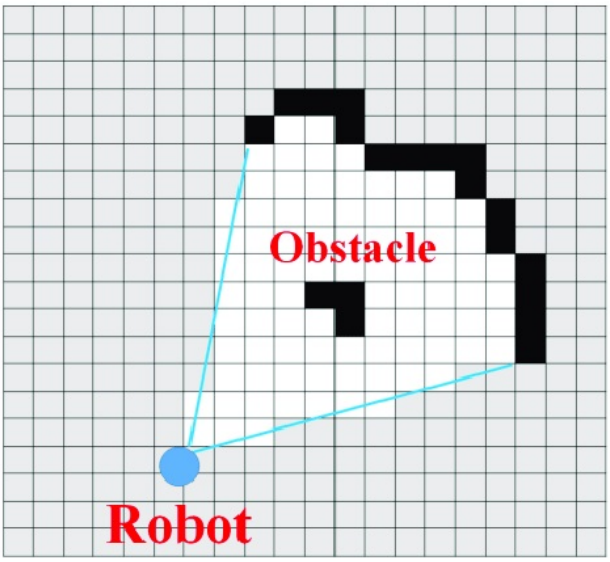

Fig. 1.2 An occupancy grid map; the world is subdivided into discrete regions. Black and white regions indicate the presence of obstacles and explored areas. Grey regions are not yet observed. (Image taken from [1])

1.2.2 Localization

Employing visual sensors, Localization is the process of tracking the position of robot within the map. This module processes the incoming visual data and outputs a belief about the current place within the map. Using the belief, system decides whether the presently encountered place is previously seen or a new place. Two similar places usually implies the same location. However, this supposition does not work when there are significant weather and illumination changes which might cause the robot to incorrectly localization within the map.

Localization can either be local or global. In global localization, the robot assumes that there is no prior knowledge and can move anywhere within the map. Such re-localization of the robot using visual information and map give rise to kidnapped robot problem. Environments within which the places are quite similar, global localization is difficult to achieve. However, the robot generally has some prior information of its current pose and map; known as local localization [20]. The prior knowledge comes from the previous states of the robot.

1.3 Visual Place Recognition (VPR)

In an outdoor environment, Global Positioning System (GPS) offers a cheap solution for localization. However, it requires satellite coverage which in some cases is intermittently

available in areas with trees or taller buildings because signals are seldomly out of reach. As an alternative, visual sensors are employed and have shown state-of-the-art performance in place recognition and SLAM based applications [21][22][23][24].

Contemporary robotic localization systems use vision sensors [25][26]. For visual localization, place recognition employing visual cameras are cost-effective and lightweight. Taking precedence from their sizes and power consumptions, their applicability can be expanded to mass-production. Secondly, the incoming visual data contains rich semantic information coupled with the texture and appearance of places. Moreover, a visual source of information can provide better understanding of the environment especially for far away landmarks including buildings structures etc. Visual place recognition matches a place from previously encountered places and allocate the current position (place) within an environment either performing single image or sequence of image matching, known as image retrieval task [27][28]. Here, a place is interpreted as a part of an environment or abstraction of a region, where a region corresponds to a two-dimensional subset of the environment [29]. Therefore, each place can be denoted as an image and localization through visual place recognition can be achieved either by employing a single image or sequence of images [30].

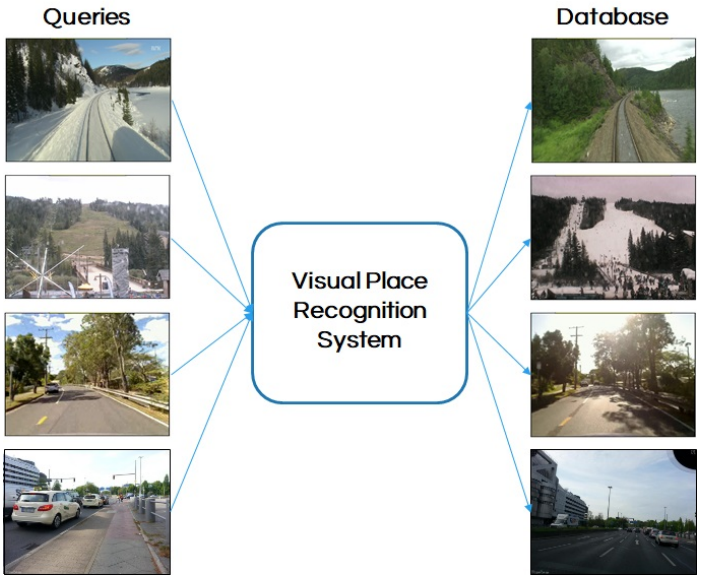

Fig. 1.3 A visual place recognition system takes query images as an input and returns the visually similar database images for localization.

In visual place recognition, captured images of the places within an environment are stored in a database. Therefore, runtime comparison of the query with the stored database is performed with closest database image which is considered as a currently visible place (location) as shown in Fig. 1.3. Recent advancements in computer vision improve the

performance of localization systems utilizing different robust feature detectors that interpret visual information efficiently.

1.4 Problem Statement and Challenges

Visual place recognition is a well-known paradigm where contemporary algorithms perform relatively well in environments with minimal dynamic instances but often challenging in complex outdoor and extreme environments. In particular, due to large-scale environmental variations experienced in the form of drastic appearance changes coupled with viewpoint variations, visual place recognition becomes difficult to achieve. This dissertation addresses the visual place recognition challenges by making sure that even under such perceptual and environmental changes, the object still correctly localize itself within the environment using the visual information. In Fig. 1.4, two cases of place recognition are presented. In (a), under day-night transition, the system still retrieves the correct match for a query image taken at the same place at night time whereas in (b), the retrieved image looks very similar to the query but geographically different, a problem known as *perceptual aliasing*. An efficient visual place recognition system should be able to correctly localize (match) a place (location) while minimizing the incorrect match.

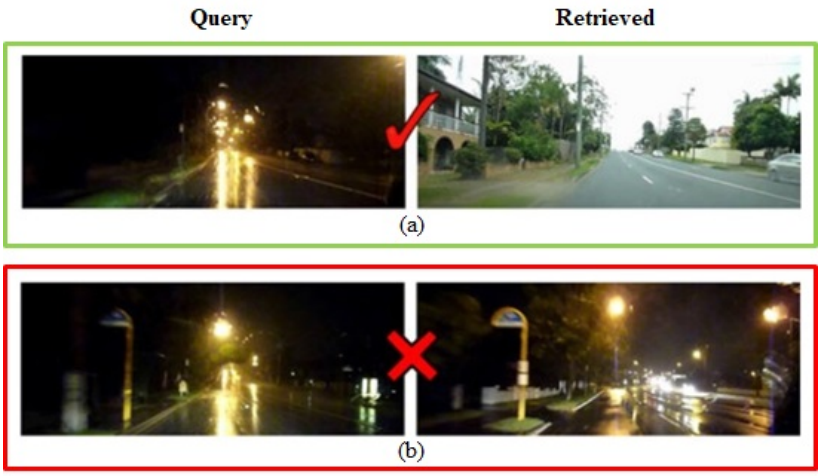

Fig. 1.4 A generic visual place recognition system must be able to successfully recognize (a) the correct place irrespective of the visual changes and (b) reject the visually similar but geographically different places (Image taken from [2]).

To differentiate two places or locations in the form of captured images, the first step is to produce a compact image representation in the form of feature descriptors. Two geographically different places may look similar due to the surrounding atmosphere and

environment, such as, a bar, an office etc. Therefore, to recognize the place correctly, it is important to identify and employ distinguishing and meaningful feature representations. In layman terms, we can say that the visual place recognition (localization) performance depends upon the feature descriptors being used for manipulating the visual information. Widely used feature descriptors include hand-crafted techniques (such as, Scale-Invariant Feature Transform [31] and Speeded-Up Robust Features [32]), deep learning techniques and 3D-based information (will discuss in more detail in Section 2.2). Similarly, techniques involved in feature matching vary and a better matching approach can improve the overall retrieval accuracy. Two widely employed recognition techniques include single image matching and sequence-based image matching. The most vital components involved for both feature description and feature matching are computation time and resource utilization. The mobile robots are usually battery operated and should be able to localize in real-time environment with minimum resource utilization.

1.5 Thesis Contributions

This thesis works around lightweight deep learning techniques coupled with single image matching for visual place recognition under changing appearance and viewpoint. The proposed approaches are tested and evaluated on several publicly available benchmark place recognition datasets, where the final outcome of this dissertation is to have a real-time lightweight visual place recognition system. The main contributions of this thesis are as follow:

1. The first contribution is associated with the exploration of regions-based Convolutional Neural Networks (CNNs) techniques that can be effective for place recognition under uncertain perceptual and environmental variations. Two places captured from the same location can appear differently due to the changing viewpoints and conditions. Therefore, identifying the common regions across places can improve the VPR performance. I have proposed a novel place-centric, region-finding approach employing convolutional layers of the CNN models.

2. The second contribution addresses the scalability and runtime matching performance. Once the robot experiences a new environment, it captures, stores and updates the reference map. It is important to have compact representation of the stored information so that runtime feature matching can be fast. The identified regional feature representations are encoded into Vector of Locally Aggregated Descriptors (VLAD). Precisely, regional features are quantized to the pre-trained dictionary clusters and their

accumulated residuals are concatenated to retrieve a VLAD representation, followed by their cosine matching such that the database image with highest score is treated as the final match.

3. Deep learning techniques are computationally expensive, and it becomes important to determine the runtime resource and memory utilization. Thus, it is critical to make sure that the computation and memory requirements are economical which makes the proposed place recognition frameworks suitable for resource-constraint mobile robots. Instead of deep neural networks, we have employed shallow CNN models to meet the real-time demand while improving the recognition performance at low memory and time cost. We have experimentally shown that our proposed region-based features extracted from less-layered CNNs can still deliver better results against state-of-the-art deep CNN-based place recognition contemporary techniques.

1.6 Thesis Structure

The rest of the thesis is divided into four chapters:

In **Chapter 2**, existing approaches for visual place recognition are reviewed. Depending upon the feature describing methodologies, the visual place recognition techniques are classified into: techniques which employ global descriptors, local descriptors, and techniques that are a combination of both the global and local descriptors. Other approaches employ deep CNNs and 3D information for visual localization. Similarly, feature matching can either be single image or sequence of images based matching, as discussed further in this chapter.

As presented in Section 2.2.5, region-based CNN techniques are capable of improving the VPR performance under changing environmental conditions. Taking inspiration from the regional techniques, a novel regions finding approach is proposed in **Chapter 3**. To reduce the memory footprint and time consumption, shallow CNN (AlexNet) pre-trained on scene-centric Place365 dataset [33] is employed. CNN-based regional features are coupled with VLAD encoding for single image matching. The proposed framework is evaluated on several challenging benchmark datasets and achieves boostup in matching speed and accuracy over state-of-the-art contemporary VPR algorithms in terms of area computed on precision-recall curves.

Rich semantic CNN-based regional features have shown robustness against severe visual changes along with moderate viewpoint variations, as illustrated in section 2.2.5. **Chapter 4** presents a multi-scale attention-based CNN approach for environment invariant visual place recognition. The presented technique is tested on publicly available datasets with environment experienced by real-world robots. The multi-layer context-aware framework

is employed on shallow HybridNet (pre-trained on place recognition-centric SPED [4]) and achieves better performance than deep neural network based VPR approaches at lower memory footprint and computational resource utilization.

In **Chapter 5**, a summary of the achieved outcomes and discussion of their relevance to the current research and future work in visual place recognition is presented.

1.7 List of Publications

Following contributions were made during this course of study:

1. Khaliq, A., Ehsan, S., Chen, Z., Milford, M., and McDonald Maier, K. (2019)., "A Holistic Visual Place Recognition Approach using Lightweight CNNs for Significant ViewPoint and Appearance Changes". Accepted and published as a short paper in IEEE Transactions on Robotics (T-RO).

2. Zaffar, M., Khaliq, A., Ehsan, S., Milford, M., and McDonald-Maier, K. (2019b)., "Levelling the Playing Field: A Comprehensive Comparison of Visual Place Recognition Approaches under Changing Conditions". Accepted in IEEE International Conference on Robotics and Automation (ICRA 2019) workshop.

3. Zaffar, M., Khaliq, A., Ehsan, S., Milford, M., Alexis, K., and McDonald-Maier, K. (2019a)., "Are State-of-the-art Visual Place Recognition Techniques any Good for Aerial Robotics?". Accepted in aerial robotics workshop at IEEE International Conference on Robotics and Automation (ICRA 2019).

Chapter 2

Literature Review

This chapter presents an overview of the relevant research work in the area of visual place recognition. It starts with the core components, namely: describe a place, store a place representation and recognize a revisited place, followed up by their pertinent works respectively.

2.1 Overview

With the availability and accessibility of economical cameras which provide rich visual information, vision-driven localization and place recognition are getting more and more attention [34][35][36][37]. In the context of place recognition, visual information of a place/location is stored in the form of the captured image. A vehicle or robot localizes itself within an environment by identifying and recognizing the location previously encountered through an image retrieval mechanism [2]. However, the recognition techniques should be robust such that even when the estimated metric position of the robot is inaccurate, it can still localize the robot.

The contemporary research challenge in visual place recognition is to deal with the uncertainty of the changing environment [38][39] because the appearance of the same place changes abruptly at multiple times of the day and months, coupled with viewpoint changes induced by the different viewing positions of the robot [40]. Fig. 2.1 illustrates the basic mechanism of vision-aware place recognition with components described as follows.

1. Image processing module: The module takes the visual data as an input, coming from the visual sensors (such as, camera) and processes the rich semantic visual data, followed by the identification and extraction of useful information in the form of feature descriptors [2].

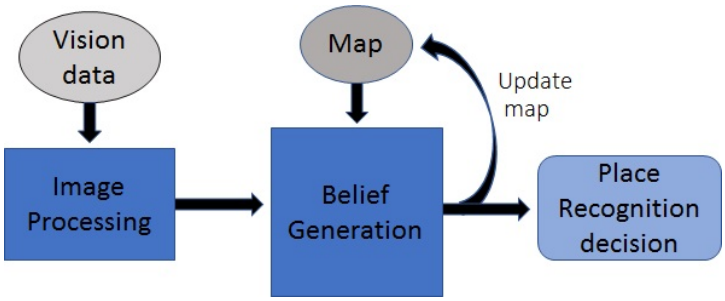

Fig. 2.1 Incoming visual data is processed by the image processing module and its description is stored in the place mapping framework. The belief generation module takes the decision by matching the current location with the stored places.

2. Place mapping framework: A map of the surrounding environment is stored in the form of the feature descriptors. It updates the map once encountered with the new places. Depending upon the purpose, a places map can be divided into following categories:

   (a) Database-centric: It is a simplest way to remember a particular environment by storing the visual information in the form of captured frames. Therefore, recognition is entirely based on how similar the visual places are, often treated as image retrieval problem [41]. However, valuable information including relative pose is not stored which makes place recognition less precise but computationally efficient. This thesis is focused on database-centric place remembering approach.

   (b) Topological maps: This category of remembering places captures the relative information of the location in an environment [37][27]. It can be a collection of images captured in sequence.

3. Belief generation module: Within this module, the query/current visual information is compared with the stored information and the system retrieves the best matched database image (place recognition decision). Generally, if two places are captured at the same or different location then their descriptors similarity is determined by the matching score. Matching with single image and sequence of images are the two conventional place recognition approaches [42]. However, sequenced-based matching is more robust as it reduces the false-positive scenarios but this thesis focuses on single image matching for place recognition as it is computation efficient.

To perform the task of visual localization via appearance-aware place recognition, it is necessary to extract the useful visual information in the form of feature descriptors. Under uncertain environmental variations, the performance of the place recognition system depends upon the approaches employed for processing the visual information. Existing visual feature

description approaches differ on the basis of visual data processing. Therefore, these can be: local features, global features, a combination of both the local and global features, 3-Dimensional features and deep learnt CNN features. Similarly, recognizing approaches either use single image of a place or collection of images captured in a sequence for matching purpose.

In this chapter, we will discuss the contemporary techniques proposed for place recognition based visual localization. The rest of the chapter is organized as follows: Section 2.2 discusses the multiple place describing approaches. Section 2.2.6 presents the matching techniques used for place recognition. Section 2.3 highlights the contemporary state-of-the-art VPR techniques. Section 2.4 and 2.5 discuss the benchmark place recognition datasets and evaluation criteria. Section 2.6 summarizes this chapter.

2.2 Visual Place Recognition Methodologies

A place in an environment is a distinctive location and should be describe in such a way that it should be scalable and efficiently recognized whenever revisited. In the literature, many techniques for place recognition have been proposed which differ in the process of extracting visual information. Some approaches select regions that are in some way meaningful and notable; and others that process all the visual data with no region selection block. Similarly, there are techniques that use the 3-dimensional information and state-of-the-art neural network based techniques in which an image is passed into the input layer, and responses at some certain layer(s) are pooled and employed as feature descriptors.

2.2.1 Approaches based on Local features

Techniques employing local features first analyze the visual data (image) and then the detection of the meaningful keypoints at various spatial locations is carried out, named as local features [43][44]. The detection is based on the distinctive pixel patterns coupled with the description at that spatial location. Within each location, concatenation of the neighbour pixels is performed for retrieving final multi-dimensional floating-point feature vector or bit strings [45].

Local features are typically invariant to affine transformations including image scaling and camera rotation. Therefore, in places with similar environment and scenes, conventional local feature descriptors can be utilised. Research work in the context of place recognition is still a growing research domain, as evidenced by citation analysis and several workshops and IEEE conferences including IEEE International conferences on Robotics and Automation

Fig. 2.2 Example of local and global descriptors for place recognition (Image taken from [2]).

and Computer Vision and Pattern Recognition. It all started from the development of Scale-Invariant Feature Transform (SIFT) detector [31] and Speeded-Up Robust Features (SURF) [32], used for visual localization shown in Fig. 2.2. In Fig. 2.2 (a), an image is processed by local SURF feature detector, with circles denote the keypoints. For global descriptor as in Fig. 2.2 (b), the image is subdivided into grids and each block is separately processed. Place recognition with local descriptors is usually robust under viewpoint changes whereas global approaches are efficient under conditional changes.

Recent development of local binary feature detectors enables the research community to improve the description and recognition of places [46]. Such detectors are quite efficient and invariant to monotonic gray-scale changes. A typical Local Binary Pattern (LBP) feature detector is employed in [47], and coupled with support vector machine (SVM) based model for place recognition. Other local detectors include Binary Robust Independent Elementary Features (BRIEF) [48], oriented BRIEF (ORB) [49], Binary Robust Invariant Scalable Keypoints (BRISK) [50], Local Difference Binary (LDB) [51], KAZE [52] and Fast Retina Keypoints (FREAK) [53], which were usually employed in earlier place recognition systems.

Local feature descriptors possess high discriminative power which results into better recognition performance. However, they are computationally inefficient and suffer from higher dimensionality of the features. Bag-of-Words (BoW) [54], is an efficient feature quantizing technique where local features are assigned to particular centroid (words) of a trained vocabulary, as shown in Fig. 2.3. Thus, BoW results in a compact representation of the place with low dimensional vector or binary string. It ignores the geometric information of the place and performs recognition regardless of the topology of the deatures, technically termed as pose invariance. However, it is very sensitive to extreme conditional variance [55] and underperforms under lightening and seasonal changes. It is because the identified local feature descriptors are less distinctive in such kind of visual changes which leads to incorrect place matching.

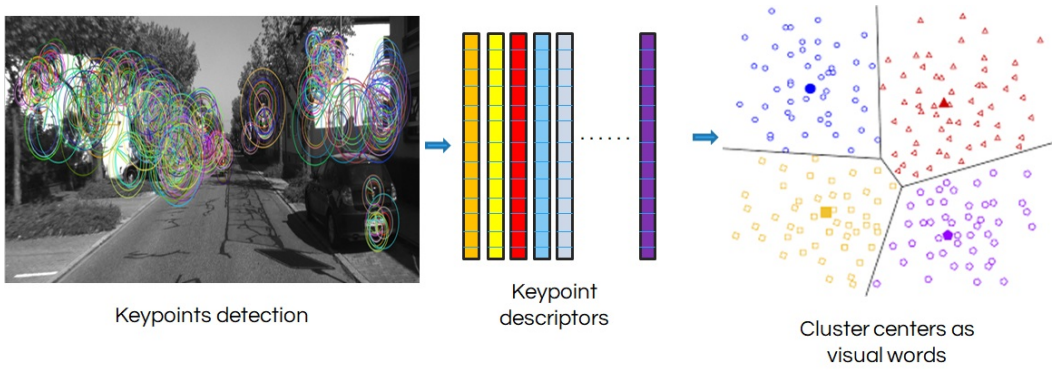

Fig. 2.3 K-means clustering of the feature descriptors with each cluster center treated as a visual word.

2.2.2 Approaches based on Global features

The previous section reviewed methods that put more focus on sub-regions/parts of the image. Such feature extraction techniques are quite efficient under partial occlusions and viewpoint variance, but they do not generally consider the whole scene or the structure of the place. In comparison, this section discusses the global feature detectors which process and describe the whole image for place recognition.

Global features are quite fast and robust under uncertain environmental variations. For place recognition, widely used global descriptors include color histograms [37] and Histogram of Oriented Gradient (HOG) [56]. Ulrich et al. in [37] used six one-dimensional color histogram from HLS and RGB color spaces coupled with nearest neighbor scheme in their topological map for retrieval. Their whole image-based place recognition system successfully matched 87% of the images. In [56], a vision-aware navigation system was proposed within which HOG employed as feature representations, followed-up by the descriptors comparison for determining the similarity.

Later, Winters et al. in [57] have utilized an omni-dimensional camera for creating the topological map. Principal Component Analysis (PCA) was employed for image compression and dimensionality reduction with appearance-aware localization determines the global topological position of the robot. A popular global descriptor GIST [58], introduced for scene recognition has been employed in [59] for place recognition. It used Gabor filters at different orientations and frequencies for features extraction, resulting into a compact vector representing the overall scene.

Concatenation of local features is also studied and represented as global image representation. Using omni-directional cameras, Lamon et al. in [60] used edges, corners and color patches, combined them as features for rotation-invariant place recognition. A whole image based technique WI-SURF used SURF detector on the whole image with investigations

claiming better matching performance for long-term localization. Similarly, combination of the BRIEF-GIST [61] detector for place recognition found to be resource efficient without the need of trained vocabulary.

Other approaches include grid-based image partitioning (illustrated in Fig. 2.2(b)), followed by the concatenation of each grid feature to retrieve the final descriptor. Lategahn et al. in [62] proposed an illumination robust feature over small image regions and normalized filter responses were used with results confirming better loop closure performance. Generally, global features are convenient to determine and scalable for large-scale place recognition. These approaches are invariant to conditional changes but less effective under viewpoint variations.

2.2.3 Approaches based on a Combination of Local and Global features

Both local and global feature descriptors have shown increasing level of integration for early visual place recognition techniques. Taking precedence of their individual advantages, the research community has proposed several frameworks based on a combination of feature descriptors for place recognition [38].

Murillo et al. in [63] used omni-directional camera and proposed a three-step hierarchical architecture for localization. Color based global detector was applied, followed by the line features description to retrieve the visually similar image employing pyramidal matching of their line supported regions. Using omni-directional camera, Goedemé et al. in [64] extracted the vertical segments of each image coupled with the description of ten different descriptors. The local descriptors were clustered and then inserted into a kd-tree structure for efficient retrieval. For each query image, same descriptors were employed on the vertical image segments and later used for possible loop candidates retrieval. For accurate matching, column segments' matching distance was applied between the candidates and query image.

A real-time appearance-aware place recognition system was proposed in [65] which combined Features from Accelerated Segment Test (FAST) and Complete Center-symmetric Local Binary Patterns (CSLBP). BoW and SVM were integrated with results showing robust and high real-time place classification. A combination of local and global descriptor based techniques, Hull Census Transform (HCT) proposed in [66]. It first filters the SURF features which are composed of convex hull then computed their relative magnitude, resulting into a set of binary vectors efficient for scene detection. A similar approach in [67] combined local features, edges and colour histograms such that Harris detector was used for regions of interest and edges, followed up with SIFT description.

2.2.4 Approaches based on 3D information

In addition to rich semantic 2D visual data, metric information transforms it into three-dimensional information. Stereo cameras are employed as a source of metric information whereas monocular cameras with structure-from-motion techniques such as MonoSLAM [68], LSD-SLAM [69] and ORB-SLAM [70] can also provide the required metric information.

In the literature, several works have been proposed employing three-dimensional information for visual place recognition. Cummins et al. in [27] extended the Fast Appearance Based Mapping (FAP-MAP) by adding spatial distribution of 3d-based visual words. A similar strategy was proposed in [71] that coupled the 3D metric information with stereo image sequences for place localization. In [72], Morioka et al. have proposed a SLAM based navigation system which extracted 3D Position Invariant Robust Feature (PIRF) from the sequential images with results demonstrating effective outdoor localization. A 3D point cloud and depth image based descriptor which is a variant of Surface Entropy for Distinctive 3D Features (SURE) is showed in [73] discussing the applicability of SURE features coupled with BoW approach for indoor place localization.

Maddern et al. in [74] proposed Continuous Appearance-based Trajectory SLAM (CAT-SLAM), an appearance-based place recognition system that filtered the local metric pose for improving the frequency and reliability of loop closure. Its extension integrated appearance with local odometry information, named as CAT-Graph later introduced in [75]. With a large-scale loop closure detection, it demonstrated the recall boostup in performance from FAB-MAP by a factor of 3 at 100% precision. Cadena et al. in [76] introduced a stereo vision-based recognition architecture which employed BoW feature encoding for retrieving loop closure candidates, coupled with Conditional Random Fields-Matching (CRF-Matching) for verification. However, the matching approach was found to be more robust than using epipolar geometry only because it used 3D information from stereo images.

Cadena et al. in [71] proposed a SLAM-based place recognition system by considering both the visual and geometric information coming from the camera. Loop closure hypothesis was evaluated by coupling the appearance based approach with BoW model. Results in both indoor and outdoor environments claimed zero false-positives (full precision) for a few false-negatives (high recall). Sensors including RGB-D cameras [77] were used and coupled the depth information with visual data and further improved place recognition system/performance [78].

2.2.5 Approaches based on Deep CNN features

Visual Place recognition techniques based on handcrafted features are generally sensitive towards simultaneous conditional and viewpoint variations. Therefore, their performance in uncertain extreme environments relies on the type of feature detector employed. FAB-MAP, a combination of SURF-BoW was found to be robust in dealing with viewpoint changes whereas SEQSLAM performed better in changing conditions due to its whole image-based approach. However, such techniques have shown inferior performances under simultaneous changes observed in illumination, conditional or different camera position. Thanks to deep convolutional neural networks which have shown stronger generalization power to describe the places [79] with robustness to conditional and viewpoint variations.

In 1989, LeCun et al. in [80] for the first time introduced the idea of a Convolutional Neural Network (CNN). The proposed architecture consisted of multi-layered network, trained on human annotated datasets and automatically learned features at multiple scales through classification-based training. In comparison with conventional handcraft-based feature detectors, CNNs have shown state-of-the-art performance for image/object classification and recognition [81][82]. However, collection of the labelled large-scale datasets for training the neural network and resources for computation are the early limitations faced by the research community. The recent advent of GPUs encouraged and allowed the research community to perform resource intense calculation, such as, back-propagation [83].

Encouraged from the initial boom of deep learning, Chen et al. in [84] presented a visual place recognition system employing powerful CNN features coupled with spatial and sequential filter. Evaluating the proposed framework with a 70km Eynsham dataset claimed 85.7% recall at 100% precision. Later, [40][85] coupled external landmark detectors with pretrained object-centric CNNs employing locality-sensitive hashing and optimization approach for region-based real-time place recognition. The regional CNN features were found to be robust under strong viewpoint and conditional changes. Further investigation in [86][4] demonstrated that middle convolution layers emphasize upon edges and colours, thus, efficient under conditional variations.

Arandjelovic et al. in [3] added a Vector of Locally Aggregated Descriptor (VLAD) layer inside the CNN architecture (shown in Fig. 2.4). Using back-propagation, it trained the model on a newly collected large scale urban dataset containing dynamic objects, coupled with appearance and viewpoint changes. The framework was evaluated on Tokyo, Oxford and Paris datasets with results claiming superior performance of deep CNN features. Gomez-Ojeda et al. in [87] trained a CNN model for recognizing revisited places under strong conditional variations. Images were mapped onto the lower dimension with euclidean distance calculation for place similarity. To deal with the appearance variance, the authors

used triplet of images for training the neural network such that one image presents the same place with different conditions and other with different place exhibiting similar environment.

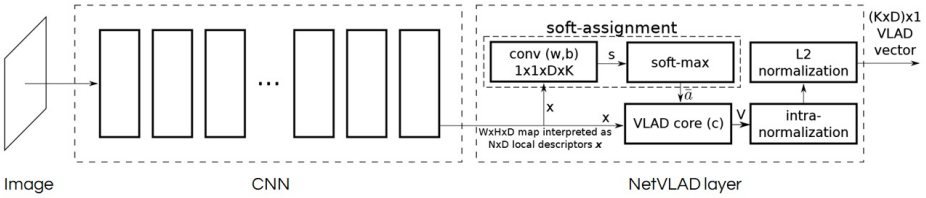

Fig. 2.4 Convolutional Neural Network (CNN) with the NetVLAD layer (image taken from [3]).

Pooling responses from convolutional layers have been an area of research interest for computer vision and robotic communities. With the advent of several feature pooling techniques including Sum-Pooling [88], Max-Pooling [10], Spatial Max-Pooling [89] and Cross-Pooling [90] employed in deep CNNs have demonstrated performance boost in tasks requiring image classification/recognition and object detection/retrieval [10][90]. Applying a specific sized window over convolutional layer's feature maps and picking responses either based on the max value, termed as Max-Pool [10], adding all the values - Sum/Average-Pool [88] or mapping the window's responses into the previous layer, known as Cross-Pool [90].

All such response pooling techniques have shown performance boost in vision-based image retrieval tasks where the image is majorly covered with a single object. Responses within feature maps are non-uniformly distributed and finding single or fewer regions of interest becomes relatively easier. However, such image retrieval tasks are different in nature from the VPR systems where recognizing a place which undergoes diverse changes due to illumination, winter to summer transitions or viewpoint variance added by different capturing angles is quite challenging. It is because the same place appears differently thus making it harder to identify the common regions. Even when the above mentioned pooling techniques are integrated on external tasks based pre-trained CNNs for the VPR problem, the convolutional layers feature maps focus on the trained objects such as vehicles, pedestrians and other non-salient objects which are not suitable for place recognition [5]. Therefore, a generic Visual Place Recognition system capable of efficiently dealing with simultaneous viewpoint and condition variations remains an open challenge.

Using an Archive of Many Outdoor Scenes (AMOS), a 2 Million diverse Specific Places Dataset (SPED) was collected in [4]. SPED contains thousands of images captured at each place throughout the year. Each place is treated as a label and with millions of places allowing classification-based fine-tuning (HybridNet) of the object-centric CaffeNet [91] and training from scratch (AMOSNet). Spatial Pyramid Pooling (SPP) for feature extraction

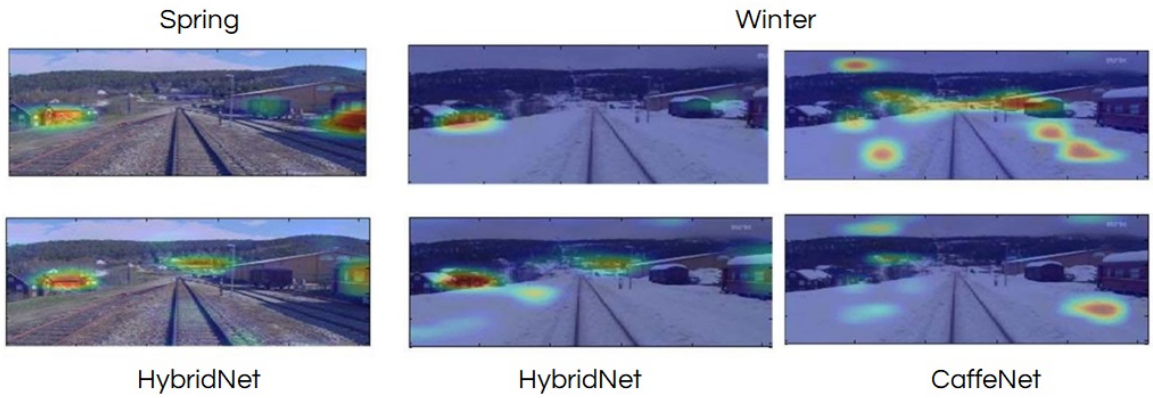

Fig. 2.5 Images from the spring season are matched against the winter season. Feature heat maps in the first and second columns are taken from HybridNet [4]. The third column exhibits the heat maps of CaffeNet.

was employed for picking responses from the convolutional layers. SPP picks activations from convolutional layers at multiple scales e.g. [1 2 3 4] by dividing each feature map into scale based cells and apply max-pooling on each cell to pool the responses. By evaluation on publicly available benchmark datasets, fine-tuned HybridNet claimed performance boost over AMOSNet, CaffeNet and AlexNet. Furthermore, deep investigation on HybridNet showed that middle convolutional layers focused on corners, edges and colours where higher layers captured semantically meaningful regions under strong conditional changes, such as, building structures which means the network has learned appearance invariant features as shown in Fig. 2.5. It is evident that HybridNet fires at semantically meaningful regions whereas CaffeNet's responses are less meaningful due to object-centric CNN training.

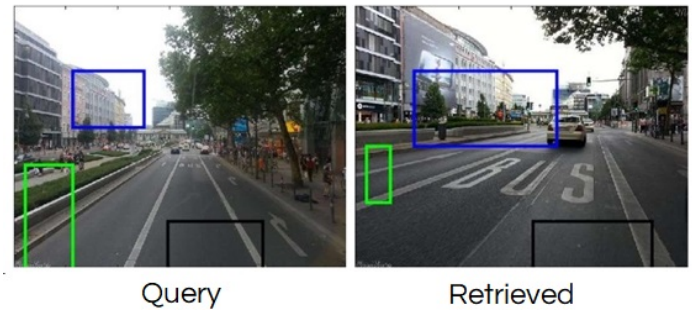

Fig. 2.6 A place recognition system utilising deep VGG-16 pre-trained on ImageNet under strong viewpoint and moderate conditional changes (Image taken from [5]).

Motivated from [92][90] which employed a cross-convolutional technique for image recognition and classification, Chen et al. in [5] proposed a region-based VPR system, illustrated in Fig. 2.6. Particularly, salient CNN-based regional representations were identified (shown with different colours) for recognizing places under simultaneous viewpoints

and conditional changes. Employing deep object-centric VGG-16 and a cross-convolution based regional approach was used such that group of connected activations from the late convolutional layer are filtered on the basis of their energies followed up by their mapping onto the previous convolutional layer for CNN-based regional extraction. Based on the regional features, places are recognized using BoW encoding approach.

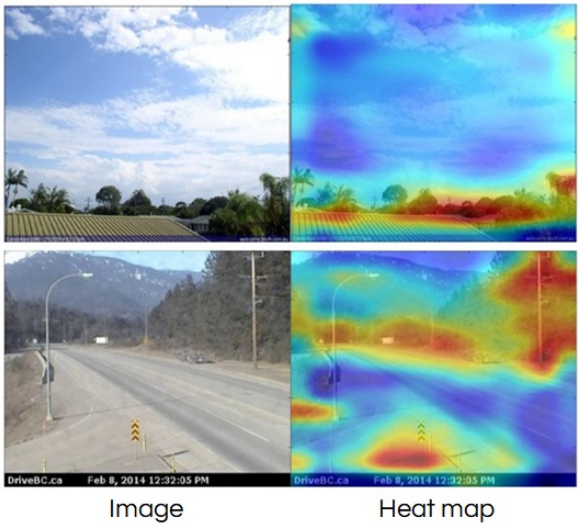

Fig. 2.7 A deep neural network based place recognition framework focusing on context-flexible attentions. Two exemplars with their heat maps are shown here (image taken from [6]).

Recently, the authors in [93][94][6] have demonstrated that fused features from multiple convolutional layers can improve place recognition under visual changes. Jin Kim et al. in [94] proposed a fixed context-aware attention model that captures the manually defined rectangular shaped most contributing distinctive patterns efficient for visual localization. Work in [4][93] demonstrated that different convolutional layers capture different semantic information. Chen et al. in [6] fine-tuned the deep object-centric VGG-16 on SPED dataset and employed fused multi-scale features for place recognition. A context flexible block is integrated within the late convolutional layers which automatically learns context flexible attentions upon fine-tuning on SPED dataset. Evaluations based on fused multi-layer attentions were found to be efficient under severe conditional changes coupled with moderate viewpoint changes. However, the efficiency of the proposed approach may be compromised if there is a simultaneous severe viewpoint and conditional variation. Moreover, performance and efficient resource usage have become two important aspects to be looked into for real-world VPR applications.

Image retrieval tasks which either rely on handcrafted features, such as, local SIFT and SURF features [32][31] or combining these with convolutional and fully connected layers of

deep/shallow CNNs [95][96][84], Bag-of-Words (BoW) or Support Vector Machine (SVM) [97] are employed for classification, detection and recognition [10][90] purposes. As an alternative for BoW feature encoding scheme, several other approaches including Fisher vector [98] and Vector of Locally aggregated descriptor (VLAD) have shown promising results with smaller visual words vocabularies [99]. To perform instance level image retrieval where objects from the same category are to be separated, [96] suggested to combine the rich spatial middle convolutional layers' features with VLAD encoding. Jin Kim et al. in [100] employed MSER [101] for regions identification, followed by the detection of SIFT features within the identified regions and described each region/bundle as a fixed size VLAD, named as PBVLAD. 2D-based localization methods generally offer efficient database management at lower accuracy cost whereas 3D-based techniques are computationally complex but more reliable in localization. Sattler et al. in [102] refuted this notation by combining 2D-based approaches with SfM-based post-processing and have shown better performance than structure-based methods. Merrill et al. [103] trained a convolutional auto-encoder in an unsupervised manner. The objective of auto-encoder based VPR was to re-create the HOG descriptor of original image using a distorted version of the original image as input.

Recently, Teichmann et al. in [104] trained the landmark detectors [105][10] with a newly introduced $1.2M$ Google Landmark Dataset (GLD). It contains $15k$ landmark categories including buildings, monuments and bridges annotated by human. They have proposed a technique which retrieves the normalized regional residuals, termed as R-VLAD. Thus, it down-weights all the regional residuals and stores a single aggregated regional descriptor per image. Custom landmark detectors including ASMK [106], RMACB [105], RMAC [10] and selective search [107] are incorporated for the regional search and coupled with the proposed R-VLAD on deep CNNs.

2.2.6 Image Matching

Matching approaches for visual place recognition can be classified into two categories; 1) using one single frame and 2) using two or more captured frames in the sequence for recognition purpose.

Single Image Matching

In this category, visual place recognition is performed with a single image of a query place. An image database is pre-collected and stored for runtime matching with the query location. By employing a VPR, the most similar database image is claimed as the place currently encountered by the robot or vehicle. However, it is important for two places to be visually

Fig. 2.8 VPR using single image matching.

similar to the extent that they can be claimed as representing the same physical location. Therefore, in single image matching approach (shown in Fig. 2.8), the query place is captured once in a single frame, compared with all the stored places and a visually similar image is retrieved as the localized place.

Various VPR algorithms have employed a single image of the current scene and matched with the stored image collection [2]. FABMAP was the first to employ single frame matching approach; detected SURF keypoints on the visually captured appearance of the place and combined it with BoW model for single query matching with the reference frames. During training, the employment of the BoW encoding with SURF descriptors allowed to determine the distinctiveness of each feature/word. Chow Liu tree was employed for measuring the probabilities of the visual words carried out by determining the maximum-weight spanning tree of a directed graph of co-occurrence between the visual words of training dataset. Multiple places exhibiting similar environment were efficiently tackled by FAB-MAP using common visual words approach such that the less common words highlight difference in places and vice-versa.

Knopp et al. in [108] performed single image-based place recognition using BoW encoding mechanism. They demonstrated that single image matching approach is easier and simpler to deploy. However, under large scale extreme environments, place recognition with single image matching is sensitive towards lightning changes and dynamic objects including vehicles and pedestrians. Similarly, neural network based VPR approaches also employed single matching technique for retrieval purposes [4][5][6]. These techniques pass down a single frame of a location and extract middle and late convolutional layers as features representation. This is then followed-up by post feature matching with reference image descriptors such that the maximum matching score is claimed as the currently visible location.

Sequence-based Image Matching

Earlier VPR techniques used an assumption that the appearance of places would not experience any changes in indoor environments. However, robot localization in extreme and uncertain outdoor environments refute to this assumption and under such visual changes, appearance aware place recognition techniques sometimes underperform with single image approach [2]. Therefore, in the presence of lightning and weather changes, instead of employing single place similarity, a sequence of images is preferred to be matched for VPR. A sequence can be a combination of two or more frames, therefore, by utilising sequences even from different times of the month or year can still lead to recognize successful VPR, as illustrated in Fig. 2.9. Each test sequence needs to match with the specific reference sequence captured in a different time interval.

Sequence-based Simultaneous Localization and Mapping (SeqSLAM) was the first sequence based place recognizing technique introduced in [42]. Particularly, the matrix of image similarities between the current image sequence and reference image sequence was computed, followed up by the sum of absolute difference of their enhanced contrast for image similarity. The maximum sum of normalized similarity was treated as the recognizing score measured through a matrix over predefined constant velocity sequence paths.

2.3 State-of-the-Art Visual Place Recognition techniques

Wide range of VPR techniques are proposed by the research community with some focused-on conditional- or appearance-variance while others emphasis upon the viewpoint changes. Under common-ground, Zaffar et al. in [109] assess and tested multiple VPR approaches on three challenging place-recognition datasets exhibiting seasonal, conditional and viewpoint vairations. The work provides a comprehensive evaluation of 10 state-of-the-art contemporary

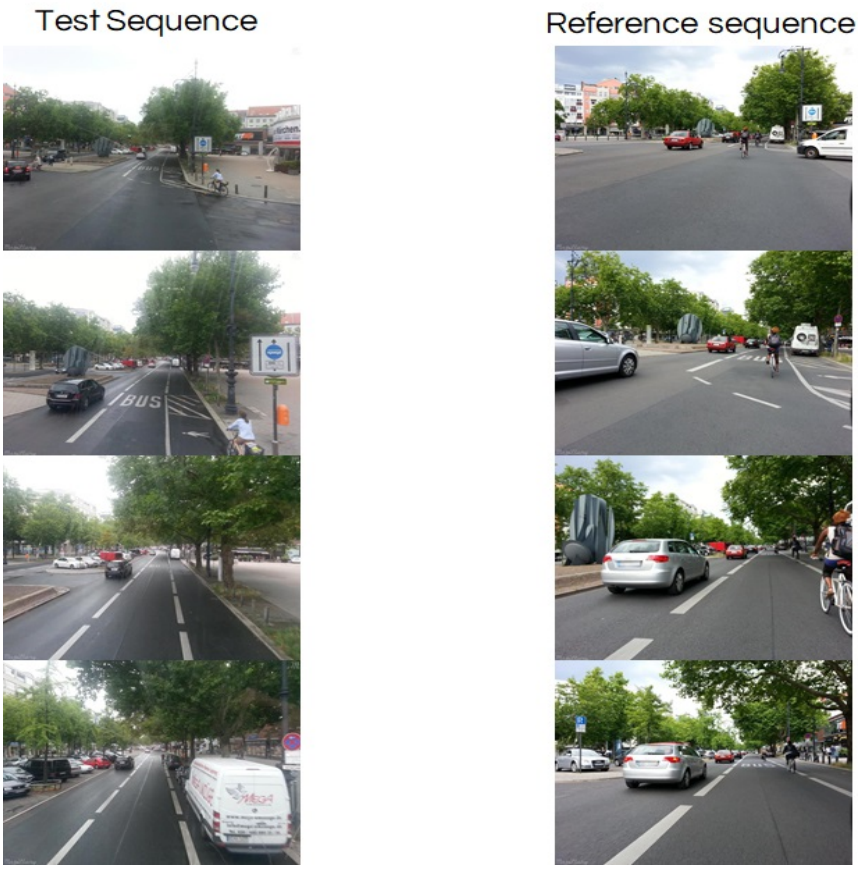

Fig. 2.9 VPR using sequence-based image matching.

VPR techniques in a chronological order, shown in Fig. 2.10. It includes HOG, SeqSLAM, AlexNet, NetVLAD, AMOSNet, HybridNet, Cross-Region-BoW, RMAC, Region-VLAD and CALC. Tthe expected increase in VPR performances have not been observed in a chronological order but investigations based on matching performance, retrieval time and memory requirement claim that deep neural network based VPR techniques are better under severe seasonal, illumination and viewpoint variations but at the cost of memory whereas handcraft-based VPR frameworks have shown inferior results at low resource utilization.

Another work by Zaffar et al. in [7] evaluates the applicability of ground-based VPR techniques for aerial robotics. They employed two recently open-sourced aerial robotics datasets [110] exhibiting 6-DOF viewpoint variation and evaluated 8 state-of-the-art ground-based VPR approaches. It gives a bird-eye view of applicability of the VPR systems that work quite well under lateral viewpoint variation shown in Fig. 2.11. Therefore, performance analysis of these techniques under 6-DOF is carried out while considering run-time processing power and memory consumption in real-time aerial robotic application. The results showed that NetVLAD again outperformed other regions-based techniques under

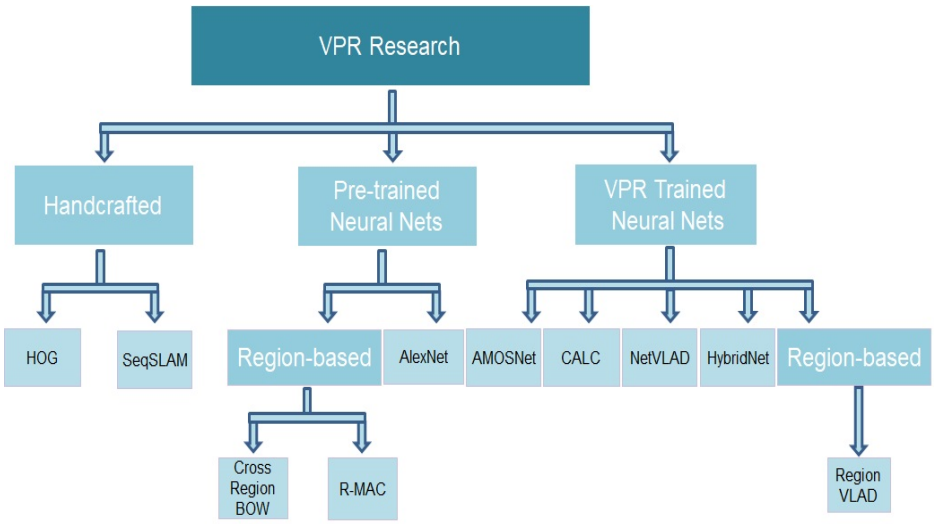

Fig. 2.10 Hand-crafted and neural network-based contemporary VPR techniques.

6-DOF viewpoint variation. However, most neural network based VPR techniques are not scalable for resource-constrained platforms like aerial robots. Cross-Region-BoW leads to highest power consumption due to its intense matching scheme, as illustrated in Table 2.1.

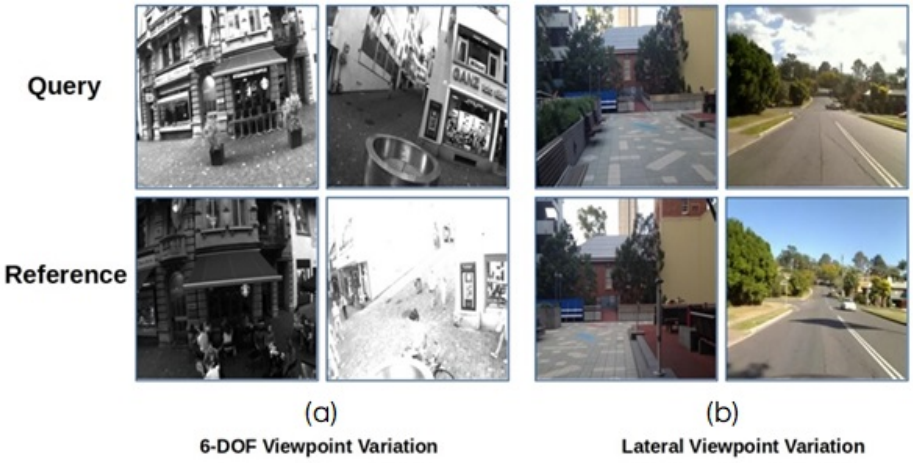

Fig. 2.11 Examples of 6-DOF (degree of freedom) and lateral viewpoint variations. (image taken from [7])

2.4 Benchmark Visual Place Recognition Datasets

In the context of place recognition, datasets proposed by the research community vary in terms of seasonal, viewpoint and illumination variations. Some datasets exhibit stronger viewpoint changes coupled with moderate conditional variations, others are captured under

Table 2.1 Computational power requirements (taken from [7])

Techniques	**CPU Utilization**		**Time (sec)**		**Power Consumption (Ah)**
	Encoding	**Matching**	**Encoding**	**Matching**	
	Intel(R) Xeon(R) Gold 6134 CPU @ 3.20GHz with 32 cores, 64GB RAM				
AlexNet	0.734	0.0312	0.666	3.222	0.3128
NetVLAD	0.656	0.036	0.77	0.0374	0.2688
AMOSNet	0.437	0.03	0.359	0.614	0.0931
HybridNet	0.437	0.03	0.357	0.584	0.0921
Cross-Region -BoW	0.32	0.1	0.834	1199.04	63.836
RMAC	0.5	0.371	0.478	0.254	0.1768
Region-VLAD	0.25	0.031	0.463	0.899	0.0764
CALC	0.781	0.0312	0.027	0.974	0.0272

stronger conditional and adequate viewpoint variations. For all the datasets, two traverses along the same route are taken at multiple times of the day/year. More specifically, the datasets employed in this thesis include *Berlin Halenseestrasse* [40], *Berlin A100* [40], *Berlin Kudamm* [40], *Gardens Point* [84], *Synthesized Nordland* [23], *Query247* [8], *St. Lucia* [4] and *SPEDTest* [6]. All the Berlin datasets are captured in urban environment and have been introduced and employed for evaluating VPR approaches [40][4]. Crowd-sourced geotagged photo-mapping platform *Mapillary* [111] is used for gathering the Berlin datasets. Traverses of the same route are uploaded by different users exhibiting viewpoint and conditional variations among the same places. Berlin Halenseestrasse and Berlin Kudamm datasets exhibit strong viewpoint variations. Other datasets including Berlin A100, Gardens Point, Synthesized Nordland, Query247 and SPEDTest exhibit strong illumination and seasonal changes. Gardens Point dataset was captured at QUT, Brisbane campus with one traverse taken during daytime on left side walk and the other traverse was recorded in right side walk at night time [4]. The Synthesized Nordland dataset was recorded on a train with one traverse taken in winter and reference traverse was recorded in spring. Viewpoint variance was added to Synthesized Nordland by cropping frames of summer traverse to keep 75% resemblance [6]. Under urban environment, Query247 [8] captured images of 365 places in day, evening and night times. The St.Lucia dataset was captured in the sub-urban route at multiple day times with sufficient viewpoint- and condition-variation. The SPEDTest [6] is the newly introduced dataset which contains very diverse scenarios captured with surveillance cameras during different times of the year. Sample images of both the traverses of the datasets are shown in Fig. 2.12. Severe conditional and viewpoint variations can be seen across the same places.

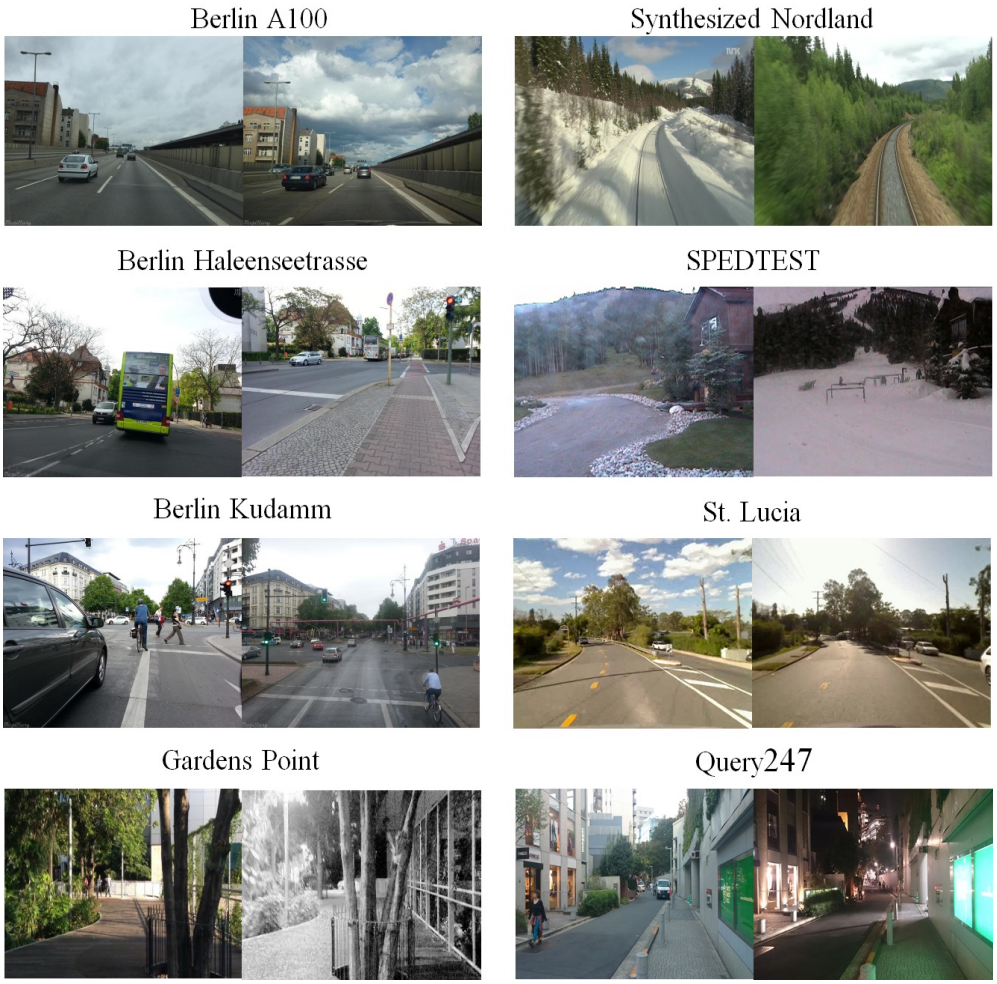

Fig. 2.12 Strong viewpoint and conditional variations can be observed across the same places. Left and right column frames of each dataset are taken from the test and reference traverses.

In Chapter 3, challenging benchmark VPR datasets *Berlin A100*, *Berlin Halenseestrasse* and *Berlin Kudamm*, *Gardens Point* and *Synthesized Nordland* are employed to evaluate the proposed VPR framework. One traverse is used as R reference database and the other traverse is employed as T test database (see TABLE 2.2) where R' represents the reduced reference traverse which contains T' matched test images (will discuss in section 3.4.3). For *Berlin A100*, *Berlin Halenseestrasse* and *Berlin Kudamm* datasets, geotagged information is used for ground truth with 0 to ± 2 frame tolerance. For *Gardens Point* and *Synthesized Nordland* datasets, the ground truth data is obtained by parsing the frames and maintaining place level resemblance with 0 to ± 3 and 0 to ± 2 frame tolerance.

For evaluating the proposed VPR system in Chapter 4, three severe condition variant place recognition datasets are targeted (please see Table 2.3). The first traverse is used for

Table 2.2 Benchmark visual place recognition datasets employed in Chapter 3

Dataset	Environment	Variation		T	R	T'	R'
		Viewpoint	Condition				
Berlin A100	urban	moderate	moderate	81	85	70	64
Berlin Haleenseetrasse	urban, suburban	very strong	moderate	67	157	50	138
Berlin Kudamm	urban	very strong	moderate	222	201	166	151
Gardens Point	campus	strong	strong	200	200	152	150
Synthesized Nordland	train	moderate	very strong	1622	1622	1221	1217

testing and the second traverse is served as reference frames. The original GPS annotation with the *St.Lucia* dataset employed to build place and frame level correspondence, used as Ground truth. Each test image in *SPEDTEST* resembles with three known reference images provided with the dataset. For *Synthesized Nordland* dataset, frame and place level resemblance is used as ground truth.

Table 2.3 Benchmark visual place recognition datasets employed in Chapter 4

Dataset	Traverse		Environment	Variation	
	Test	Reference		Viewpoint	Condition
St. Lucia	1249	1249	Suburban	Adequate	Moderate
SPEDTEST	607	1821	Diverse	Moderate	Strong
Synthesized Nordland	1622	1622	Train journey	Moderate	Very Strong

2.5 Evaluation Criteria

In the majority of VPR techniques available in literature [84][86][40][5][6], the authors have employed AUC under Precision Recall curves (AUC-PR curves) [112] as an evaluation metric. They assumed that all the queries have matched reference frames and the system must return a matched database image.

$$Precision = TP/(TP+FP) \tag{2.1}$$

$$Recall = Sensitivity = TP/(TP+FN) \tag{2.2}$$

TP (true-positive) and FP (false-positive) refer to correctly and incorrectly retrieved images where FN (false negative) represents the images which should have been retrieved

but the system missed those images. There can be such scenarios where few queries might be new places that have not been previously observed. By nature, PR curves do not cater such scenarios i.e., True Negative cases (TN, correctly missed the non existing events/classes) and are only concerned with the correct prediction using the scores/probabilities. On the other hand, Receiver Operating Characteristic (ROC) curves [113] are employed when each prediction class has equal number of observations with True Positive and False Positive rates, such that:

$$Specificity = TN/(TN+FP) \tag{2.3}$$

$$False\ positive\ rate = FP/(FP+TN) = 1 - Specificity \tag{2.4}$$

$$True\ positive\ rate = Recall = TP/(TP+FN) \tag{2.5}$$

ROC curve considers True Negative scenarios in the false positive rate which do care of the predicted positive class when the actual outcome is negative. Work in [114][115] suggested that while performing image retrieval tasks with imbalanced datasets, employing ROC might be deceptive in interpreting specificity which is the inverted false positive rate. If the proportion of positive to negative instances changes in the test cases, ROC will not change since true and false positive rates do not depend upon the class distribution. Increasing 2x the number of positive sample in the test set would increase TP and FN by 2x, hence, no change in true positive rate at any threshold and same goes for false positive rate. The employed datasets in this thesis are imbalanced i.e. against every query image, there exists at-least one matched reference image along with more mismatched scenarios. Therefore, we have employed AUC-PR curves for evaluating the proposed VPR systems and always match each query with any of the reference frames irrespective of its TP or FP nature.

2.6 Summary

This chapter presented a detailed review of visual place recognition research work. Various approaches to address the issue of poor visibility coupled with the conditional, viewpoint and seasonal variations are studied. However, there is still significant room of improvement for visual place recognition under simultaneous viewpoint and conditional changes. The contributions of this dissertation are encouraged by short-comings of the reviewed literature. Particularly, the novel work presented in this thesis put more emphasis upon regional CNNs-

based place recognition system yielding performance boost in uncertain visual and viewpoint changes.

Chapter 3

Shallow CNN-based Regional-approach for Visual Place Recognition

In this chapter, a regions-based CNN approach for visual place recognition is proposed. The proposed regional approach is employed on shallow CNN models combined with VLAD feature encoding scheme for single image matching. The CNN-based regional features are found to be robust against strong viewpoint and conditional changes including dynamic objects. In comparison to other deep neural network based visual place recognition techniques (such as, Cross-Region-BoW, RMAC), the presented approach is computationally efficient and scalable for visual place recognition. Extensive experiments on several benchmark place recognition datasets demonstrate better results in terms of precision-recall curves against 10 state-of-the-art contemporary visual place recognition and image retrieval tasks.

3.1 Introduction

Using a pre-trained CNN for VPR, there are three standard approaches to produce a compact image representation: (a) the entire image is directly fed into the CNN and responses from convolutional layers are extracted [84]; (b) CNN is applied on user-defined regions of the image and prominent activations are pooled from the layers representing those regions [40]; (c) the entire image is fed into the CNN and salient regions are identified by directly extracting distinguishing patterns based on convolutional layers responses [5][6]. Generally, global image representations retrieved from category (a) are not robust against strong viewpoint variations and partial occlusion. Image representations emerging from category (b) usually handle viewpoint changes better but are computationally intensive. Image representations

resulting from category (c) address both the appearance and viewpoint variations. In this paper, we focus on category (c).

As presented in Chapter 2, the work by [5] and [6] are considered state-of-the-art in identifying prominent regions by directly extracting unique patterns based on convolutional layers' responses. Despite its good AUC-PR performance, the method proposed in [5] has some shortcomings. A common strategy for improving CNN accuracy is to make it deeper by adding more layers (provided sufficient data and strong regularization). However, increasing network's size results in increased computation and using more memory both at time of training and testing (such as, for storing outputs of intermediate layers and for storing parameters) is not ideal for resource-constrained robots that are usually battery-operated. Using $10k$ BoW dictionary for regions-based feature encoding (extracted from late convolutional layers of deep VGG-16) followed up with their cross-matching degrades the matching performance. Secondly, employment of object-centric deep VGG-16 model results in a system attempting to put more emphasis on objects rather than the place itself. This reflects on the regions-based pooled feature and leads to failure cases. Also, the regional approach proposed in [5] hinders the identification of individual static place-centric regions that can be more effective under condition and viewpoint variations.

To bridge these research gaps, a holistic approach is proposed that targets a CNN architecture comprising a small number of layers pre-trained on a scene- [33] and place recognition-centric [4] image databases to reduce the memory consumption and computational cost. The proposed method detects novel CNN-based regional features and combines them with VLAD [99] adapted specifically for VPR based localization problem. The motivation behind employing VLAD comes from its better performance in various CNN-based image retrieval tasks utilizing a smaller visual word dictionary [99][116] compared to BoW [54]. To the best of our knowledge, this is the first work that combines novel lightweight CNN-based regional features with VLAD encoding adapted for computation-efficient VPR.

As opposed to [5] which uses object-centric VGG-16 architecture and employs a cross-convolution based regional extraction approach (resembles [90]), the proposed VPR technique here, is different both in identification and extraction of regional features (discussed in detail, in section 3.2.2). The approach presented in this paper showcases enhanced accuracy by employing middle convolutional layer of the 8-layered CNN architecture. Evaluation on several viewpoint- and condition-variant benchmark place recognition datasets shows an average performance boost of 13% over state-of-the-art VPR algorithms in-terms of AUC computed under Precision-Recall curves. In Fig. 3.1, for a query image (a), the proposed system retrieved image (c) from the stored database. (b) and (d) highlight the salient regions which the proposed framework identified under severe viewpoint- and condition-variation.

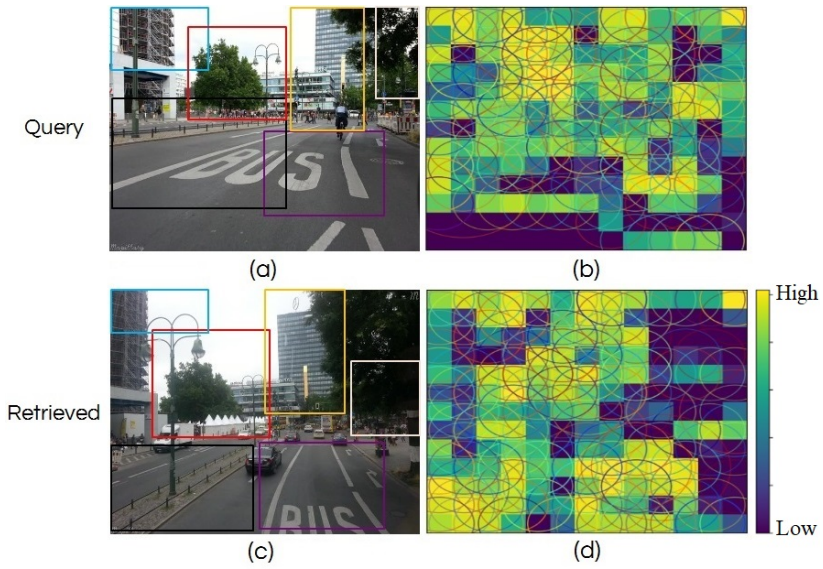

Fig. 3.1 For a query image (a), the proposed Region-VLAD approach successfully retrieves the correct image (c) from a stored image database under significant viewpoint- and condition-variation. (b) and (d) represent their CNN based meaningful regions identified by our proposed methodology.

The remainder of the chapter is organized as follows. In Section 3.2, the proposed methodology is presented in detail. Section 3.3 and 3.4 illustrate the implementation details and results achieved on several benchmark datasets. Section 3.5 presents the conclusion.

3.2 Proposed Region-VLAD VPR framework

In this section, the key steps of the proposed methodology are described in detail. It starts by stacking activations of feature maps for retrieving local descriptors, followed up with the identification of distinguishing regional patterns. It then illustrates the aggregation of local feature descriptors lying under those identified salient regions. Finally, it shows how to retrieve the compact VLAD representation using the extracted CNN-based regional features, later used for determining a match between two images. The workflow of the proposed methodology is shown in Fig. 3.2. The query VLAD run-time matching with the pre-stored reference VLADs reduces the retrieval time.

3.2.1 Stacking of Convolutional Activations for making Descriptors

Given an image I as an input to the CNN model, at a certain convolutional layer, the output is a $3D$ tensor M of $X \times Y \times K$ dimensions. K denotes the number of feature maps, X and

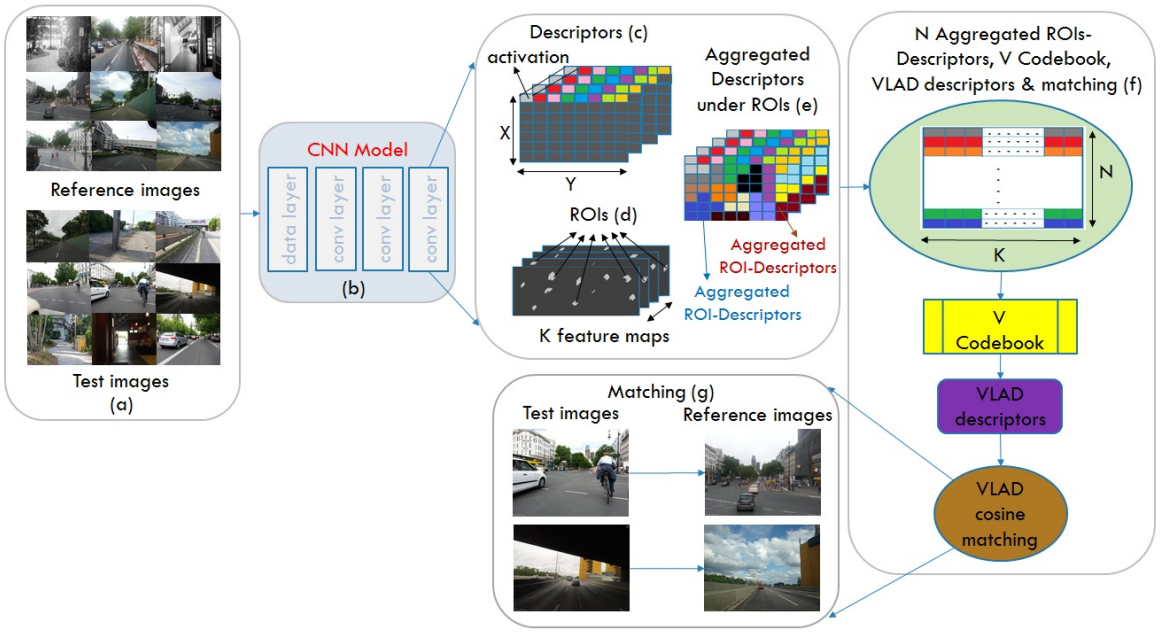

Fig. 3.2 Workflow of the proposed VPR framework is shown here. Test/reference images are fed into the CNN model, Region-of-Interests (ROIs) are identified across all the feature maps of the convolution layer and their compact VLAD representation is stored for image matching.

Y represent the width and height of feature map / channel. We can also interpret it as M^k being a set of $X \times Y$ activations / responses for k^{th} feature map where $k = \{1, 2, ..., K\}$. For K feature maps in the convolution layer, we stack each activation at some certain spatial location into K dimensional local feature as shown with different colors in Fig. 3.2 (c). D^L in (3.1) represents the K dimensional d_l feature descriptor(s) at L^{th} convolutional layer of m_c model.

$$D^L = \{d_l \in M^K \ \ \forall\, l \in \{(i, j) \mid i = 1, ..., X; j = 1, ..., Y\}\} \tag{3.1}$$

3.2.2 Identification of Regions of Interest

To extract region-based CNN features, the most prominent regions need to be identified. Two or more activations are considered to be connected and represented as a region if they are neighbours and have approximately the same value. For K feature maps, each region is denoted by G_h, $\forall\, h \in \{1, ..., H\}$ where H is the total identified regions at L^{th} convolution layer, visualized in Fig. 3.2(d)/Fig. 3.4.

The mean energy of each G_h region is calculated by averaging all a_h activations lying under the region. In (3.2), a_h^f represents the f^{th} activation lying under the G_h region and E^L

denotes the calculated mean energies of H regions. Based on the sorted E^L energies, top N energetic ROIs (with their bounding boxes) are picked in (4.3), denoted as R^L novel regions at L^{th} convolution layer.

$$E^L = \{\frac{1}{|G_h|}\sum_f a_h^f,\ \forall\ a_h^f \in G_h\} \tag{3.2}$$

$$R^L = \{G_t\ \forall\ t \in \{1,...,N\}\} \tag{3.3}$$

Fig. 3.3 illustrates the top $N = \{50, 200, 400\}$ novel R^L regions identified by our proposed regions-based VPR system. Our novel CNN based identified regions strongly concentrate on the static objects including buildings, trees and road signals. D^L local descriptors in (3.1) which fall under the bounding boxes of R^L regions in (3.3), aggregated in (3.4) to retrieve CNN-based regional features. Intuitively, each regional feature is $1 \times K$ dimensional f_t vector where q be the R_t^L region under which D_q^L descriptors fall. For N novel regions, (3.5) represents $N \times K$ dimensional F^L region-based CNN features representing an image at L^{th} convolutional layer (intuitively shown in Fig. 3.2 (e) and Fig. 3.2 (f)).

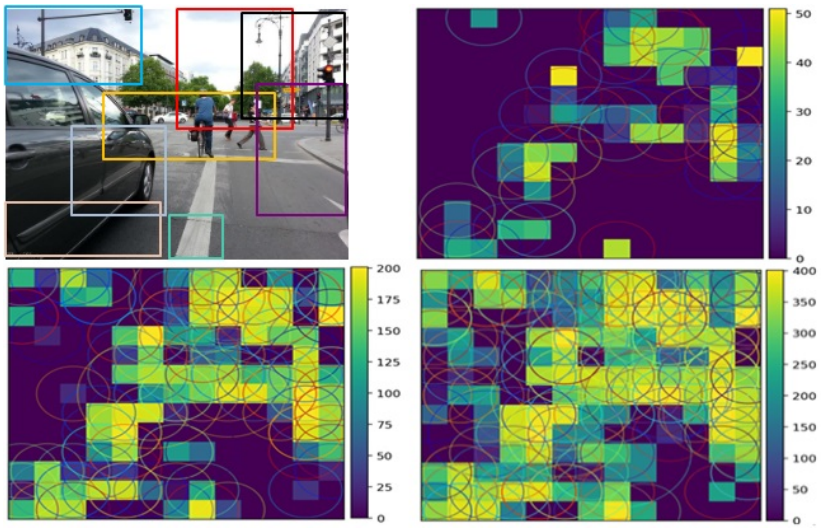

Fig. 3.3 Sample images of top 50, 200 and 400 Regions-Of-Interest (ROIs) identified by the proposed approach.

$$f_t = \sum_{q \in R_t^L} D_q^L \tag{3.4}$$

$$F^L = \{f_t\ \forall\ t \in \{1,...,N\}\} \tag{3.5}$$

In comparison, authors in [5] first identified regions, calculated their mean energies and selected $N = 200$ energetic regions. Precisely, N regional activations at L^{th} convolution layer were mapped onto the $L-1^{th}$ convolutional feature maps and aggregation of modified cross-mapped regions-based local descriptors at $L-1^{th}$ convolution layer was carried out for feature extraction. Note that depending upon the quantity of activations per ROI(s) at L^{th} convolution layer and receptive field of the filter (e.g. 3×3, 5×5) for cross-mapping of L^{th} convolution layer regions at $L-1^{th}$ layer, the bounding box (area) per cross-mapped regional feature varies for [5].

Furthermore, Fig. 3.4 illustrates that the identified ROIs from two feature maps (M^1 and M^2) at L^{th} convolutional layer with Region-VLAD and Cross-Region-BoW [5] are different in quantity and size/activations per region(s). Thus, the computed regional mean energies of [5] are different from the mean energies of regions identified by our approach. Our approach identifies 36 and 40 ROIs from feature map M^1 and M^2, shown with different colors. Later, based on their computed mean energies, top N energetic regions are selected from H identified regions at L^{th} convolutional layer, shown in Fig. 3.3. The 8-connected component-based regional approach in Cross-Region-BoW [5] identifies 6 and 4 yellow colored ROIs for feature map M^1 and M^2. As explained above, N energetic regional feature extraction for [5] is carried out by first selecting N energetic regions at L^{th} layer (Fig. 3.4) followed up with their mapping at $L-1^{th}$ convolution layer and aggregation of cross-mapped regions-based local descriptors at $L-1^{th}$ convolution layer (not shown in the figure). Exemplars exhibiting the identified regions by Cross-Region-BoW [5] and with our proposed Region-VLAD framework are shown in Fig. 3.5. We observe that regional patterns covering more areas similar to [5] hinder the identification of individual place-centric instances vital in recognizing places under changing conditions and viewpoints.

3.2.3 Regional Vocabulary and Extraction of VLAD for Image Matching

VLAD adopts K-means [54] based vector quantization, accumulates the residues of features quantized to each dictionary cluster and concatenates those accumulated vectors into a single feature representation. A separate dataset of $2.6k$ images is collected and afore-described regions-based feature extraction is employed for generating a regional vocabulary. To learn a diverse vocabulary, we employed 1125 place-recognition centric images of 365 places from Query247 [8] (taken at day, evening and night times). Other images include a benchmark place recognition dataset St.lucia [4] with $1k$ frames of two traverses captured in suburban environment at multiple times of the day. The left over images consist of multiple viewpoint-

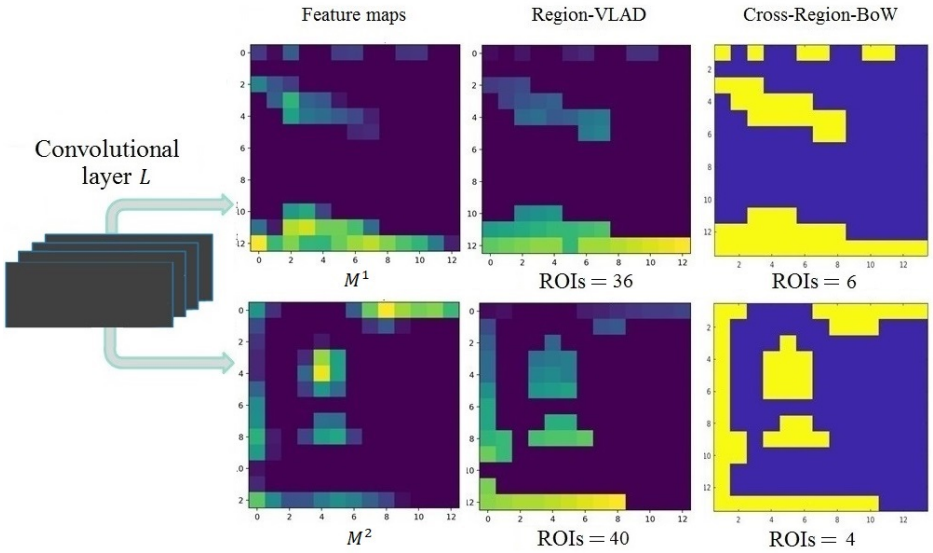

Fig. 3.4 Employing two features maps M^1 and M^2, sample images of ROIs identified by Region-VLAD and Cross-Region-BoW [5] are shown here. Note that feature maps (1^{st} column) illustrate the intensities of a activations. However, regardless of the intensity, each identified G_h region per feature map for Region-VLAD (2^{nd} column) is indicated with a different color i.e. 36 and 40 colored regions for feature map M^1 and M^2. For Cross-Region-BoW (3^{rd} column), all the regions are denoted as yellow patterns i.e. 6 and 4 ROIs for M^1 and M^2 feature maps.

and condition-variant traverses of urban and suburban routes collected from *Mapillary* (previously employed by [40] and [5] for capturing place recognition datasets). K-means is employed for clustering the $2600 \times N \times K$ dimensional regional features into V regions such that o_u in (3.6) represents the u^{th} region of C^L codebook.

$$C^L = \{o_u \, \forall \, u \in \{1, ..., V\}\}, \; V \in \{64, 128, 256\} \tag{3.6}$$

Using the learned codebook, F^L regions of benchmark test / reference traverses are quantized in (3.7) to predict the clusters or labels Z^L, where α is the quantization function. Employing regions-based F^L feature, predicted labels Z^L and regional codebook C^L, summed residue v corresponding to each u^{th} region can be retrieved using (3.8).

$$Z^L = \alpha(F^L) \tag{3.7}$$

In (3.8), for all the F^L regional features that fall in u^{th} region of the C^L codebook, the residues of F^L_u regions and C^L_u codebook's region center are summed. Sometimes, few regions/words appear more frequently in an image than the statistical expectation known as visual word burstiness [117]. Standard techniques include power normalization [118] is

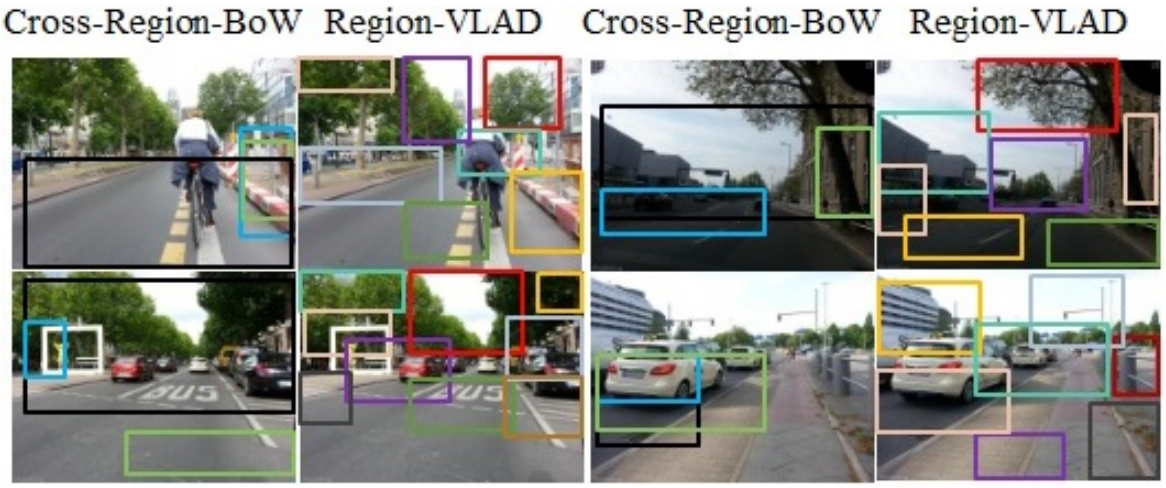

Fig. 3.5 Sample images of ROIs identified with Cross-Region-BoW [5] and Region-VLAD are shown here. Our regional approach subdivides each image into large number of most contributing regional blocks.

performed in (3.9) to avoid it where each $1 \times K$ dimensional residue v_u undergoes non-linear transformation γ. In (3.10), power normalization is followed by l_2 normalization. For each image, l_2 normalized residues corresponding to V regions are stored in (3.11) to get final $V \times K$ dimensional VLAD representation S^L.

$$v_u = \sum_{F_u^L : Z_u^L = C_u^L} F_u^L - C_u^L \tag{3.8}$$

$$v_u := sign(v_u)\|v_u\|^{\gamma} \tag{3.9}$$

$$v_u := \frac{v_u}{\sqrt{v_u^T v_u}} \tag{3.10}$$

$$S^L = \{v_u \; \forall \; u \in \{1, ..., V\}\} \tag{3.11}$$

To match a test image “A" against the reference image “B" in (3.12), the dot/scalar product of their u^{th} regional VLAD components $S_u^{L^A}$ and $S_u^{L^B}$, each with dimension $1 \times K$ reaches to an individual regional matching score $j_u^{A,B}$, visualized in Fig. 3.7 (h).

$$j_u^{A,B} = \frac{(S_u^{L^A}).(S_u^{L^B})}{\|(S_u^{L^A})\|\|(S_u^{L^B})\|} \tag{3.12}$$

All the scalar $j_u^{A,B}$ scores for all V regions are summed up in (3.13) to get final single $J^{A,B}$ matching score. For each test image “A", the cosine matching in (3.12) is performed against

Fig. 3.6 First and second column present Query247 images [8]. Images in the third column are taken from the suburban datasets collected from *Mapillary* where forth column showcases St.lucia traverses [4].

all the reference images and at the end, reference image “X" with the highest similarity score is picked as a matched image using (3.14).

$$J^{A,B} = \sum_{u=1}^{V} j_u^{A,B} \tag{3.13}$$

$$P^{A} = \arg\max_{X} J^{A,X} \tag{3.14}$$

3.3 Setup and Implementation Details

The proposed VPR framework is implemented in Python 3.6.4 and the system average runtime over 5 iterations is recorded with 1125 images. AlexNet pre-trained on Places365 dataset is employed as a CNN model for region-based features extraction with 256×256

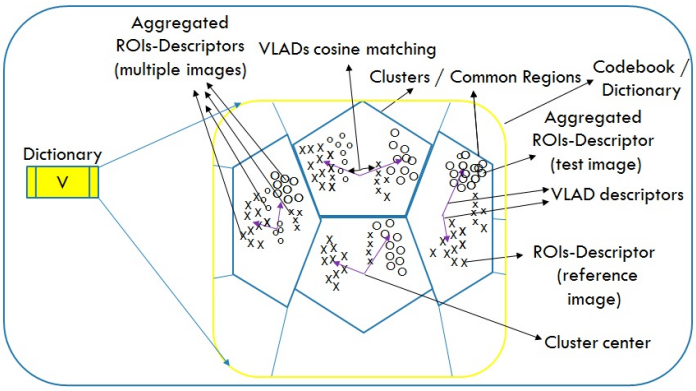

Fig. 3.7 Pictorial view of the regional vocabulary illustrating mapping of the ROIs-Descriptors of test and reference images for VLAD retrieval.

input image size. AlexNet is a light-weight CNN model that contains five convolutional and three fully connected layers. Convolutional layers contain rich spatial information as compared to the fully connected layers. Middle convolutional layers have more generic features (edges and colors) which are found to be efficient under conditional variations and late convolutional layers focus on higher semantic information including shapes, objects and buildings [4] robust against viewpoint changes. For all the baseline experiments, middle *conv3* convolutional layer is used only due to its better performance in various VPR approaches [40][85] but other convolutional layers including *conv2*, *conv4* and *conv5* can also be used for regional features extraction.

For a single image, a forward pass using Caffe takes around an average $0.305ms$ on NVIDIA P100 and $15.57ms$ on Intel Xeon Gold 6134 @3.2GHz. N ROIs-Descriptors are extracted and aggregated with total time comparable with the state of the art methods [5] and other techniques (see Table 3.1). The VLAD representations are retrieved and matched using N ROIs-Descriptors mapped on V clustered dictionary C^L (trained on N ROIs-Descriptors of the $2.6k$ dataset). For direct comparison with [5], $N = 200$ with $V = 128$ are kept. The results are also reported for $N = 400$ with $V = 256$. Table 3.1 shows that for both the regional settings, the average VLAD matching times are $100x$ and $58x$ faster than [5]. The feature encoding times (including forward pass) for RMAC and SPP are comparable with our proposed framework but for NetVLAD, it is approximately twice. It is because passing an image into the VGG-16 model, and then retrieving the VLAD representation from the last layer increase the computation time.

In VPR-based robotic applications which include robotic agricultural devices, autonomous infrastructure, environmental monitoring equipment or other agriculture based use-cases, with exploration of new places, the size of the database can grow unbounded and scalabil-

ity becomes an important factor to be considered [119]. Under both the regional settings, employing NVIDIA P100 for forward pass and Intel Xeon Gold 6134 @3.2GHz for both feature extraction and VLAD encoding, the overall times for retrieving a single query VLAD are 396ms and 447ms. Whereas, Titan X Pascal in [5] takes 408ms for feature encoding per query. In Fig. 3.8 further confirms that the proposed system consumes an average 0.07ms ($N = 200$) and 0.12ms ($N = 400$) for matching VLAD representations of a single query and reference image. Therefore, the total retrieval times per query against $R = 750$ reference images are approximately around 446.405ms and 533.245ms. In comparison, Cross-Region-BoW [5] takes 7ms for matching features of one test and one reference image. The overall retrieval time against $R = 750$ reference images is 5.658s which is 12x and 11x more than our proposed approaches and practically inappropriate for large scale VPR applications. Our Region-VLAD VPR technique can store the encoded VLAD representations of all the reference frames whereas Cross-Region-BoW needs to perform run-time cross matching of given query regions against all the reference frames' regions, and mutually matched regional features are picked.

Furthermore, Fig. 3.9 evaluates our proposed system's run-time performance when more places are added in test and reference traverses. For each PR-curve, we employed T test and R reference images. Their VLAD representations are retrieved followed up by their cosine matching and in parallel, we record down the system's performance. We can see that as the size of test and reference traverses increases, the AUC-PR curves remain higher where "Time" represents the overall matching period for a single test image against R reference traverse. This shows that the system is capable to handle large number of reference/database images while maintaining performance both in accuracy and retrieval time. It should be noted that [5] used MATLAB which is practically slower than Python but we have employed Intel Xeon Gold 6134 @3.2GHz in comparison to [5] which used Titan X Pascal.

3.4 Results and Analysis

In this section, a comparison of the proposed method with 10 state-of-the-art VPR and image retrieval algorithms has been conducted. The section ends with analysis and results on correctly matched and mismatched scenarios of the proposed Region-VLAD framework along with a discussion on the same.

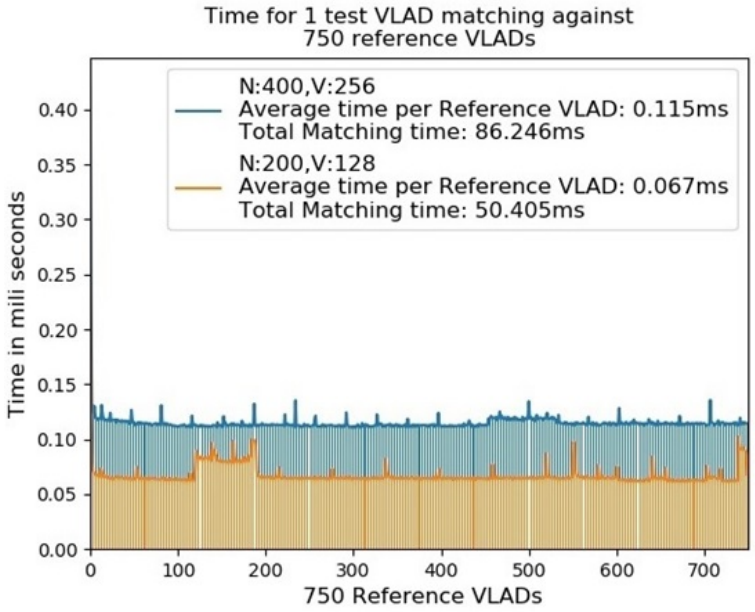

Fig. 3.8 Matching times for 1 test VLAD against 750 reference VLADs are presented.

3.4.1 Comparison Techniques

To show the dominance of our novel place-centric regions finding approach, we replaced VGG-16 with AlexNet365 in [5] (open-source MATLAB code [120]) is integrated, and combined the regional features with the VLAD and BoW encodings, named as Cross-Region-VLAD and Cross-Region-BoW [120]. For a fair comparison, using $2.6k$ dataset, a separate regional vocabulary is trained employing *conv4* for regions identification and *conv3* for features extraction. Keeping $N = 200$, we used $V = 128$ for Cross-Region-VLAD and $V = 2.6k$ for Cross-Region-BoW. Furthermore, results are also reported for HybridNet with Spatial Pyramidal Pooling (SPP) [4] employed on convolutional layers as features representation. We also integrated AlexNet365 and HybridNet with RMAC [10]. Similar to [5], mutual regions are filtered using cross matching, their scores are summed up and maximum matching score is considered for retrieval. Although feature encoding time of NetVLAD [3] (reported in Table 3.1) is approximately twice than our proposed approach but to make a fair comparison, we have evaluated the technique using employed benchmark datasets.

PR-curves across all other image retrieval approaches including Cross-Pool, Max-Pool, Sum-Pool, Whole and state-of-the-art VPR approaches FABMAP and SEQSLAM are taken from [5]. Authors in [5] employed *conv5_2* of deep object centric VGG-16 as features representation. However, Cross-Region-BoW [5] with deep VGG-16 model used *conv5_3*

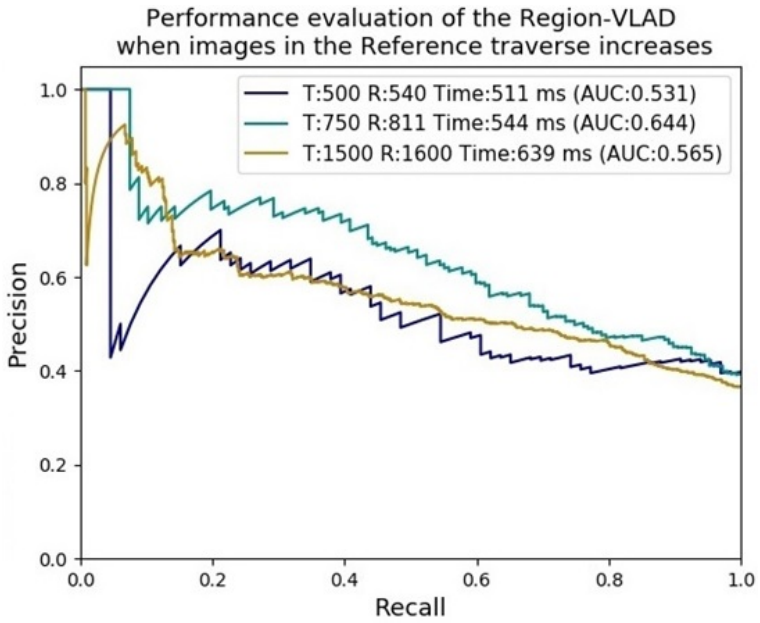

Fig. 3.9 AUC-PR performance and retrieval time of Region-VLAD are reported while adding more images in T test and R reference traverses.

for landmarks identification and *conv5_2* for feature extraction. Standard FABMAP implementation and three sequential frames configuration for SEQSLAM were used by [5].

We have also employed convolutional layers of the shallow HybridNet with separately trained regional vocabularies using 2.6*k* images. Fig. 3.18 illustrates the ROIs identified with our proposed approach utilizing AlexNet365 and HybridNet as CNN models. It is observed that HybridNet combined with our region extraction delivers similar but sometime more place-centric regions as compared to AlexNet365 (which does includes dynamic objects such as clouds due to scene centric training). We employed late convolutional layers of HybridNet for severe viewpoint variant datasets exhibiting mild condition changes and middle convolutional layers for the datasets which are more condition- and moderate viewpoint-variant.

In addition, to closely differentiate our Region-VLAD VPR approach from [5], we have also integrated HybridNet with the cross-convolutional regional approach of [5], and combined the regional features with the VLAD and BoW encoding, named as Cross-Region-VLAD and Cross-Region-BoW. For HybridNet integrated with [5], we encoded the cross-convolutional regional features of 2.6*k* dataset and trained separate layers' regional vocabularies. We employed *conv4* or *conv6* for landmark detection and *conv3* or *conv5* for features extraction. Therefore, for HybridNet, we have a choice to either integrate middle or late convolutional layers with our proposed Region-VLAD or Cross-Region-VLAD [5].

Table 3.1 Runtime performance comparison of our proposed Region-VLAD with Cross-Region-BoW [5], NetVLAD [3], RMAC [10] and SPP [4].

Methdology		Region-VLAD		Cross-Region-BoW	NetVLAD	RMAC	SPP
Model		AlexNet365		VGG-16		AlexNet365/ HybridNet	
Images		1125		1000	-		
GPU/CPU		Intel Xeon Gold 6134 @3.2GHz		Titan X Pascal GPU	Intel Xeon Gold 6134 @3.2GHz		
Forward Pass (ms)		15.57		59	-		
ROIs-Descriptors (N)		200	400	200	-		
Extraction time / Feature encoding (sec)		0.394	0.443	0.349	0.77	0.47	0.36
Regions (V)		128	256	10k Visual words	-		
Matching time (msec)	**VLAD encoding**	2.4	4.54	7	0.005	0.04	0.078
	VLAD matching	0.07	0.12				

3.4.2 Precision Recall Characteristics

In image retrieval tasks where there is a moderate to large class imbalance which means the positive class samples are quite rare as compare to negative classes, Precision Recall curves are usually employed as evaluation metric [112]. Since, the aim is to localize a robot using its previous experience so, for any query place experienced by the robot, with maximum matching score approach (section 3.2.3), it has to either match with the correct place from the previous experience or matches incorrectly with no third option. For all the benchmark datasets, we first calculate the difference in AUC-PR performance of [5] and Region-VLAD, determine their average which comes around an overall 13% performance improvement.

Berlin Halenseestrasse

Area under PR-curves in shown in Fig. 3.10 suggest that the proposed Region-VLAD methodology in the upper/top graph significantly perform better than all other state-of-the-art methods in both the settings. i.e., $N = 200$ and $N = 400$. Both the test and reference traverses of Berlin Halenseestrasse experience strong viewpoint and moderate condition changes among the places. The middle *conv4* and late *conv5* of HybridNet almost showed similar results but we consider middle *conv4* of HybridNet for evaluating other approaches.

To further investigate the reasons of improvement as shown in the middle PR-curves graphs of Fig. 3.10, we replaced the VGG-16 CNN model with AlexNet365 and HybridNet

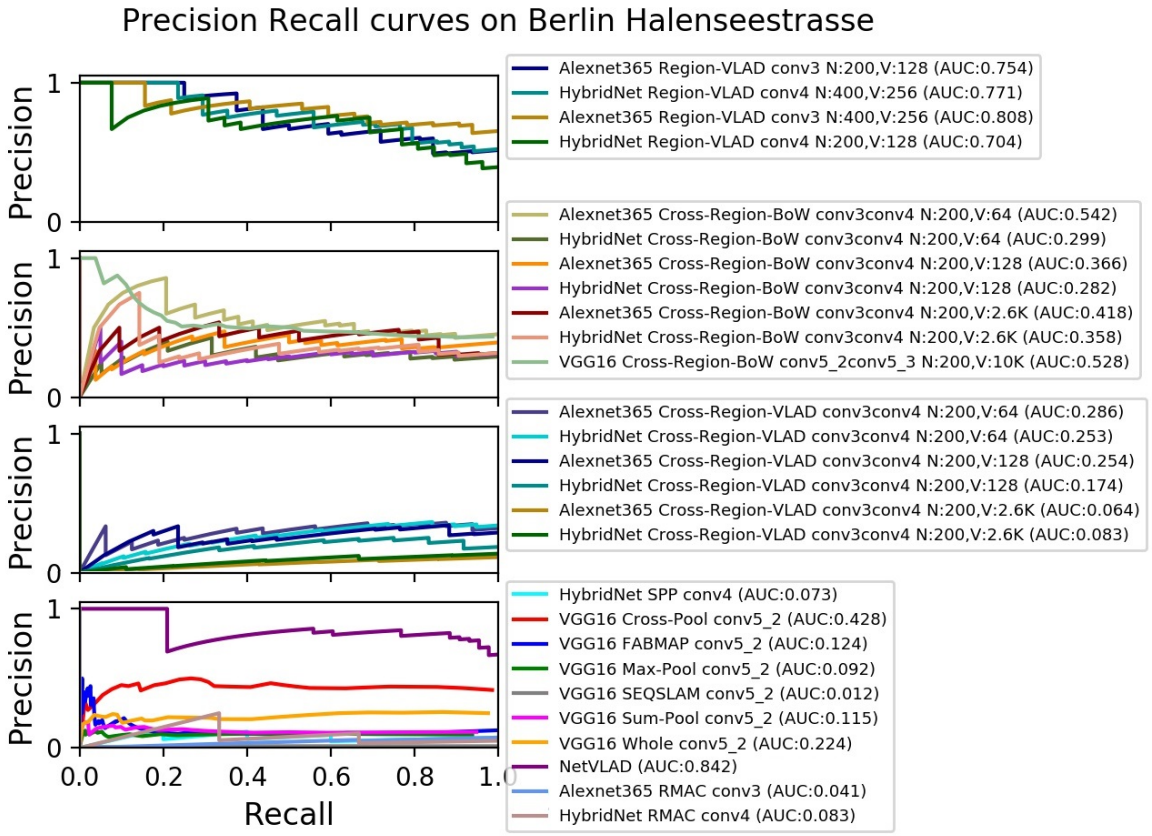

Fig. 3.10 Top: PR-curves of our proposed Region-VLAD approach. Middle: Cross-Region-BoW [5] employed on AlexNet365 and HybridNet with VLAD and BoW encodings. Bottom: Comparison with state-of-the-art VPR approaches

in [5] and reported the PR-curves.Surprisingly for both AlexNet365 and HybridNet, Cross-Region-VLAD PR-curves underperformed with a big margin. Firstly, this mimics the boost up in our Region-VLAD approach is encouraged with the use of our novel regional features. In other words, cross-convolution based regional approach in [5] combined with VLAD encoding has not worked well. We can correspond this behavior towards the difference in regions finding and extracting approach. While employing [5] regional approach, we also observed the total identified regions sometimes even lesser than $N = 250$ so we kept $N = 200$ regions only with V regional vocabulary. Cross-Region-BoW only considers the mutually matched regions where Cross-Region-VLAD needs to calculate residues (per region) of the employed dictionary, which sometimes observed to be containing more zero than non-zero residues in the VLAD representations. The performance degrades especially for HybridNet. With higher regional codebook, Cross-Region-BoW integrated with deep VGG-16 in [5] has shown similar performance as AlexNet365 and HybridNet.

It is worth noting that even with smaller regional dictionaries, our proposed Region-VLAD framework still manages to achieve better results than VGG-16 Cross-Region-BoW [5] and other methodologies. This indicates the potential of our shallow CNN based regional features robustness against strong environment variations. In the bottom PR-curves graph, PR-curves of other approaches on the dataset are presented. FABMAP despite its better viewpoint variations tackling and SEQSLAM with its sequence based whole image approach have not perform well. Cross-Pool employs a similar idea of pooling as [5], so both have achieved a similar PR-curves whereas other pooling techniques under-performed. NetVLAD [3] exhibits similar AUC-PR curve pattern as our Region-VLAD framework. RMAC [10] which is state-of-the-art in other image retrieval techniques underperformed when employed on *conv3* and *conv4* of AlexNet365 and HybridNet. PR-curve for Spatial Pyramidal Pooling (SPP) employed on *conv4* of HybridNet further confirms our proposed VPR framework superiority over state-of-the-art approaches.

Berlin Kudamm

Due to urban environment, too many dynamic and confusing objects such as vehicles, trees and pedestrians with homogeneous scenes lead to perceptual aliasing coupled with severe viewpoint changes making this a challenging dataset. Fig. 3.11 displays three places (a), (c) and (d) of Berlin Kudamm exhibiting a similar scene but captured at different locations in the reference traverse. Fig. 3.11 (b) represents our identified ROIs of (a); our novel regional approach on such places majorly concentrates on road, trees, road pathways. PR-curves with their AUC in Fig. 3.12 are shown and our proposed Region-VLAD approach on AlexNet365 and HybridNet again achieves similar and better results than other approaches. As expected from the environment variations, we observed the late *conv5* of HybridNet employed for our and [5] regional approaches performed better then middle *conv4*.

Replacing the VGG-16 model of [5] with AlexNet365 and HybridNet while employing BoW encodinge exhibit better results with $2.6k$ regional vocabulary. Pooling techniques other than Cross-Pool (which shows similar results as Cross-Region-BoW) have not perform well. RMAC [10] again underperformed both in HybridNet and AlexNet365. This is because, VPR is different from other image retrieval and recognition systems where mostly a single object majorly covers the whole image and Sum-, Max-pool and RMAC which perform relatively well in those vision based tasks actually not performed in VPR under environment changes. Due to the resembles among the places captured in sequence, Whole and SeqSLAM with their whole-image based approaches have shown better performances. At relatively higher feature encoding time, NetVLAD [3] achieved state-of-the-art performance on this dataset. One of the reasons for its supreme performance could be its pre-training on large urban

place-centric dataset (Pittsburgh) which exhibits strong lightning and viewpoint variations in the presence of dynamic and confusing objects. We also employed SPP on *conv5* of HybridNet. With higher precision at start and as recall increases, Region-VLAD PR-curves are quite similar but covering more AUC than Whole, SeqSLAM, Cross-Pool and VGG-16 Cross-Region-BoW which mimics the usefulness of our novel CNN based regional approach merged with the VLAD encoding.

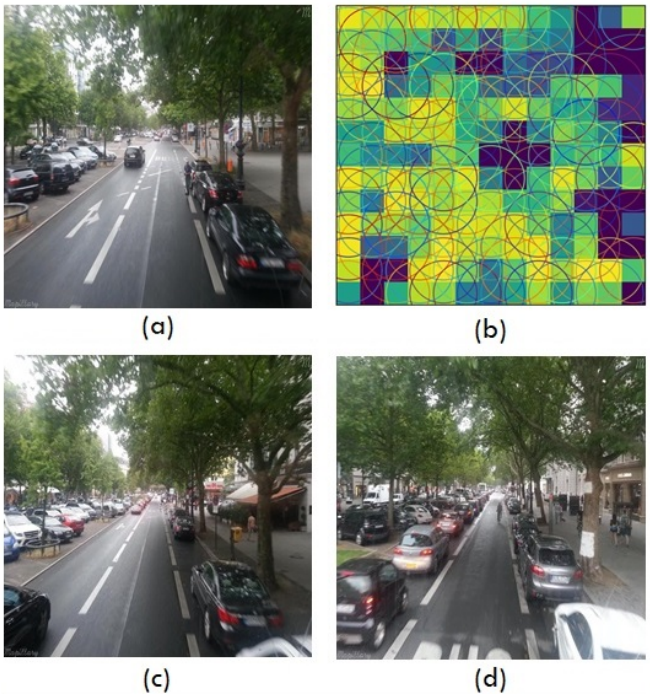

Fig. 3.11 Three different places (a), (c) and (d) of *Berlin Kudamm* exhibiting a similar scene. (b) represents the novel regions identified from (a) using our region finding approach employed on AlexNet365.

Berlin A100

The Berlin A100 dataset exhibits moderate viewpoint and moderate illumination changes. The PR-curves of the analysis are displayed in Fig. 3.13. The middle *conv4* and late *conv5* of HybridNet integrated with our region-based approach marginally show the similar PR-curves, thus, we employed *conv4* for further experiments. It is quite evident that Region-VLAD framework coupled with AlexNet365 and HybridNet achieves similar results as VGG-16 based Cross-Region-BoW [5]. Replacing VGG-16 with AlexNet365 and HybridNet in [5] achieves better results for BoW than VLAD but Cross-Region-BoW with VGG-16 still the best when comparing it with AlexNet365 and HybridNet. Against our approach, SPP on

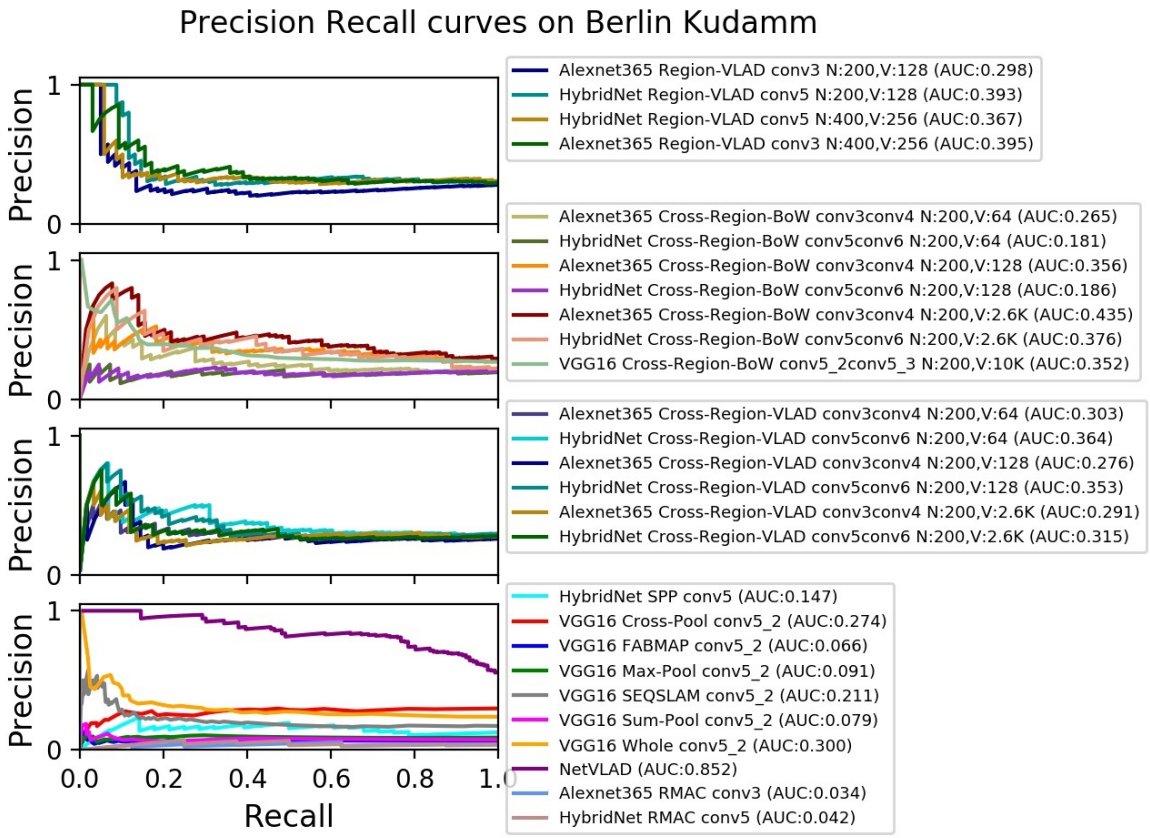

Fig. 3.12 Top: PR-curves of our proposed Region-VLAD approach. Middle: Cross-Region-BoW [5] employed on AlexNet365 and HybridNet with VLAD and BoW encodings. Bottom: Comparison with state-of-the-art VPR approaches.

HybridNet and RMAC on AlexNet365 and HybridNet also achieve comparable performances as Whole but better than FABMAP and other pooling techniques including Sum-, Max- and Cross-Pool. NetVLAD again outperformed all other VPR techniques and its PR-curve is similar to our proposed approach. Condition and viewpoint changes are not much stronger, therefore, RMAC and other approaches have also shown better performance on this dataset. A deep analysis on the datasets reveals varied time interval between consecutive captured frames which causes SEQSLAM to underperform. Overall, our proposed Region-VLAD achieved second best performance after VGG-16 Cross-Region-BoW [5].

Synthesized Nordland

The Synthesized Nordland dataset exhibits severe conditional changes and moderate viewpoint variations. Fig. 3.14 shows that our proposed approach works relatively well in comparison with all other approaches excluding RMAC which achieves state-of-the-art

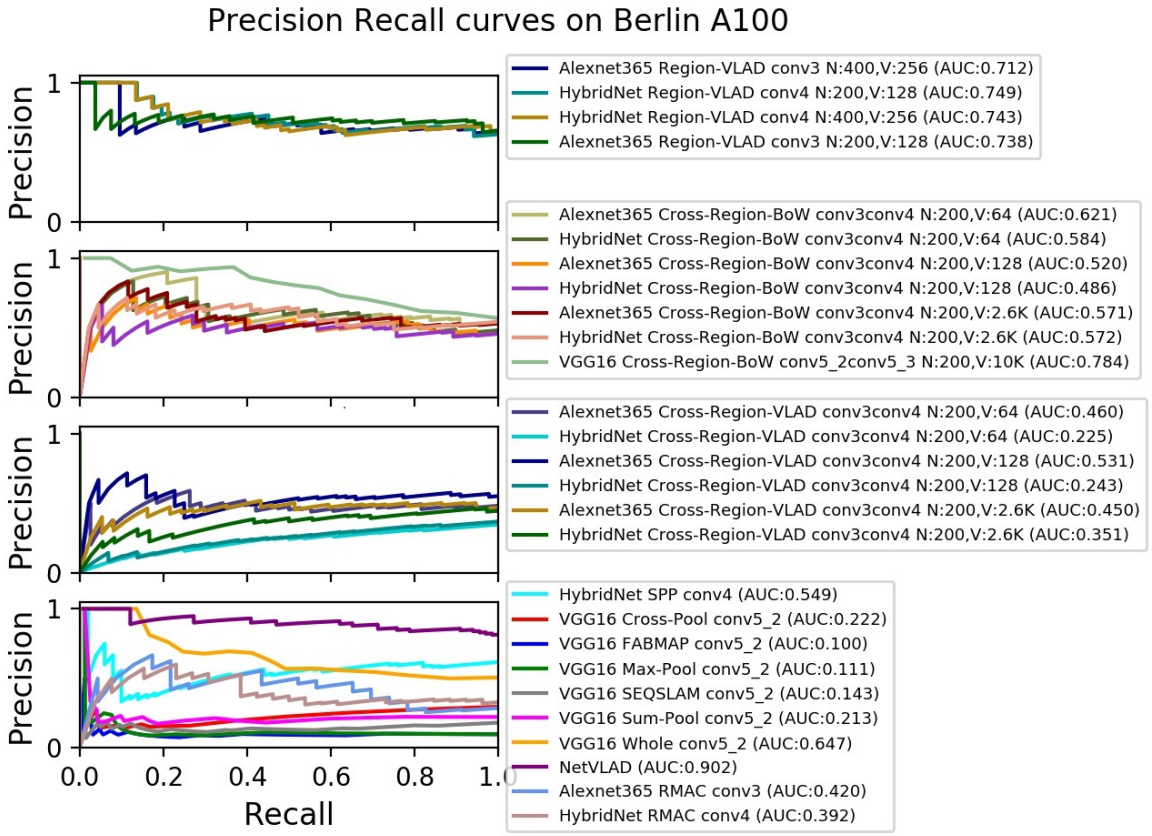

Fig. 3.13 Top: PR-curves of our proposed Region-VLAD approach. Middle: Cross-Region-BoW [5] employed on AlexNet365 and HybridNet with VLAD and BoW encodings. Bottom: Comparison with state-of-the-art VPR approaches.

performance. Approaches including Max- and Sum-Pool have not perform well on this dataset. Similar feature extracting approaches Cross-Pool and VGG-16 Cross-Region-BoW [5] and whole-image processing techniques i.e. SEQSLAM and Whole have shown similar PR-curves. Due to presence of much stronger conditional variance, middle *conv4* of HybridNet relatively shown state-of-the-art performance. NetVLAD has shown inferior results on that dataset. It might be due to the difference in training dataset and is highly sensitive under seasonal changes combined with perceptual aliasing. By nature, HybridNet is fine-tuned version of object-centric CaffeNet on SPED dataset, so condition invariance is induced into the convolutional layers feature maps which can be observed in Fig. 3.18. HybridNet integrating with RMAC and Region-VLAD has shown better performance than scene-centric AlexNet365.

Employing both AlexNet365 and HybridNet with Cross-Region-BoW [5] outperformed VGG-16 Cross-Region-BoW. Performance of HybridNet for Cross-Region-VLAD again

found not very convincing although the same model exhibits second best performance. It mimics that performance cannot be improved simply by employing different models, its the features pooling technique that makes the difference. Further investigations for Cross-Region-VLAD also suggest that it is due to different regions finding approach which cause multiple regions to cover the same areas or subset of areas of other regions, cross-convolutional regional aggregation and their mapping over the vocabulary results into non-uniform features distribution. Although, normalization is carried out but many zero regional residues get stored in the VLAD which reflects on the PR-curves. It is also observed that due to difference in regional approach, such behavior has not been observed for our Region-VLAD VPR framework.

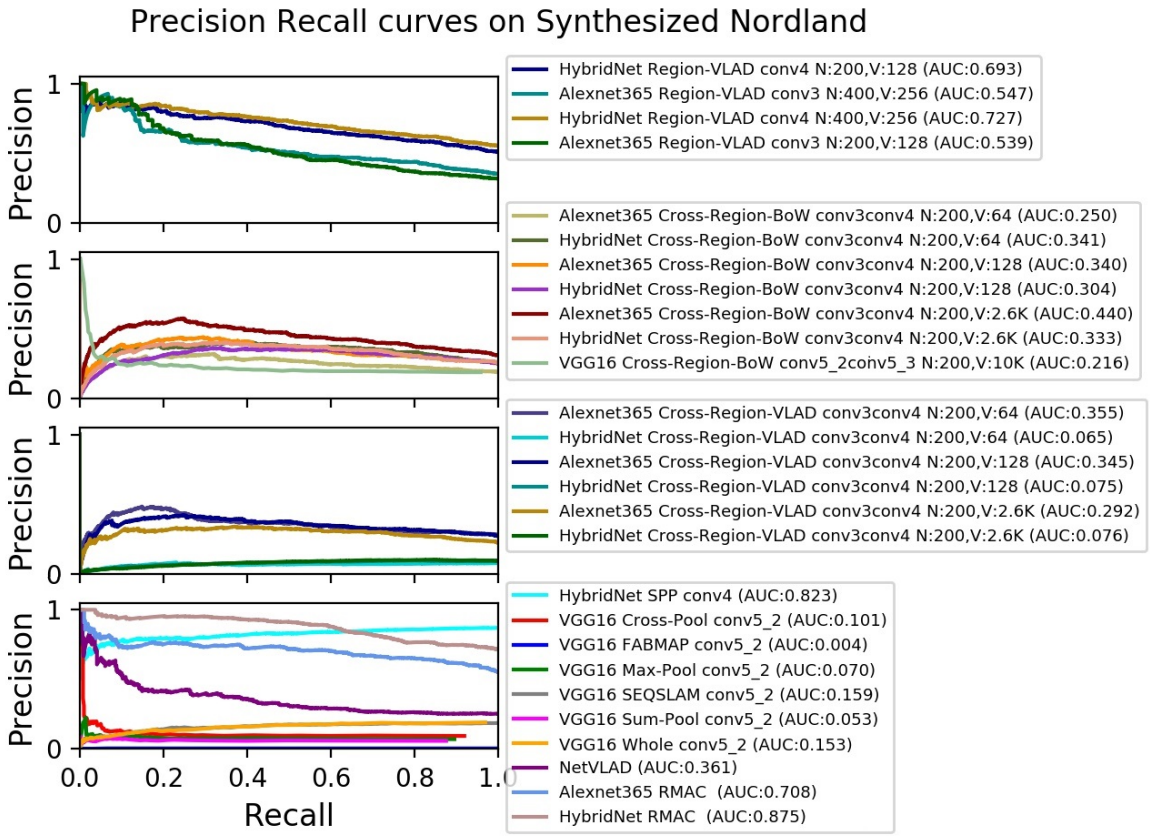

Fig. 3.14 Top: PR-curves of our proposed Region-VLAD approach. Middle: Cross-Region-BoW [5] employed on AlexNet365 and HybridNet with VLAD and BoW encodings. Bottom: Comparison with state-of-the-art VPR approaches.

Table 3.2 AUC PR-curves of Region-VLAD and Cross-Region-BoW [5] on the benchmark datasets.

Dataset	Test	Reference	AUC-PR curves			
			Region-VLAD		Cross-Region-BoW [5]	
			AlexNet365	HybridNet	AlexNet365	HybridNet
			N=400 V=256	N=400 V=256	N=200 V=2.6k	N=200 V=2.6k
Berlin A100	81	85	0.71	0.74	0.571	0.572
Berlin Haleenseetrasse	67	157	0.80	0.77	0.418	0.358
Berlin Kudamm	222	201	0.395	0.367	0.435	0.376
Gardens Point	200	200	0.726	0.668	0.683	0.558
Synthesized Nordland	1622	1622	0.54	0.727	0.440	0.333

Gardens Point

Both the Gardens Point traverses exhibit stronger viewpoint- and illumination-variance with adequate temporal coherence between the frames. AlexNet365 and HybridNet integrated with our Region-VLAD approach manages to achieve similar and better performance as Whole, RMAC, SPP and SEQSLAM which takes advantage from the sequential information. NetVLAD achieved the best performance on this dataset. Due to similar pooling approach, Cross-Pool and VGG-16 Cross-Region-BoW again exhibit similar performances but approaches including Sum-, Max-Pool and FABMAP relatively underperformed. Observing the presence of strong viewpoint changes with strong lightning change, we used late convolutional layers of HybridNet both for our approach and [5]. Combining Cross-Region-BoW [5] with AlexNet365 and HybridNet has shown comparative PR-curves as VGG-16 Cross-Region-BoW.

Overall, AUC PR-curves for Region-VLAD and Cross-Region-BoW [5] integrated with AlexNet365 and HybridNet are shown in Table 3.2. The regional configuration for Cross-Region-BoW is selected based on the AUC under PR-curves which highlight that BoW encoding worked well with $2.6k$ regional vocabulary. For our Region-VLAD approach, we have chosen $N = 400$ regional configuration due to its better overall performance. It is evident that Region-VLAD outperforms Cross-Region-BoW.

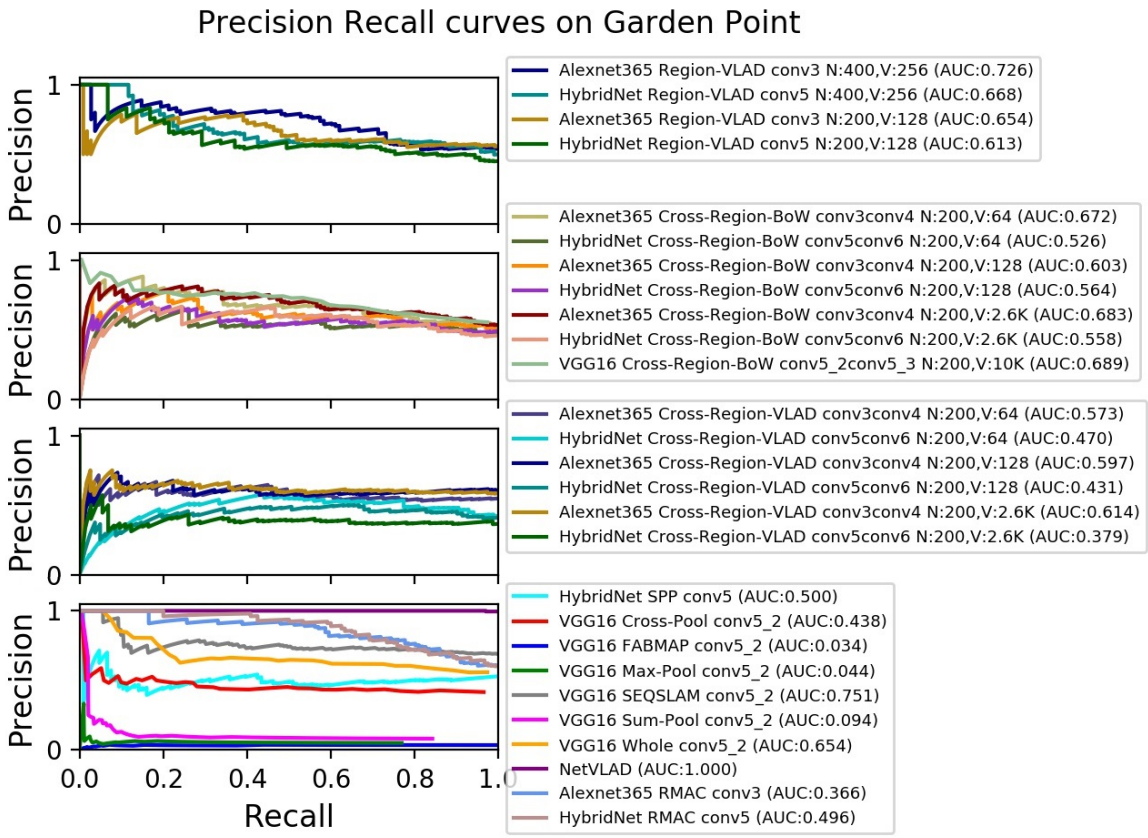

Fig. 3.15 Top: PR-curves of our proposed Region-VLAD approach. Middle: Cross-Region-BoW [5] employed on AlexNet and HybridNet with VLAD and BoW encodings. Bottom: Comparison with state-of-the-art VPR approaches.

3.4.3 Receiver Operating Characteristic (ROC) curves and Matching Score Thresholding

In majority of the VPR techniques [84][86][40][5][6], AUC under PR-curve is used as a evaluation parameter. For an input query, the proposed frameworks retrieve the reference/database image with maximum matching score criteria but what will happen if no match exists for some queries? So, to tackle such tricky situations, firstly, we set a limit on the matching score such that the maximum matching score value of the retrieved database place should be greater than the threshold value set and if that not is the case, then the system should not retrieve any match and increment the True-Negatives (TNs) counter by one. However, when the system returns the correct match but the matching score is lower than the threshold then False-Negatives (FNs) counter will increase which means that the system has discarded the correctly retrieved match due to the match score thresholding criteria. By nature, PR curves

Table 3.3 AUC ROC-curves of Region-VLAD and Cross-Region-BoW [5] on the benchmark datasets with reduced reference traverses

Dataset	T	R'	T'	AUC-ROC curves			
				Region-VLAD		Cross-Region-BoW	
				AlexNet365	HybridNet	AlexNet365	HybridNet
				N=400 V=256	N=400 V=256	N=200 V=2.6k	N=200 V=2.6k
Berlin A100'	81	64	70	0.687	0.719	0.587	0.64
Berlin Haleenseetrasse'	67	138	50	0.837	0.771	0.632	0.564
Berlin Kudamm'	222	151	166	0.657	0.65	0.617	0.591
Garden Point'	200	150	152	0.766	0.797	0.766	0.718
Synthesized Nordland'	1622	1217	1221	0.674	0.7555	0.670	0.657

do not consider the True Negative cases (correctly missed the non existing events/classes) [112]. Therefore, we employed the ROC curves [113] using T test traverse and R' reference traverse for all the datasets knowing that $T - T'$ queries can be treated as new places (see Table 2.2).

For all the modified datasets, Fig. 3.16 illustrates the ROC curves for AlexNet365 and HybridNet integrated with our Region-VLAD (top row) and Cross-Region-BoW[5] (bottom row). For a fair comparison, the same layer configurations are kept with results confirm that the overall performance improvements are better and comparable. The AUC under ROC curves for the approaches in Fig. 3.16 are presented in Table 3.3. Since, ROC curves dominate if and only if PR-curves work well [113]. So, we pass down the T queries into both the systems, knowing that only T' have matched reference images. Both the Region-VLAD and Cross-Region-BoW [5] still keep the AUC under ROC curves higher that means AUC under PR curves will also be higher too but here again, Region-VLAD outperforms Cross-Region-BoW in terms of AUC ROC curves. The shape and AUC of the ROC curves in Fig. 3.16 suggest that integrating HybridNet with Region-VLAD is more effective when the places experience severe conditional and moderate viewpoint variations (Synthesized Nordland) whereas AlexNet365 is efficient in dealing strong viewpoint and moderate condition variations (Berlin Haleenseetrasse). With moderate conditional and viewpoint variations such as Berlin A100, both the models exhibit similar performances when integrated with Region-VLAD framework. It is interesting to note that both the models when integrated with Cross-Region-BoW have different ROC curves for Berlin A100 and Synthesized Nordland which also highlights towards the difference in regions employed by our Region-VLAD and Cross-Region-BoW.

Furthermore, Fig. 3.17 visualized the deep analysis of Fig. 3.16 when Region-VLAD is integrated with AlexNet365 and employed on multiple datasets having true-negative events. Matching scores in y-axis differentiate the TP, FN, FP and TN events shown with different colored curves and length of the curves in x-axis denotes the number of images which the events contain. In the figure, left column graphs exhibit the scores distribution with no thresholding. Upon thresholding on the matching scores, the right column showcases the changes in TP, FP, TN and FN. For Berlin Halenseestrasse dataset in left graph of the top row, when reference frames are reduced and only $T' = 50$ out of $T = 67$ queries have matched reference images then Region-VLAD assigns low scores to those 17 queries. Note that the threshold is calculated by averaging the true negative scores calculated over all the modified datasets with reduced reference traverse. The right column graphs illustrate the changes when minimum matching score criteria is applied. We can see that for Berlin Halenseestrasse, Region-VLAD missed 2 correctly retrieved matched images, TP changes from 30 to 28, FP increased to 27 where TN reduce down to 10 from 17. The same behaviour is observed for Berlin A100 and Garden Point which have 70 and 150 queries with matched reference frames out of 81 and 200. It is evident that Region-VLAD performs relatively well in retrieving correct images by maximizing the TN and minimizing the FP and FN. Knowing that under scenarios when a query is a new place, and for every query, it's quite challenging to successfully retrieve the correct match while reducing incorrect retrieval. Therefore, 100% performance is impractical to achieve. However, Fig. 3.16 and Fig. 3.17 do emphasis upon the fact that our proposed Region-VLAD framework not only boost up the AUC under PR-curves but also skilful in assigning low scores to the queries against which no match should be retrieved.

3.4.4 Performance Analysis

The variations in the AUC-PR curves for the proposed Region-VLAD VPR framework across all the benchmark datasets are due to many factors. The first reason is the environment of the dataset on which the CNN model is trained. Since Place365 database [33] consists of scenes/labels, where each label contains different places exhibiting the same scene like shopping mall, restaurant, rain-forest and other indoor/outdoor scenes. It strongly influence the CNN layers responses. Region-VLAD taking advantage from the training, it strains activations which are more focused on objects of the trained label/scene. So, even under different conditions and viewpoints of the same places, the Region-VLAD focuses on the scenes by putting emphasis on place-centric regions, hence, the scenes get correctly recognized. However, we have also seen employing such trained models directly with other features pooling techniques like Sum-, Cross- and Max-Pooling have not worked well which

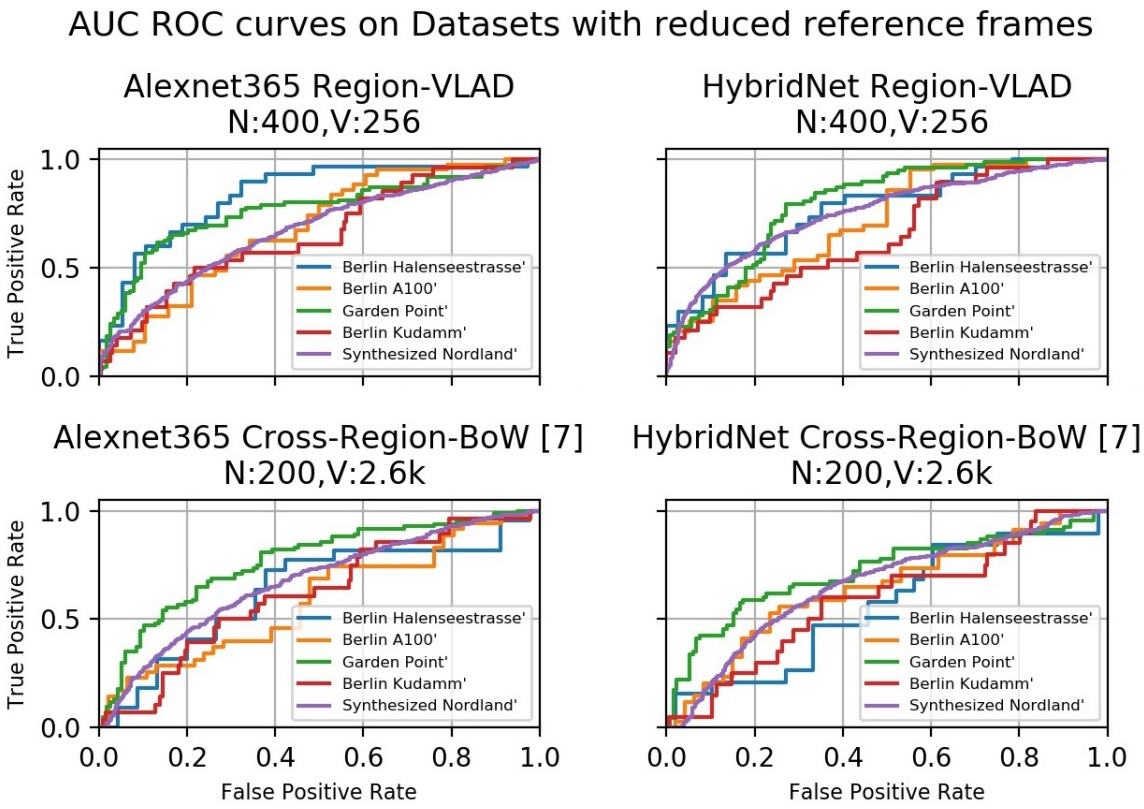

Fig. 3.16 ROC curves for datasets with true-negative scenarios for Region-VLAD and Cross-Region-BoW [5].

highlights the superiority of our novel regions extraction approach. The superior performance of NetVLAD also point towards the CNN pre-training; Pittsburgh dataset contains places captured under very strong viewpoint and lighting variations in multiple day/night/evening times of the year. However, we have also observed its worst recognition performance when there is a strong seasonal change with resemblance among the geographically different scenes.

The results suggest that HybridNet with our Region-VLAD framework found to be better than AlexNet365. It is because SPED added condition invariance in the original object-centric convloutional layers, and for datasets with severe conditional changes such as Synthesized Nordland, we can see the performance difference over AlexNet365. It would also be interesting to fine-tune scene-centric AlexNet365 with SPED dataset then to investigate the Region-VLAD performance. Nearly for all the benchmark datasets, the performance of Cross-Region-BoW [5] integrated with AlexNet365 or HybridNet is comparable but at a large regional dictionary and matching time cost. A deep analysis suggests that the Cross-Region-BoW after retrieving the cross-convolutional regional features performs cross matching

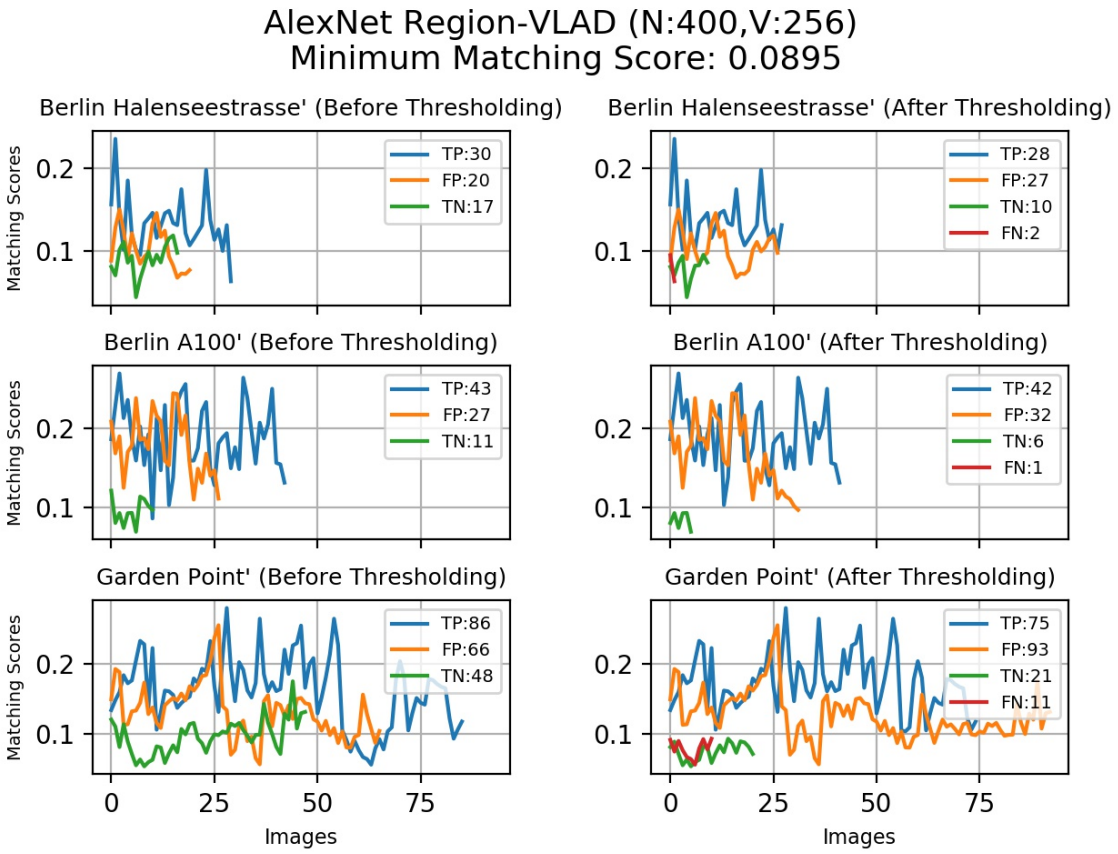

Fig. 3.17 Matching scores thresholding using Region-VLAD with true-negative cases. Each row is associated with a dataset; left graph presents TP, FP, TN and FN before thresholding and right side graph showcases the change upon thresholding.

and then only employs the mutually matched regions which reduce down from $N = 200$ to $N = 9$. Therefore, scores remain positive by neglecting the non-mutual regions. However, our Region-VLAD still outperformed Cross-Region-BoW [5] with smaller dictionary and low time cost. It is observed that the boost in performance is encouraged with our environment invariant robust novel regions because the regional approach of [5] when combined with the VLAD, denoted as Cross-Region-VLAD underperformed both with small and large regional dictionaries. Furthermore, we also evaluated the system's skill in rejecting the match for a new query place, carried out by reducing down the datasets' reference traverses followed by the AUC under ROC curves calculation. AUC under ROC curves on the reduced reference datasets demonstrated that Region-VLAD assigns low scores to the true negative scenarios which get filtered out by matched score thresholding.

Secondly, the diversity and size of the dataset employed to make the regional vocabulary also play an important role with contribution of VLAD encoding and cosine matching for

determining the regions similarities. We have also observed that picking more regions boost up the accuracy. This is apparent because in the pre-trained regional vocabulary, this might be possible that few or more clustered regions suit one dataset more as compared to others. But sometimes, inclusion of more regions also degrade the performance; each region contributes to the final matching score which might results into a wrong match if multiple reference images exhibit the similar scenes and inclusion of more but less energetic regions decay the overall final score for the correct match. For VLAD retrieval, the collected dataset for regional vocabulary contains only $2.6k$ images, whereas in Cross-Region-BoW [5], $5k$ images were employed. Bigger the dataset, more diverse the dictionary will be. However, due to our system run-time memory limitation and to load images for regions and features aggregation, we have confined ourself to $2.6k$ images. But, we have kept variety in our $2.6k$ dataset to learn diverse regional features which reflects on our results with small vocabulary size. Clustering the regions using K-means to make the regional vocabulary is also valuable; we generated the dictionary twice using the same $2.6k$ dataset. AUC-PR curves across all the benchmark datasets using both the dictionaries vary with an average marginal difference of 0.03 AUC-PR.

Lastly, employing a less layered CNN architecture found to be computation- and memory-efficient as it has lesser trainable parameters. It showed the potential to boost up the recognition performance with our proposed Region-VLAD approach for VPR under changing environment. Fig. 3.19 and Fig. 3.20 illustrate some of the matched and mismatched scenarios for AlexNet365-based Region-VLAD framework. For the correct matches, taking advantage from CNN's scene-centric training, Region-VLAD identifies the common regions shown with different colored boxes under simultaneous viewpoint and appearance changes. For the mismatched scenarios, the identified top novel regions with colored boxes (trees, lamp posts) show the areas where the system confuses in and matches the scenes but wrongly recognizes the places. The failure cases again point towards the scene- or place- centric training of the CNN; our proposed approach identifies the common regional features of geographically different places (query and retrieved frames) exhibiting the similar scene or condition and leads to places mismatch. We have seen that Cross-Region-BoW [5] when integrated with AlexNet365 showed comparable performance but at high time computation cost. However, the presented Region-VLAD still outperformed Cross-Region-BoW [5] with smaller dictionary and low retrieval time. Also, cross-regional approach of [5] when combined with the VLAD has shown inferior results which confirms the performance boost in the Region-VLAD encouraged with the novel regional approach. Datasets and results are placed at [121].

3.5 Summary

For Visual Place Recognition on resource-constrained mobile robots, achieving state-of-the-art performance/accuracy with lightweight CNN architectures is highly desirable but a challenging problem. This chapter has taken a step in this direction and presented a holistic approach targeted for a CNN architecture comprising a small number of layers. With the shallow nature of CNN model, there is reduction in memory and computational cost, thus, suitable for power constrained VPR applications.

The proposed framework detects novel CNN-based regional features and combines them with the VLAD encoding methodology adapted specifically for computation-efficient VPR. In terms of performance, it has shown state-of-the-art AUC-PR curves on severe viewpoint and moderate condition-variant place recognition datasets against the deep neural network-based VPR techniques. In the next chapter, we will present and discuss another lightweight and computation-efficient VPR technique efficient under severe seasonal and conditional changes coupled with moderate viewpoint variation.

Fig. 3.18 Sample images of identified ROIs using our Region-VLAD approach employed on AlexNet365 and HybridNet.

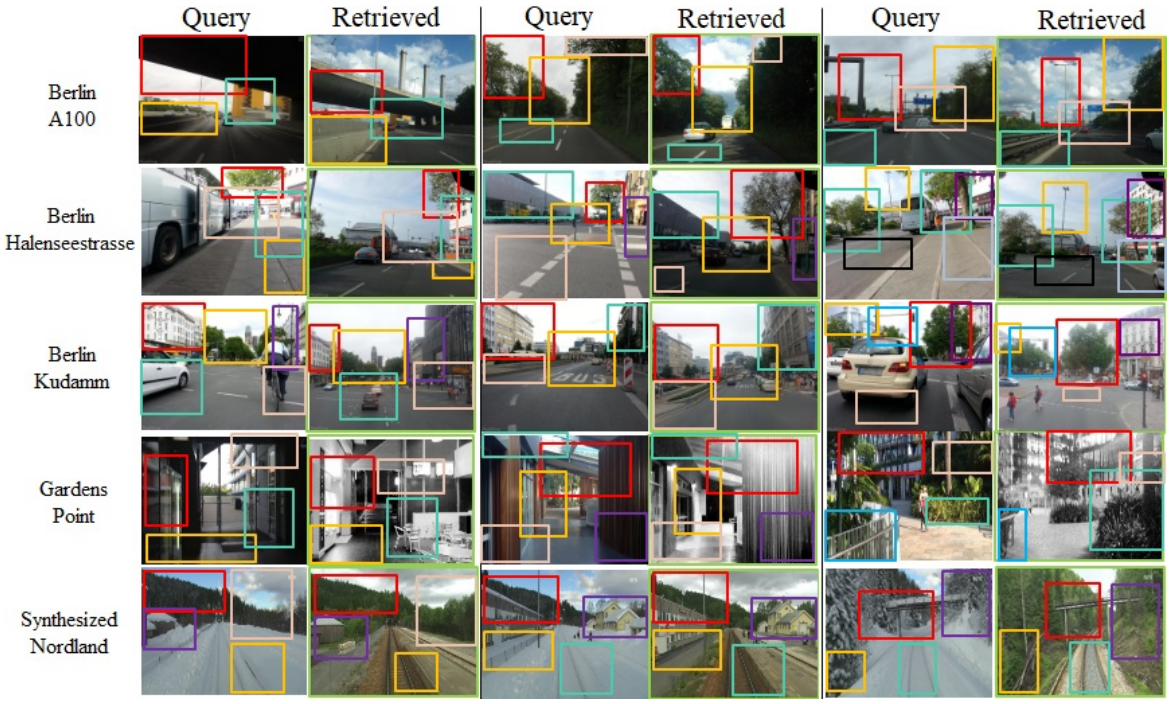

Fig. 3.19 Correctly retrieved places with the proposed Region-VLAD framework.

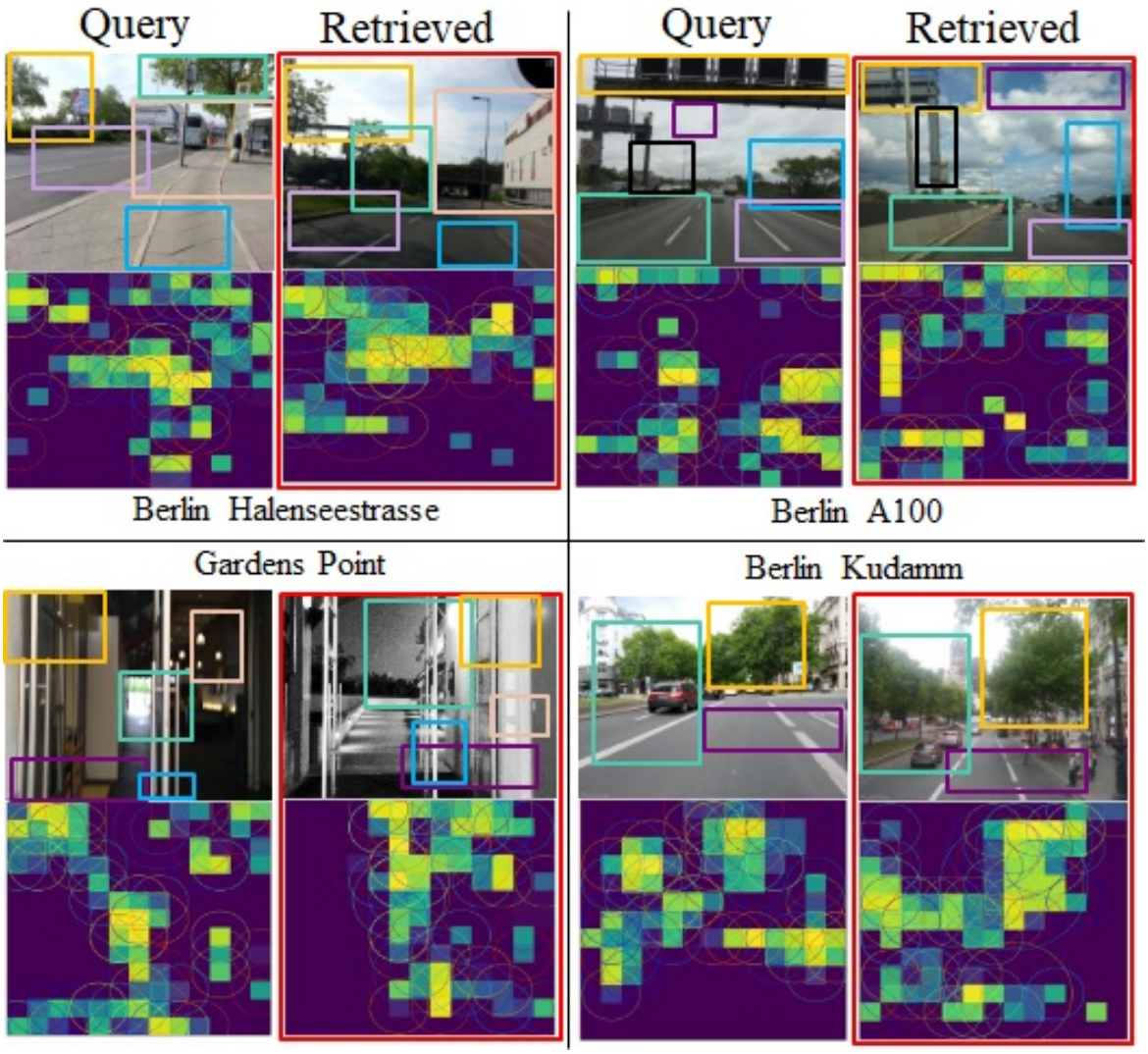

Fig. 3.20 Incorrectly retrieved places with the proposed Region-VLAD framework.

Chapter 4

Context-Aware Attention framework for Visual Place Recognition

In this chapter, another visual place recognition system is presented that combines context-aware attentions from multiple convolutional layers. Shallow place recognition-centric HybridNet is integrated with the proposed multi-scale attention framework. At low computation cost, the proposed VPR technique has shown comparable recognition performance than state-of-the-art deep neural network based place recognition and image retrieval tasks. It is found that the system focuses on most persistent place-centric regions while filtering down the dynamic instances, and significantly improves the single image matching. Publicly available challenging place recognition datasets are tested with area under precision-recall curves used as an evaluating criteria. The results confirms better and comparable AUC-PR curves against 11 state-of-the-art deep CNNs based VPR techniques.

4.1 Introduction

In VPR, focusing on dynamic entities other than static objects (such as, road signs, buildings structures) can instigate deceptive information in recognizing places. Despite a better AUC-PR performance of the framework proposed in Chapter 3 [9], sometimes, with higher regional features it suffers with the inclusion of time-changing objects in the final VLAD representations. This is due to fact that the employed CNN is pre-trained on Places365 dataset [33], which consists of millions of scenes, and within each scene/label, it contains geographically different places exhibiting the similar environment. Regional approach of [9] on scene-centric AlexNet365 considers the time-varying objects (such as pedestrians,

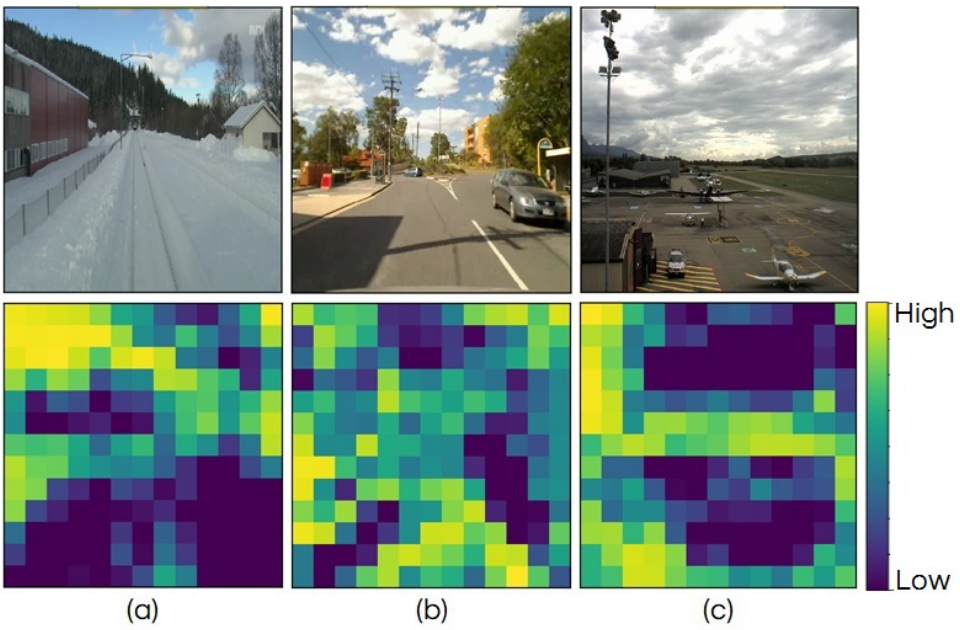

Fig. 4.1 Three exemplars are shown against which (a), (b) and (c) represent their identified novel multi-layer fused regions.

vehicles, clouds) vital in distinguishing the types of places that leads to places mismatch while matching the scenes successfully, often referred as perceptual aliasing.

To address this problem, the idea of Region-VLAD [9] is extended to multi-layered region-based approach and integrated it with the shallow SPED-centric HybridNet [4]. The proposed framework captures more powerful and rich semantic region-based attentions where the attentions' areas vary with the context. Similar to the proposed approach, the authors in [4][93][6] also attempted to learn fused multi-level regional features based on context of the places. The authors in [94] proposed a fixed context-aware attention model that captures the manually defined rectangular shaped most contributing regions efficient for localization. [4] and [93] suggested that middle and late convolutional layers capture different structural information. [6] fine-tuned the deep object-centric VGG-16 on SPED dataset to learn context-aware features in the newly added late context-flexible layer. However, improving VPR performance with deep CNNs does add computational and memory constraints in robotic applications where the response time is vital [119]. Passing down each input query into the deep neural network while utilizing late convolutional layers' features degrade the overall retrieval performance. Several experiments show that at higher performance with reduced time-computation and memory footprints, the multi-layer region-based attentions captured by the approach are robust under severe condition and appearance changes. Fig. 4.1 illustrates the novel multi-layered CNN-based regional attentions identified by the proposed lightweight

M-Region-VLAD framework on three exemplars. The proposed M-Region-VLAD VPR contemporary technique focuses on those image regions which remain static under changing environment. It is worth noticing that the time-changing place-centric objects such as, cars and clouds are strained by the system, thus, improving the overall matching performance in terms of AUC-PR curves over the state-of-the-art VPR approaches including Attentive Attention [122], Region-VLAD [9], NetVLAD [3], RMAC [10], Cross-Region-BoW [5], Cross-Pool, FABMAP [123], Fix-Context [94], Context Flexible Attention [6], Places365 [33] and SEQSLAM [42].

The remainder of the chapter is organized as follows. Section 4.2 describes the proposed framework in detail. Section 4.3 and 4.4 present the experimental setup, detailed analysis and results obtained by evaluating the proposed framework on challenging benchmark datasets. Section 4.5 ends with the conclusion.

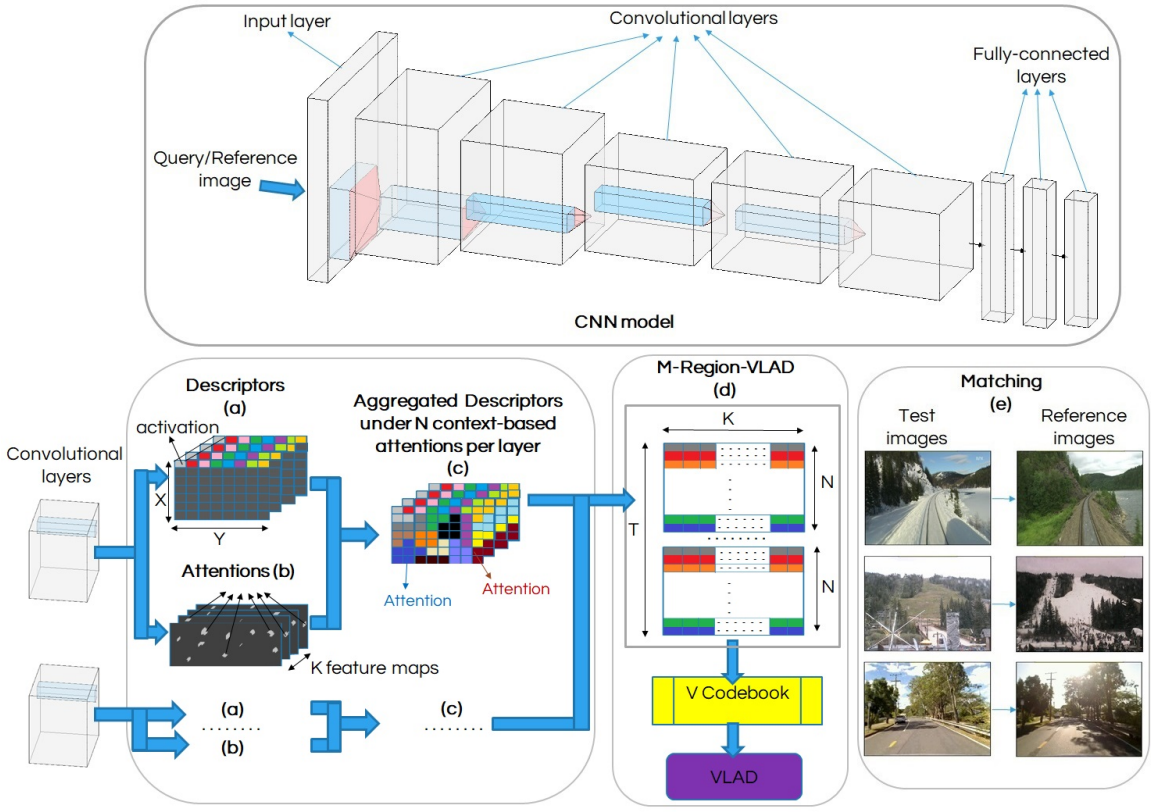

Fig. 4.2 Images are fed into the CNN model. The identified attentions from multiple convolutional layers are fused and mapped on a dictionary for VLAD retrieval.

4.2 Proposed Multi-Layer Region-VLAD VPR framework

This section will describe the proposed framework in more detail. To subdivide an image into spatial regional representations, retrieval of the local descriptors from the convolutional layer feature maps is discussed. It then demonstrate the approach of finding regional attentions from multiple convolutional layers, followed by the discussion on how to aggregate and map the regional local descriptors on a separate regional vocabulary to retrieve a compact VLAD image representation. The overall framework is shown in Fig. 4.2. N Regions of Interest (ROIs) per convolutional layer is identified, followed by regional local descriptors aggregation. T aggregated region-based attentions across the multiple layers are used as features representation. The proposed M-Region-VLAD VPR approach can be incorporated with any CNN model.

4.2.1 Stacking of Convolutional Activations for making Descriptors

In a neural network, $X \times Y \times K$ is the dimension of $3D$ convolutional layer tensor M, where X and Y represent the width and height of each channel and K is the number of channels, also termed as feature maps. In layman terms, each feature map $k = \{1, 2, ..., K\}$ corresponds to some filter being convolved with the input image I. At certain spatial location(s), stacking the activations of K feature maps is performed, and each spatially stacked activations vector is termed as a local descriptor, visually shown in Fig. 4.2(a). In (4.1), D^L denotes the K dimensional d_l local descriptors at L^{th} convolutional layer of the m_c model.

$$D^L = \{d_l \in M^K \;\; \forall\, l \in \{(i, j) \mid i = 1, ..., X; j = 1, ..., Y\}\} \tag{4.1}$$

4.2.2 Identification of Context Aware Regional Attentions

Within the convolutional layer of a CNN, certain spatial regions of the feature maps do have more intensity mimicking the presence of certain visual patterns in the image. For example, giving an image of an urban/rural road scene as an input to a CNN, one certain convolutional layer feature map might be focusing on the vehicles while others can find buildings as an important visual clue. In the context of Visual Place Recognition, focusing on time-varying objects such as pedestrians and vehicles can degrade the overall matching performance. Therefore, salient regions corresponding to static objects including road signal, buildings can help to recognize a visual place even under severe condition and viewpoint variations.

For finding context-based most contributing regions in an image of the place, a shallow pre-trained SPED-centric HybridNet [4] is integrated. Particularly, we process the feature

maps of the convolutional layer and grouped the non-zero spatially connected activations such that two or more activations couple to represent a G_h salient region if roughly have similar responses, $\forall\, h \in \{1,...,H\}$ where H is the total number of identified salient attentions from K feature maps at L^{th} convolution layer (visualized in Fig. 4.2(b)). Similar to [9], energies of all the identified regional attentions are calculated by averaging over all the a_h activations lying under each G_h attention. In (4.2), a_h^f represents the f^{th} activation lying under G_h region where E^L denotes the regional energies. In (4.3), with sorted E^L energies, R^L represents the top N energetic novel context-based ROIs.

$$E^L = \{\frac{1}{|G_h|}\sum_f a_h^f,\ \forall\, a_h^f \in G_h\} \tag{4.2}$$

$$R^L = \{G_t\ \forall\, t \in \{1,...,N\}\} \tag{4.3}$$

Considering the recognition performance and also to forbade the inclusion of time-varying objects in the final region-based features, $N = 300$ attentions per layer are captured because with the inclusion of more but less energetic regions, activations concentrated on dynamic objects do get included. Experimentation at $N = 300$ confirms negligible dynamic instances involvement in the captured regional representations. Under the identified q attention of R_N^L total attentions, D_q^L denotes the underlying regional local descriptors, aggregated in (4.4) to retrieve $1 \times K$ dimensional f_t^L context-based regional feature. F_I in (4.5) represents the concatenated $T \times K$ attention-based CNN features for an image I in L_3 and L_4 convolutional layers of the model, illustrated in Fig. 4.2(d). Fig. 4.3 illustrates the fused multi-semantic attentions captured from middle *conv*3 and late *conv*4 convolutional layers of HybridNet. It is be seen that the system focuses on static objects that remain persistent even in the presence of confusing and dynamic instances (such as, sky, cars).

$$f_t^L = \sum_{q \in R_t^L} D_q^L\ \forall\, t \in \{1,...,N\} \tag{4.4}$$

$$F_I = f_t^l\ \forall\, l \in \{L_3, L_4\} \tag{4.5}$$

4.2.3 Attentions-based Vocabulary and Extraction of VLAD for Image Matching

With smaller visual word vocabulary in tasks including image retrieval, recognition and object detection [10][90], VLAD [99] has shown state-of-the-art performances. For the VLAD retrieval, K-means [54] is acquired where features are quantized to the dictionary

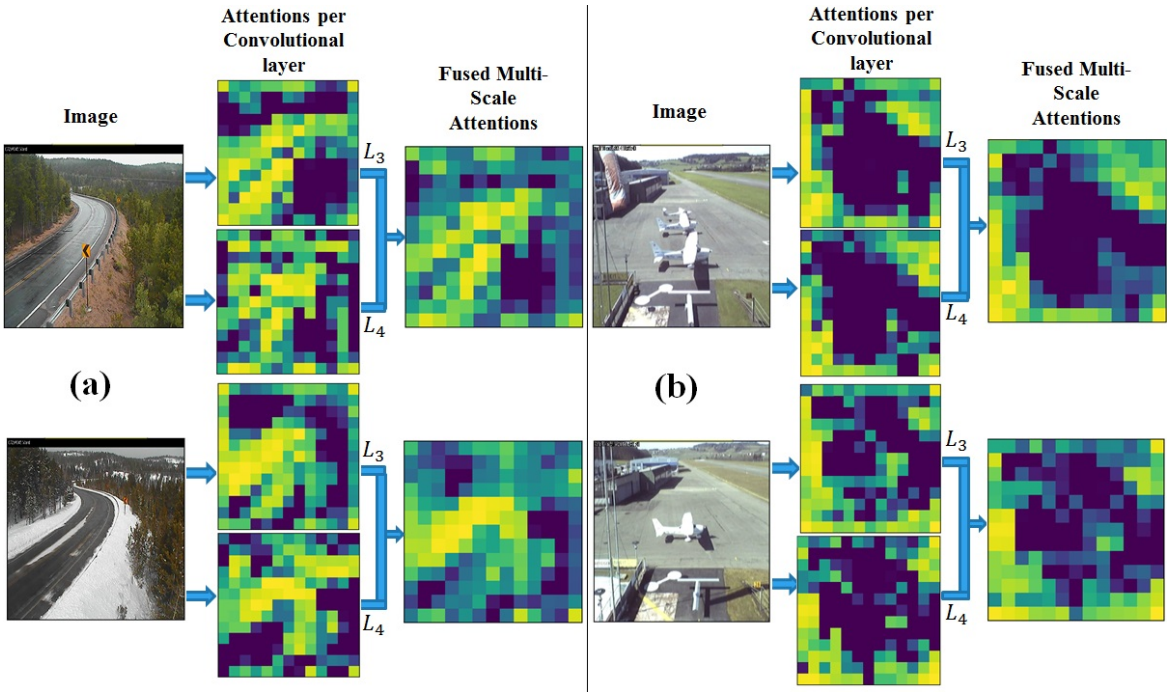

Fig. 4.3 Fused multi-scale attentions captured under strong conditional changes coupled with dynamic instance experienced by the place (a) and (b) under different times of the year

clusters and their residues are accumulated and concatenated to retrieve a single compact vector. Therefore, for the attention-based dictionary, a separate dataset of $3K$ images is collected which contains 1125 Query247 [8] images taken in day, evening and night times of 365 places. The other images consist of Garden Point [4], Eynsham [84] and multiple environment variant rural and urban road traverses captured from *Mapillary*.

The novel multi-layer attentions capturing approach is employed on the $3K$ dataset to learn a dictionary of attentions. In (4.6), K-means is used to cluster $3000 \times T \times K$ dimensional context-aware attentions into V regions such that o_u represents the u^{th} attention centre of the C codebook. For all the benchmark test and reference frames, in (4.7), their F_I attentions are quantized to predict their Z clusters/labels where the quantization function α maps all the attentions on the V clustered learned dictionary.

$$C = o_u \; \forall \; u \in \{1, ..., V\}, \; V \in \{128\} \tag{4.6}$$

$$Z = \alpha(F_I) \tag{4.7}$$

For each u^{th} attention of C codebook, using the multi-layer context-based attentions F_{I_u}, predicted labels Z_u and the C_u attention center, the VLAD component v_u with dimension

$1 \times K$ can be obtained using (4.8). Precisely, the residues of the F_{Iu} attention-based features and C_u attention center are summed.

$$v_u = \sum_{F_{Iu}:Z_u=C_u} F_{Iu} - C_u \tag{4.8}$$

The $V \times K$ dimensional S representation in (4.11) is the final VLAD representation of the image. However, visual word burstiness is the common issue being faced in many other visual words based task [117] where some or fewer visual words appears more frequently than the statistical expectation. Power normalization [118] followed by L_2 normalization is the standard techniques applied on the summed residual in (4.9) and (4.10) with γ served as a non-linear transformation function.

$$v_u := sign(v_u)\|v_u\|^{\gamma} \tag{4.9}$$

$$v_u := \frac{v_u}{\sqrt{v_u^T v_u}} \tag{4.10}$$

$$S = \{v_u \; \forall \; u \in \{1, ..., V\}\} \tag{4.11}$$

In (4.12), for a test image “A" against the reference “B", the scalar product of their u^{th} VLAD components, G_u^A and G_u^B reaches to individual attention matching score $j_u^{A,B}$. All the V attentions score are summed up in (4.13) to get final $J^{A,B}$ matching score. Against all the “X" reference images, highest $J^{A,X}$ matching score is picked with “X" being claimed as the matched image.

$$j_u^{A,B} = \frac{(S_u^A).(S_u^B)}{\|(S_u^A)\|\|(S_u^B)\|} \tag{4.12}$$

$$J^{A,B} = \sum_{u=1}^{V} j_u^{A,B} \tag{4.13}$$

$$P^A = \arg\max_X J^{A,X} \tag{4.14}$$

4.3 Setup and Implementation details

Deep learning techniques are computationally expensive which makes it indispensable to evaluate the run-time performance in order to realize the system’s deployment in robotic VPR applications. The presented VPR framework is implemented in Python 3.6.4 and the system average run-time over 3 iterations with 3244 images is recorded. For all the baseline

experiments, we employed HybridNet and used middle *conv*3 and late *conv*4 convolution layers to capture rich semantic context-aware regional features. For an image, the forward pass takes an average $M_f = 13.85\ ms$ using Caffe on **Intel Xeon Gold 6134 @3.2GHz**. Other parameters including $N = 300$ attentions per layer with $V = 128$ clustered vocabulary for VLAD encoding. Extraction of T context-aware attentions per image takes around $M_e = 140.5\ ms$ with VLAD encoding and (two VLADs) matching takes $M_v = 2.68\ ms$ and $M_m = 0.07\ ms$ [9]. Therefore, let say with $R = 1622$ reference VLAD representations, the total retrieval time M_q for a single query against R stored database VLADs can be calculated using (4.15), comes around 270.57 *ms*. 128×384 dimensional VLAD representation per image consumes around $393KBytes$ memory.

$$M_q = M_f + M_e + M_v + M_m * R \tag{4.15}$$

In comparison, memory and time computation for NetVLAD, RMAC, Region-VLAD, Cross-Region-VLAD and SeqSLAM are higher than the proposed M-Region-VLAD framework, as reported in [109]. Employing **Intel Xeon Gold 6134 @3.2GHz** for feature encoding, NetVLAD takes an average $0.77s$, RMAC takes $0.47s$ as reported in Table 4.1. Employing **Titan X 1080**, state-of-the-art Context Flexible Attention [6] is evaluated on 1101 images and takes around $M_f + M_e = 14.1\ ms$ $(M_v = 0)$ per image for features extraction. The $512 \times 14 \times 14$ dimensional feature vector consists of multi-scale fused attentions, consumes $401KBytes$ memory. Using Python 3.6.4, feature matching is performed by flattening the 3D vector, followed up with cosine distance matching further takes an average $M_v = 0.63\ ms$ employing Intel Xeon Gold 6134 @3.2GHz. Therefore, an overall retrieval time for matching a single query against $R = 1622$ reference images takes around 1035.96 *ms*. It should be noted that Context Flexible Attention [6] employed Titan X 1080 for feature encoding and NetVLAD [3] takes significantly higher feature encoding time using Intel Xeon Gold 6134 @3.2GHz whereas our M-Region-VLAD framework achieves comparable performance at low computation and resource utilization, as reported in Table 4.1.

In our experimentations, we have not considered the runtime memory consumed when loading the CNN model but we observed that a small layered CNN model (AlexNet365, HybridNet) has less number of trainable parameters (weights) as compared to the deep neural networks such as, VGG-16 [4]. Moreover, rather than preloading all the weights of CNN model, we load the parameters til the feature extraction layer which is in our case is middle convolutional layer(s). In comparison, state-of-the-art techniques including NetVLAD, RMAC, Context Flexible Attention and Cross-Region-BoW, deep VGG-based features are extracted from the late convolutional layer(s) (conv5_2 and conv5_3), thus needs to preload all the layers' weights. Therefore, we can expect that the overall memory required

by our proposed technique should be lesser than the state-of-the-art deep CNN-based VPR techniques.

Table 4.1 Feature encoding and matching times of the VPR aproaches.

Techniques	Feature Encoding (ms)	Feature Matching (ms)	**Techniques**	Feature Encoding (ms)	Feature Matching (ms)
Intel Xeon(R) Gold 6134 @ 3.20GHz with 32 cores, 64GB RAM					
SeqSLAM	0	1.5	*Cross-Region -BoW*	830	160
NetVLAD	770	0.005	*RMAC*	470	0.04
Region-VLAD	460	0.12	*M-Region-VLAD*	157.03	0.07
Techniques	Feature Encoding (ms)		Feature Matching (ms)		
	Titan X 1080		Intel Xeon(R) Gold 6134 @ 3.20GHz with 32 cores, 64GB RAM		
Context Flexible Attention	14.1		0.63		

4.4 Results and Analysis

This section compares the context-based attentions identified by the presented approach with state-of-the-art VPR and image retrieval techniques [6].

4.4.1 Comparison Techniques

To make a fair comparison, we also reported the performance of other VPR approaches evaluated in [6] that includes Attentive Attention [122], Cross-Pool, FABMAP, Fix-Context [94], Context Flexible Attention, Places365 [33] and SEQSLAM. Particularly, for state-of-the-art Attentive Attention approach and VPR-based Fix-Context framework, authors have fine-tuned these models on SPED dataset while removing the geometric verification layer. For Cross-Pool [90], the late convolutional layer is employed to generate a fixed attention mask, used as features representations. For handcraft-based VPR approaches such as, FABMAP and SEQSLAM, the authors employed their official implementations [123][42]. Places365 is a CNN model pre-trained on *2 Million* diverse scenes. The authors used responses of the late fully-connected convolution layer as features representation.

Furthermore, other CNN-based VPR algorithms such as NetVLAD, RMAC, Cross-Region-BoW and Region-VLAD are also evaluated. For Region-VLAD, $N = 200$ regions are employed from *conv3* of AlexNet365 with $V = 128$ clustered vocabulary for VLAD retrieval [9]. All other approaches used VGG-16 pre-trained on object-centric ImageNet. Their layers configuration are kept same as in [109]; *conv5_2* is used for RMAC, with power- and l2-normalization on the regional features. For Cross-Region-BoW, *conv5_2* and *conv5_3* are employed with $10k$ BoW dictionary. For both the techniques, cosine matching is performed for filtering the mutual regions and their scores are summed and database image with highest score considered as matched place. Given an image, NetVLAD outputs a feature descriptor and cosine matching of the feature descriptors is performed with scores summation and reference image with highest score represents the currently encountered place.

4.4.2 Precision Recall Characteristics

For all the benchmark place recognition datasets, Area under Precision-Recall curves [112] (AUC-PR curves) is used for evaluating the proposed place recognition framework, state-of-the art image retrieval and VPR-based contemporary approaches (mentioned in 4.4.1).

More area the PR-curve covers, better the performance of the technique. Fig. 4.4 displays the AUC-PR curves for the St.Lucia datasets on the employed approaches. It is quite evident from Fig. 4.4 and Fig. 4.5 that for St.Lucia and Synthesized Nordland datasets, our proposed VPR approach has shown the best performance. Comparing from other datasets, St.Lucia exhibits moderate appearance change coupled with an appropriate viewpoint variation. A closer look at the results confirm that the system identified and captured context-based salient regions and boostup the overall retrieval performance, as illustrated in Fig. 4.7.

For St.Lucia dataset, Region-VLAD, Cross-Region-BoW and RMAC have shown similar performance as M-Region-VLAD. However, their performance degrades for SPEDTest (please see Fig. 4.6) and Synthesized Nordland as both the datasets experience strong seasonal and conditional changes. It suggests that under moderate conditions, all these regions-based techniques focus on place-recognition centric regions which results into better recognition performance. However, under severe conditional changes, employment of multiple convolutional layers found to be very productive. NetVLAD showcases nearly the similar PR-characteristic for St.lucia as Fixed Context and Context Flexible Attention. Better and comparable performance of M-Region-VLAD on all the dataset highlights the usefulness and generalization power of shallow attentions over deeply learned context flexible salient representations [6].

In Fig. 4.6, it is worth noticing that NetVLAD which underperformed under Synthesized Nordland, has shown state-of-the-art performance on SPEDTest. Although both the datasets

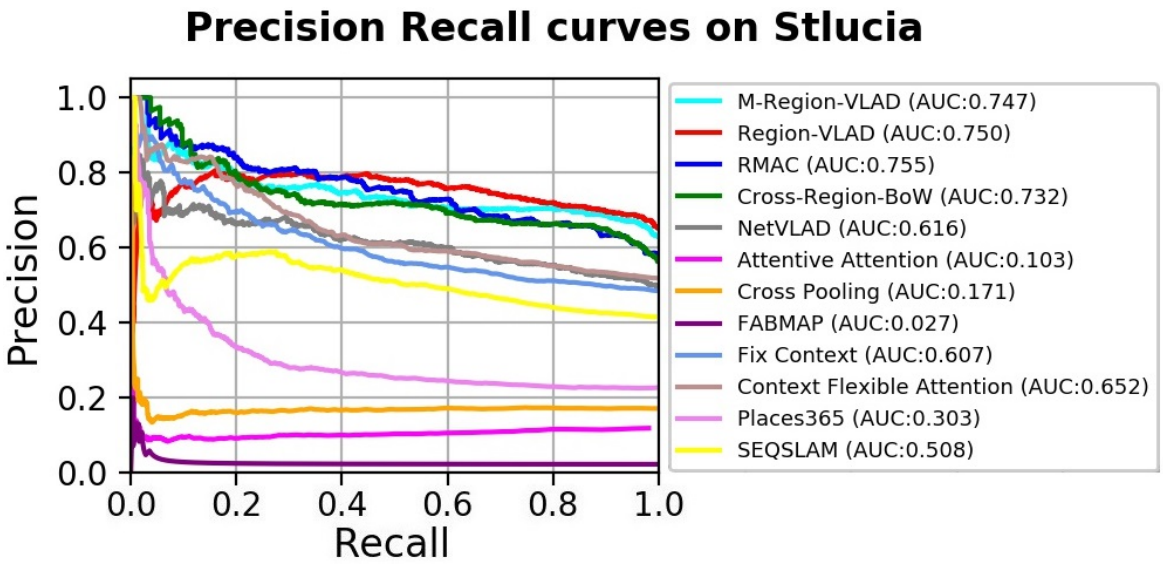

Fig. 4.4 Area under Precision-Recall curves for *St.Lucia* dataset on contemporary VPR techniques.

exhibit severe condition-variation among the traverses. One of the reason could be the existence of perceptual aliasing in Synthesized Nordland i.e. much resemblance among the sequentially captured frames. For SPEDTest, the environment of the test images is very diverse and each has only three matched images in the reference traverse. Majority of the techniques perform well on this dataset. In comparison, our proposed M-Region-VLAD achieves comparable AUC under PR curve against deep Context Flexible Attention, RMAC and Fixed Context frameworks. Cross-Region-BoW has shown an average performance both on Synthesized Nordland and SPEDTest. It is observed that due to ImageNet-centric training of VGG-16, the cross-convolutional regions-based approach concentrates more on objects. As expected, Region-VLAD which is integrated with AlexNet365 exhibits a comparable performance for Synthesized Nordland and SPEDTest. It is probably because the model is pre-trained on scene-centric Place365 dataset and with novel region finding approach, it sometime considers dynamic instances e.g. sky as a valuable region for distinguishing the scene which leads to place mismatch, also shown in Fig. 4.7.

SPEDTest dataset is a subset of SPED [4] but has not been used to train the models. A deep analysis suggests that although HybridNet [4] and Context Flexible Attention [6] models are fine-tuned on SPED dataset but training parameters such as, learning rates are kept different; dual learning rates approach was employed in [6]. The weight decay and iteration also differ from the values set for HyridNet and SPED-centric VGG-16. Moreover, employing three convolutional layers, the deep multi-scale features of Context Flexible Attention [6] can be more robust against condition-invariance and hence, exhibits better

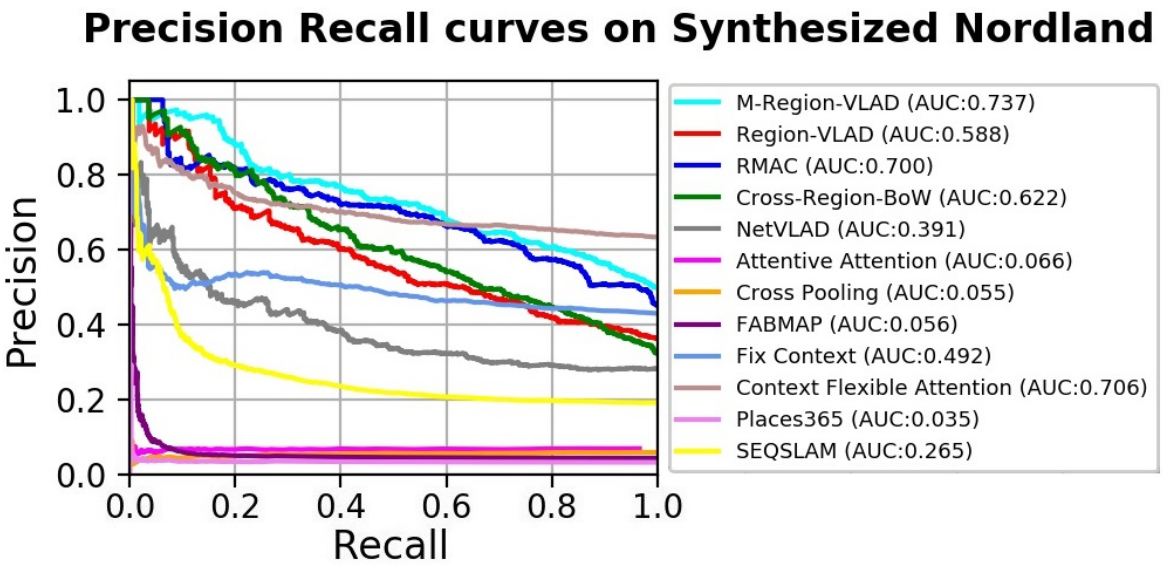

Fig. 4.5 Area under Precision-Recall curves for *Synthesized Nordland* dataset on contemporary VPR techniques.

performance for this datasets. However, when the places experience seasonal changes and perceptual aliasing (Synthesized Nordland), the performance degrades which indicates that the system is sensitive under such variations. It should be noted that our proposed M-Region-VLAD approach is employed only two convolutional layers of HybridNet and still delivers a comparable performances across all the datasets and mimics the generalization power at low computation and memory needs.

It is visible that the worst performance of FABMAP is consistent throughout the datasets. It is because FAPMAP used viewpoint-invariant SURF feature detector which is sensitive under condition and appearance changes. It is interesting that SEQSLAM with its better appearance tackling and whole image-based matching approach shown inferior performance under SPEDTest. It is probably due to the fact that the places exhibit diverse environment and sequence-based matching requirement is violated.

Cross-Pool and Attention Attentive approach, despite their better performances in other vision-based tasks under-performed in St.Lucia and Synthesized Nordland. This highlights the difference in other image retrieval/classification systems from place recognition where convolutional layers' responses are non-uniformly distributed and the place is subdivided into multiple contributing salient regions. However, their better performances under SPEDTest point towards the importance of CNN training. Fixed context and Places365 exhibit better results for St.Lucia and SPEDTest only. This implies that both the approaches are sensitive under perceptual aliasing experienced in Synthesized Nordland.

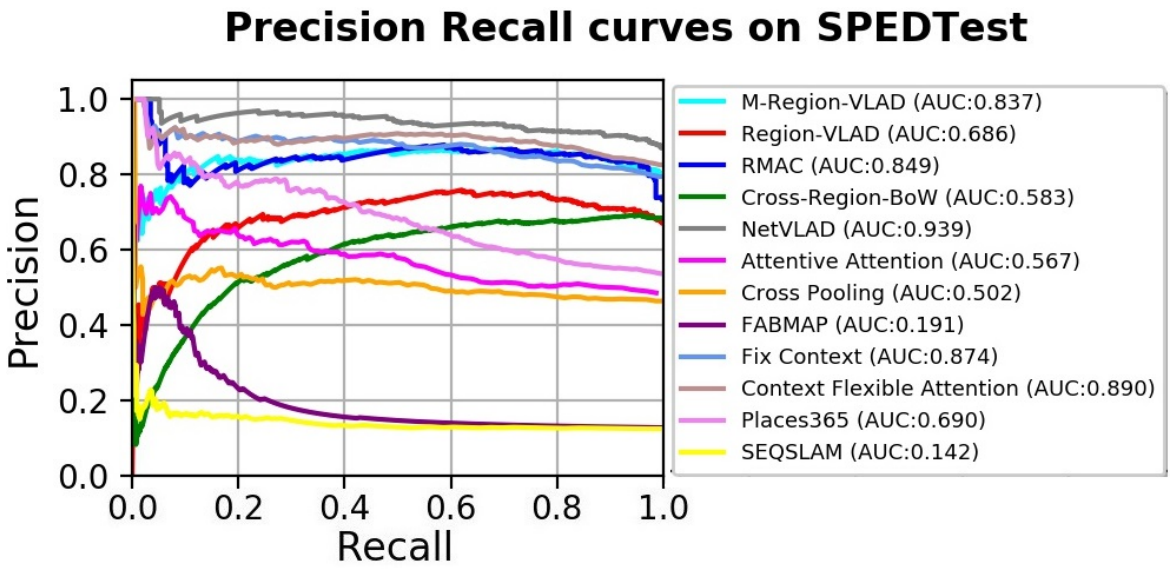

Fig. 4.6 Area under Precision-Recall curves for *SPEDTest* dataset on contemporary VPR techniques.

Furthermore, to analyse and differentiate the multi-semantic attentions captured by our proposed M-Region-VLAD framework against state-of-the-art Context Flexible Attention [6] and Region-VLAD [9], Fig. 4.7 shows some of the sample places with their corresponding salient regions. Both Context Flexible Attention and M-Region-VLAD emphasis upon most distinguishing structures, such as, houses, street lights while filtering out the confusing areas including clouds, vehicles etc. It is evident that Region-VLAD sometime includes sky and other dynamic instances as vital regions. It is worth noticing that M-Region-VLAD captures meaningful and place-centric spatial regions from a shallow CNN architecture against long-term condition and seasonal variations. Thus, it reduced down the overall memory and computational cost.

4.5 Summary

Despite the recent state-of-the-art performances of D-CNNs for VPR, the high computation and memory cost limit their practical deployment for battery-operated mobile robots. Achieving superior performance with shallow CNN architectures is thus desirable, but a challenging problem. In this chapter, a multi-scale context-aware attention approach is presented that combines salient regions from multiple convolutional layers of a place-recognition centric CNN architecture. The proposed approach captures persistent regional features under changing conditions and viewpoints while filtering down the confusing instances including sky,

moving objects etc. Evaluation on several challenging benchmark datasets confirms the dominance over state-of-the-art algorithms in terms of area under precision-recall curves.

In future, we will incorporate the proposed multi-scale attention block in a shallow feed forward neural network and fine-tune the CNN model on a large-scale place recognition dataset. It should reduce the feature encoding time and the system learns image regions invariant to strong viewpoint and condition variations.

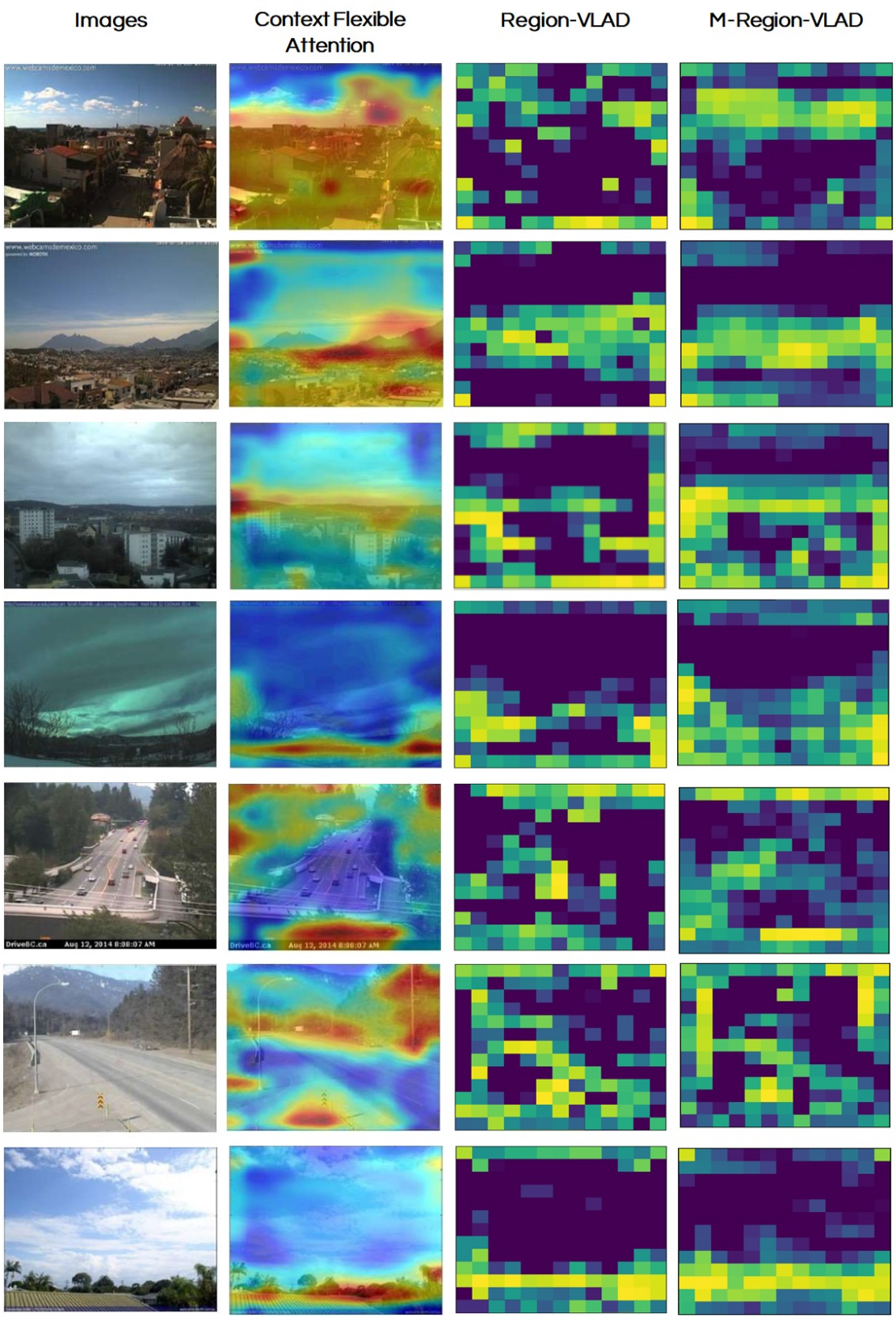

Fig. 4.7 Sample context-based regional attentions identified by Context Flexible Attention [6], Region-VLAD [9] and our proposed M-Region-VLAD.

Chapter 5

Conclusions and Future Directions

This thesis investigates the use of shallow neural networks for lightweight VPR under simultaneous condition and viewpoint variations. Outcome of this reputable research is directly affiliated with many real world applications including robot-centric agriculture devices, autonomous infrastructure inspection, environmental monitoring equipment with other transportation and security-based use cases.

This chapter summarizes the presented research work in this thesis with potential future direction. Section 5.1 outlines the research and contributions presented in this thesis. Section 5.2 presents the potential future plans and direction in the field of VPR.

5.1 Contributions Summary

Considering real world scenarios where a single place can suffer extreme visual changes triggered by the environmental transitions, such as, summer to winter and day to night changeovers coupled with viewpoint and weather variations. These uncertain circumstances altogether make the task of VPR extremely challenging. Chapter 1 starts with a background and provides an introduction to SLAM and VPR, followed-up with the research challenges and objectives of this dissertation. Chapter 2 discusses the methodologies employed for VPR either using hand-crafted feature descriptors or deep neural networks.

The first contribution in Chapter 3 presented a lightweight regions-based VPR technique. The scene-centric CNN-based regional features exhibit robustness against changes caused by different camera viewpoints coupled with environmental day-evening or day-night transitions. The proposed VPR framework has shown better place matching performance in terms of area under precision recall curves (AUC-PR curve) against the state-of-the-art hand-crafted and deep neural work based VPR approaches at low computation demand. Chapter 4 presented a multi-layer attention framework for improving place recognition under severe conditional

and moderate viewpoint changes. The proposed mult-scale approach is integrated with a shallow place recognition centric CNN model. The multi-scale context-aware attentions focus on persistent place-centric regions and filter out the dynamic and confusing objects, such as, vehicles and pedestrians. Evaluation on benchmark place recognition datasets exhibits superior AUC-PR curve against the contemporary deep neural network based VPR and other vision-based image retrieval approaches at low memory and resource utilization.

5.2 Future Directions

This thesis has shown that how performance of a VPR system can be improved by employing the proposed frameworks (presented in Chapter 3 and Chapter 4) at low computation demand. Despite the fast ongoing and encouraging growth in the field of robotic vision, there is still room for improvements as explored in this subsection. Overall, following are the proposed futuristic ideas for improving the performance of VPR:

1. Fusion of multi-model CNN features for tackling visual changes resulting from viewpoint, seasonal and illumination variations coupled with dynamic objects in appearance-aware place recognition.

2. Fine-tuning of viewpoint invariant AlexNet365 (pretrained on Places365) on SPED dataset with an addition of a convolutional block that learns and captures context-aware attentions robust under uncertain visual variations.

3. Evaluate the performance of place recognition system explicitly as a function of visual changes severity rather than the size of benchmark datasets.

4. Training CNN models for place recognition by exploiting high-level semantic information either by employing object detection or image segmentation techniques to improve the end-to-end attention-based place recognition system.

5. Visual Place Recognition using GANs and Capsule Networks

CNN models trained for place recognition are either condition-invariant (HybridNet) or viewpoint-invariant (NetVLAD). The authors have evaluated their proposed systems employing suitable benchmark datasets exhibiting strong conditional or viewpoint variations. The use of deep neural networks improves the matching performance but increases the retrieval time which is impractical for many real world robotic applications. The proposed VPR framework in Chapter 3 employed novel region finding approach on a shallow scene-centric AlexNet365 for VPR. Evaluation was performed with results claiming better performance

under strong viewpoint variation along with moderate conditional changes. However, the system suffers with the inclusion of dynamic instances at higher regional configuration. Place365 dataset contains collection of scenes, such as, bars, restaurants, offices, thus, the system taking precedence of the scene-centric CNN training manages to find the common regions of the same place from different viewpoints. The proposed multi-semantic attention framework in Chapter 4 can be added as a block within scene-centric AlexNet365 and fine-tuned on SPED dataset. The original scene-invariant convolutional layers will learn seasonal/illumination invariance, thus, both the severe viewpoint and appearance variations can be tackled simultaneously at minimum resource utilization. Secondly, multi-model approach is another area of research which has not been touched for place recognition; fusion of rich condition-invariant HybridNet features with rich viewpoint-invariant AlexNet365 features can improve the performance of place recognition under changing viewpoint and environment.

Evaluation of 10 SOTA contemporary VPR approaches, discussed in Chapter 2 showed that no universal VPR technique exists. The results confirm that neural network based NetVLAD and region-based approaches (Cross-Region-BoW, RMAC, Region-VLAD and so on) worked well for VPR under severe viewpoint, seasonal and illumination changes. However, the best matching performance comes at higher computational cost like for NetVLAD, the feature extraction time is quite high in comparison to regions-based CNN methods which consume more memory at run-time. Similar results have been observed in another research work (presented in Chapter 2) within which ground-based VPR techniques are evaluated on dataset with 6-DOF viewpoint variation. The results claim deployment trade-off between matching performance and resource utilization. Since, NetVLAD is pre-trained on urban place-centric dataset (Pittsburgh 30K) which exhibits strong conditional and viewpoint variations coupled with dynamic and confusing instances. This is in contrast to the training datasets of Imagenet-centric VGG-16 and SPED-centric HybridNet. ImageNet is an object detection dataset and is intrinsically not good for place recognition, while SPED does not contain dynamic objects observed in urban road scenes. Region-based CNN approaches integrated on diversely trained CNNs can be a worthwhile research direction for place recognition.

The VPR research community either employed middle convolutional layers for severe appearance changes and late convolutional layers for viewpoint changes. The standardized widely used benchmark place-recognition datasets do not provide any information or metric to determine the severity of changes experienced in day-night and summer-winter transitions other than to observe the datasets visually to determine the amount of conditional or viewpoint variation. Therefore, another worth while research idea is to introduce such datasets and

then evaluate the system providing quantitative information on how strong the viewpoint and conditional changes it can handle in average. Recent works suggested that the employment of multiple convolutional layers as feature representation can help in focusing on persistent instances and filtering of confusing objects. Features from lower convolutional layers of the CNN models generally respond to low level image features, such as corners or edges, thus, efficient against conditional variations. While features from higher layers focus on structures that are more semantically meaningful (such as, human faces or buildings) and robust against viewpoint changes. Training of the CNN model employing semantically meaningful late convolutional layers coupled with object detection and segmentation based approaches can improve the robustness of regions-based place recognition techniques.

The employment of Generative Adversarial Networks (GANs) for VPR suggested the possibility of generating multiple appearance of the place under different environment. Experimentations have shown that the encoded features of visually similar images are very closer in the descriptor domain. However, storing images' translation in various conditions is impractical, thus, focusing on the semantic information using GANs can be a potential research direction for VPR. Moreover, the GANs-based semantic maps can be used for generating different pose of the same place catering viewpoint invariance efficiently. Recently, for unmanned aerial vehicle (UAV), Capsule Networks (CapsNet) have shown performance improvement over state-of-the-art CNNs in depth estimation-based SLAM problems. Moreover, ground-to-aerial cross-view image geo-localization also encouraged the use of Capsule Network, named GeoCapsNet. It's a two-branch Siamese network-based architecture and takes a pair of cross-view images as an input. With soft-margin triplet loss in the capsule layer, the architecture captures high-level semantic information from ResNetX and capsule layers. It would be interesting to investigate the applicability of CapsNets for VPR-based SLAM tasks.

References

[1] T. Nam, J. Shim, and Y. Cho, "A 2.5 D map-based mobile robot localization via cooperation of aerial and ground robots," *Sensors*, vol. 17, no. 12, p. 2730, 2017.

[2] S. Lowry, N. Sünderhauf, P. Newman, *et al.*, "Visual place recognition: A survey," *IEEE Transactions on Robotics*, vol. 32, no. 1, pp. 1–19, 2016.

[3] R. Arandjelovic *et al.*, "NetVLAD: CNN architecture for weakly supervised place recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 5297–5307.

[4] Z. Chen, A. Jacobson, N. Sünderhauf, B. Upcroft, L. Liu, C. Shen, I. Reid, and M. Milford, "Deep learning features at scale for visual place recognition," in *IEEE International Conference on Robotics and Automation*, 2017, pp. 3223–3230.

[5] Z. Chen, F. Maffra, I. Sa, M. Chli, and other, "Only look once, mining distinctive landmarks from convnet for visual place recognition," in *IEEE International Conference on Intelligent Robots and Systems*, 2017, pp. 9–16.

[6] Z. Chen, L. Liu, I. Sa, Z. Ge, and M. Chli, "Learning context flexible attention model for long-term visual place recognition," *IEEE Robotics and Automation Letters*, vol. 3, no. 4, pp. 4015–4022, 2018.

[7] M. Zaffar, A. Khaliq, S. Ehsan, M. Milford, K. Alexis, and K. McDonald-Maier, "Are state-of-the-art visual place recognition techniques any good for aerial robotics?" *arXiv preprint arXiv:1904.07967*, 2019.

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

[9] A. Khaliq, S. Ehsan, Z. Chen, M. Milford, and K. McDonald-Maier, "A holistic visual place recognition approach using lightweight cnns for significant viewpoint and appearance changes," *IEEE Transactions on Robotics*, pp. 1–9, 2019.

[10] G. Tolias, R. Sicre, and H. Jégou, "Particular object retrieval with integral max-pooling of cnn activations," *Proc. International Conference on Learning Representations*, 2016.

[11] S. Lowry, N. Sünderhauf, P. Newman, *et al.*, "Visual place recognition: A survey," *IEEE Transactions on Robotics*, vol. 32, no. 1, pp. 1–19, 2016.

[12] R. Bishop, *Intelligent vehicle technology and trends*, 2005.

[13] S. Chen, L. Huang, J. Bai, H. Jiang, and L. Chang, "Multi-sensor information fusion algorithm with central level architecture for intelligent vehicle environmental perception system," SAE Technical Paper, Tech. Rep., 2016.

[14] M. Milford and G. Wyeth, "Persistent navigation and mapping using a biologically inspired slam system," *The International Journal of Robotics Research*, vol. 29, no. 9, pp. 1131–1153, 2010.

[15] J. Biswas and M. M. Veloso, "Localization and navigation of the cobots over long-term deployments," *The International Journal of Robotics Research*, vol. 32, no. 14, pp. 1679–1694, 2013.

[16] S. Thrun *et al.*, "Robotic mapping: A survey," *Exploring artificial intelligence in the new millennium*, vol. 1, no. 1-35, p. 1, 2002.

[17] H. Moravec and A. Elfes, "High resolution maps from wide angle sonar," in *Proc. IEEE International Conference on Robotics and Automation*, vol. 2, 1985, pp. 116–121.

[18] O. Khatib, "Real-time obstacle avoidance for manipulators and mobile robots," in *Autonomous Robot Vehicles*. Springer, 1986, pp. 396–404.

[19] K. Konolige, J. Bowman, J. Chen, P. Mihelich, M. Calonder, V. Lepetit, and P. Fua, "View-based maps," *International Journal of Robotics Research*, vol. 29, no. 8, pp. 941–957, 2010.

[20] A. Diosi and L. Kleeman, "Fast laser scan matching using polar coordinates," *International Journal of Robotics Research*, vol. 26, no. 10, pp. 1125–1153, 2007.

[21] M. Douze, H. Jégou, H. Sandhawalia, L. Amsaleg, and C. Schmid, "Evaluation of gist descriptors for web-scale image search," in *Proc. International Conference on Image and Video Retrieval*. ACM, 2009, p. 19.

[22] M. J. Milford, I. Turner, and P. Corke, "Long exposure localization in darkness using consumer cameras," in *IEEE International Conference on Robotics and Automation*, 2013, pp. 3755–3761.

[23] P. Neubert, N. Sünderhauf, and P. Protzel, "Appearance change prediction for long-term navigation across seasons," in *IEEE European Conference on Mobile Robots*, 2013, pp. 198–203.

[24] A. Ranganathan, S. Matsumoto, and D. Ilstrup, "Towards illumination invariance for visual localization," in *IEEE International Conference on Robotics and Automation*, 2013, pp. 3791–3798.

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

[26] M. J. Milford, G. F. Wyeth, and D. Prasser, "Ratslam: a hippocampal model for simultaneous localization and mapping," in *Proc. IEEE International Conference on Robotics and Automation*, vol. 1, 2004, pp. 403–408.

[27] M. Cummins and P. Newman, "FAB-MAP: Probabilistic localization and mapping in the space of appearance," *International Journal of Robotics Research*, vol. 27, no. 6, pp. 647–665, 2008.

[28] E. Johns and G.-Z. Yang, "Dynamic scene models for incremental, long-term, appearance-based localisation," in *IEEE International Conference on Robotics and Automation*, 2013, pp. 2731–2736.

[29] B. Kuipers, "The spatial semantic hierarchy," *Artificial intelligence*, vol. 119, no. 1-2, pp. 191–233, 2000.

[30] E. Garcia-Fidalgo and A. Ortiz, "Vision-based topological mapping and localization methods: A survey," *Robotics and Autonomous Systems*, vol. 64, pp. 1–20, 2015.

[31] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[32] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," in *Proc. European Conference on Computer Vision*, 2006, pp. 404–417.

[33] B. Zhou, A. Lapedriza, *et al.*, "Places: A 10 million image database for scene recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[34] Z. Chen, A. Jacobson, U. M. Erdem, M. E. Hasselmo, and M. Milford, "Multi-scale bio-inspired place recognition," in *IEEE International Conference on Robotics and Automation*, 2014, pp. 1895–1901.

[35] J. Collier, S. Se, and V. Kotamraju, "Multi-sensor appearance-based place recognition," in *IEEE International Conference on Computer and Robot Vision*, 2013, pp. 128–135.

[36] P. Newman, D. Cole, and K. Ho, "Outdoor slam using visual appearance and laser ranging," in *IEEE International Conference on Robotics and Automation*, 2006, pp. 1180–1187.

[37] I. Ulrich and I. Nourbakhsh, "Appearance-based place recognition for topological localization," in *Proc. IEEE International Conference on Robotics and Automation Symposia*, vol. 2, 2000, pp. 1023–1029.

[38] E. Johns and G.-Z. Yang, "Feature co-occurrence maps: Appearance-based localisation throughout the day," in *IEEE International Conference on Robotics and Automation*, 2013, pp. 3212–3218.

[39] N. Sünderhauf, P. Neubert, and P. Protzel, "Are we there yet? challenging seqslam on a 3000 km journey across all four seasons," in *Proc. IEEE International Conference on Robotics and Automation*, 2013.

[40] N. Sünderhauf, S. Shirazi, A. Jacobson, F. Dayoub, E. Pepperell, B. Upcroft, *et al.*, "Place recognition with convnet landmarks: Viewpoint-robust, condition-robust, training-free," *Proc. Robotics: Science and Systems Conference*, 2015.

[41] G. Schindler, M. Brown, and R. Szeliski, "City-scale location recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*. Citeseer, 2007, pp. 1–7.

[42] M. J. Milford and G. F. Wyeth, "SeqSLAM: Visual route-based navigation for sunny summer days and stormy winter nights," in *IEEE International Conference on Robotics and Automation*, 2012, pp. 1643–1649.

[43] S. Brahnam, L. C. Jain, L. Nanni, A. Lumini, *et al.*, *Local binary patterns: new variants and applications*. Springer, 2014.

[44] X. Yang, C. Huang, and K.-T. T. Cheng, "libldb: A library for extracting ultrafast and distinctive binary feature description," in *Proc. International Conference on Multimedia*. ACM, 2014, pp. 671–674.

[45] J. Gaspar, N. Winters, and J. 'Santos-Victor, "Vision-based navigation and environmental representations with an omnidirectional camera," *IEEE Transactions on robotics and automation*, vol. 16, no. 6, pp. 890–898, 2000.

[46] A. Gil, O. Reinoso, O. M. Mozos, C. Stachniss, and W. Burgard, "Improving data association in vision-based slam," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2006, pp. 2076–2081.

[47] T. Ojala, M. Pietikäinen, and T. Mäenpää, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, no. 7, pp. 971–987, 2002.

[48] M. Calonder, V. Lepetit, C. Strecha, and P. Fua, "Brief: Binary robust independent elementary features," in *European Conference on Computer Vision*. Springer, 2010, pp. 778–792.

[49] E. Rublee, V. Rabaud, K. Konolige, and G. R. Bradski, "Orb: An efficient alternative to sift or surf." in *IEEE International Conference on Computer Vision*, vol. 11, no. 1. Citeseer, 2011, p. 2.

[50] S. Leutenegger, M. Chli, and R. Siegwart, "Brisk: Binary robust invariant scalable keypoints," in *IEEE International Conference on Computer Vision*, 2011, pp. 2548–2555.

[51] R. Arroyo, P. F. Alcantarilla, L. M. Bergasa, J. J. Yebes, and S. Bronte, "Fast and effective visual place recognition using binary codes and disparity information," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2014, pp. 3089–3094.

[52] P. F. Alcantarilla, A. Bartoli, and A. J. Davison, "Kaze features," in *European Conference on Computer Vision*. Springer, 2012, pp. 214–227.

[53] A. Alahi, R. Ortiz, and P. Vandergheynst, "Freak: Fast retina keypoint," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 510–517.

[54] J. Sivic *et al.*, "Video google: A text retrieval approach to object matching in videos," in *IEEE International Conference on Computer Vision*, 2003, p. 1470.

[55] A. Glover, E. Pepperell, G. Wyeth, B. Upcroft, and M. Milford, "Repeatable condition-invariant visual odometry for sequence-based place recognition," in *Proc. Australasian Conference on Robotics and Automation*, 2015.

[56] J. Kosecka, L. Zhou, P. Barber, and Z. Duric, "Qualitative image based localization in indoors environments," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2, 2003, pp. II–II.

[57] N. Winters, J. Gaspar, G. Lacey, and J. Santos-Victor, "Omni-directional vision for robot navigation," in *Proc. IEEE Omnidirectional Vision Workshop*, 2000, pp. 21–28.

[58] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *International Journal of Computer Vision*, vol. 42, no. 3, pp. 145–175, 2001.

[59] N. Sünderhauf and P. Protzel, "Brief-gist-closing the loop by simple means," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2011, pp. 1234–1241.

[60] P. Lamon, I. Nourbakhsh, B. Jensen, and R. Siegwart, "Deriving and matching image fingerprint sequences for mobile robot localization," in *Proc. IEEE International Conference on Robotics and Automation*, vol. 2, 2001, pp. 1609–1614.

[61] H. Badino, D. Huber, and T. Kanade, "Real-time topometric localization," in *IEEE International Conference on Robotics and Automation*, 2012, pp. 1635–1642.

[62] H. Lategahn, J. Beck, B. Kitt, and C. Stiller, "How to learn an illumination robust image feature for place recognition," in *IEEE Intelligent Vehicles Symposium*, 2013, pp. 285–291.

[63] A. Murillo, C. Sagüés, J. J. Guerrero, T. Goedemé, T. Tuytelaars, and L. Van Gool, "From omnidirectional images to hierarchical localization," *Robotics and Autonomous Systems*, vol. 55, no. 5, pp. 372–382, 2007.

[64] T. Goedemé, M. Nuttin, T. Tuytelaars, and L. Van Gool, "Markerless computer vision based localization using automatically generated topological maps," in *European Navigation Conference GNSS, Rotterdam*, 2004, pp. 219–236.

[65] H. Lu, X. Li, H. Zhang, and Z. Zheng, "Robust place recognition based on omnidirectional vision and real-time local visual features for mobile robots," *Advanced Robotics*, vol. 27, no. 18, pp. 1439–1453, 2013.

[66] M.-L. Wang and H.-Y. Lin, "A hull census transform for scene change detection and recognition towards topological map building," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2010, pp. 548–553.

[67] J. Wang and Y. Yagi, "Robust location recognition based on efficient feature integration," in *IEEE International Conference on Robotics and Biomimetics*, 2012, pp. 97–101.

[68] A. J. Davison, I. D. Reid, N. D. Molton, and O. Stasse, "Monoslam: Real-time single camera slam," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, no. 6, pp. 1052–1067, 2007.

[69] D. Caruso, J. Engel, and D. Cremers, "Large-scale direct slam for omnidirectional cameras," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2015, pp. 141–148.

[70] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardos, "ORB-SLAM: a versatile and accurate monocular slam system," *IEEE Transactions on Robotics*, vol. 31, no. 5, pp. 1147–1163, 2015.

[71] C. Cadena, D. Gálvez-López, J. D. Tardós, and J. Neira, "Robust place recognition with stereo sequences," *IEEE Transactions on Robotics*, vol. 28, no. 4, pp. 871–885, 2012.

[72] H. Morioka, S. Yi, and O. Hasegawa, "Vision-based mobile robot's slam and navigation in crowded environments," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2011, pp. 3998–4005.

[73] T. Fiolka, J. Stückler, D. A. Klein, D. Schulz, and S. Behnke, "Distinctive 3d surface entropy features for place recognition," in *IEEE European Conference on Mobile Robots*, 2013, pp. 204–209.

[74] W. Maddern, M. Milford, and G. Wyeth, "CAT-SLAM: probabilistic localisation and mapping using a continuous appearance-based trajectory," *International Journal of Robotics Research*, vol. 31, no. 4, pp. 429–451, 2012.

[75] W. Maddern, M. Milford, *et al.*, "Towards persistent indoor appearance-based localization, mapping and navigation using CAT-Graph," pp. 4224–4230, 2012.

[76] C. Cadena, D. Gálvez-López, F. Ramos, J. D. Tardós, and J. Neira, "Robust place recognition with stereo cameras," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2010, pp. 5182–5189.

[77] F. Endres, J. Hess, N. Engelhard, J. Sturm, D. Cremers, and W. Burgard, "An evaluation of the rgb-d slam system." in *IEEE International Conference on Robotics and Automation*, vol. 3, no. c, 2012, pp. 1691–1696.

[78] R. Finman, L. Paull, and J. J. Leonard, "Toward object-based place recognition in dense rgb-d maps," in *IEEE International Conference on Robotics and Automation Workshop Visual Place Recognition in Changing Environments, Seattle, WA*, 2015.

[79] A. Sharif Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "Cnn features off-the-shelf: an astounding baseline for recognition," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2014, pp. 806–813.

[80] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural computation*, vol. 1, no. 4, pp. 541–551, 1989.

[81] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, "Residual attention network for image classification," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3156–3164.

[82] P. Wang, L. Liu, C. Shen, Z. Huang, A. van den Hengel, and H. Tao Shen, "Multi-attention network for one shot learning," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2721–2729.

[83] R. Hecht-Nielsen, "Theory of the backpropagation neural network," in *Neural networks for perception*. Elsevier, 1992, pp. 65–93.

[84] Z. Chen, O. Lam, A. Jacobson, and M. Milford, "Convolutional neural network-based place recognition," *Australasian Conference on Robotics and Automation*, 2014.

[85] P. Panphattarasap and A. Calway, "Visual place recognition using landmark distribution descriptors," in *Asian Conference on Computer Vision*. Springer, 2016, pp. 487–502.

[86] N. Sünderhauf, S. Shirazi, F. Dayoub, B. Upcroft, and M. Milford, "On the performance of convnet features for place recognition," in *IEEE International Conference on Intelligent Robots and Systems*, 2015, pp. 4297–4304.

[87] R. Gomez-Ojeda, M. Lopez-Antequera, N. Petkov, and J. Gonzalez-Jimenez, "Training a convolutional neural network for appearance-invariant place recognition," *arXiv preprint arXiv:1505.07428*, 2015.

[88] A. Babenko and V. Lempitsky, "Aggregating local deep features for image retrieval," in *IEEE International Conference on Computer Vision*, 2015, pp. 1269–1277.

[89] M. Jaderberg, K. Simonyan, *et al.*, "Spatial transformer networks," in *Annual Conference on Neural Information Processing Systems*, 2015, pp. 2017–2025.

[90] L. Liu, C. Shen, and A. van den Hengel, "Cross-convolutional-layer pooling for image recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 11, pp. 2305–2313, 2017.

[91] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Annual Conference in Neural Information Processing Systems*, 2012, pp. 1097–1105.

[92] L. Liu, C. Shen, and A. van den Hengel, "The treasure beneath convolutional layers: Cross-convolutional-layer pooling for image classification," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 4749–4757.

[93] X. Chu, W. Yang, W. Ouyang, C. Ma, A. L. Yuille, and X. Wang, "Multi-context attention for human pose estimation," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1831–1840.

[94] H. J. Kim, E. Dunn, and J.-M. Frahm, "Learned contextual feature reweighting for image geo-localization," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2136–2145.

[95] L. Zheng, Y. Yang, and Q. Tian, "SIFT meets CNN: A decade survey of instance retrieval," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 5, pp. 1224–1244, 2018.

[96] J. Yue-Hei Ng, F. Yang, and L. S. Davis, "Exploiting local features from deep networks for image retrieval," in *IEEE Conference on Computer Vision and Pattern Recognition Workshop*, 2015, pp. 53–61.

[97] A. F. M. Agarap, "A neural network architecture combining gated recurrent unit and support vector machine for intrusion detection in network traffic data," in *Proc. International Conference on Machine Learning and Computing*, 2018, pp. 26–30.

[98] J. Sánchez *et al.*, "Image classification with the fisher vector: Theory and practice," *International Journal of Computer Vision*, vol. 105, no. 3, pp. 222–245, 2013.

[99] H. Jégou, M. Douze, C. Schmid, and P. Pérez, "Aggregating local descriptors into a compact image representation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2010, pp. 3304–3311.

[100] H. Jin Kim, E. Dunn, and J.-M. Frahm, "Predicting good features for image geo-localization using per-bundle vlad," in *IEEE International Conference on Computer Vision*, 2015, pp. 1170–1178.

[101] J. Matas, O. Chum, M. Urban, and T. Pajdla, "Robust wide-baseline stereo from maximally stable extremal regions," *Image and Vision Computing*, vol. 22, no. 10, pp. 761–767, 2004.

[102] T. Sattler, A. Torii, J. Sivic, M. Pollefeys, and other, "Are large-scale 3d models really necessary for accurate visual localization?" in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1637–1646.

[103] N. Merrill and G. Huang, "Lightweight unsupervised deep loop closure," *Robotics: Science and Systems*, June 2018.

[104] M. Teichmann, A. Araujo, M. Zhu, and J. Sim, "Detect-to-retrieve: Efficient regional aggregation for image search," *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[105] A. S. Razavian *et al.*, "Visual instance retrieval with deep convolutional networks," *ITE Transactions on Media Technology and Applications*, vol. 4, no. 3, pp. 251–258, 2016.

[106] G. Tolias, Y. Avrithis, and H. Jégou, "Image search with selective match kernels: aggregation across single and multiple images," *International Journal of Computer Vision*, vol. 116, no. 3, pp. 247–261, 2016.

[107] R. Tao *et al.*, "Locality in generic instance search from one example," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 2091–2098.

[108] J. Knopp, J. Sivic, and T. Pajdla, "Avoiding confusing features in place recognition," in *European Conference on Computer Vision*. Springer, 2010, pp. 748–761.

[109] M. Zaffar, A. Khaliq, S. Ehsan, M. Milford, and K. McDonald-Maier, "Levelling the playing field: A comprehensive comparison of visual place recognition approaches under changing conditions," *arXiv preprint arXiv:1903.09107*, 2019.

[110] F. Maffra, Z. Chen, and M. Chli, "Viewpoint-tolerant place recognition combining 2d and 3d information for uav navigation," in *IEEE International Conference on Robotics and Automation*, 2018.

[111] "Mapillary platform," https://www.mapillary.com/, 2013.

[112] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve." *Radiology*, vol. 143, no. 1, pp. 29–36, 1982.

[113] J. Davis *et al.*, "The relationship between precision-recall and roc curves," in *International Conference on Machine Learning*, 2006, pp. 233–240.

[114] T. Fawcett, "Roc graphs: Notes and practical considerations for researchers," *Machine learning*, vol. 31, no. 1, pp. 1–38, 2004.

[115] T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets," *PloS one*, vol. 10, no. 3, p. e0118432, 2015.

[116] R. Arandjelovic and A. Zisserman, "All about VLAD," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 1578–1585.

[117] H. Jégou, M. Douze, *et al.*, "On the burstiness of visual elements," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 1169–1176.

[118] T.-T. Do, T. Hoang, D.-K. L. Tan, H. Le, T. V. Nguyen, and N.-M. Cheung, "From selective deep convolutional features to compact binary representations for image retrieval," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 15, no. 2, 2019.

[119] C. Cadena, L. Carlone, H. Carrillo, *et al.*, "Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age," *IEEE Transactions on Robotics*, vol. 32, no. 6, pp. 1309–1332, 2016.

[120] "Cross-Region-BoW source code," https://github.com/scutzetao/IROS2017_OnlyLookOnce.

[121] "Results and datasets," https://github.com/Ahmedest61/CNN-Region-VLAD-VPR/.

[122] D. Yu, J. Fu, T. Mei, and Y. Rui, "Multi-level attention networks for visual question answering," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4187–4195.

[123] M. Cummins, "Highly scalable appearance-only slam-fab-map 2.0," *Proc. Robotics: Sciences and Systems (RSS), 2009*, 2009.