Visual Place Recognition for Autonomous Robots

Mubariz Zaffar

School of Computer Science and Electronic Engineering
University of Essex

The thesis is submitted for the degree of
Master by Dissertation

October 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.

Mubariz Zaffar
October 2020

Acknowledgements

First and foremost, I would like to thank Allah for giving me the strength, ability and opportunity to undertake this study and complete it successfully. Without Allah's blessings, it would not have been possible to conduct this extensive research work. Over the past 25 years of my life, Allah has taken me through an immensely steady and an incrementally soaring journey. Some times it gets difficult to understand and inter-link how I could have even made it from where I started to where I am now, without having those small incremental pushes, supports and blessings along the way. I am excited to see what's in store for the next (potentially) 25 years of my life, but I do wish/pray/hope that these next years also enable me to significantly uplift, motivate, help and support all those around me, all those who I wanted to but I couldn't back then and all those who may ever need that in any capacity.

I am thankful to my supervisors Dr. Klaus McDonald-Maier, Dr. Shoaib Ehsan and my collaborator Dr. Michael Milford, who have provided me with excellent guidance and feedback on my research, that has led today to this thesis. I hold immense respect, gratitude and admiration for these individuals in both personal and professional capacity and I wish them the best for their futures. I am immensely thankful to my MSD committee, who have supported and encouraged me during my MSD. I have had a number of interesting, lengthy and fruitful discussions with Dr. Shoaib over the past 2 years about Visual Place Recognition (VPR), and we have worked very hard to establish our group's name among the VPR community. I wish him the best with his future endeavours and I hope that this group soars to new heights and joins the league of some of the top-most robotic vision groups in the world.

I am thankful to my parents, my siblings, my cousins in the UK and my extended family who have supported me through-out and who have given me all the love, affection and motivation to continue pursuing my ambitions. To enable me to get to this stage, my family has given me almost everything that they could, irrespective of the challenges that it may have caused them. I will continue to make you proud and I hope that I have the strength, abilities and character to continue to do so.

I am thankful to all my teachers, mentors, managers and supervisors over the past many years, whose continued efforts enabled me to get to this stage. I am thankful to my tuition

teacher who tutored me for 12 years and who has probably played the most significant role in the early years of my life. I hold sincerest respect for having the opportunity to work directly with Dr. Farrukh Kamran (CTO at Skyelectric), from whom I have learned the ability to approach any problem and the confidence to undertake any research/development task. In his words (to which I completely agree): An engineer and a researcher should never be afraid of undertaking new challenges, that is what differentiates him from the rest of the crowd.

My colleagues at Essex and my friends have been wonderful to me and have stood by me at all times, to whom I extend my sincerest affection and I hope that we continue to remain in contact. I am thankful to my closest, sincerest and brother-like friend Ahmad Khaliq. We have had a wonderful time together at Essex and the past 2 years have been amazing. I wish him the best for his future and I hope that we get the chance to work together again in future. My friends and colleagues in our lab, including but not limited to Zeba Khanam, Sangeet Saha, Bilal Aslam, Arun, Eduardo Watcher, Server Kasap and many other friends made these 2 years bearable, enjoyable and a happy time to look back to.

I am thankful to the UK-EPSRC, UKRI, UK-National Center for Nuclear Robotics, who have financially supported me in my research journey. There is immense talent and commitment among the youngsters of my country, but they are limited by the lack of financial support, as was I prior to 2018. These funding agencies have enabled me to explore my research ambitions without worrying about the financial aspects and I hope that more and more people from my country get the chance to pursue their research dreams without having the financial weight on their shoulders. I am thankful to the University of Essex and to the School of Computer Science and Electronics Engineering for providing an inclusive, hospitable and exciting environment for me to undertake my Master's degree.

Abstract

Autonomous robotics has been the subject of great interest within the research community over the past few decades. Its applications are wide-spread, ranging from health-care to manufacturing, goods transportation to home deliveries, site-maintenance to construction, planetary explorations to rescue operations and many others, including but not limited to agriculture, defence, commerce, leisure and extreme environments. At the core of robot autonomy lies the problem of localisation, i.e, knowing where it is and within the robotics community, this problem is termed as place recognition. Place recognition using only visual input is termed as Visual Place Recognition (VPR) and refers to the ability of an autonomous system to recall a previously visited place using only visual input, under changing viewpoint, illumination and seasonal conditions, and given computational and storage constraints.

This thesis is a collection of 4 inter-linked, mutually-relevant but branching-out topics within VPR: 1) What makes a place/image worthy for VPR?, 2) How to define a state-of-the-art in VPR?, 3) Do VPR techniques designed for ground-based platforms extend to aerial platforms? and 4) Can a handcrafted VPR technique outperform deep-learning-based VPR techniques? Each of these questions is a dedicated, peer-reviewed chapter in this thesis and the author attempts to answer these questions to the best of his abilities.

The worthiness of a place essentially refers to the salience and distinctiveness of the content in the image of this place. This salience is modelled as a framework, namely memorable-maps, comprising of 3 conjoint criteria: a) Human-memorability of an image, 2) Staticity and 3) Information content. Because a large number of VPR techniques have been proposed over the past 10-15 years, and due to the variation of employed VPR datasets and metrics for evaluation, the correct state-of-the-art remains ambiguous. The author levels this playing field by deploying 10 contemporary techniques on a common platform and use the most challenging VPR datasets to provide a holistic performance comparison. This platform is then extended to aerial place recognition datasets to answer the 3rd question above. Finally, the author designs a novel, handcrafted, compute-efficient and training-free VPR technique that outperforms state-of-the-art VPR techniques on 5 different VPR datasets.

Contents

List of Figures

List of Tables

Abbreviations

AUC	*Area Under the Curve*
COVID	*Corona Virus Disease*
DARPA	*Defense Advanced Research Projects Agency*
DOF	*Degrees Of Freedom*
GPS	*Global Positioning System*
LiDAR	*Light Detection and Ranging*
PR Curves	*Precision Recall Curves*
RGB	*Red Green Blue*
RGB – D	*Red Green Blue Depth*
SAD	*Sum of Absolute Differences*
SLAM	*Simultaneous Localization and Mapping*
VPR	*Visual Place Recognition*

Chapter 1

Introduction

In the words of Nicola Tesla, “In the twenty-first century, the robot will take the place which slave labor occupied in ancient civilization.” The realisation of this quote has crucial implications for the survival, evolution, comfort and collective growth of the human civilisation. Robot autonomy is not a dream anymore, but in-fact, we now see real-world examples of autonomous robotics in the forms of self-driving cars for example. Robot autonomy has now become even more critical in today’s world and the COVID-19 pandemic has proven this much needed revolution to transform the health-care, transportation, commerce, agriculture and many other businesses and industries.

To achieve any task autonomously, a mobile robot equipped with some limited sensors and actuators, needs to be able to plan its motion, execute and control this motion, localise itself while performing this motion and potentially perform any required manipulation, if required. Within robotics these are the major domains of research, usually referred to as, motion planning, robot control, robot localisation and manipulation, respectively. All of these research domains then branch-out based on the type of sensory input or the combination of sensory inputs. Researchers working within robotic vision investigate these domains based on visual information as the primary sensing modality. The author’s area of interest and the scope of this thesis is localisation given only visual input.

This chapter introduces the problem of localisation within autonomous robotics. The author discusses Visual Place Recognition (VPR) specifically, but also relates it to and distinguishes it from closely related concepts of visual-SLAM, visual-localisation and the image-correspondence problem. The various challenges within VPR, research gaps, applications and the theory behind VPR are presented.

1.1 Background

The ability of a robot to autonomously perform various tasks has been the main focus of the robotics community over the past few decades. The inherent interest of mankind in this domain and the sheer amount of applications of robot autonomy has led to a large number of academic institutions, research groups, industries and entrepreneurs to investigate the various sub-domains of autonomous robotics. This is evident by the fact that almost every major university now has a robotics and autonomous systems program in their curriculum.

One of the most investigated application of autonomous robotics has been in self-driving cars, which has made significant progress over the past 2 decades. Not only have industry joints like Google, Facebook, Tesla, Uber, Lyft, General Motors, Ford and others have invested in driver-less cars, but academic institutions like Carnegie Mellon University (CMU), Stanford, MIT, Harvard, Oxford and others have also designed modified versions of driver-less cars for research purposes. Additionally, several spin-offs from academia have also targeted directly or indirectly the self-driving cars industry, for example within UK, Oxbotica originated from Oxford university, SLAMCore originated from Imperial College London and Wayve from Cambridge University. There are many other small- and medium-scale enterprises (SMEs) working in this domain, which have not been mentioned but that continue to advance the state-of-the-art in autonomous robotics.

Some of the major breakthroughs as seen in the self-driving cars paradigm (or autonomous robotics in general) have resulted from the DARPA (Defense Advanced Research Projects Agency) Grand Challenge held in 2004, followed-up in 2005 and then the Urban challenge held in 2007. The objective was to autonomously traverse a track of a few hundred kilometers in challenging environments. All the participating cars were fully-autonomous and were equipped with a range of different sensors including LiDARs, cameras, ultrasonic sensors, radars and GPS. Figure 1.1 encloses the winners of the competition.

The autonomy of a robot can be broken down into its abilities to perceive, plan and control a required task. Within perception, a key research challenge is to localise (identify current robot location) given visual information and usually with no GPS (Global Positioning System) prior. However, localisation is not the only requirement, because for an autonomous robot to localise, it needs a map to localise itself within it. Often times, this map is not available, especially when the underlying problem is related to exploration. Thus, creating a map of the environment is also an essential element. This problem of simultaneously creating the map and localising itself within it, is termed as SLAM (Simultaneous Localisation and Mapping) and has been a major topic of research within the robotic vision community.

However, because any localisation estimates performed within a SLAM system are prone to errors, which get accumulated over time as the robot explores an environment, there is the

Stanley: DARPA Grand Challenge Winner 2005 Tartan: DARPA Urban Challenge Winner 2007

Fig. 1.1 The winners of the DARPA Grand Challenge 2005 (left) and the DARPA Urban Challenge 2007 (right) are enclosed here.

need for a mechanism to correct these error drifts. This has been addressed by 'loop-closure', where a robot is required to recall a previously visited place (when it revisits it) in its map and then correct any error drifts accumulated over time. Within SLAM, the essence of loop-closure is Visual Place Recognition and it has its corresponding challenges. Both SLAM and VPR are further discussed in the following 2 sub-sections, respectively.

1.2 Simultaneous Localization and Mapping (SLAM)

Simultaneous Localization and Mapping (SLAM) refers to the ability of a robot to simultaneously map an environment and localise itself within it. The existence of a map is crucial for an autonomous robot due to several applications including path planning, obstacle avoidance and navigation. This map can be built in various forms, including appearance-only, metric-only, topological and topo-metric maps.

The essence of all of these maps are vertices and edges, where vertices define physical points in the world and edges define the relations between them. An appearance-only map contains visual information (either monocular, omni-directional, RGB-D or stereo images) as the vertices of the map and edges define the topological relation (connectivity) between them. A metric map (e.g, occupancy grid maps) consists of metrically precise locations/landmarks as the vertices and edges define the exact metric distance between them. A topological map contains only the topological relation between edges and does not contain any metric information about either the vertices or the edges. A topo-metric map, however, contains metric information about the edges, while the metric information about the vertices is

unknown. For all of these maps, some level/type of landmark information is extracted from the environment and associated with the edges of the map.

As the robot navigates through the environment it adds newer landmarks to its map. However, as it is important for the robot to localise itself within the map (to account for dead-reckoning errors), it needs to extract transitional relations between the landmarks as it moves through the map. This extraction of transitional information (e.g, moved 1 meters forward etc) from co-observed landmarks/features in the map is termed as localisation. Thus, as the robot moves through the environment, it adds newer information into the map (mapping) and also localises itself within the map (localisation) simultaneously, achieving SLAM as shown in Fig. 1.2. When the added information is vision-based and the transitional-information is extracted using co-observed visual features between consecutive images/sequences, SLAM becomes visual-SLAM.

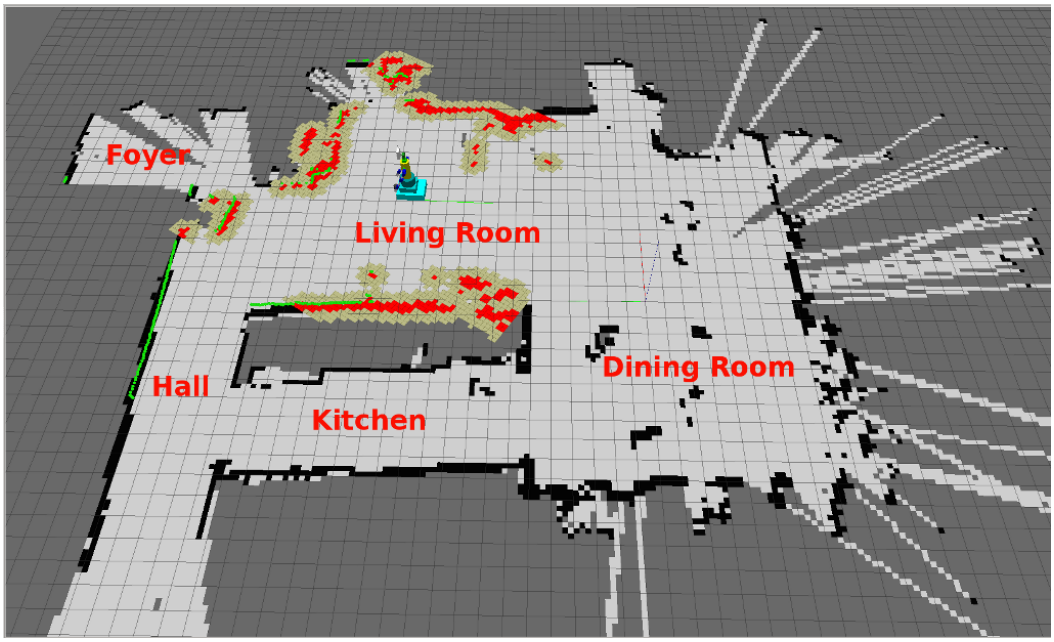

Fig. 1.2 An example of a robot creating a map and localising itself within it.

The localisation element of visual-SLAM is visual-localisation (also termed as the image correspondence problem in computer vision community), which should not be confused with Visual Place Recognition (VPR). Visual-localisation is not related to the revisiting of a place as in VPR, but only refers to the extraction of transitional information between consecutive images for motion estimates. However, because motion estimates extracted using visual-localisation are prone to errors and these errors accumulate over time due to the iterative nature of the SLAM mechanism, there is the need for correcting this accumulated

error. This can be achieved if while exploring, the robot revisits a place/landmark that it had previously seen, the exact location information of which is already available in the map and can thus be used to correct errors. When this ability to recognise a previously visited place is required from a robot equipped with only a camera, Visual Place Recognition becomes the topic of research and has its own associated challenges, as discussed in the following sub-section.

1.3 Visual Place Recognition (VPR)

Visual Place Recognition is a key research problem within the computer vision community and refers to the ability of a system to match images of a place under changing viewpoint and illumination conditions. This ability to recognise previously visited places has several applications in various domains. As previously discussed, a major application of VPR is in loop-closure for SLAM systems [10]. However, VPR is not limited just to SLAM systems but can also be used for improved representations [11], asset management using aerial imagery [12], location-refinement given human-machine interfaces [13], query-expansion [14], image-search based on visual content [15], 3D-model creation [16] and vehicular navigation [17].

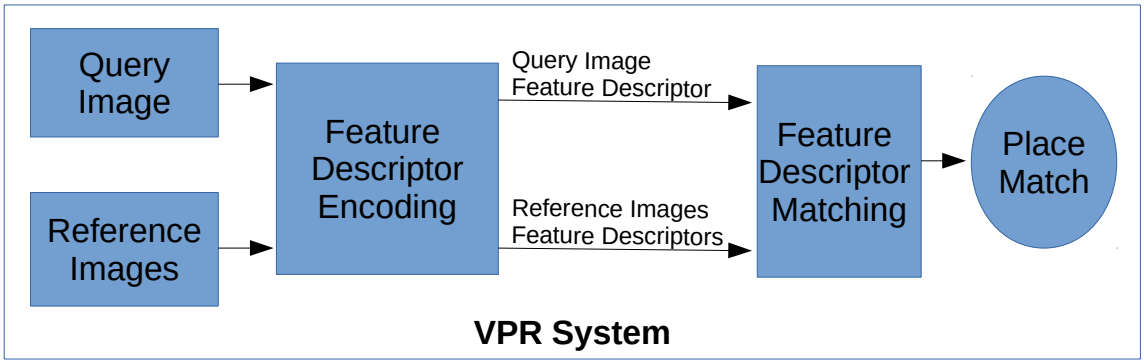

Fig. 1.3 A block diagram of a typical VPR system is shown here.

The usual architecture of any VPR system is shown in Fig. 1.3. The input of a VPR system consists of a query image and reference images. A feature encoding block then computes the feature descriptors of these query and reference images, either using handcrafted feature encoding techniques or deep-learning-based techniques. Usually, the map (reference images) is already available and thus the reference images' feature descriptors are pre-computed to save time. A feature matching technique is then used to match the feature descriptor of a

query image with those of reference images. This feature matching can be done for all the reference images (linear search) or for some of the reference images depending upon the underlying search technique. The best matched image is then selected as the correct place match. The place matching confidence is usually associated with the image matching score (e.g, cosine-matching, L1-matching, L2-matching), such that the 2 images with the highest matching score are considered to be the most likely place matches.

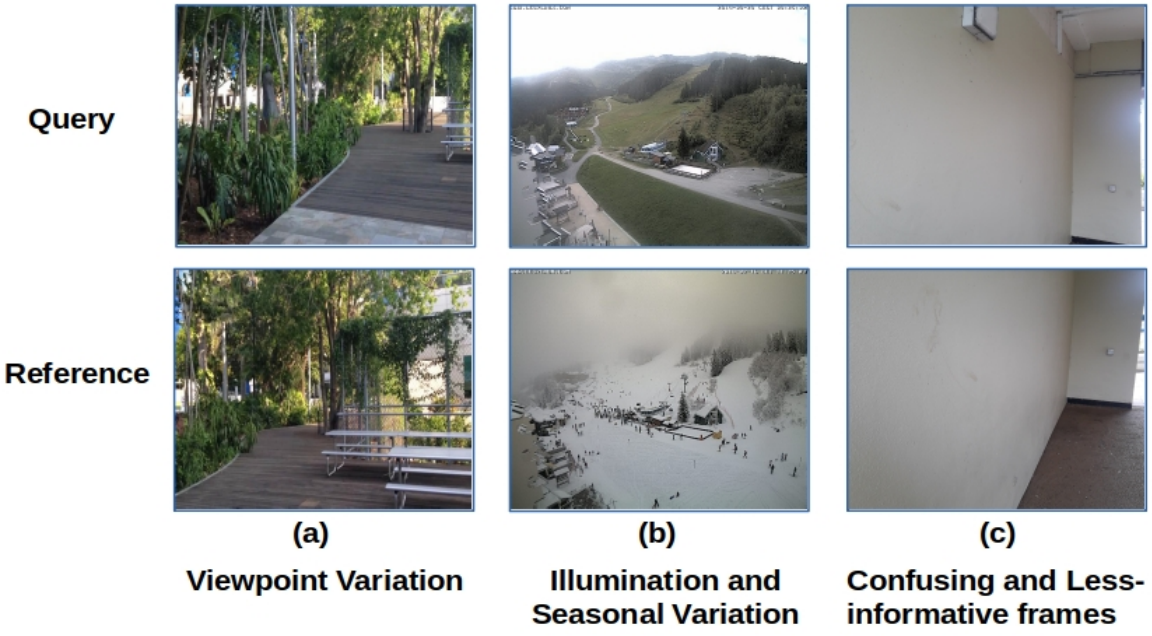

Fig. 1.4 Examples of places that a VPR system is expected to correctly match.

The objective of an ideal VPR system is to retrieve correct place matches under extreme viewpoint, illumination, weather and seasonal variations as shown in Fig. 1.4, while requiring minimum computational power and memory storage. Thus, the typical performance metrics used to estimate the performance of a VPR technique are related to the accuracy and precision of place matching, feature descriptor encoding time, feature descriptor matching time and feature descriptor size (in Bytes). The exact details of these metrics are further explained later in this thesis. The author would also like to clarify some common beginner-level misconceptions about VPR for the reader of this thesis, as enlisted below.

1. VPR does not necessarily need to be a sub-module of a SLAM system, but can in-fact be used as the primary localisation system for an autonomous robot [18]. However, loop-closure for SLAM is one of the major applications for VPR.

2. While viewpoint-invariance is a key requirement for a VPR system in general, there are cases where some viewpoint-variance may be required instead of invariance. For example, when VPR is the primary localisation system, viewpoint-invariance may actually lead to errors in position estimates.

3. Unlike object-detection or other similar topics in deep-learning for computer vision, where the number of output classes/labels are constant, the number of classes/places is usually not constant for VPR and hence an end-to-end classifier cannot be designed. However, if the environment/map is fixed and does not need to incorporate new places, the earlier statement becomes invalid [19].

4. No universal state-of-the-art VPR technique exists in literature and each technique has its strengths and weaknesses.

5. VPR has matured significantly as a field over the past 10-15 years, but there are still a number of challenges yet to be addressed and it has huge potential for future research. For example, controllable viewpoint-variance, salience-based ensemble of VPR techniques, evaluation metric design, application of other deep-learning-based techniques (transformers, deep reinforcement learning etc.) to VPR and many others.

1.4 Problem Statement and Challenges

Visual Place Recognition contains several interesting research challenges as a domain. This sub-section identifies some of the most crucial challenges, which have also been addressed in the later chapters of this thesis.

1.4.1 Perceptual Aliasing

The baseline requirement of a VPR system is to successfully recognise a previously visited place, which for a monocular camera-based system essentially translates to matching the visual content in multiple RGB images. Without any geographical prior, matching of images that come from semantically similar areas (e.g, different road crossings at night time, different car parkings at noon, natural scenery etc.) can be very challenging, because 2 images of geographically different car parks at noon may have more in common than 2 images of the same car park at noon and midnight times. This problem of different places that look similar under certain conditions, is usually termed as perceptual aliasing.

Aliasing in general is the result of lack of sufficient information to distinguish two models/functions/data from each other, and in RGB-based VPR refers to the lack of semantic

context, lack of geographical prior and depth information of the scene. Fig. 1.5 presents this challenge visually. Chapter 3 of this thesis is dedicated to addressing this challenge of perceptual aliasing.

Fig. 1.5 (a) Places that are geographically similar may look very different, while (b) places that are geographically different may look very similar. Images are taken from GardensPoint dataset [1] and ESSEX3IN1 dataset [2].

1.4.2 Viewpoint and Conditional Variations

One of the most commonly attributed challenges for VPR systems is the extreme viewpoint and conditional variations that a VPR system may be required to handle. The viewpoint variation could be lateral/2D (2-Dimensional) as for cars changing lanes, 3D for human-like motion and/or 6-DOF (Degrees-of-Freedom) for aerial platforms. Each of these variations has its associated challenges and different VPR techniques have varying levels of invariance to these viewpoint variations. Examples of these variations are shown in Fig. 1.6.

Fig. 1.6 Different types of viewpoint changes that a VPR system may be required to handle. Images are taken from GardensPoint dataset [1], Berlin Halenseestrasse dataset [3] and ESSEX3IN1 dataset [2].

Conditional variations, on the other hand, could results from changes in times of the day, weather conditions or seasonal changes. Fig. 1.7 show some of the challenges presented by conditional variations to VPR systems. Chapter 6 of this thesis is dedicated to addressing this challenge of perceptual aliasing.

1.4.3 Computational and Storage Needs

While successfully matching different places under changing viewpoints and conditions remains the top requirement from a VPR system, computational and storage needs should also be considered to reflect the practical deployment of a VPR technique. The performance of a VPR system in real-time depends on its feature encoding time and descriptor matching time; the higher these timings, the lesser the deployment practicality of the technique on a resource-constrained platform. Additionally, because the feature descriptors computed by a VPR technique are iteratively stored in the map as the robot explores an environment, memory footprint also becomes an important criterion. Storage requirements scale linearly with the number of images in the map, thus, lower memory footprint is desirable for VPR techniques. Chapter 6 of this thesis is dedicated to addressing this challenge of perceptual aliasing.

Fig. 1.7 Different types of conditional changes that a VPR system may be required to handle. Images are taken from Nordland dataset [4], GardensPoint dataset [1] and SPEDTest dataset [3], respectively.

1.4.4 Establishing State-of-the-Art Technique

One of the major challenges that was identified and the author attempted to address through this thesis (Chapter 4 and 5), is that of establishing which VPR technique out of the many proposed over the past 15-20 years is state-of-the-art. This problem arises due to the apparent zoo of VPR techniques, datasets and evaluation metrics available for VPR evaluation. As a result, this makes comparison between different VPR techniques difficult, as some datasets may or may not have a particular challenge. Moreover, the limited comparison with contemporary VPR techniques leads to ambiguity of the correct state-of-the-art technique.

We show in Fig. 1.8 in a chronological order, the matching performance of various VPR techniques. It is clear that the variation trends in between techniques and datasets is not as expected and therefore, there is a need to evaluate the contemporary VPR techniques on a common platform.

1.5 Thesis Contributions

The contributions of this thesis expand around addressing the challenges and limitations as identified in sub-section 1.4. These can be primarily broken down into the below few points.

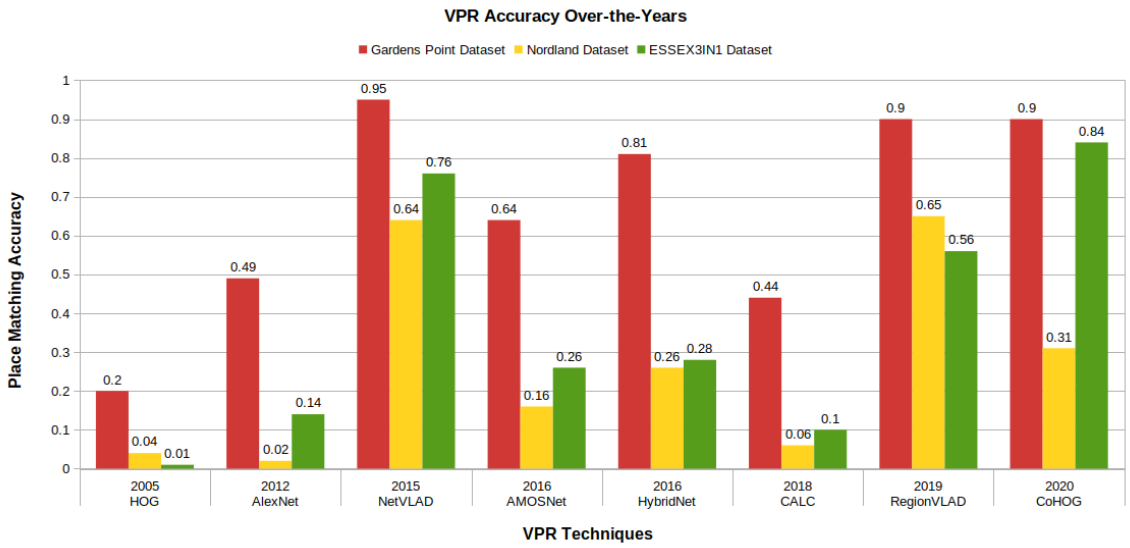

Fig. 1.8 The matching performance of VPR techniques is presented in a chronological order to reflect the ambiguity in establishing a state-of-the-art VPR technique.

1. The first contribution of this work is a framework designed to estimate perceptual aliasing for VPR. This framework, namely 'Memorable Maps', attempts to classify images based on the salience of their content and therefore, potentially enables the prediction of false-positives. The efficacy of this framework is evaluated on challenging VPR datasets and a performance boost is reported for a number of contemporary VPR techniques, when used in conjunction with our framework.

2. The second contribution of this work is an extensive evaluation of 10 state-of-the-art VPR techniques on some of the most challenging VPR datasets. This evaluation is performed on a common platform and different evaluation metrics are used to report the performance on a levelled playing field.

3. The third contribution of this work is the analysis of VPR state-of-the-art for aerial platforms. Such aerial platforms have 6 Degrees-of-Freedom (DOF), which makes the viewpoint-invariance element of VPR techniques very challenging. We provide several useful insights into this area.

4. The fourth contribution of this work is a novel, handcrafted, training-free VPR technique that achieves state-of-the-art place matching performance per compute unit against 10 VPR techniques on 5 challenging VPR datasets. This technique uses regions-of-interest (ROIs) extraction and regional-convolutional matching of Histogram-of-

Oriented-Gradients (HOG) descriptors to tackle perceptual-aliasing, viewpoint and conditional variations.

1.6 Thesis Structure

The rest of the thesis is divided into the following 6 chapters:

In **Chapter 2**, we provide a detailed literature review of the existing research in the domain of Visual Place Recognition, but also in the closely related topics of SLAM, image matching and semantic mapping.

In **Chapter 3**, we present a cognition-inspired agnostic framework for building a map for Visual Place Recognition. This framework draws inspiration from human-memorability, utilises the traditional image entropy concept and computes the static content in an image; thereby presenting a tri-folded criteria to assess the memorability of an image for VPR. A dataset namely ESSEX3IN1 is also created, composed of highly confusing images from indoor, outdoor and natural scenes for analysis. When used in conjunction with state-of-the-art VPR methods, the proposed framework provided significant performance boost to these techniques, as evidenced by results on ESSEX3IN1 and other public datasets.

Chapter 4 builds upon the fact that in recent years there has been significant improvement in the capability of VPR methods, building on the success of both hand-crafted and learnt visual features, temporal filtering and usage of semantic scene information. The wide range of approaches and the relatively recent growth in interest in the field has meant that a wide range of datasets and assessment methodologies have been proposed, often with a focus only on precision-recall type metrics, making comparison difficult. Therefore, in this chapter, we present a comprehensive approach to evaluating the performance of 10 state-of-the-art recently-developed VPR techniques, which utilises three standardized metrics:(a) Matching Performance b) Matching Time c) Memory Footprint. Together this analysis provides an up-to-date and widely encompassing snapshot of the various strengths and weaknesses of contemporary approaches to the VPR problem.

In **Chapter 5**, we propose that the existing VPR evaluations (including our Chapter 4) are performed for ground-based mobile platforms and cannot be generalized to aerial platforms. The degree of viewpoint variation experienced by aerial robots is complex, with their processing power and on-board memory limited by payload size and battery ratings. Therefore, in this chapter, we collect state-of-the-art VPR techniques that have been previously evaluated for ground-based platforms and compare them on recently proposed aerial place recognition datasets with three prime focuses: a) Matching performance b) Processing power consumption c) Projected memory requirements. This gives a birds-eye

view of the applicability of contemporary VPR research to aerial robotics and lays down the the nature of challenges for aerial-VPR.

In **Chapter 6**, we present a novel, compute-efficient and training-free approach based on Histogram-of-Oriented Gradients (HOG) descriptor for achieving state-of-the-art Performance-per-Compute-Unit (PCU) in VPR. The inspiration for this approach (namely CoHOG) is based on the convolutional scanning and regions-based feature extraction employed by Convolutional Neural Networks (CNNs). By using image entropy to extract regions-of-interest (ROI) and regional-convolutional descriptor matching, our technique performs successful place recognition in changing environments. The author has used viewpoint- and appearance-variant public VPR datasets to report this matching performance, at lower RAM commitment, zero training requirements and 20 times lesser feature encoding time compared to state-of-the-art neural networks. The author also discusses the image retrieval time of CoHOG and the effect of CoHOG's parametric variation on its place matching performance and encoding time.

Finally in **Chapter 7**, the author provides his concluding remarks and summarises the contributions of this thesis. The future directions of research and the limitations in contributions of this thesis are highlighted.

1.7 List of Publications

The following contributions were made during the course of this Masters by Dissertation:

1. Zaffar, M., Ehsan, S., Stolkin, R. and Maier, K.M., 2018, August. Sensors, SLAM and Long-term Autonomy: A Review. In 2018 NASA/ESA Conference on Adaptive Hardware and Systems (AHS) (pp. 285-290). IEEE.

2. Zaffar, M., Ehsan, S., Milford, M. and Maier, K.M., 2020. "Memorable Maps: A Framework for Re-defining Places in Visual Place Recognition". IEEE Transactions on Intelligent Transportation Systems.

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

4. Zaffar, M., Khaliq, A., Ehsan, S., Milford, M., Alexis, K. and McDonald-Maier, K., 2019. "Are State-of-the-art Visual Place Recognition Techniques any Good for Aerial

Robotics?". IEEE International Conference on Robotics and Automation (ICRA 2019), Workshop on Aerial Robotics.

5. Zaffar, M., Ehsan, S., Milford, M. and McDonald-Maier, K., 2020. "CoHOG: A Light-Weight, Compute-Efficient, and Training-Free Visual Place Recognition Technique for Changing Environments". IEEE Robotics and Automation Letters, 5(2), pp.1835-1842.

Chapter 2

Literature Review

This chapter presents a detailed overview of the existing literature within Visual Place Recognition. A breakdown of techniques is performed into feature-less techniques, handcrafted feature descriptors-based techniques and deep learning-based techniques, which are then each discussed in depth. Research within domains that are closely related to VPR is also presented briefly. The author also presents detailed insights into the datasets and evaluation metrics available for VPR. In summary, this chapter attempts to provide a strong literature base for the contributions presented in later chapters, but also covers closely-related research to provide a good overview for the reader.

2.1 Overview

Visual Place Recognition (VPR) has developed as an independent research domain over the past 15-20 years, primarily due to its wide-spread applications in various fields of autonomous systems, including but not limited to construction, surveying, guidance, agriculture, industry, mining, transportation, human-machine interfaces and/or any other autonomous system application that requires localisation estimates. A detailed survey of VPR has been conducted by Lowry et al. [8], which should serve as a good point of start for any VPR researcher. The survey of [8] develops the theory behind VPR, how it's related to SLAM and other domains, the challenges within VPR and future areas of investigation.

While the survey of [8] should be referred to for the theoretical aspects of what a Place is and how the conceptual understanding of VPR can be broken down; this section attempts to extensively cover more practical elements of VPR and also tries to bridge the gap between VPR research performed within the years of 2015 and 2020. The rest of the Chapter is divided as follows: In sub-section 2.2, the author provides a brief coverage of research works done within SLAM systems. To set the tone for Chapter 3, we briefly discuss the research in

semantic mapping in sub-section 2.4. Next, an extensive review of VPR techniques proposed to date is presented in sub-section 2.5. The author provides details about the datasets available for VPR in sub-section 2.6 and their strengths/weaknesses are listed. Finally, a summary of evaluation metrics utilised for VPR is presented in sub-section 2.7.

Although this thesis presents a literature overview of both SLAM and VPR, a disclosure is that VPR and SLAM are both different yet similar. VPR can be used as a sub-module of a SLAM system [8], but a SLAM system can have other methodologies to achieve loop closure as well [10]. On the other hand, VPR does not necessarily have to be a sub-module of SLAM, but can also act as an independent primary localisation system [20].

2.2 A Brief Overview of SLAM Research

Simultaneous Localisation and Mapping (SLAM) has been one of the most active field of research within robotics over the past couple of decades. Due to the wide coverage of the SLAM problem by the robotics community, Cadena et al. [10] presented the most detailed survey of SLAM systems. This survey paper touches both the software and hardware aspects of research within SLAM, because SLAM is highly dependent on the sensing modality (or modalities). The author of this thesis also linked the research between the hardware and software aspects of SLAM and how they affect the long-term autonomy of the system [21].

The core research goals in SLAM have been efficient mapping topologies [22], feature extraction and matching [23], location estimation [24] and loop closure techniques [25]. Interestingly, research in each of these areas has mostly been fuelled by the underlying sensor technology. Acoustic sensors are some of the earliest, low-cost, compact, range-measurement sensors, which have been widely used in solving the SLAM problem. An early implementation of this is [26]. In [27], the authors show an implementation of Acoustic-SLAM using moving microphone array and surrounding speakers. Assuming an Omni-directional acoustic sensor and receiver, [28] presents Echo-SLAM with a co-located microphone and acoustic source. Using landmarks as nodes of a sensor network, authors in [29] have shown a range-only SLAM system working in conjunction with sensor networks.

Light Detection and Ranging (LIDAR) has driven a lot of research in range-based SLAM systems. LIDAR provides depth point-clouds of the environment, as shown in Fig. 2.1, at a very high latency and resolution. An early work utilising LIDAR in conjunction with Rao-Blackwellized particle filters is presented in [30]. Authors in [31] present an approach for finding interest regions in the data coming from Laser Sensors. Using occupancy grid-maps for mapping, [32] shows a scalable SLAM system with full-estimation of 3D pose. In [33], real-time loop-closure is achieved with a LIDAR.

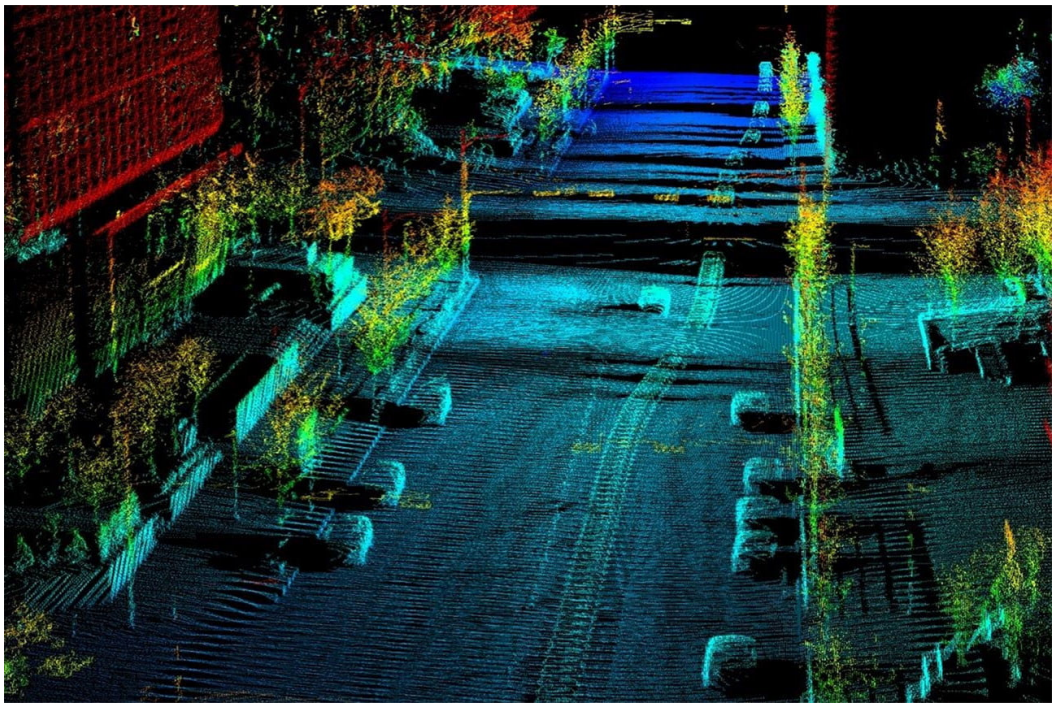

Fig. 2.1 A point-cloud map created by a LIDAR scan is enclosed here.

One of the prime sensors used for SLAM is a monocular camera, primarily inspired by its low-cost, wide-availability, ease-of-use and information quality. The first-time a fully capable SLAM system was demonstrated based on a monocular camera was the pioneer work of [34]. A real-time SLAM system based on a single camera is also presented in [35]. A SLAM system resulting from the fusion of monocular cameras and inertial sensors is proposed in [36]. A comparison of monocular SLAM and Stereo SLAM is presented by authors in [37]. A semantic SLAM system for a monocular camera is proposed in [38]. Other than traditional monocular cameras, omni-directional cameras have also been used for SLAM due to their wide field-of-view. An early implementation of SLAM with an omni-directional camera can be observed in [39]. Authors in [40] combine particle filters with a SIFT feature extractor for images obtained from Omni-directional camera. An extensive review of SLAM based on Omni-directional camera is presented by [41].

Depth measuring RGB cameras, in the form of RGB-D sensors and Stereo cameras, introduced a new dimension to the SLAM problem and have drawn significant interest from the community due to their similarity with biological visual cognition. Authors in [42],[43] and [44] present an evaluation of RGB-D SLAM. A real-time, large-scale dense SLAM system is developed in [45] using RGB-D sensors and an application of RGB-D SLAM to aerial systems is shown by [46]. While RGB-D cameras are active, power-hungry and

expensive, stereo cameras are passive, cheaper and closer to biological vision. Authors in [47], present an implementation of Stereo-SLAM using particle filters. To perform SLAM in large indoor and outdoor environments, [48] presents a 6-DOF SLAM using hand-held stereo camera. Using iterative closest point algorithm, [49] shows a robust 3D stereo camera SLAM. In [50], the ability of stereo cameras to provide depth information in addition to the conventional multi-view disparity-based depth calculation is exploited.

Recently, event cameras led to a new branch of SLAM systems, due to their very high dynamic range, no motion blur and a latency in the order of microseconds. These event cameras output pixel-level intensity changes instead of standard RGB-frames in a sequential asynchronous manner. Fig. 2.2 shows the difference between an RGB image and an event camera image. Parallel Tracking and Mapping (PTAM) [51] is one of the major SLAM techniques and [52] shows an implementation of PTAM for event cameras. Although event cameras are an excellent choice for dynamic scenes but in static scenes, they give little-to-no information. Thus, [53] combines event cameras and monocular cameras to achieve an ultimate SLAM system. Authors in [54] present a complete continuous-time event-based SLAM system in conjunction with inertial measurements.

Fig. 2.2 Difference between an RGB image and an event-camera image is shown here, where the latter only focuses on the dynamics in the image, thereby avoiding redundant static information. Picture courtesy: Prophesee France.

2.3 Visual-SLAM, Visual-Localisation and VPR

This brief sub-section attempts to identify the differences between the closely related concepts and terms of Visual-localisation, Visual-odometry, Visual-SLAM, Visual Place Recognition, Image Matching and the Correspondence Problem for the clarity of the reader. These terms are so similar that without any prior knowledge of each of the domains, it is too difficult to distinguish which term may refer to which domain. Because there is a significantly large literature published in each of these domains, which is out of the scope of this thesis, the objective of this sub-section is to primarily distinguish between the domains, explain all of these terms and to provide reference to surveys conducted in each of these domains.

Visual-SLAM refers to a Simultaneous Localisation and Mapping System, where the source of information is a camera (usually monocular but can also be other variants like omni-directional, stereo etc.). The objective of a Visual-SLAM system is two-fold: (a) Create a map of a previously unknown environment, (b) Localise the robot within this map. A good survey of Visual-SLAM has been presented in [55]. The localisation element is further sub-divided into 2 categories: (a) Visual-localisation, (b) Visual Place Recognition for loop-closure. Visual-localisation, which may also be referred to as Visual-odometry estimates the motion of a robot by using overlapping information between consecutive frames and has been covered in [56]. While Visual-localisation is an application-specific term, Image Matching or the Correspondence Problem abstracts away the underlying problem for identifying and locating overlapping information between 2 camera frames. A good discussion about Image Matching is presented by [57]. Visual Place Recognition (more abstractly Image Retrieval) is different from each of these problems and refers to identifying a previously visited place under changing viewpoint and appearance conditions [8]. VPR finds its major application in loop-closure for Visual-SLAM systems but also has many other computer vision applications [20]. Being the main area-of-interest for this thesis, VPR has been discussed in depth in this literature review in the following sub-sections.

2.4 Semantic Mapping

Semantic mapping refers to the creation of maps where the nodes of the map have a semantic attribute. This semantic attribute could be based on object classifiers, scene segmentation, place salience and any other. In general, semantic mapping techniques for summarizing a robot's experience are surveyed by Kostavelis et al. [58]. The author's objective in this thesis is to discuss semantic mapping but with focus on VPR and how semantic mapping is associated with VPR.

VPR requires a pre-known map of the environment in the form of images representing places. Traditionally, places have been described by camera frames, where a place is selected from multiple video frames based on either time-step, distance or distinctiveness. Most of the VPR datasets (discussed in sub-section 2.6) are time-based, as frames are selected given a fixed FPS (frames per second) rate of a video camera. However, time-based place selection assumes a constant non-zero speed of the robotic platform and is thus impractical in many situations. To cater for variable speed, distance-based frame selection is used where a frame is picked every few metres to represent a new place [59] [60]. Both time- and distance-based approaches lead to huge database sizes and frequently sample visually identical frames as different places; thus leading to inaccuracies and impracticality for long-term autonomy.

Different research works have tried to overcome these intrinsic limitations of image sampling by proposing image selection based on visual distinctiveness. Chapoulie et al. [61] use a customised algorithm that detects change point for segmentation between different topological places in both indoor and outdoor scenes. Image sequence partitioning for creating sparse topological maps is presented by Korrapati et al. [62], where sequences of images are divided into nodes/places using four descriptors namely GIST, Optical Flow, Local Feature Mapping and Common-Important Words. In [63], a thematic approach is adapted to evaluate the novelty of an incoming image by co-relating it with the redundancy of visual features/topics. Bayesian surprise is adapted with immunity to sensor type, for extracting landmarks to create a sparse topological map in [64]. Online topic modeling with visual surprise calculation is done by Girdhar et al. [65] for under-water explorations. An incremental unsupervised place discovery scheme is adopted by Murphy et al. [66] which fuses information over time to find visually distinct places.

Authors in [67] present both offline and online solutions for finding images that best summarize a given sequence. The score for every incoming image is related to the difference of posterior distribution from prior distribution using bayesian surprise or set theoretic surprise. In [68], coresets are used to pre-cluster input image stream and then topic-based image representation is used followed with graph-based incremental clustering. A place detection scheme is proposed by Karaoguz et al. [69] based on bubble-space representation. A new place is checked for informativeness based on surface deformation and variance in a time-window of coherent images. The authors in [70] use region proposals in spatio-temporal context instead of low-level features to represent input frames and then based on region-adjacency-graph detect visually distinct places. A human-augmented change point detection scheme is presented by Topp et al. [71] where a change stimuli could either be pointed out by the robot or its operator. The authors propose the change as a structural ambiguity, which

can be pointed out either by the robot or a human operator during a guided tour. Detection of change point is also targeted by Ranganathan [72] with a Bayesian probabilistic model.

One common element to all these works is that they focus on map compression, video segmentation or experience summarisation, but do not discuss if the resulting compressed/summarized map is actually composed of good matchable images of places. These methods define the distinctive nature of images based on their visual difference from previously seen images. Resultingly, such visually different images may come from grassy plains, natural scenery, dynamic objects or low-textured places leading to poor VPR performance.

2.5 Visual Place Recognition Techniques

This section presents the various VPR techniques that have been proposed over the past many years. These techniques have been further sub-classified based on their prominent technology/mechanism.

2.5.1 Feature-less VPR Techniques

VPR essentially unravels into an image matching problem, where the objective is to distinguish images of the same place from images of different places. Under no viewpoint and conditional variations, VPR is a straight-forward process of matching the pixel-level intensities e.g, by using Sum-of-Absolute-Differences (SAD) on grayscale or RGB images. This, however, is almost never the case, because revisiting a place always has some level of viewpoint and/or conditional variation, and there is the need to have a mechanism where these variations can be handled.

If there is no conditional variation and only lateral (perpendicular to the axis of movement) viewpoint variation is present, it may still be possible to achieve VPR by segmenting the query and reference images into multiple crops and using an All-to-All matching (SAD) of these crops to achieve some level of viewpoint invariance. These rather simplistic approaches do not work well in real-world because there is always some sort of conditional variation resulting from different times of the day, different seasons and weather conditions and/or dynamic objects. However, Milford et al. (SeqSLAM [5]) showed that it is still possible to use feature-less VPR for handling conditional variations by using sequence of images even under challenging day-to-night and extreme weather conditions. This sequential matching is implemented as a confusion matrix of templates as shown in Fig. 2.3, which are then searched across at different angles for prospective ‘sequence’ matches. The original SeqSLAM algorithm did not accommodate variable speed of the robot platform, so Pepperell et al.

[59] extended SeqSLAM by accommodating odometry information and employing sky filtering in SMART. This was further extended to show the effects of various parameters in [73]. A detailed analysis of the various parameters of SeqSLAM has been shown in [74], where the author performs stress testing of the SeqSLAM algorithm. An open-source detailed toolbox based on SeqSLAM is presented in [75]. This sequence-based VPR is then combined with a biologically-inspired SLAM system (RAT-SLAM) in [76]. More recently, a sequence-inspired approach is presented in [77], where change across sequence is used for descriptor computation, namely delta-descriptors. In general, these sequence-based feature-less techniques do not perform well under significant conditional variations and especially under viewpoint variations. Therefore, viewpoint- and condition-invariant feature extraction, description and matching have been the key areas of research within SLAM, as discussed in the following sub-sections.

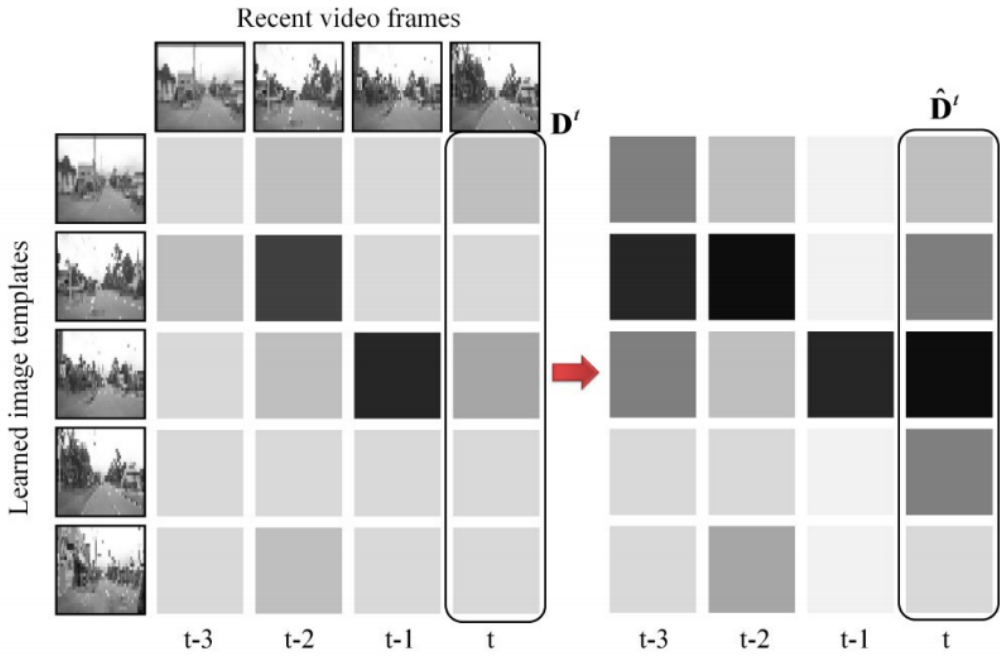

Fig. 2.3 The sequential matching of templates, followed by local contrast enhancement as in SeqSLAM is enclosed here. Darker shades mean stronger matches. Figure taken from [5].

2.5.2 Handcrafted VPR Techniques

The traditional computer vision research which usually refers to the pre-deep-learning era (generally pre-2012) was focused on designing handcrafted feature descriptors that were

resilient to viewpoint and conditional variations. These descriptors were not designed specifically for any particular task such as VPR, but were general purpose and found usage in almost all computer vision applications, including object-detection, pose-estimation, localisation, structure-from-motion, scene segmentation and many others. The viewpoint and conditional invariance requirements for handcrafted feature descriptors actually lead to two separate classes of descriptors: local feature descriptors and global feature descriptors, as visually differentiated in Fig. 2.4.

Scale Invariant Feature Transform (SIFT [78]) and Speeded Up Robust Features (SURF [6]) are two of the most widely used local descriptors. These local techniques extract keypoints from an image, describe these keypoints by an underlying low-level gradient-based descriptors. They have been applied to the VPR problem in [79] [80] [81] [82] [83]. FAB-MAP (Frequent Appearance Based Mapping [84]) is a probabilistic visual-SLAM algorithm that represents places as visual words and uses SURF as the underlying interest point detector. An open-source implementation of FAB-MAP is presented in [85]. Furthermore, an extension to FAB-MAP is presented by utilising odometry information in CAT-SLAM [86]. Center Surround Extremas for real-time feature detection and matching (CenSurE [87]) has been used for VPR in [88]. FAST [89] is a high-speed corner detector for real-time image processing that has been used for SLAM by Mei et al. [90], coupled with SIFT descriptor. One common drawback to all these keypoint based approaches is the extensive matching requirements, which has been addressed by Bag of visual Words (BoW [91]) approach. BoW collects visually similar features in dedicated bins (pre-defined or learned by training a visual-dictionary) without topological consideration, enabling direct matching of BoW descriptors. Different research works have used BoW for VPR, including [92] [93] [94] [95]. While local feature descriptors have viewpoint invariance, they suffer from conditional changes as there is no underlying mechanism to handle this.

Global feature descriptors like Gist [7] use Gabor filters to create the signature of an entire image and have been used for VPR with panaromic images by Murillo et al. [96] and Singh et al. [97]. BRIEF [98] descriptor due to its lower encoding requirements and faster matching time is combined with Gist by Sünderhauf et al. [99] to perform large scale visual-SLAM. Whole-Image SURF (WI-SURF) is a global variant of SURF and has been used for visual localization by Badino et al. [100]. McManus et al. [101] have proposed an approach where scene signatures are extracted and described by dedicated HOG descriptors. Global feature descriptors can handle moderate illumination changes by normalisation-techniques because the change is usually uniform and global, however, suffer from all levels of viewpoint change.

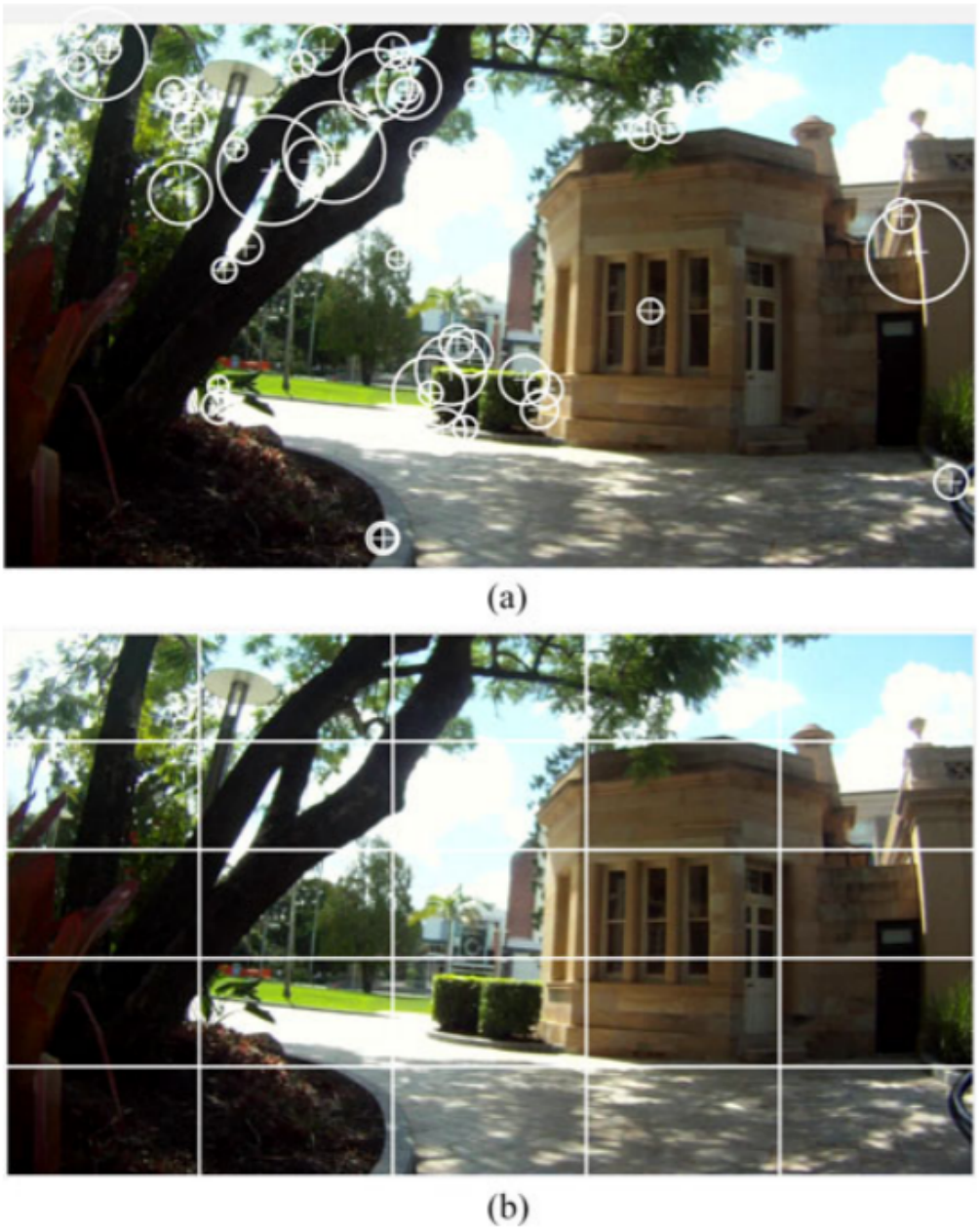

Fig. 2.4 The usual approach towards handcrafted feature descriptors is shown here: (a) Local feature descriptor SURF [6] extracts keypoints from an image, (b) Global feature descriptor Gist [7] divided image into fixed-size portions and computes global descriptor. Picture taken from [8].

2.5.3 Deep Learning-based VPR Techniques

Because handcrafted feature descriptors cannot handle extreme viewpoint variations and conditional changes, the use of deep learning was explored for VPR by the robotic vision community, which led to some successful works. Since 2014, deep learning has been the most prominent and successful go-to area of research for VPR. Within deep learning, Convolutional Neural Networks (CNNs) have been the most investigated for VPR. These CNNs convolutionally scan for features in an image, spanning both low-level features (in earlier layers) and high-level semantics (in later layers). A typical architecture is shown in Fig. 2.5.

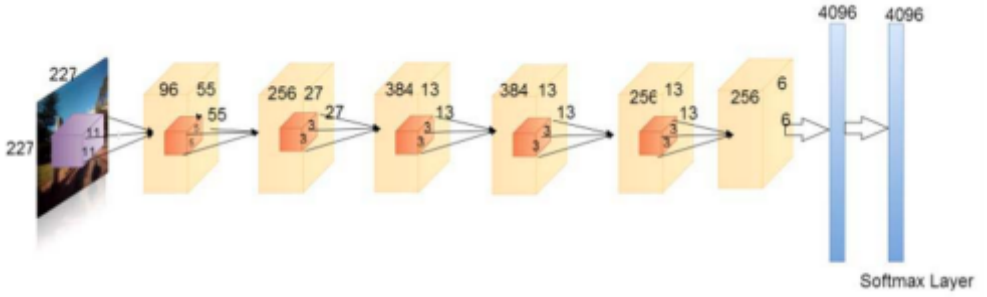

Fig. 2.5 A Convolutional Neural Network (CNN) architecture is shown here. Each Yellow box represents a layer in the CNN architecture consisting of several convolutional filters (orange boxes). The output of these convolutional layers is combined in a Softmax layer to obtain a final descriptor. Figure taken from [102]

Features extracted from CNNs showed promising results on condition- and viewpoint-variant datasets, leading to a paradigm shift in VPR research from traditional handcrafted feature descriptors to neural network activations-based descriptors. Chen et al. [103] used features from all layers of Overfeat Network [104] and integrated it into the spatial filtering scheme of Seq-SLAM. Improving upon CNN-based VPR, Chen et al. [102] trained two neural networks on Specific Places Dataset (SPED), namely AMOSNet and HybridNet. AMOSNet was trained from scratch on SPED while HybridNet initialized weights from top-5 convolutional layers of Caffe-Net. Different off-the-shelf feature encoding methods have been used to create the signature of an image from CNN activations; including cross-pooling [105], holistic pooling [106] and multi-scale pooling [102]. However, in Net-VLAD [107], authors introduce a new VLAD (Vector-of-Locally-Aggregated-Descriptors [108]) layer into the CNN architecture for end-to-end VPR-specfic training, achieving excellent results.

Recently, CNN-based description of images/places using only regions of interest (ROI) showed enhanced performance compared to whole-image description. The work in [109],

namely R-MAC (Regions of Maximum Activated Convolutions) uses max-pooling on cropped areas in CNN layers' features to extract ROI. Chen et al. [3] in Cross-Region-BoW used the CNN layers behaving as high-level feature extractors to identify salient regions in an input image which were subsequently described by low-level feature encoding convolutional layers. This work was followed-up with a flexible attention-based model for region extraction [110]. Khaliq et al. [111] combine VLAD with ROI-extraction to show significant robustness to appearance and viewpoint variation. Authors in [112] propose a multi-scale, non-rigid, pyramidal fusion of local features as extracted from CNNs for VPR.

A convolutional auto-encoder network is trained in an unsupervised fashion by Merrill et al. [113], utilising HOG-descriptors of images and synthetic viewpoint variation. Authors in [114] present a compact neural-architecture for VPR, which has been shown to work well for previously-observed but conditionally changed (weather, seasons, time-of-day etc.) traverses. An ensemble-based approach to VPR is shown by [115], where a framework is designed to sequentially use the place matching proposals from 6 different VPR techniques.

An interesting approach is adopted by [116] for VPR-based localisation in underground tunnel environments, where sequential inter-image and intra-image similarity is employed for localisation. This work is further extended in [18], where now a front-facing camera and an upward looking camera are collectively used for accurate localisation. Authors improve the accuracy of their front-camera-based coarse localisation system by training a neural-network for homography estimation in the sequential images obtained from the upward-facing camera. This work is then followed-up with [117], by creating mosaics of images obtained from the upward-facing camera and using these mosaics for much accurate localisation.

While most of these works have been explored specifically for VPR, some recent techniques including SuperPoint [118] and D2-net [119] propose generic, deep-learnt, sparse descriptors that are robust across various conditional changes. Authors in [120] [121] have formulated VPR as a two-stage process: 1) global matching-based, less-intensive place matching candidates selection 2) local features-based, intensive final candidate selection with focus on spatial constraints. Other interesting approaches to place recognition have also been adopted, including semantic segmentation-based VPR (as in [122] [123] [124]) and object proposals-based place recognition [125]. For images containing repetitive structures, Torii et al. [126] proposed a robust mechanism for collecting visual words into descriptors. Synthetic views are utilised for enhanced illumination invariant VPR in [127], which shows that highly condition variant images can still be matched if they are from the same viewpoint.

2.6 Visual Place Recognition Datasets

A key reason for the rapid growth of VPR as a field is the widely available and open-sourced datasets. A large number of VPR datasets have been proposed over the past many years. These datasets have their own associated challenges and this sub-section is dedicated to discussing all of these VPR datasets. The author has summarised all of these datasets in Table 6.1, while Fig. 2.6 shows samples of images from these datasets. Most of these datasets are available in the author's GitHub repository[1] for the convenience of VPR researchers. The basic template of these datasets is two folders (a query images folder and a reference images folder) containing multiple images of the same place but under different viewpoint and/or appearance conditions. Sometimes the indices of the images in the two folders represent the same place, but this may not always be the case and therefore usually associating ground-truth information is provided. The number of images in each dataset vary from one usage to another, because sometimes subsets of these datasets are used by researchers. Therefore, in this sub-section the author has referenced the originally proposed datasets and avoided specifying the exact number of images in each of these datasets.

The Gardens Point dataset (introduced in [1]) is one of the most widely employed datasets in VPR for the purpose of testing. This dataset consists of three traverses performed through the Garden Point campus of Queensland University of Technology. Two of the traverses are performed during day time and have lateral viewpoint variation. The third traverse is performed during night time and serves as a good challenge for condition-invariant VPR techniques. The 24/7 Query dataset was introduced in [127] and contains a pair of three images of the same place under different conditions and correspondingly many such pairs of places. This difference in conditions includes 6-DOF extreme viewpoint variation, different times of the day (noon, evening and night) and a large number of dynamic objects. Synthia dataset is a large-scale dataset of various environments traversed in a simulated world environment and has been presented in [128]. This dataset consists of traverses through different types of outdoor environments and within each environment during different seasons. The dataset also has simulated dynamic objects and some lateral viewpoint variation, when occasionally driving lane is changed.

The Cross-Seasons correspondence dataset [129] is built on top of the CMU Visual Localisation dataset [130] and the Oxford RobotCar dataset [131], which each consist of conditional changes resulting from different seasons and times of day. A small-scale indoor dataset of a Corridor consisting of very low-resolution images is presented in [74]. This Corridor dataset has lateral viewpoint variation and no appearance change, but serves as

[1] https://github.com/MubarizZaffar/VisualPlaceRecognitionDatasets

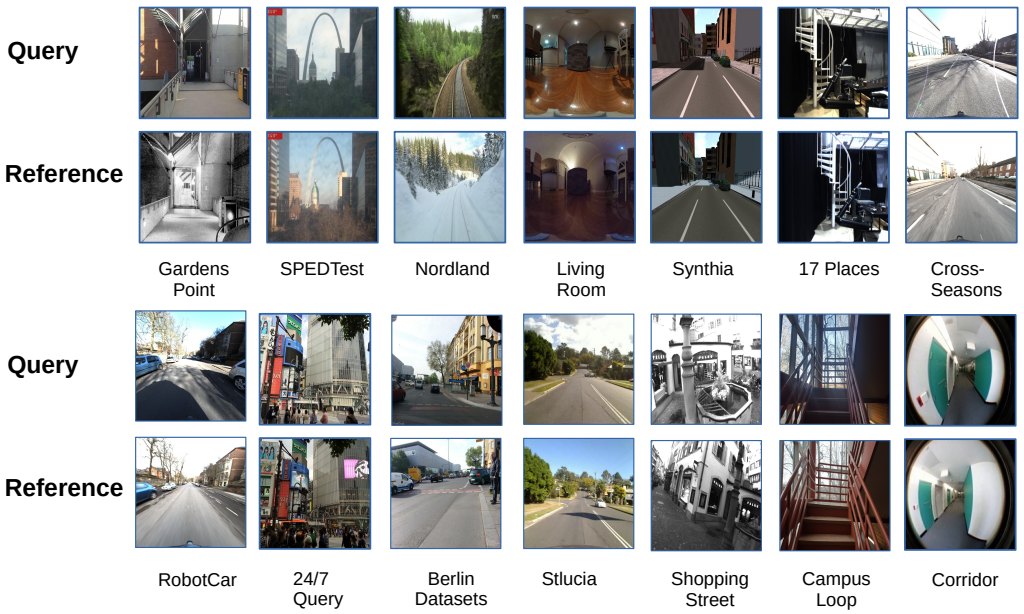

Fig. 2.6 Exemplar images from all the datasets discussed in this thesis are presented here.

a challenge for VPR given low-resolution images. The 17-Places dataset is introduced by [132], which consists of several different indoor scenes, ranging from office environment to labs, hallways, seminar rooms, bedrooms and many other. Authors in [133] introduced the Living Room dataset, consisting of high-resolution, wide field-of-view images obtained from a home service robot. This dataset introduces the world appearance from a different perspective of a close-to-ground mounted camera rather than the usual egocentric viewpoint. This dataset exhibits both viewpoint and illumination variations. The Nordland dataset [4] consists of a train journey through a natural environment in Norway performed during 4 different seasons. The original dataset does not have any viewpoint variation but some usages have employed manual cropping to introduce synthetic lateral viewpoint variation [113] [134].

The Specific Places Dataset (SPED) was introduced in [102] and consists of images of various places as seen through the eyes of CCTV cameras taken under different seasonal conditions. This dataset does not have any viewpoint variation, but serves as a good challenge due to the randomness in conditional changes, i.e, the conditional changes vary significantly in comparison to other datasets which usually have a few particular appearance conditions (e.g, day/night, summer/winter etc). The SPEDTest dataset was introduced by the same author later in [110] and consists of previously unseen images of the same nature, but limited in number for the purpose of VPR evaluation. StLucia dataset consists of a car-journey in

Table 2.1 Visual Place Recognition Datasets

Dataset Name	Environment	Viewpoint Variation	Conditional Variation
Gardens Point	Mixed	Lateral	Day-Night
24/7 Query	Outdoor	6-DOF	Day-Eve-Night
Synthia	Synthetic Urban	Lateral (Occasional)	Seasonal
Cross-Seasons	Outdoor	Lateral	Time and Weather
Corridor	Indoor	Lateral	None
17 Places	Indoor	Lateral	Illumination
Living Room	Indoor	Lateral	Illumination
Nordland	Natural	None	Seasonal
Oxford Robotcar	Urban	Lateral	Time and Weather
SPED	Outdoor (CCTVs)	None	Time, Seasonal and Weather
Berlin Datasets	Urban	Extreme	Time and Weather
Campus Loop	Mixed	Lateral	Seasonal
SPEDTest	Outdoor (CCTVs)	None	Time, Seasonal and Weather
Stlucia	Outdoor	Lateral (Occasional)	Times of Day
Shopping Street	City Center	6-DOF	Illumination and Dynamic Objects

a sub-urban environment during five different times of the day and has been presented in [9]. Campus Loop dataset is a small-scale dataset of a university campus traversed during different seasons (summer and winter) and has been presented in [113]. This dataset also contains lateral viewpoint variation in between the traverses. The Berlin datasets (three different datasets) were introduced in [3] and have been captured from crowd-sourced photo-mapping platform *Mapillary*[2]. The traverses exhibit extreme viewpoint variation compared to all available VPR datasets and also contain conditional changes. Due to its urban nature, dynamic objects such as vehicles and pedestrians are observed in most of the captured frames. The Shopping Street datasets contains the interesting 6-DOF viewpoint change, which has been utilised and serves as the basis for Chapter 5 of this thesis. This viewpoint change has been introduced by mounting the camera on a 4 meter long rod such that the motion of camera imitates the flying behavior of a drone. This dataset also contains significant illumination variation and temporal appearance change.

2.7 Visual Place Recognition Evaluation

In this section, the author presents the different evaluation metrics that have been used within VPR, but also sets the motivation for the evaluation-based work done in Chapter 4 and Chapter 5 of this thesis. Because a large number of techniques have been proposed over the

[2]https://www.mapillary.com/

past many years and due to the apparent zoo of VPR datasets to choose from, establishing state-of-the-art in VPR has been a challenge. Prior to this thesis, no performance evaluation-based work existed for VPR and therefore, it was difficult to compare the strengths and weaknesses of techniques with each other.

In general Precision-Recall curves (PR curves) have been the most utilised mode of statistical evaluation for VPR [8]. These precision-recall curves comprise of finding the precision and recall values of the system at different cut-offs of the matching confidence. The precision and recall of a system at a particular confidence threshold can be calculated as:

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives} \tag{2.1}$$

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives} \tag{2.2}$$

Where, true-positives are the images/places that are correctly matched by a VPR technique based on ground-truth information and false-positives are the images that are incorrectly matched by the VPR technique based on ground-truth information. False-negatives are images that were correctly matched by a VPR technique, but their matching score was lower than the confidence threshold and hence these matches were incorrectly discarded. Usually in VPR, the descriptor matching score (e.g, cosine-matching, L1-Match, L2-Match etc) has been employed as the confidence score, but theoretically it can be another criterion and is an open area of research within VPR. By varying the confidence-threshold (descriptor matching score) from minimum to maximum value, different values of precision and recall are computed, which are then plotted against each other to yield PR-Curves.

Using the PR-Curves, Area-Under-the-Curve (AUC) has mostly been the go-to metric for giving a single quantitative performance value to the corresponding VPR technique. Most of the VPR techniques (e.g, [102] [111] [3] [110] [8] [113]) have used this metric for performance evaluation. Given the different values of Precision and Recall at various confidence thresholds (the author used confidence thresholds at all matching scores of query images), AUC can be computed as below:

$$AUC = \sum_{i=1}^{N-1} \frac{(p_i + p_{i+1})}{2} \times (r_{i+1} - r_i) \tag{2.3}$$

$$\begin{aligned} where;\ N &= No.\ of\ Query\ Images \\ p_i &= Precision\ at\ point\ i \\ r_i &= Recall\ at\ point\ i \end{aligned}$$

Recall at 100% Precision has also been employed as a relevant metric for VPR, especially when false-positives are unacceptable for the system, but [8] proposes that the availability of false-positive prediction systems advocates for the use of AUC rather than Recall at 100% Precision. More recently, [135] provide a counter argument against the usage of AUC as an evaluation metric and present a new metric, namely 'Extended Precision (EP)'.

$$EP = \frac{P_{Rmin} + R_{P100}}{2} \tag{2.4}$$

In equation 2.4, R_{P100} is the maximum value of recall at which precision is equal to 1, i.e, Recall at 100% Precision, while, P_{Rmin} is the value of precision at the minimum value of recall and generally represents the highest possible value of precision. Please note that the minimum value of recall in a PR-curve may or may not be zero.

In addition to the metrics focused on the matching performance of a VPR technique, computational and storage needs are also important for real-world applications. For VPR, the most discussed computational metric has been the image retrieval time of a technique, i.e, how quickly can this VPR technique retrieve a match given a query image. This retrieval time is then further divided into feature encoding time (t_e) and descriptor matching time (t_m). The feature encoding time represents the time taken by a particular VPR technique to compute the feature descriptor of a query image and the descriptor matching time is the time taken by this technique to match this query descriptor with a pre-computed reference descriptor.

Additionally, run-time memory (RAM) consumption of a VPR technique and the dependency on GPU are also computationally-relevant. The feature descriptor footprint/size is also an important metric for storage needs, because it scales linearly with the map-size.

2.8 Summary

This chapter presented a detailed review of the research performed for localising an autonomous robot and with exclusive focus on Visual Place Recognition-the subject of this thesis. In summary the below few points were covered.

1. An overview of research in Simultaneous Localisation and Mapping (SLAM) from both software and hardware perspectives.

2. Co-relating and distinguishing closely related concepts within localisation, including Visual-SLAM, VPR, visual-odometry, image matching, visual-localisation and the correspondence problem.

3. An overview of research within Semantic Mapping-the ground-work for **Chapter 3** of this thesis.

4. A detailed literature survey of Visual Place Recognition (VPR), the VPR techniques and datasets proposed to date and the evaluation metrics utilised within VPR. This serves as the basis for **Chapters 4 and 5** of this thesis.

5. Identification of the explicit challenges within VPR for the various classes of VPR techniques and their corresponding strengths. These challenged serve as the motivation and the gap addressed by **Chapter 6** of this thesis.

   The existing literature in VPR is primarily focused on proposing new VPR techniques that achieve state-of-the-art performance on some datasets for a given choice of metrics and within a custom suite of comparison techniques. Most of the newly proposed techniques are deep-learning-based and therefore environment-specific and computationally-expensive. A very limited comparison of VPR techniques exists in literature and the focus has almost always been on retrieving a correct match for a VPR technique, rather than evaluating the quality of the query image for VPR before any such retrieval is performed. Therefore, in this thesis, the author has tried to bridge these gaps and the following chapters present how these existing gaps have been covered with due diligence of the respective limitations.

Chapter 3

Memorable Maps: A Framework for Redefining Places in Visual Place Recognition

One of the critical challenges for VPR systems, as identified in Chapter 1, is the perceptual aliasing and non-salience of images or content within images. The ability to model and compute this non-salience can have applications for VPR systems and prevent potential false-positives. Therefore, in this chapter, the author presents a cognition-inspired agnostic framework for building a 'memorable map' for Visual Place Recognition (VPR). This framework draws inspiration from human-memorability, utilises the traditional image entropy concept and computes the static content in an image; thereby presenting a tri-folded criteria to assess the 'memorability' of an image for VPR. A dataset namely 'ESSEX3IN1' is created, composed of highly confusing images from indoor, outdoor and natural scenes for analysis. When used in conjunction with state-of-the-art visual place recognition methods, the proposed framework provides significant performance boost to these techniques, as evidenced by results on ESSEX3IN1 and other public datasets. This chapter presents all the details of the developed framework and an extensive analysis of this framework.

3.1 Background

Chapter 1 discussed that VPR is a well-defined, albeit a highly challenging module of a Visual-SLAM based autonomous system [8] and that a VPR system needs to be robust to viewpoint and conditional changes. Chapter 2 then presented a number of research works that have handled these challenges. However, perceptual aliasing and VPR-specific salience

of the content within an image has mostly been unexplored, which sets the stage for the research presented in this chapter.

Generally, VPR can either be used as a stand-alone vehicle localization system in an appearance-only topological and/or topometric map or it can be combined with metric SLAM techniques to perform loop closure [66]. The scope of this chapter and the evaluations is limited to the former, however, it is possible to adopt the combination of this work and VPR within SLAM systems for loop-closure. Some key advances in SLAM research can be broken down into semantic mapping (surveyed in [58]) and VPR (surveyed in [8]), where the latter can be annexed into the former [58].

Traditionally, for VPR, 'Places' have been selected/sampled based on time-interval [136], distance [60] or distinctiveness [61] in different approaches. Most of these methods attempt to reduce the size of robot's map and do not quantify if a sampled/sub-sampled image is a good representation of a place; thereby has a greater chance of matching upon revisiting. The quality of image selection mechanism restricts the performance of a VPR system, both in the short-term and long-term. Due to limited number of images being stored in the map, it is critical to select those images that can be matched successfully upon repeated traversal-the motivation for this research.

Therefore, in this chapter, the author looks at image selection from a semantic point of view and draws inspiration from images memorable to a human-cognition system. A Convolutional Neural Network [137] is used to compute the memorability of an incoming camera frame. However, while objects like vehicles and pedestrians in an image are subjectively-memorable; they are intrinsically not good for VPR as these dynamic objects are rarely re-observed. The author thus performs object detection to compute the staticity of an image and mask memorability of dynamic content. In addition to being memorable and static, an image should be content-rich thus the entropy map is calculated.

The contribution of this chapter is a semantically coherent framework (Fig. 3.1) that filters an input image through a tri-folded criteria. Hence, ensuring that every image to be inserted against a place in robot's map is a good representation of the said place and highly recognizable. To analyze the effectiveness of this framework, a dataset 'ESSEX3IN1' is created from indoor, outdoor and natural environments. Unlike existing VPR datasets, ESSEX3IN1 mimics a robot exploring an environment instead of traditional path-following and is thus composed of highly confusing images from all three environments. The author shows how these confusing images lead to poor performance of current VPR systems. The final results show the effectiveness of proposed framework in segregating these 'confusing' images from 'good' images, thereby increasing VPR precision and reducing database size. The framework is also evaluated on other public VPR datasets to show that this performance

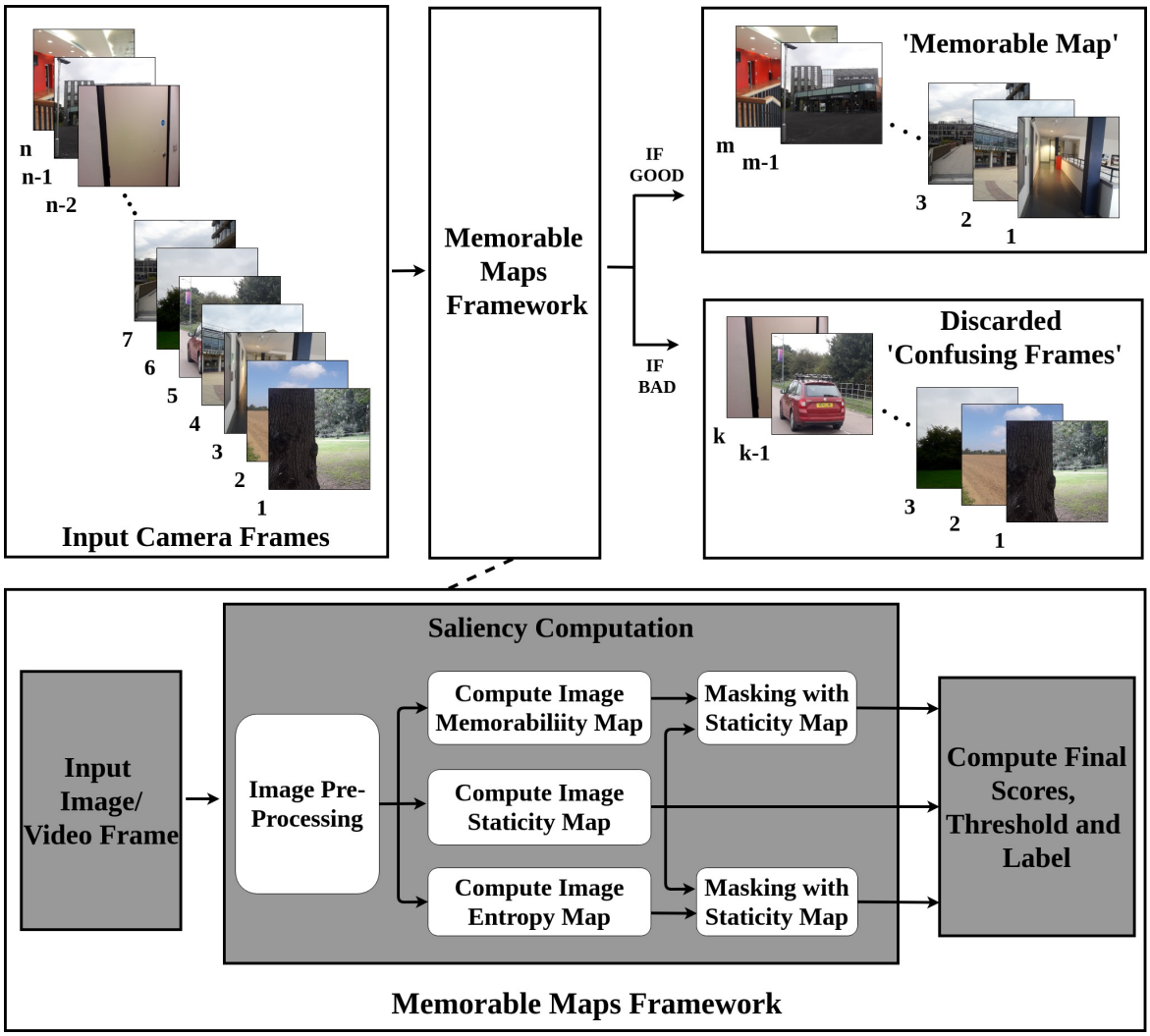

Fig. 3.1 A block-level overview of the proposed memorable maps framework is shown here.

enhancement can be generalized. Due to its agnostic nature, any VPR technique can obtain a performance boost by stacking the presented framework as an additional layer in the VPR pipeline.

In addition to the semantic mapping literature review in Chapter 2, the author discusses two works that have similar motivation to the proposed approach. The interesting work by Hartmann et al. [138] proposes a random forest classifier of 5 decision trees trained on a dataset of 455 outdoor images. The objective of this random forest is to find keypoints in an image with low matchability and subsequently discarding them. This technique is computationally intensive in comparison to the proposed methodology as we compute a single matchability (memorability) score against an image instead of scores against each

image keypoint. Moreover, in VPR, features coming from dynamic objects and low-textured scenes are usually not re-observable/matchable (as shown later in this Chapter) but have not been examined in [138]. Although vegetation is considered to belong to non-matchable category, results show features coming from trees as being classified as matchable in [138]; which usually in VPR contribute negatively to the distinctiveness of a place (as shown in Fig. 3.3). More recently, a CNN able to classify input frames as stable/unstable is trained by Dymczyk et al. [139] for long term visual place recognition. Similar to [138], this work also proposes that vegetation in outdoor scenes is not suitable, but does not consider outdoor dynamic objects like cars, pedestrians, animals etc. Also, informativeness of stable frames in terms of extracted features and predicted matchability is not inspected given that walls are selected as stable elements of an image. Therefore, to the best of author's knowledge, this work combines for the first time all three of these criteria namely memorability, staticity and entropy to create memorable maps.

3.2 Methodology

This section presents in depth the framework developed in this work. A sub-section is dedicated to each of the three criteria (i.e., memorability, staticity and entropy) adopted by the framework. The author also discusses the integration of this framework with VPR techniques as a final sub-section.

For the purpose of evaluation and analysis, the author has used AMOS-Net [102], Hybrid-Net [102] and Region-VLAD [111] as the VPR techniques throughout this chapter. The details of these techniques are given in Section 3.3.2.

3.2.1 Memorability

Why Memorability?

The human-cognition system is powerful in evaluating what images are useful to be stored in the brain's memory fragments [140] [141]. We usually remember concrete structures like buildings, streets, squares etc. However, more natural scenes like fields, forests, grassy plains and far out sceneries are less memorable. This 'memorability' concept is also intuitive as it is easy to confuse different natural scenes with each other compared to concrete structures. In reference to VPR, there are two further reasons for the non-salience of trees, vegetation and natural scenery: 1) They are highly appearance-variant compared to concrete structures, 2) Local features coming from trees and vegetation etc. are usually mismatched, as explored in the works of [142] [143] [144]. In order to explain (1), the author has shown samples of

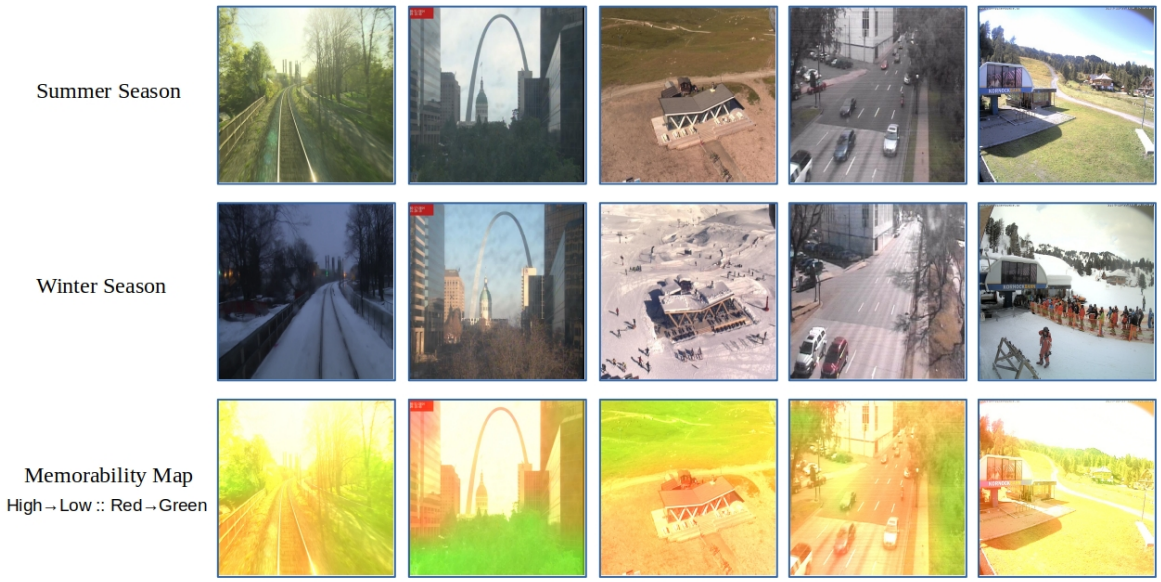

Fig. 3.2 Concrete structures appear similar under seasonal changes while non-memorable elements like trees, vegetation and natural scenery appear very different. The memorability maps (in last row) show the effectiveness of proposed memorability implementation (sub-section 3.2.1) in segregating concrete structures from these appearance variant regions.

appearance changes in Fig. 3.2, along with the memorability maps created (methodology explained in sub-section 3.2.1) by this framework. All example images in Fig. 3.2 have been utilised from the Nordland dataset [4] and SPEDTest dataset [110] to ensure consistency with the evaluation mechanism. In Fig. 3.3, the author has also shown how non-memorable scenes are mismatched by state-of-the-art VPR systems leading to false-positives.

Memorability Implementation

Inspired from human-memorability, the author applies the work done originally for marketing and advertising in [137] to VPR problem. A Convolutional Neural Network namely Hybrid-CNN [145] which was originally trained on Places365 dataset [145] for deep learning-based scene recognition, has been fine-tuned on LaMem dataset by Khosla et al in [137]. The authors in [137] have named this fine-tuned Hybrid-CNN as 'MemNet'. The LaMem dataset (introduced by [137]) is composed of 60,000 images covering multiple scenarios ranging from natural scenery, indoor scenes, outdoor scenes and distinctive objects. The ground-truth human-memorability provided in LaMem dataset has been computed for each of the images using an interactive game played by multiple human subjects. Images are shown to players in a sequence and are repeated after a random interval where a human has to identify/recall

Fig. 3.3 Natural places mismatched by VPR methods due to confusing features coming from trees, grass and plains are shown here. Red boundary represents an incorrect match while green boundary represents a correct match. All images in this figure were found static and information-rich, i.e., human-memorability is the only criterion that can discard these images.

a previously seen image. By using this LaMem dataset, the authors [137] fine-tuned the Hybrid-CNN achieving a high co-relation (0.64) with human memorability. Resultingly, the output of this fine-tuned Hybrid-CNN (MemNet) is a human-memorability score m for each input image in the range of $0-1$, with $m=1$ being the most memorable. However, for the proposed framework, a memorability map is required (as in Fig. 3.6) against every image instead of a single memorability score as output by MemNet. The motivation for this memorability map is to cater for highly memorable but dynamic objects as discussed later in this sub-section and utilised in sub-section 3.2.4.

The CNN input layer size is set to $W1 \times H1$. The author re-sizes every incoming image to $W2 \times H2$.

$$where; \quad W2 = a \times W1, \;\; H2 = b \times H1$$

Then this rescaled-image is split into C (*where*, $C = a \times b$) non-overlapping crops of size $W1 \times H1$ each and which are sequentially fed as inputs to CNN. This in turn yields the memorability matrix M as shown below.

$$M = \begin{bmatrix} m_{11} & m_{12} & \dots & m_{1b} \\ m_{21} & m_{22} & \dots & m_{2b} \\ \vdots & \vdots & \ddots & \vdots \\ m_{a1} & m_{a2} & \dots & m_{ab} \end{bmatrix}$$

Where, m_{ij} is the memorability of each $W1 \times H1$ cropped image. To create a memorability map, the author rescales the matrix M from $a \times b$ to $W2 \times H2$ with bilinear interpolation. Some examples of memorability maps overlayed on images are shown in Fig. 3.2 and Fig. 3.6. The author has employed $C = 5 \times 5$ through-out this work and a parametric variation of this is shown later in sub-section 3.4.6. It can be seen (in Fig. 3.2 and Fig. 3.6) that vegetation, natural scenery and trees are identified as less-memorable which is consistent with the motivation in sub-section 3.2.1. However, for human cognition (and therefore for [137]), objects such as faces, animals and vehicles are memorable. But, such dynamic objects are not re-observable and therefore, they are not salient for VPR; the author caters for this in the following sub-section.

3.2.2 Staticity

Why Staticity?

The previous sub-section shows how memorability is a good evaluation criterion for a camera frame to be used in VPR. However, one limitation is the fact that objects like cars, pedestrians, buses, animals and bicycles in an image are all classified as highly memorable but are not re-observable (for VPR problem). Resultingly, images that may be memorable but have high dynamic content will fail to match upon repeated traversal. Fig. 3.4 shows some of these images mismatched by VPR techniques [102] [111].

Staticity Implementation

To cater for highly dynamic images, the author performs image segmentation into static and dynamic pixels. All input images are re-sized to $W2 \times H2$. An object detector is used [146] that can detect 80 different classes of objects in an image. Out of these 80 classes, 21

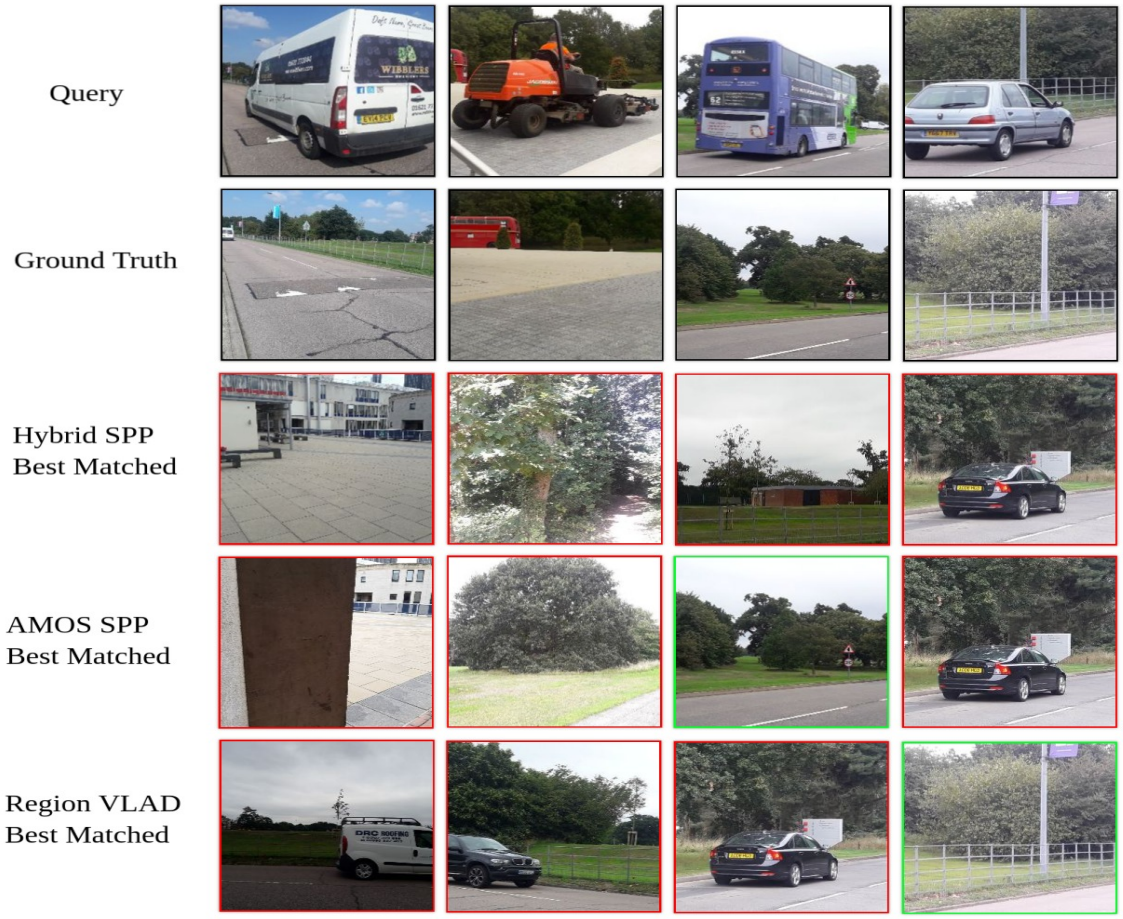

Fig. 3.4 Dynamic places mismatched by state-of-the-art VPR systems. Features coming from vehicles are not re-observable in addition to the occlusion caused by them in different scenes. Red boundary represents an incorrect match while green boundary represents a correct match.

correspond to highly-dynamic, commonly-observed objects. These dynamic objects include cars, pedestrians, buses, trucks, animals etc. The author, therefore, only considers proposals of bounding boxes coming from objects of interest, i.e., dynamic objects. Please note that the default parameters of YOLO [146] were used in this work.

Since the staticity map is computed for each pixel in the image, it can be represented as a staticity-matrix S of size $W2 \times H2$ as below.

$$S = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1W_2} \\ s_{21} & s_{22} & \cdots & s_{2W_2} \\ \vdots & \vdots & \ddots & \vdots \\ s_{H_2W_2} & s_{H_2W_2} & \cdots & s_{H_2W_2} \end{bmatrix}$$

$$\begin{aligned} \textit{where}; \quad & \{s_{ij} \in \mathbb{Z}_2 \mid \mathbb{Z}_2 = [0,1]\} \\ & s_{ij} = 1 | \textit{Pixel} = \textit{Static} \\ & s_{ij} = 0 | \textit{Pixel} = \textit{Dynamic} \end{aligned}$$

Fig. 3.6 shows the typical staticity map computed in our framework. However, although an image containing low-textured scenes (walls/door/pillars) can be classified as concrete (memorable) and static but it does not have distinguishable features and hence, it is not distinct. This limitation is accommodated in the following sub-section.

3.2.3 Entropy

Why Entropy?

An input camera frame containing a room/lift door is commonly observed by a robot navigating indoors. Such a frame is classified as memorable and static, but has little to no information differentiating it from other doors in the building, thus leading to false positives. The same can be extended to any other frame with occlusion resulting from walls, pillars etc. Examples of such confusing frames are shown in Fig. 3.5.

Entropy Implementation

To avoid less informative or occluded frames, the information content of an image is evaluated by computing its local entropy against every image pixel. This local entropy corresponds to the number of bits required to encode the local gray-scale distribution in an image. Based on standard boolean algebra, the number of bits required to represent any positive integer value can be computed by $log_2(\textit{Numerical\_Value})$. A circular window of r pixels radius is used as the local neighbourhood to get the entropy map of an incoming camera frame against each pixel. The total number of histogram bins used for entropy computation are 256 corresponding to $0-255$ gray-scale intensity values. The generic algorithm for entropy map computation is shown below and adapted from [147].

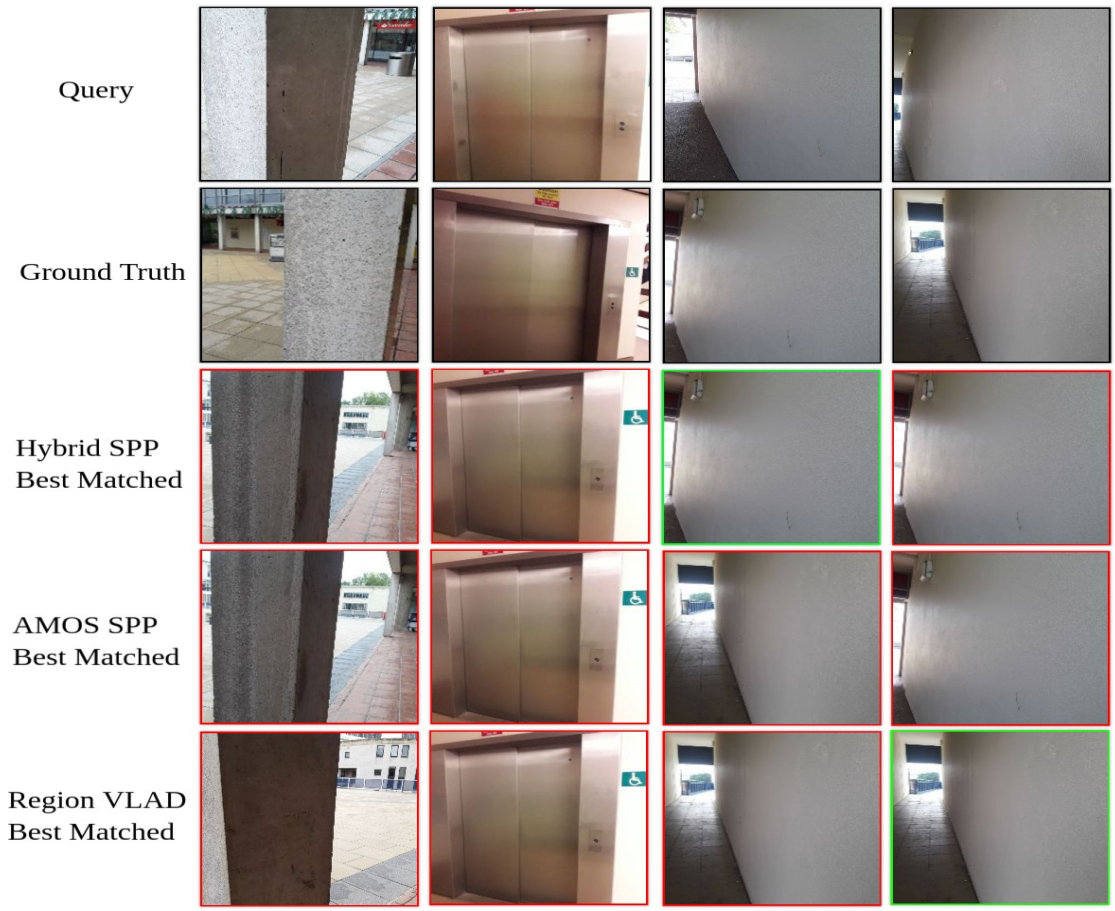

Fig. 3.5 Low-entropy places mismatched by state-of-the-art VPR methods can be commonly observed in indoor robot navigation datasets. Along with intrinsically less-informative images of doors/walls, static occlusion can also lead to poorly defined places. Red boundary represents an incorrect match while green boundary represents a correct match.

Algorithm Computing entropy map

Create a Histogram of 256 *Bins*
for all *Local Neighbourhoods in Image* **do**
  for all *Pixels in Current Neighbourhood* **do**
    if *Current_Pixel lies in BinX* **then**
      $Items\_in\_BinX = Items\_in\_BinX + 1$
    end if
  end for
  $Local\_Entropy = log_2\,(No.\,of\,Filled\,Histogram\,Bins)$
  Clear all Histogram Bins
end for

This algorithm gives the entropy map represented as matrix E of size $W2 \times H2$. Local circular regions of images containing texture-less doors/walls have a small range of intensity gradients within the region and thereby have lower entropy value. The maximum value of entropy is computed from equation (1) and equals 8, given that the maximum number of filled histogram bins is 256. Fig. 3.6 shows examples of entropy maps computed by the proposed framework. The author has used $r = 5$ in this work, where the reasons for this selection and parametric variation are shown in sub-section 3.4.6.

$$MaxEntropy = log_2\left(No.\,of\,Histogram\,Bins\right) \tag{3.1}$$

$$E = \begin{bmatrix} e_{11} & e_{12} & \dots & e_{1W_2} \\ e_{21} & e_{22} & \dots & e_{2W_2} \\ \vdots & \vdots & \ddots & \vdots \\ e_{H_2W_2} & e_{H_2W_2} & \dots & e_{H_2W_2} \end{bmatrix}$$

$$where; \quad \{e_{ij} \in K \mid K \subseteq \mathbb{R} \wedge K = \{0,\dots,8\}\}$$

Fig. 3.6 The three types of image maps created by the proposed framework for evaluating the content of an input image. Concrete structures like buildings and roads are memorable in comparison to grassy plains and trees [Top]. Cars, pedestrians and other dynamic objects are detected and evaluated for the amount (approximate) of pixels they occupy [Middle]. Uniform and texture-less scenes, sky portions have low-entropy compared to feature rich structures [Bottom].

3.2.4 Computing Scores and Thresholding

After acquiring all three maps of an image, the author masks memorability map and entropy map with staticity map. This ensures that the decision to select an image based upon memorability and entropy is immune to the information coming from dynamic objects. Next, the author computes the memorability score (MS) of an image as the average value of memorability map and compares it with a memorability threshold (MT) to evaluate if this image/frame is memorable enough for use in VPR. Secondly, we compute the percentage of static pixels in our staticity map to get a staticity score (SS). This is then contrasted with staticity threshold (ST) to decide if an incoming frame has enough static content to be inserted into the map. Thirdly, the author calculates the average value of entropy map and scales it with the maximum value of entropy to get the percentage of information content. This percentage dubbed as the entropy score (ES) is compared with the entropy threshold (ET) to settle if an input frame has enough information. Please note that for computing the average values of staticity-masked memorability-map and entropy-map, the normalisation is done over the original map-size (i.e. $W2 \times H2$), because masking actually leads to a lack of information and should therefore be penalised.

Finally, a tri-input AND criteria is used to select images that are memorable, static and information-rich to be inserted into the memorable map.

Algorithm Image Selection For Memorable Map

for all *Incoming Images* **do**

Compute All Three Image Maps

$$MS = \frac{1}{W2 \times H2} \sum_{i,j=1,1}^{W2,H2} m_{ij}$$

$$SS = \frac{1}{W2 \times H2} \sum_{i,j=1,1}^{W2,H2} s_{ij}$$

$$ES = \frac{1}{W2 \times H2 \times 8} \sum_{i,j=1,1}^{W2,H2} e_{ij}$$

if $MS \geq MT$ & $SS \geq ST$ & $ES \geq ET$ **then**

Insert into Map

else

Dicard Image

end if

end for

3.2.5 Integration of Memorable Maps and VPR Techniques

The integration of proposed framework with VPR techniques is seamless and straight-forward. The core component that all VPR techniques require to operate is a reference image database, using which the VPR techniques propose a place-match (image-retrieval) given an input query image. The creation of this reference database by employing the memorable maps framework instead of the traditional time-based or distance-based approaches is what brings our framework together with state-of-the-art VPR techniques. This integration of memorable maps framework with the VPR methods can be in an online or an offline fashion.

In an offline approach, where *a priori* knowledge of the environment is available in the form of images, memorable maps framework can take this knowledge (images) and output a memorable map as depicted in Fig. 3.1. In this case, the 'Input Camera Frames' block of Fig. 3.1 represents the input knowledge where each image is indexed in a sequential manner and evaluated by our framework yielding a memorable map. The contemporary VPR techniques

can then use this memorable map instead of the original time-based, distance-based or distinctiveness-based reference image database, achieving place matching performance boost and map-size reduction as reported later in Section 3.4.

Before discussing the integration of proposed framework in an online manner, it is important to understand that every query image in an online VPR system becomes a reference image at the next time step and is stored in the reference image database. Thus, for every input query image two operations are traditionally performed: 1) It is input to a VPR technique to search for a prospective place match, 2) If it matches to a previously known place, it is stored as an additional representation of the place and if it does not match to a previously known place, it is stored in the map as a 'new place'. Given this understanding, the memorable maps framework can easily be integrated into an online VPR system, where the input query image is first evaluated for its salience by our framework. If it is largely memorable, static and information-rich, only then it is used for VPR and subsequent storage in the reference map. For the online case, images in Fig. 3.1 would represent query images such that their indices represent time-stamps.

3.3 Experimental Setup

This section discusses the datasets, VPR techniques and evaluation metric used in the analysis of this chapter. Although these have been previously touched very briefly in the literature review, the author has explained each of these further in detail here for ease of reproducibility and to avoid any utilisation confusions. A new dataset ESSEX3IN1 is presented, which is publicly available [1]. Additionally, the author further discusses three previously introduced public datasets used for reporting the framework's performance. The VPR techniques used for the results and analysis are then summarized. The author uses AUC for reporting results which is a well-established performance metric for VPR techniques as previously discussed.

3.3.1 Evaluation Datasets

This sub-section introduces the 4 datasets that have been used in this work to discuss and analyse the performance of memorable maps framework. Please note that none of these datasets were used for training the 3 VPR techniques employed in this work.

[1] https://github.com/MubarizZaffar/ESSEX3IN1-Dataset

ESSEX3IN1 Dataset

From sub-section 2.6, it is clear that most of the Visual Place Recognition datasets have been created from a pre-planned path traversal. Thus, these datasets do not contain confusing images that an exploration robot may come across. Also, these datasets focus on a single type of environment either indoor or outdoor. To evaluate and challenge the memorable maps framework, the author has created a new dataset ESSEX3IN1 which is composed of images from indoor, outdoor and natural scenes.

The dataset was created in two stages using a human-held mobile phone camera at the University of Essex (Colchester Campus) and contains 210 query images and 210 reference images with viewpoint variations. In the first stage, the objective was to take images from all sorts of environments that were either 'confusing' or didn't qualify the definition of a 'distinct Place', where this indistinctness of a place refers to perceptual aliasing. Two-thirds of the images in ESSEX3IN1 are from this first stage. The second stage, consists of images that were not confusing and could be defined as 'distinct places'. One-third of the total images are from this second stage. Some images from these stages are shown in Fig. 3.7. The ground-truth data provides information about a single correct reference image against every query image. This ground-truth is created manually by looking at individual images such that the ground-truth pair of query and reference image represent the same geographic location in the world.

It is important to note that none of these images were used in tuning the three thresholds and were not seen prior by the proposed framework. The collection of dataset in this two-staged manner was useful for analysis in Section 3.4.

Nordland Dataset

The Nordland dataset [4] comprises of a train journey through Norway and is collected in four different seasons with frame-to-frame ground-truth correspondence. The author uses a subset of this dataset which consists of 1622 query images and 1622 reference images. The query images are from the traversal performed in summer whereas the reference images are from winter. Although this dataset does not provide any viewpoint variation, but has significant conditional variation. A retrieved image n is considered true-positive if the original ground-truth is between $n-1$ to $n+1$, i.e., each query image has 3 ground-truth references.

St. Lucia Dataset

The St. Lucia dataset was first introduced in [9]. It was recorded in the surroundings of University of Queensland's St. Lucia campus during multiple times of the day. This dataset

Fig. 3.7 Sample images from ESSEX3IN1 dataset. The first stage [on the left hand side] images contain occlusions, dynamic objects, information-less frames and non-memorable content like plains, natural scenery, vegetation and trees. In contrast, the second stage [on the right hand side] contains semantically identifiable and distinguishable images of various places from University of Essex (Colchester campus).

consists of moderate viewpoint and illumination variation. The dataset also contains dynamic objects and scene variation. The ground-truth is derived manually from GPS data such that each query image has three reference images as true-positives. The total number of query images is 1261 and the total number of reference images is 1317.

SPEDTest Dataset

The SPEDTest dataset was introduced in [110] and is a sub-set of the original Specific Places Dataset [102]. It consists of 607 query images coming from a variety of scenes and environments. Frame-to-frame correspondence is available as the ground-truth.

3.3.2 VPR Techniques

The author has used three state-of-the-art VPR techniques (namely AMOS-SPP, Hybrid-SPP and Region-VLAD) [102] [111] that have shown promising results in recent research. AMOS-Net is a modified Caffe-Net [148] with all parameters trained on SPED dataset. Hybrid-Net is another modified version of Caffe-Net with weights for top 5 convolutional layers initialized from Caffe-Net. The author has used Spatial Pyramidal Pooling as a feature descriptor for both AMOS-Net and Hybrid-Net since it shows excellent results as compared to other feature encoding methods. Features are extracted from 'conv5' layer in case of both Amos-Net and Hybrid-Net. The third VPR technique, Region-VLAD, uses features extracted from selected/interesting regions of an AlexNet pre-trained on Places365 dataset [145]. Vector-of-Locally-Aggregated-Descriptors [108] is subsequently used for encoding the extracted features. In case of Region-VLAD, the author uses features from 'conv4', number of regions-of-interest as 400 and a visual dictionary size of 128. Evaluation of VPR techniques on existing datasets is an offline process, therefore the integration of the memorable maps framework with these techniques is in accordance to the discussion for an offline VPR system in sub-section 3.2.5.

3.3.3 Evaluation Metric

The extensive review of VPR research performed by Lowry et al. in [8] and the VPR research community [110] [3] [107] [103] [102] [134] [149] in general agree that a highly precise VPR system with high recallability is required, which serves as the author's motivation to adopt AUC as an evaluation metric. This chapter ensures consistency and fair comparison of AUC scores for different VPR methods on all datasets by computing and reporting results only using equation 2.3.

3.4 Results and Analysis

This section presents the results and analysis in a sequential manner. The author first shows that images collected from the first stage of ESSEX3IN1 actually lead to poor performance of VPR systems and are not good for insertion into a robot map. Secondly, the author shows the segregation performance of proposed framework on these 'confusing' images and 'good' images. Thirdly, the AUC improvement of different VPR systems is presented when plugged with the proposed framework on all datasets discussed in sub-section 3.3.1. This is followed-up with a sub-section dedicated to qualitative analysis showing sample images selected and discarded from all datasets. The author then highlights the contribution of each

Fig. 3.8 VPR false positives upon evaluation on ESSEX3IN1 stage: 1. Images with cars, trees and natural scenes are mismatched. Additionally, images with low information and memorability are almost indistinguishable for even human cognition.

framework criterion qualitatively and quantitatively. Next, the author reports the effect on VPR performance by sweeping framework parameters within possible range. The author shows in the next sub-section, how this framework leads to reduced map size and place matching time. Finally, the integration of Spatio-Temporal filtering with proposed framework is presented to avoid large image gaps for localization.

3.4.1 Contemporary VPR Systems on ESSEX3IN1 Stage 1

The majority of VPR false positives against ESSEX3IN1 are from the first stage of dataset collection. This is due to the confusing images of fields, trees, doors, cars etc that lead to perceptual aliasing. Some of these false positives are shown in Fig. 3.8.

The author shows the AUC performance of VPR systems separately on Stage 1 and Stage 2 in Fig. 3.9.

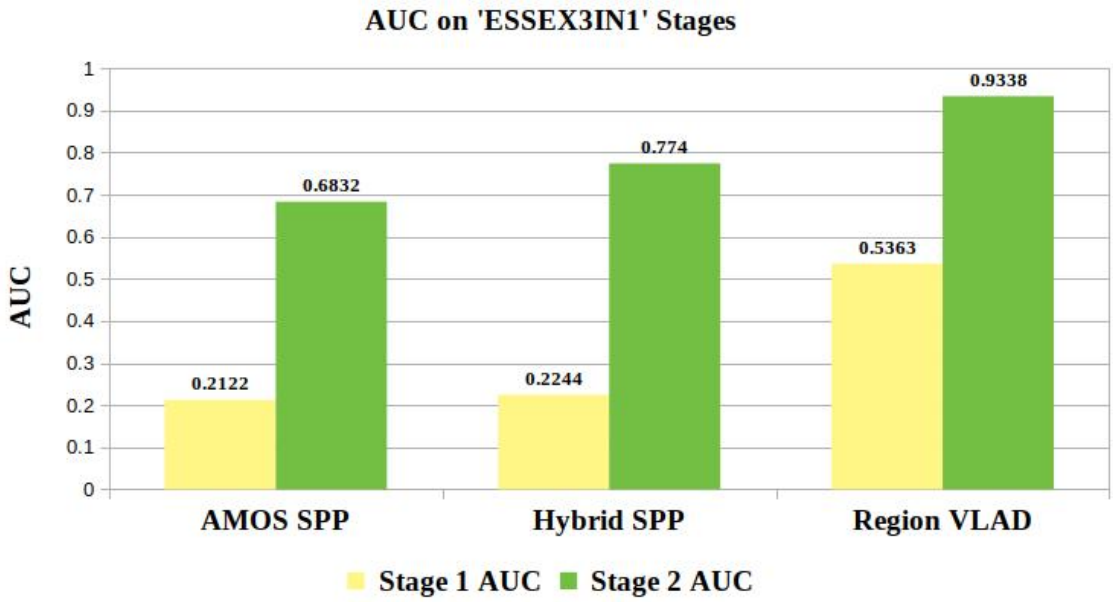

Fig. 3.9 Separate evaluation of VPR methods on each of the ESSEX3IN1 stages reveals the challenge faced by contemporary VPR techniques for matching low-entropy, low-memorability and dynamic images.

3.4.2 Segregation Performance of Proposed Framework

For this sub-section, the author applies the proposed memorable maps framework on complete and randomized ESSEX3IN1 dataset. The below thresholds are used to segregate and discard 'confusing' images from 'good' images.

$$\text{Memorability-threshold} = 0.5$$

$$\text{Staticity-threshold} = 0.6$$

$$\text{Entropy-threshold} = 0.4$$

These values for the thresholds were selected from analysis on pre-existing public VPR datasets. Increasing these thresholds reduces the number of images inserted into the memorable map. This is shown in Fig. 3.10 by varying each threshold from $0-1$, while setting the other two equal to 0, i.e, inactive. The manual selection of these particular values is based on the detailed analysis provided in sub-section 3.4.6. Briefly, these particular values were employed for 3 reasons: 1) Agnostic performance boost across all 4 datasets (refer sub-section 3.4.3)., 2) Reasonable number of 'good' images are left in the database as loop-closure candidates (refer sub-section 3.4.7)., 3) Deviating significantly from these

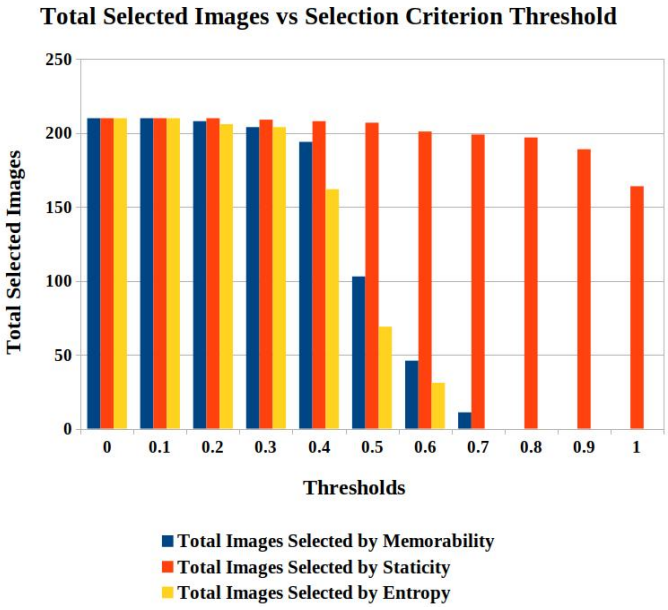

Fig. 3.10 The decrease in total selected images as each selection criterion is increased can be observed here for the whole ESSEX3IN1 dataset. Majority of the ET/MT based image selection is done between $0.4 - 0.7$. Purely static images (without vehicles, human and animals) exist in the dataset which is why $ST = 1$ does not reduce map size to zero.

values could lead to zero or negative changes in the AUC (refer sub-section 3.4.6). Setting any of the 3 thresholds equal to 0 will disable the corresponding criterion, e.g., in a continuous highly crowded scene, the ST can be disabled or the value of MT can be decreased for a continuous agricultural/natural environment. Increasing the thresholds towards 1 will result in decreased no. of images in the database, which will have higher salience.

The new database created by presented framework consists of memorable, static and informative images, thus dubbed as a memorable map. The author shows in Fig. 3.11, how many of the total images selected by presented framework are from which stage of the dataset.

3.4.3 AUC Improvement of VPR Systems

By selecting images that are memorable, static and have a higher entropy, the memorable maps framework gives performance boost to state-of-the-art VPR techniques. Here, the author used fixed thresholds, as in previous sub-section 3.4.2, but an AUC sweep across these thresholds is presented later in sub-section 3.4.6. AUC evaluation is performed on the entire (both stages combined randomly) ESSEX3IN1 dataset along with the three public VPR

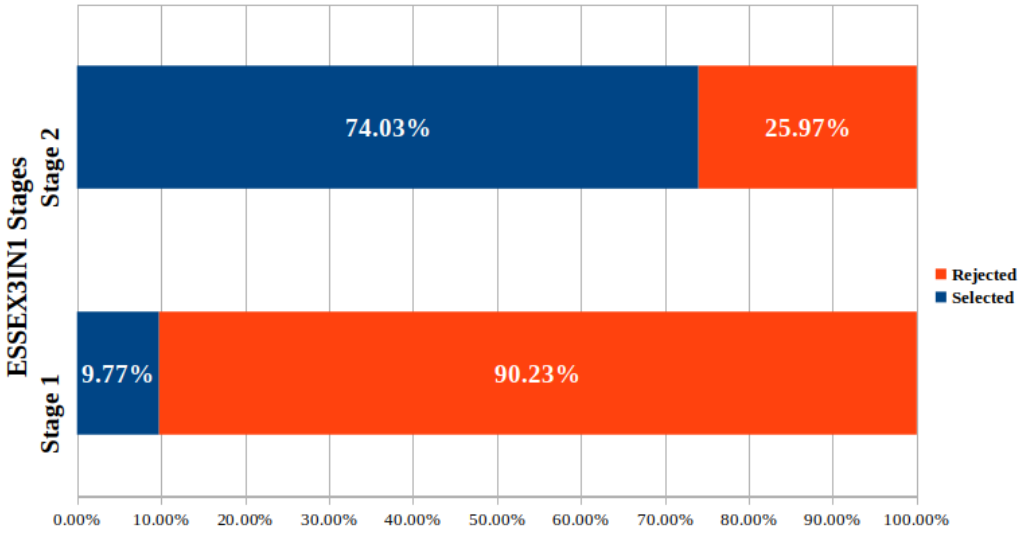

Fig. 3.11 The objective of memorable maps framework is to sample good frames and discard confusing frames. This objective achievement is presented by showing the contribution in memorable map from each ESSEX3IN1 stage.

datasets. It is important to note that bad/confusing images found by proposed framework are not removed from the reference database when evaluating AUC, but are treated as true negatives. This ensures that AUC boost reported here is not due to reduction of reference database size. For comparison with proposed framework, the author also shows the AUC performance for each technique by only employing static images on all datasets. Please note that because SPEDTest, Stlucia and Nordland datasets are largely static, the performance boost by just employing static images is only evident for ESSEX3IN1 dataset. This further validates the utility of the new proposed dataset ESSEX3IN1 for VPR, while simultaneously advocating for the efficacy of memorability and entropy criteria.

Fig. 3.12 depicts the AUC increase by employing proposed framework on ESSEX3IN1, St. Lucia, Nordland and SPEDTest dataset, respectively. The author uses the same values for MT, ST and ET as in sub-section 3.4.2 for ESSEX3IN1, Nordland and SPEDTest dataset. However, for St. Lucia we reduce each of the 3 selection thresholds by 0.05 to get a non-zero map size. This performance increase for all techniques on all datasets advocates for the utility, generalisability and agnostic nature of proposed framework. Reference database size remained the same for all AUC evaluations by treating confusing images as true-negatives.

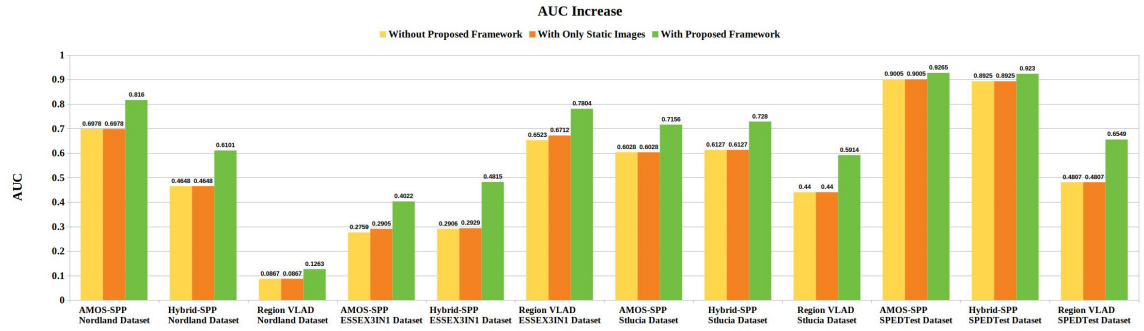

Fig. 3.12 Increase in AUC by using the proposed framework in combination with VPR techniques on all 4 datasets employed in this chapter is presented here. Please note that ESSEX3IN1 is the only dataset with highly dynamic content and therefore the AUC boost for employing only static images is not evident on other datasets.

3.4.4 Selected vs Discarded Images

In this sub-section, the author shows some images from all 4 datasets that were selected or discarded by proposed framework. This gives a qualitative insight into the working of this framework in different environments/datasets. Since the memorable maps framework evaluates both the query images and reference images, the images in Fig. 3.13 are impartial to such distinction. Selected images from ESSEX3IN1 are pre-dominantly of buildings with distinctive patterns and are largely static, while discarded images consist of far out natural scenes, dynamic objects or have low-entropy. Selected images in Stlucia dataset contain road signs, squares and houses. On the other hand, discarded images comprise of far out road scenes with trees and large portions of sky. Selected images from Nordland dataset consist of either appearing tunnels or bridges which contribute to their distinctiveness, while discarded images consist of vegetation or have low information. Staticity does not play any role in Nordland dataset due to the absence of dynamic objects. Selected images in SPEDTest dataset are from CCTVs covering buildings or distinctive locations, while discarded images consist of far out natural scenes and dynamic objects.

The author also reports the distribution of memorable images over the trajectories of Stlucia and Nordland datasets in Fig. 3.14. Because the ground-truth information for these datasets does not contain the exact inter-frame distance/time, the distribution in Fig. 3.14 is shown over image indices, which is very close to a constant distance-based distribution, as the speed of camera platform is mostly constant over the respective trajectories. ESSEX3IN1 and SPEDTest datasets are not trajectory-based, therefore, this distribution of memorable images over trajectory is not shown for these datasets.

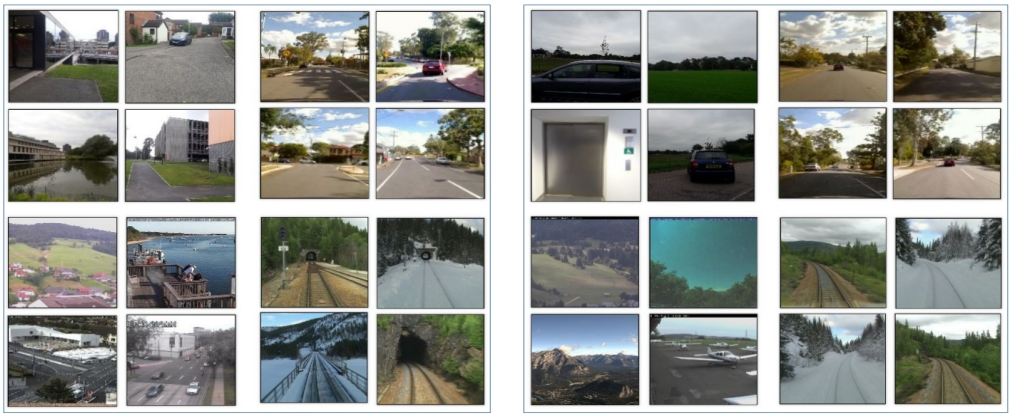

Fig. 3.13 Examples of images selected and discarded by the memorable maps framework from all 4 datasets are shown here. Top-left pairs of 4 images in each selected and discarded division are from ESSEX3IN1 dataset, followed-up with pairs from Stlucia dataset, Nordland dataset and SPEDTest dataset in clockwise manner.

3.4.5 Criterion Contribution Analysis

Each criterion in the memorable maps framework contributes to AUC boost. This subsection is dedicated to giving an insight into this individual contribution. The author uses ESSEX3IN1 for this purpose as it contains confusing images from all three (memorability, staticity and entropy) paradigms. For the AUC evaluation on ESSEX3IN1, the author shows the contribution of each criterion in Fig. 3.15. The analysis is performed based on the number of images that were mismatched by a VPR technique and were also discarded by at least one of the memorable maps framework criterion. Additionally, the author also shows in Fig. 3.16, a qualitative holistic view into cases where each criterion fails and others are used to cater for this failure, thereby, explaining intuitively why each of the criterion in proposed framework has its individual significance.

While Fig. 3.15 suggests that each of the three criteria are useful; the % contribution is linked to (and can vary with) the number of non-memorable, dynamic and information-less images in the dataset. (refer Fig. 3.10)

3.4.6 Parametric Variation

In this subsection, the author presents the variation in Visual Place Recognition performance with strictening framework criteria on ESSEX3IN1. The author sweeps each of the three criteria from 0-1 (Step size: 0.1), while keeping the other two inactive (i.e., set equal to zero).

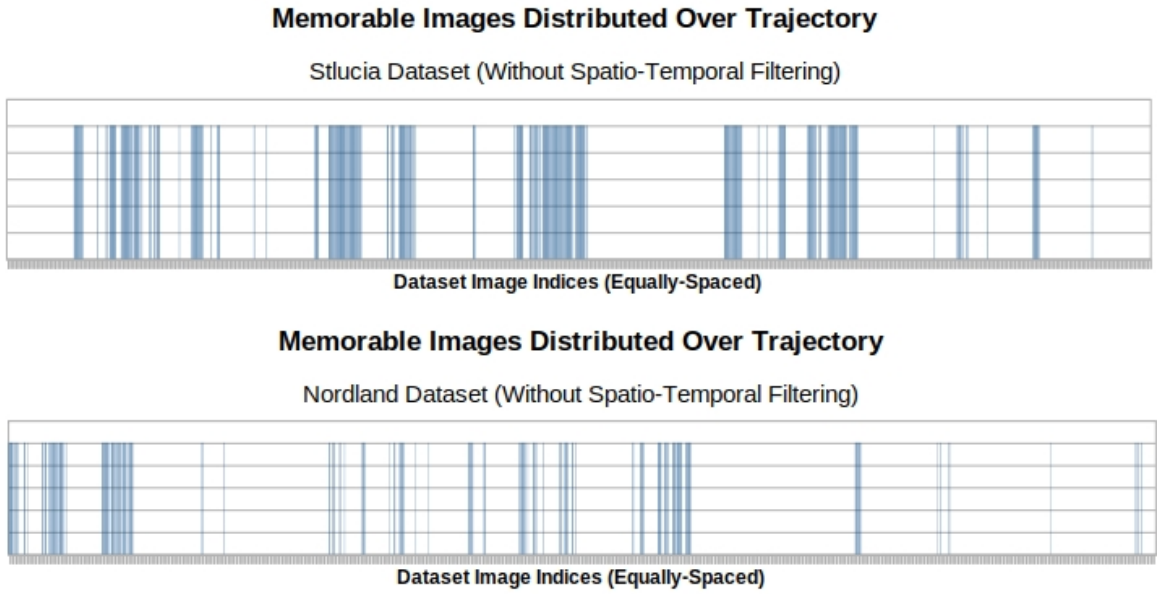

Fig. 3.14 Images selected as memorable over the trajectories of Stlucia [9] and Nordland datasets [4] are shown here. The horizontal axis represents the discrete, positive and equally-spaced indices of all the images in respective dataset. Each vertical bar represents an image selected as memorable by proposed framework. Because Spatio-Temporal filtering has not been utilised for this analysis, the selection of images is not uniform.

The data points for memorability and entropy thresholds have an upper-bound after which the total number of selected images equals to 0 (refer Fig. 3.10).

Fig. 3.17 shows that increasing entropy-threshold and memorability-threshold increases the AUC Performance for all three VPR techniques and follows a direct-relationship. On the other hand, the variation in AUC with increasing staticity-threshold follows a different trend. Firstly, the increase in AUC with ST is comparatively lower compared to MT/ET; which is due to the less number of dynamic images in the dataset compared to non-memorable and low-entropy images. Secondly, the variation in AUC with ST for Region-VLAD is higher compared to AMOS-SPP/Hybrid-SPP. This associates with the fact that AMOS-SPP/Hybrid-SPP have been trained on SPED (Specific Places Dataset) and discourage features coming from vehicles. While the analysis/results reveal that Region-VLAD extracts and positively matches features coming from cars in different places (See Fig. 3.4). Thirdly, there is an evident decrease in AUC as ST goes above 0.9. This decrease is expected as images with very low dynamic content can still be matched by contemporary VPR-techniques and discarding such images leads to the observed decline in VPR-performance. Please note that the best AUC results in Fig. 3.17 are higher than the results reported in Fig. 3.12.

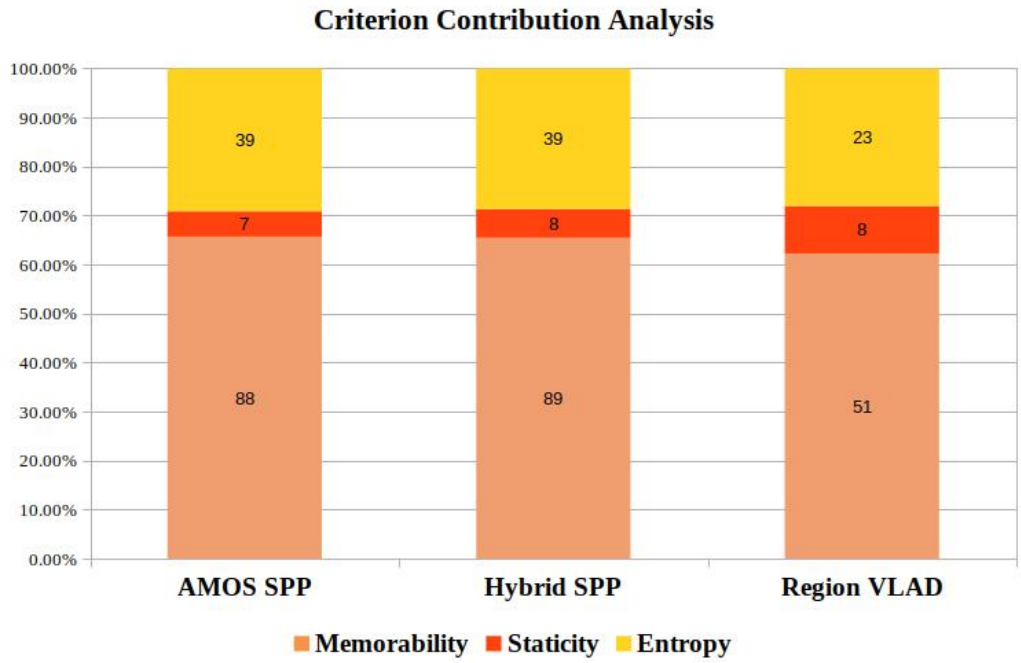

Fig. 3.15 Percentage contribution of each criterion into AUC increase is shown for the ESSEX3IN1 dataset. This contribution is directly linked with the type of environment being explored. In a highly dynamic environment, the contribution of staticity will be more significant than suggested by this chart and such.

This trend needs to be seen in co-relation with the reduction in map size as reported in Fig. 3.10. Increasing the three thresholds results in highly salient images stored in the map leading to higher AUC, however, it also reduces the absolute number of place recognition (loop-closure) candidates in the map and therefore, the framework thresholds need to be selected accordingly. The presented trends give a general idea for setting thresholds, thus to maintain a good balance between VPR performance and a salient representation of the world in a metric/topological/topo-metric map.

The author also shows the effect of varying the value of C from sub-section 3.2.1 in Fig. 3.18 for the reader's understanding. Changing this parameter within the range shown in Fig. 3.18 does not have any effect on the AUC performance of all techniques on SPEDTest dataset, suggesting that proposed framework is not sensitive to this parameter. The effect on entropy map and entropy score (ES) by varying the local circular neighbourhood (r) in sub-section 3.2.3 is reported in Fig. 3.19. The entropy score (ES) is dependent on this local circular neighbourhood r, such that increasing the value of r reduces the resolution of entropy map and increases the entropy score ES. This effect is similar to low-pass filtering and is explained as: Increasing the value of r increases the number of pixels to be added to

Fig. 3.16 Examples of images that are selected/discarded based on various combinations of memorable maps framework criteria are shown here. Please note that the understanding for 'informative' and 'memorable' nature of an image is subjective and in this work, it is expressed by the practical/implementation nature of the term. For-example, an image of a bush (top-left) is called informative because it has several edges, corners and contours for computer-vision feature descriptors and thereby has a high entropy. Similarly, memorability is explained by its cognitive perception, i.e., the work of [137].

the histogram, where the larger the radius of the circle, the greater will be pixel intensity divergence and hence higher is the log_2 score, leading to higher ES. This therefore, requires to affix the value of r to a value where coupled with ET, the framework can successfully distinguish between low and high informative images. The author is also interested in having high-resolution entropy maps instead of low-resolution entropy maps due to the salience of low-level features (like edges, corners etc) to the VPR problem.

3.4.7 Reduced Map Size and Computational Time

In addition to the increase in AUC, the developed framework helps in reducing the robot's map size which has been the motivation for semantic mapping research reviewed in Chapter 2. This size reduction also leads to lesser computational overhead for VPR. The reduction in map size for the thresholds presented in Section 3.4.2 is shown in Fig. 3.20. Because the map-size reduction by discarding non-memorable images can also lead to the reduction of absolute number of true-positives, the author shows this trend in Fig. 3.21. It can be seen in Fig. 3.21 that using the memorable maps framework does result in the decrease of true-positives, however, the proportion of discarded false-positives is greater than true-positives, which leads to the AUC boost reported previously in sub-section 3.4.3.

The computational performance is reported by calculating the time required to match a query image with all the reference images (having pre-computed feature descriptors) in

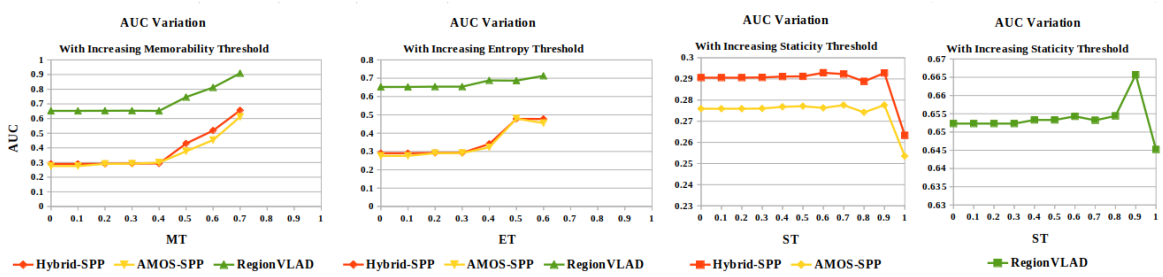

Fig. 3.17 Variation in VPR AUC performance by changing each of the memorable maps framework thresholds within their full range on ESSEX3IN1 is presented. Memorability and entropy continuously increase AUC until the total number of selected images equals to zero; suggesting that images with higher memorability and entropy are well-matched by VPR methods. On the contrary, since images with low dynamic content should/can still be matched, variation in staticity threshold does not lead to a continuous AUC increase. AUC change with ST is not at the same scale as MT/ET so it is shown separately.

both a conventional map and a memorable map. This offline matching time is elaborated in Table 1, where a memorable map having lesser number of reference images (see Fig. 3.20) achieves better matching time. The end-to-end time required in author's implementation to compute the salience of an image for memorable map is around 5 sec. Because the current implementation utilises a sequential combination of different research works, i.e., YOLO and MemNet, the timing is bottle-necked by the sum of individual timings of each of these works. The author believes that there is room to improve the time required to compute these maps by employing a different suit of CNNs (object detectors and memorability maps), improving software implementation, utilising hardware advances and by parallelizing the map computation by exploiting the independence of the three maps from each other.

Table 3.1 Matching Time Per Query Image

System Specs	Intel(R) Xeon(R) Gold 6134 CPU @ 3.20GHz, 64GB Physical Memory					
Framework	Without Memorable Maps			With Memorable Maps		
VPR Methods	AMOSNet	HybridNet	RegionVLAD	AMOSNet	HybridNet	RegionVLAD
ESSEX3IN1 (sec)	10.2	9.9	0.14	4.1	3.9	0.05
Nordland (sec)	78.7	76.4	1.1	9.1	8.7	0.12
St. Lucia (sec)	63.9	62.1	0.88	14.7	14.2	0.21
SPEDTest (sec)	29.5	28.6	0.41	7.7	7.5	0.11

3.4.8 Spatio-Temporal Filtering with Proposed Framework

A natural extension to the memorable maps framework is to define an upper bound on the maximum distance and/or time travelled within which a best image (or Top-N images) from

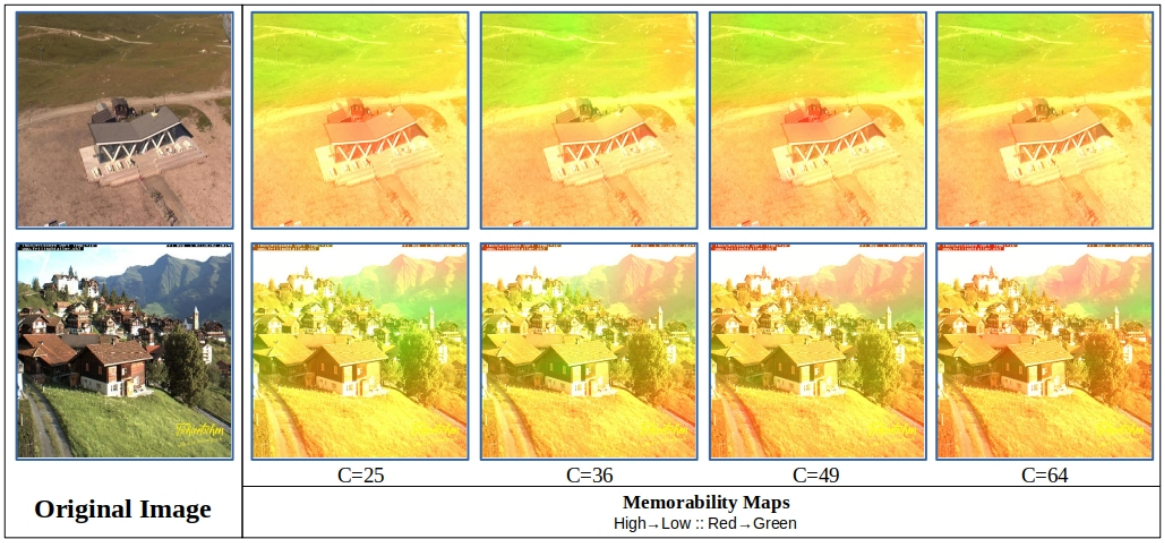

Fig. 3.18 Changing the value of C for computing memorability map does not result in any major change, as shown here. Images employed are from the SPEDTest [110] dataset and no change in AUC for this dataset was observed for the range of C used in this figure.

the traversed sequence should be be selected, even if the said image does not fully satisfy the proposed criteria. This can also be accommodated using a hysteresis-mechanism, where if a scene is continuously non-salient, the values for thresholds can be reduced to select relatively-less salient images. Depending on the size of employed hysteresis, it can be ensured that salient images distributed through-out the trajectory are selected. Thus, in a long traversal where the depicted scenery may not be memorable, static and/or highly-informative through-out the sequence, spatio-temporal filtering will ensure that the most memorable, static and informative image within the sequence is selected. This image can then be flagged as a 'low-quality' image in the memorable map and depending on the under-lying VPR system can either be treated differently (e.g., use longer matching sequences in [5]), avoided for use in loop-closure or treated as a potential false-positive prediction [150].

Because employing such a mechanism can lead to changes in AUC, the author has reported this analysis of AUC boost with and without the spatio-temporal filtering in Fig. 3.22 for Nordland dataset. The selection of filtering methodology is hysteresis-based, such that if in a sequence of 20 consecutive frames, none of the images satisfy the criteria thresholds of sub-section 3.4.2, the thresholds are reduced by 0.03 for the respective sequence. It can be clearly seen in Fig. 3.22 that allowing less-salient images into the map does lead to lesser AUC boost. The author also shows the changes in distribution of a total of 412 memorable images over the Nordland trajectory by employing such hysteresis-based spatio-temporal filtering in Fig. 3.23.

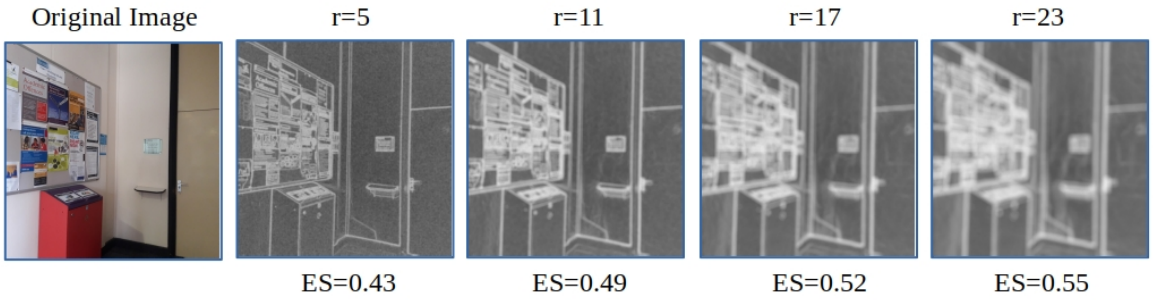

Fig. 3.19 Variation in the entropy map and the entropy-score (ES) are shown here for different values of local circular neighbourhood (r) given fixed image size ($W2 \times H2$). The larger the radius, the lower the resolution of entropy map. Increasing r also increases the value of ES due to increased no. of pixels for grayscale histogram that results in higher pixel intensity deviation.

3.5 Summary

This chapter presented a cognition-inspired generalized framework for creating 'memorable maps'. This framework evaluates an incoming camera frame for its memorability, staticity and entropy to decide a frame's insertion into the robot's map. By using 'ESSEX3IN1', the author has shown how images that are confusing and indistinct lead to perceptual aliasing and are also mismatched by contemporary VPR systems. The application of proposed framework in detecting these confusing images and subsequently improving VPR performance is presented. The author generalises the applicability of proposed framework by reporting results on multiple public datasets. Due to its agnostic nature, memorable maps framework can be plugged into any VPR technique giving performance boost.

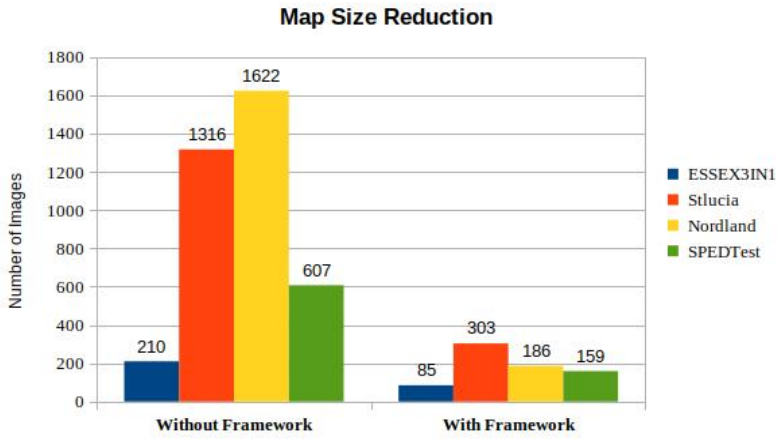

Fig. 3.20 Reduction in topological map size given similar or better VPR results is of prime importance for an autonomous robot to efficiently map/explore an environment. As depicted here, memorable maps framework intrinsically reduces map size while giving AUC boost to contemporary VPR systems.

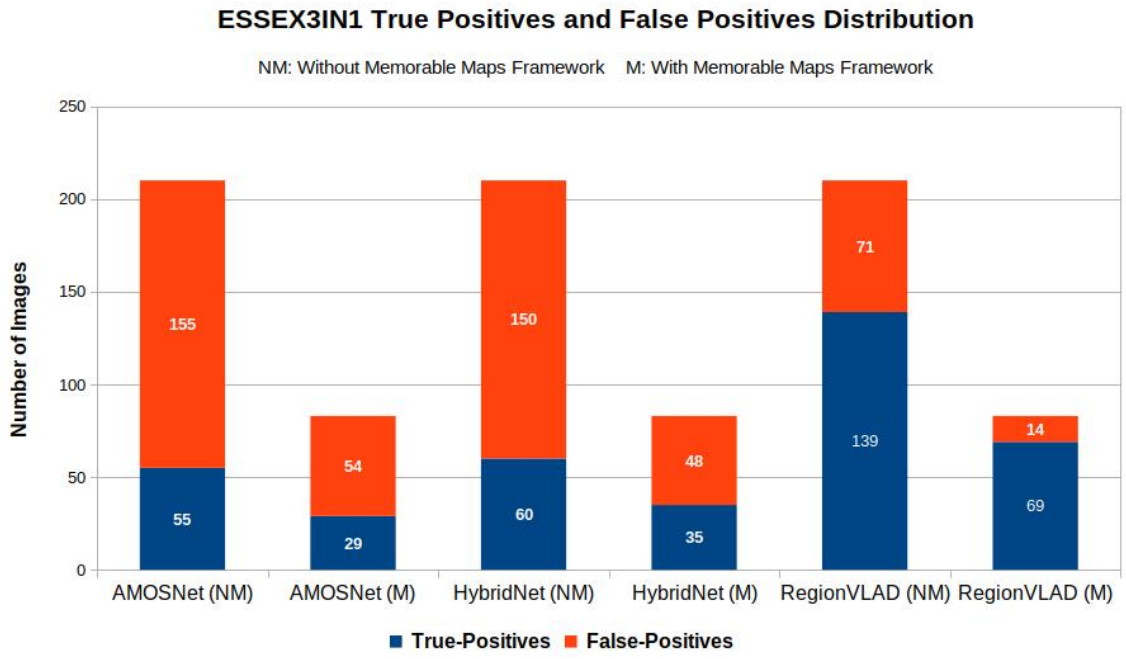

Fig. 3.21 The absolute decrease in true-positives and false-positives by using the memorable maps framework is shown here for all techniques on the ESSEX3IN1 dataset.

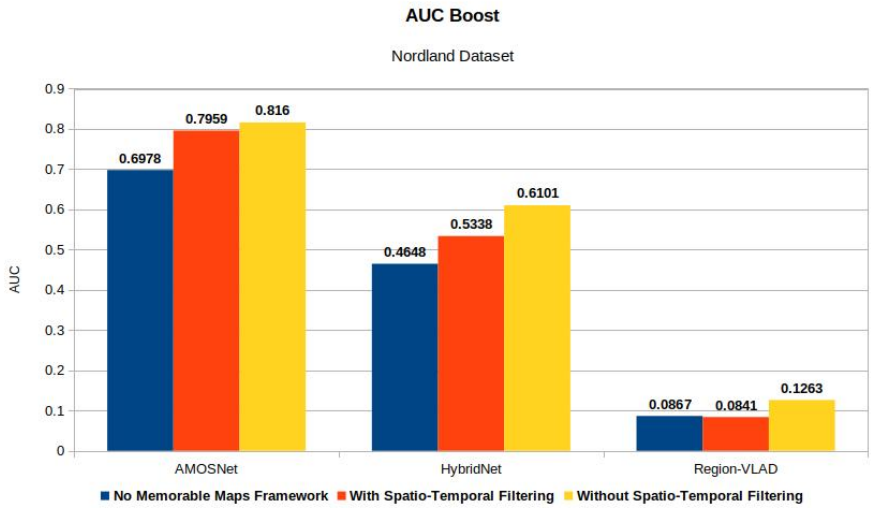

Fig. 3.22 Employing spatio-temporal filtering within the memorable maps framework to avoid large geographical gaps between salient images leads to lesser AUC boost as less salient images are added to the map. Using the proposed framework without spatio-temporal filtering leads to the highest AUC, followed by the proposed framework with spatio-temporal filtering and lastly without the memorable maps framework. Region-VLAD has significantly less number of true-positives through-out the trajectory, therefore AUC boost with spatio-temporal filtering is not evident for this technique.

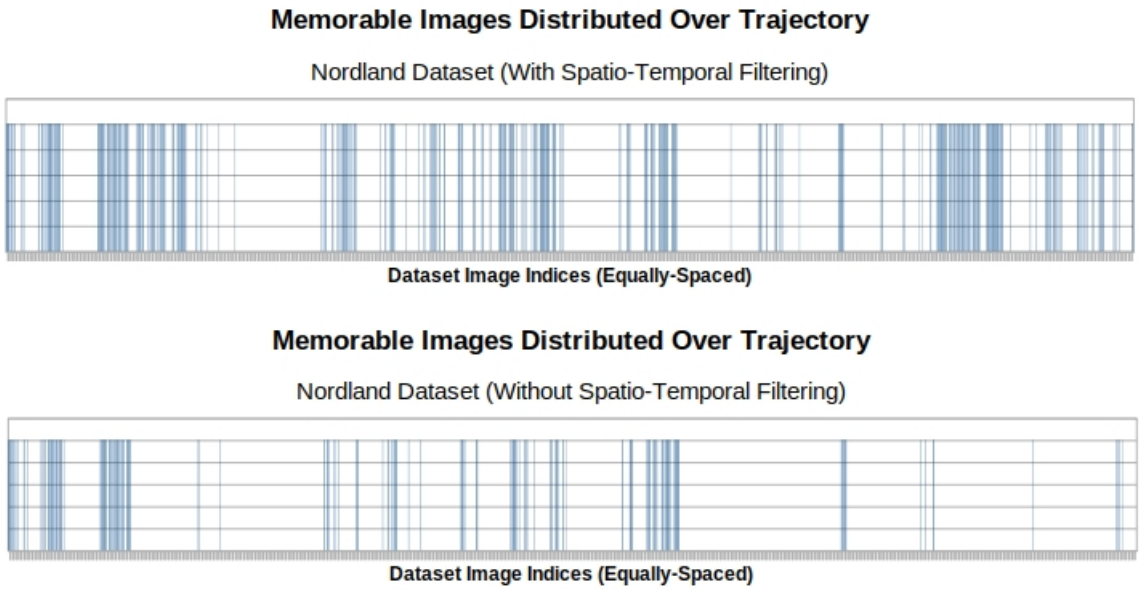

Fig. 3.23 Changes in memorable images selected over the Nordland trajectory by employing hysteresis-based spatio-temporal filtering within the memorable maps framework are shown here. Depending on the width of hysteresis, image gaps can be further reduced at the cost of reduced map salience.

Chapter 4

A Comprehensive Comparison of VPR Approaches under Changing Conditions

As discussed in this thesis, recent years have seen a significant improvement in the capability of Visual Place Recognition (VPR) methods, building on the success of both hand-crafted and learnt visual features, temporal filtering and usage of semantic scene information. The wide range of approaches and the relatively recent growth in interest in the field has meant that a wide range of datasets and assessment methodologies have been proposed, often with a focus only on precision-recall type metrics, making comparison difficult. In this chapter, the author presents a comprehensive approach to evaluating the performance of 10 state-of-the-art recently-developed VPR techniques, which utilises three standardized metrics: (a) Matching Performance b) Matching Time c) Memory Footprint. Together this analysis provides an up-to-date and widely encompassing snapshot of the various strengths and weaknesses of contemporary approaches to the VPR problem. The aim of this chapter is to help move this particular research field towards a more mature and unified approach to the problem, enabling better comparison and hence more progress to be made in future research.

4.1 Background

By now, VPR is a well-understood problem and acts as an important module of a Visual-SLAM based autonomous system [8]. However, VPR is highly challenging due to the significant variations in appearance of places under changing conditions. Throughout VPR research over the past years, we see 4 such variations in appearances of places, which have been widely discussed and tackled by different novel VPR techniques. **Seasonal Variation:** Appearance of places change drastically from summer to winter or spring to autumn posing

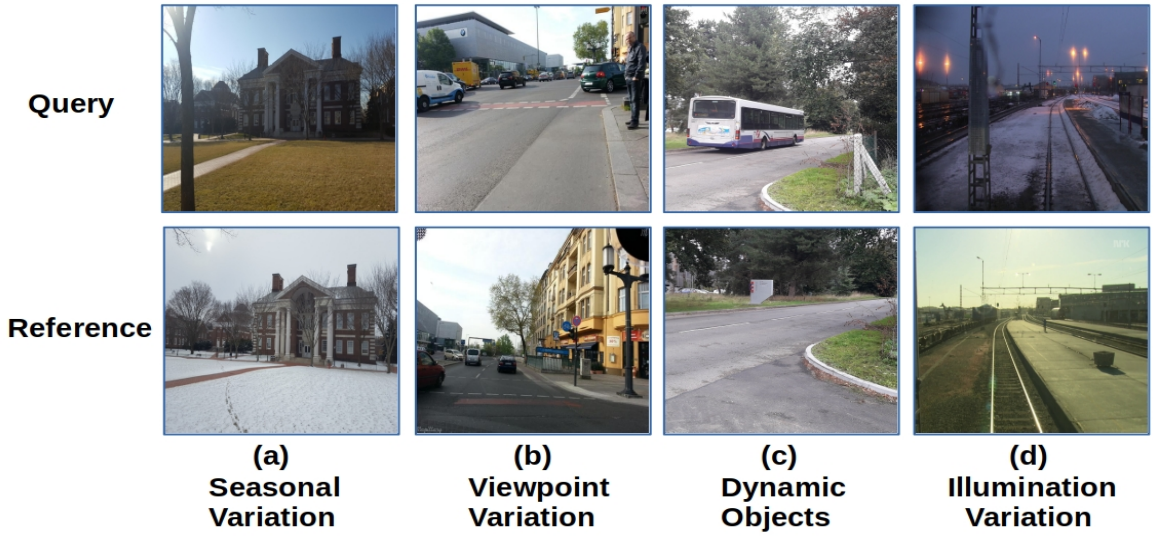

Fig. 4.1 Variations in the appearance of places are illustrated where; (a) Seasonal variation observed from summer to winter in the same place (b) Change in camera viewpoint leading to drastic change in observed structures (c) Commonly seen dynamic objects in urban scenes (d) Appearance change as a result of day-to-night transition.

challenges for state-of-the-art VPR techniques [151][152]. **Viewpoint Variation:** Images of the same place may look very different when taken from different viewpoints [153]. This viewpoint variation could be a simple lateral variation or a more complex angular variation coupled with changes in focus, base point and/or zoom during repeated traversals [154]. **Illumination Variation:** This is the result of daylight changes and intermediary transitions during different times of the day/night, which make place recognition difficult to perform [155][5]. **Dynamic Objects:** Objects such as cars, people, animals etc. can also change the appearance of a scene and thus a VPR technique should be able to suppress any features coming from these dynamic objects [156] [124]. The author has shown all these variations in Fig. 4.1 for a quick recap of reader.

While many different techniques have been proposed to tackle each (or a combination) of the 4 variations, a thorough and holistic comparison of these techniques is needed for an up-to-date review. In this chapter, the author takes up the task to evaluate 10 novel VPR techniques on the most challenging public datasets using the same platform, evaluation metric and ground truth data. While presenting the matching performance and (more recently) matching time has been common in VPR research; the author additionally enlists the memory footprint of these VPR techniques which is an essential factor at deployment. The novel contributions of this chapter are as follows:

1. The techniques compared in this chapter encompass years of VPR research and a comparison of such magnitude has not been reported previously.

2. VPR performance is highly sensitive to the choice of evaluation datasets, computational platform, evaluation metric and ground truth data. By keeping all of these variables constant, the author brings VPR techniques to a common ground for evaluation.

3. Memory footprint for map creation at deployment time is a critical factor and thus, the author reports the feature vector size for all 10 VPR techniques.

4.2 Experimental Setup

This section first presents the operational details of the VPR techniques that are compared in this chapter. The author then specifies the details of the datasets used for evaluation. Finally, the evaluation metrics considered for comparison in this chapter are quickly introduced.

4.2.1 VPR Techniques

HOG Descriptor

Histogram-of-oriented-gradients (HOG) is one of the most widely used handcrafted feature descriptor. While it does not perform nearly well to any other VPR technique, it is a good starting point for a comparison such as ours. Author's motivation to select HOG is also based upon its performance as shown by McManus et al. [101] and the utility it offers as an underlying feature descriptor for training a convolutional auto-encoder in [113]. A cell size of 8×8 and a block size of 16×16 with total number of histogram bins equal to 9 is used. HOG descriptors of two images are subsequently compared using cosine similarity.

Seq-SLAM

Seq-SLAM showed excellent immunity to seasonal and illumination variations by using sequential information to its advantage. The proposed implementation has been open-sourced in MATLAB and ported to Python. The author has used a sequence size of 10, minimum velocity of 0.8 and max velocity of 1.2 for evaluating Seq-SLAM.

AlexNet

Sünderhauf et al. studied the performance of features extracted from AlexNet and found *conv3* to be the most robust to environmental variations. The activation maps are encoded

into feature descriptors by using Gaussian random projections. The author's implementation of AlexNet is similar to the one presented by authors in [113].

NetVLAD

The author has employed the Python implementation of NetVLAD open-sourced in [157]. The model selected for evaluation is VGG-16 which has been trained in an end-to-end manner on Pittsburgh 30K dataset [107] with a dictionary size of 64 while performing whitening on the final descriptors.

AMOSNet

AMOSNet has been trained from scratch on SPED dataset and the model weights have been open-sourced by authors in [102]. The author has therefore implemented spatial pyramidal pooling on pre-trained AMOSNet and used activations from *conv5* to extract and describe features. L1-difference is subsequently used to match features descriptors of two images.

HybridNet

Similar to AMOSNet, model parameters for HybridNet trained on SPED dataset have also been open-sourced. However, the weights of top-5 HybridNet convolutional layers are initialized from CaffeNet trained on ImageNet dataset. The author has employed spatial pyramidal pooling on activations from *conv5* layer of HybridNet. Feature descriptors of two images are then matched using L1-difference.

Cross-Region-BOW

The author has employed the [158] open-source MATLAB implementation for experimentation; VGG-16 [159] pre-trained on ImageNet dataset is used while employing *conv5_3* and *conv5_2* for identification and extraction of regions respectively. Image comparison is carried out by finding the best mutually matched regions and describing these regions using a $10k$ BoW dictionary.

R-MAC

The MATLAB implementation for R-MAC is available at [160]. The author has used *conv5_2* of object-centric VGG-16 for regions-based features. For a fair comparison, the author has removed the geometric verification block while performing power and L2 normalization

Table 4.1 Benchmark Place Recognition Datasets

Dataset	Traverse		Environment	Variation	
	Test	Reference		Viewpoint	Condition
Nordland	172	172	Train journey	strong	very strong
Berlin Kudamm	222	201	Urban	very strong	strong
Gardens Point	200	200	University campus	strong	very strong

on the retrieved R-MAC representations. The retrieved R-MACs are mutually matched, followed by aggregation of the mutual regions' cross-matching scores.

Region-VLAD

The author has employed *conv4* of HybridNet for evaluating the Region-VLAD VPR approach. The employed dictionary contains 256 visual words used for VLAD retrieval. Cosine similarity is subsequently used for descriptor comparison.

CALC

Merrill et al. [113] trained a convolutional auto-encoder for the first-time in an unsupervised manner for VPR, where the objective of auto-encoder was to re-create the HOG descriptor of original image given a distorted version of the original image as input. Authors have open-sourced their implementation and this chapter uses model parameters after $100,000$ training iterations for comparison.

4.2.2 Evaluation Datasets

A number of datasets have been proposed for evaluating VPR techniques over the years, as discussed in Chapter 2 of this thesis. These datasets comprise of different types of variations ranging from viewpoint, seasonal and illumination variations to a combination of these. In order to challenge and put all the VPR techniques presented in sub-section 4.2.1 to their limits, the author selects 3 datasets with the most extreme variations. This sub-section is dedicated to introducing these 3 datasets. The author also summaries the qualitative and quantitative nature of datasets in Table 6.1.

Fig. 4.2 Berlin Kudamm dataset sample images are shown here. The query and reference traverses exhibit extreme viewpoint variation. This dataset contains recurring and upfront dynamic objects which is uncommon to any other VPR dataset.

Berlin Kudamm Dataset

This dataset was introduced in [3] and has been captured from crowd-sourced photo-mapping platform called *Mapillary*[1]. Both the traverses exhibit strong viewpoint and conditional changes as visualized in Fig. 4.2. Due to its urban nature, dynamic objects such as vehicles and pedestrians are observed in most of the captured frames. Ground truth is obtained using GPS information to build place-level correspondence. A retrieved image against a query is considered as a correct match if it is either of the 5 closest frames in ground-truth. Thus, for a query image q and its ground-truth image n in the reference database, images $n-2$ to $n+2$ also serve as corresponding correct matches.

Gardens Point Dataset

This dataset is constructed at the Gardens Point Campus of Queensland University of Technology (QUT), with the first traverse captured during the day and the reference traverse taken at night with laterally changed viewpoint. Variations in the dataset are shown in Fig. 4.3. The ground truth is obtained by frame- and place-level correspondence. A retrieved image against a query is considered as a correct match if it is either of the 5 closest frames in ground-truth. That is, for a query image q and its ground-truth image n in the reference database, images $n-2$ to $n+2$ also serve as corresponding correct matches.

[1] https://www.mapillary.com/

Fig. 4.3 Gardens Point dataset sample images are presented here. The query and reference traverses are highly illumination variant and accompanied with lateral viewpoint variation.

Nordland Dataset

A train journey is captured in this dataset with the first traverse taken during winter and the second traverse during summer. While this dataset contains strong seasonal changes as shown in Fig. 4.4, the author has introduced lateral viewpoint variation by manually cropping images. The ground truth consists of frame-level correspondence with a retrieved image against a query considered as a correct match if it is either of the 3 closest frames in ground-truth. Thus, for a query image q and its ground-truth image n in the reference database, images $n-1$ to $n+1$ also serve as corresponding correct matches.

4.2.3 Evaluation Metrics

For evaluating the place matching performance, the author has continued the use of AUC as the evaluation metric, as in equation 2.3. Moreover, for real-time autonomous robotics, matching time is an important factor to be considered at deployment. Thus, for all 10 VPR techniques, the author reports the matching time of a query image given pre-computed feature descriptors of reference images. This matching time (reported in seconds) includes the feature encoding time for an input query image and the descriptor matching time for R number of reference images.

Generally, the deployment use of VPR is coupled with map creation in SLAM. Therefore, the size of reference image descriptors is an important factor to be considered for the practicality of a VPR technique. While this has not been previously discussed, the author

Fig. 4.4 Nordland dataset sample images are presented in this figure. This dataset is one of the highly seasonally variant dataset and has manually introduced lateral viewpoint variation.

enlists the size in bytes of a reference image feature descriptor for all 10 VPR techniques. This gives a good idea about the scalability of a technique for large-scale VPR.

4.3 Results and Analysis

This section is dedicated to the performance evaluation of all VPR techniques. The author presents the image matching performance on benchmark VPR datasets, followed by matching time and memory footprint, each discussed in their respective subsections. All evaluations are performed with an Intel(R) Xeon(R) Gold 6134 CPU @ 3.20GHz with 64GB physical memory running a Ubuntu 16.04.6 LTS.

4.3.1 Matching Performance

This sub-section reports the AUC performance of all 10 VPR techniques on each of the 3 datasets. The author also shows exemplar image matches from all three datasets in Fig. 4.9. While some exemplar images have been matched by most state-of-the-art VPR techniques, examples of images mismatched by all VPR techniques have also been presented.

Berlin Kudamm Dataset

Fig. 4.8 shows that NetVLAD achieves state-of-the-art performance on Berlin Kudamm dataset, while Region-VLAD and Cross-Region-BoW follow-up with relatively poor performance. AMOSNet and HybridNet with SPP also achieve nearly similar performance to

regions based approaches and suffer due to the extreme viewpoint variation not catered by SPP. It is important to note that due to urban scenario, both the traverses in Berlin Kudamm dataset include dynamic and confusing objects such as vehicles, pedestrians and trees; as illustrated in Fig. 4.2. These confusing objects and homogeneous scenes lead to perceptual aliasing which coupled with extreme viewpoint variations makes Berlin Kudamm highly challenging for all VPR techniques.

SeqSLAM being velocity dependent has shown inferior results due to the varying speed of camera platform and significant viewpoint variation. One of the reasons for state-of-the-art performance of NetVLAD could be its training on large urban place-centric dataset (Pittsburgh 250K) which exhibits strong lightning and viewpoint variations along with dynamic and confusing objects. This is in contrast to the training datasets of VGG-16 (ImageNet) and HybridNet (SPED). 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.

Gardens Point Dataset

Although this dataset exhibits strong illumination and viewpoint variations, majority of the VPR approaches perform relatively well. This is due to the distinctive structures captured in both the traverses. Cross-Region-BoW achieves state-of-the-art results while Net-VLAD, HybridNet and AMOSNet also perform nearly well on this dataset.

Nordland Dataset

Nordland dataset exhibits strong seasonal variation and synthetic viewpoint change, as illustrated in Fig.4.4. Region-VLAD achieves state-of-the-art performance with Net-VLAD and Cross-Region-BOW also giving comparable results. HybridNet performs better than AMOSNet due to its weights being initialized from the weights of CaffeNet that have been exposed to a variety of scenes available in the ImageNet dataset.

4.3.2 Matching Time

In real-time VPR systems, matching time is an important factor that needs to be considered when comparing a query image against a large number of database images. The author shows in Fig. 4.5, the feature encoding time for all VPR techniques given a single query image. SeqSLAM does not extract features from an image but directly uses patch-normalized camera frames for comparison. As expected, CNNs take significantly more time to encode an input image compared to handcrafted feature descriptors. However, convolutional auto-encoder in

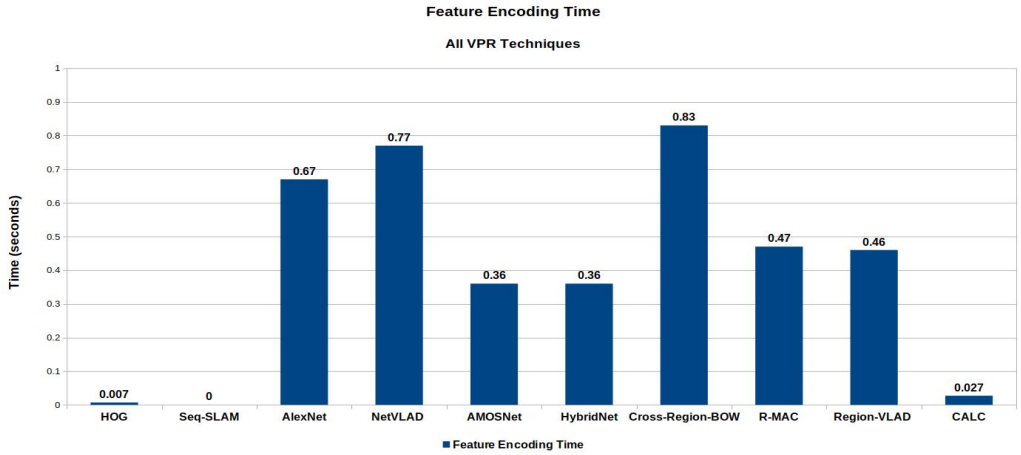

Fig. 4.5 Feature encoding time of all VPR techniques are shown in this figure. As expected, neural network based techniques have higher encoding time compared to handcrafted techniques. Although, the matching performance of CALC is lower compared to some of the neural network based VPR techniques, the significantly low encoding time of CALC promises the possibilities of real-time highly accurate VPR in future.

CALC takes significantly lower time to encode features in comparison to other CNN based VPR techniques. This is because the architecture of CALC is designed specifically for VPR as compared to off-the-shelf CNN architectures employed in other VPR techniques.

While the feature encoding time is independent of the number of reference images, feature descriptor matching time scales directly with the total number of reference images. Thus, the author also shows the time taken to match feature descriptors of a query and a reference image in Fig. 4.6. Please note that Fig. 4.6 uses logarithmic scale on horizontal-axis for clarity. This descriptor matching time can be directly multiplied with the total number of reference images in the database. It is interesting to note that although Cross-Region-BOW achieves good AUC performance on different datasets, it suffers from a significantly higher descriptor matching time. This can be associated with the one-to-many nature of Cross-Region-BOW which finds the best matched regions between a query and a reference image.

4.3.3 Memory Footprint

The size of feature descriptors plays an important role when considering the practicality of a VPR technique for deployment in real-world scenarios. Due to limited storage capabilities, compact representations of image descriptors are needed. Thus, while matching performance can be improved by increasing the size of feature descriptor (or number of regions where applicable), it limits the deployment feasibility of such a VPR technique. The author reports

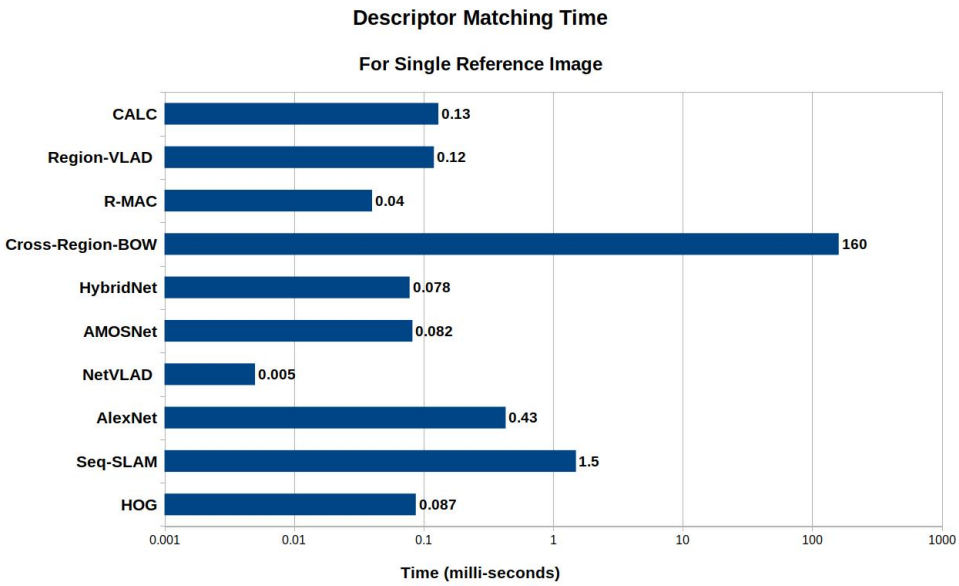

Fig. 4.6 Descriptor matching time of all VPR techniques are compared here. Please note that the horizontal axis is in logarithmic scale due to the high variance in-between matching times of different techniques.

the feature vector size of all VPR techniques in Fig. 4.7. Cross-Region-BOW and Region-VLAD notably have a large memory footprint compared to other VPR techniques. For Cross-Region-BOW, this can be associated with the number of regions (200) that have to be stored, where each region has a descriptor dimension equal to the depth (512) of convolutional layer. While in Region-VLAD, the employed VLAD dictionary size is 256 with each word in the dictionary having a dimension (depth of convolutional layer) of 384.

4.4 Summary

This chapter presented a holistic comparison of 10 VPR techniques on challenging public datasets. The choice of evaluation datasets, ground truth data, computational platform and comparison metric is kept constant to report the results on a common-ground. In addition to the matching performance and matching time, the author reported the feature vector size as an important factor for VPR deployment practicality. While neural network based techniques out-perform handcrafted feature descriptors in matching performance, they suffer from higher matching time and larger memory footprint. The performance comparison of neural network based techniques with each other also identifies their lack of generalisability from one evaluation dataset to another. While some VPR techniques can achieve better

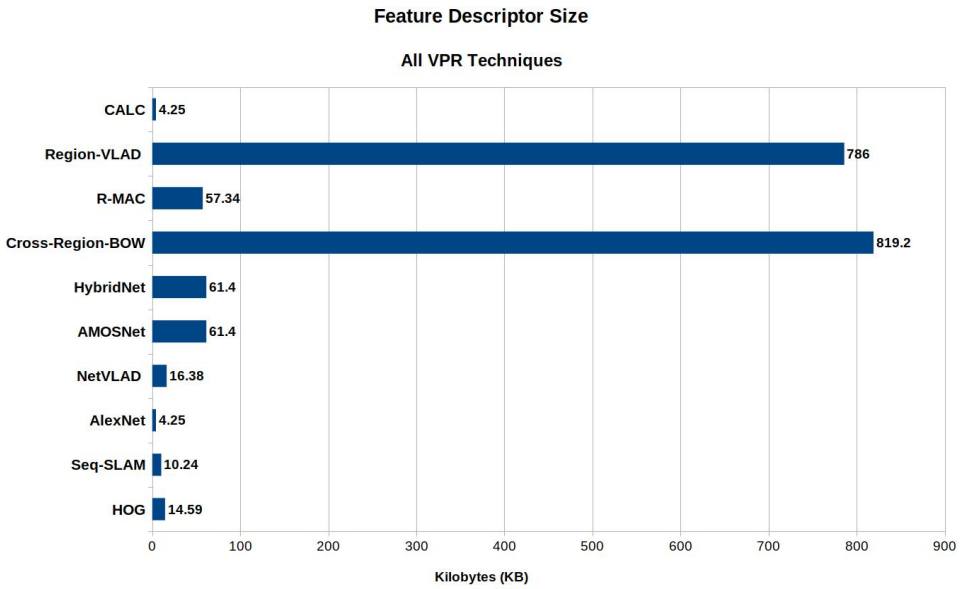

Fig. 4.7 Feature descriptor sizes of all VPR techniques are shown here. While this metric has been rarely discussed in VPR literature, it is highly significant for resource-constrained platforms and can hinder the deployment of a VPR technique in field. Thus, highly compact feature descriptors that are encoded in real-time, are condition invariant, repeatable and distinct should be the output of an ideal VPR system.

matching performance in contrast to others, there may be a trade-off between matching performance and computational requirements (i.e. higher matching time and/or memory demand). It is worth noticing that contrary to expectations, increase in VPR performance (for the author's choice of parameters and datasets) is not observed in a chronological order.

However, the evaluations performed in this chapter have been limited only to ground-based platforms, but VPR finds significant applications in aerial robotics as well. In the next chapter, this analysis of VPR techniques is extended to the aerial domain.

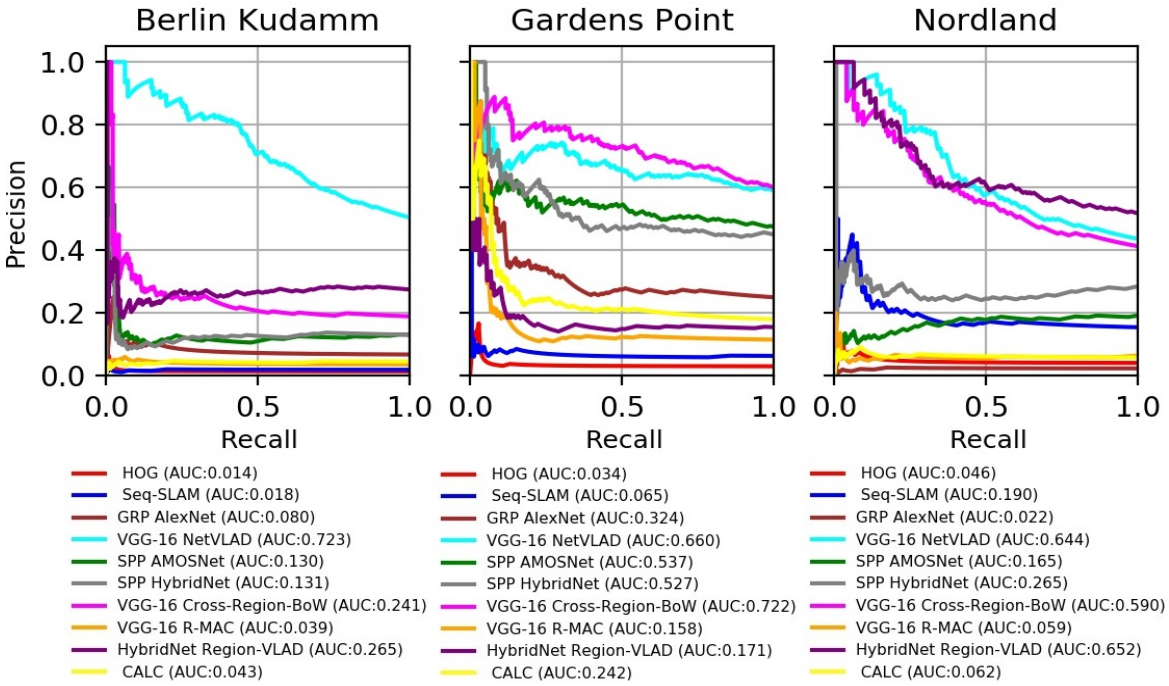

Fig. 4.8 AUC under PR curves on the benchmark datasets of all VPR techniques.

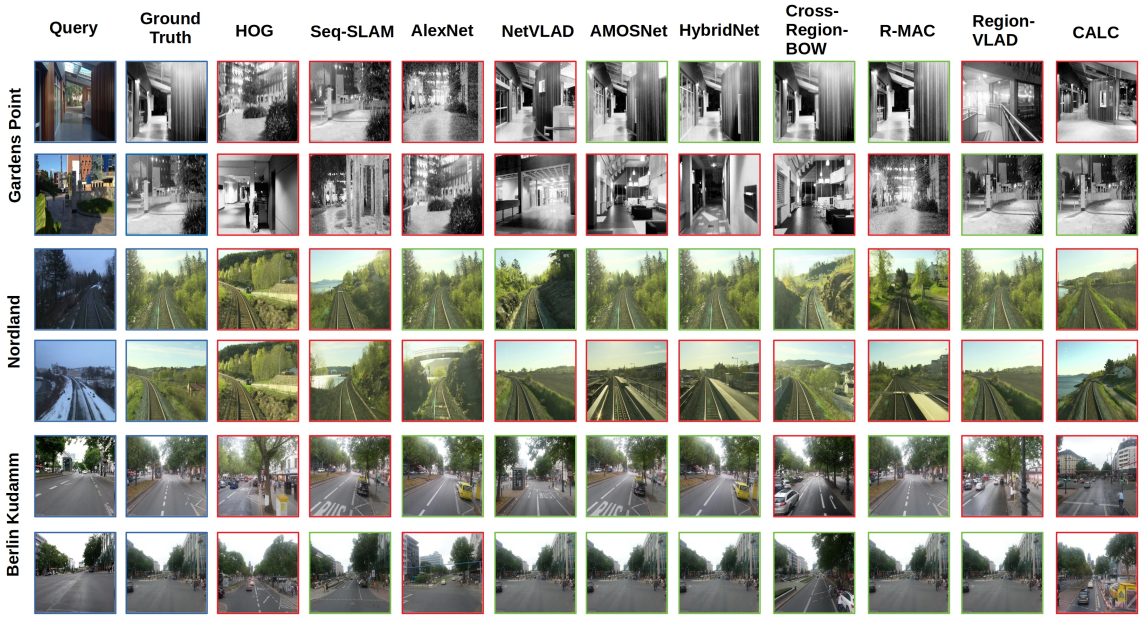

Fig. 4.9 Samples of images matched/mismatched by different VPR techniques on all three datasets are presented. The first two columns show the query and ground-truth reference images respectively, followed by images retrieved by each of the 10 VPR techniques.

Chapter 5

Are State-of-the-art VPR Techniques any Good for Aerial Robotics?

In the previous Chapter 4, evaluations are performed for ground-based mobile platforms, which cannot be generalized to aerial platforms. The degree of viewpoint variation experienced by aerial robots is complex, with their processing power and on-board memory limited by payload size and battery ratings. Therefore, in this chapter, the author uses the previously discussed state-of-the-art VPR techniques that have been evaluated for ground-based platforms and compares them on 2 recently proposed aerial place recognition datasets with three prime focuses: a) Matching performance b) Processing power consumption c) Projected memory requirements. This gives a birds-eye view of the applicability of contemporary VPR research to aerial robotics and lays down the nature of challenges for aerial-VPR.

5.1 Background

The existing research in VPR has been focused on ground-based mobile platforms and the datasets used for evaluation contain planar viewpoint changes. However, aerial platforms like drones introduce a third dimension (vertical) to viewpoint change. This added dimension, coupled with 6-degrees of freedom of aerial platforms, limited computational payload, limited sensing payload, limited power/energy, high velocity, difficulty of local motion estimation, restrained storage and run-time memory, make VPR challenging for aerial robotics.

While most of the datasets used for evaluating VPR techniques, as explored in Chapter 2, have been created using cameras mounted on cars, bicycles or hand-held setups during walk; Maffra et al. [161] recently introduced the Shopping street datasets targeted for aerial place recognition. Therefore, in this paper, the author takes up the task to evaluate 8 contemporary

Fig. 5.1 The viewpoint variation challenge for aerial platforms is shown in comparison to ground based platforms. On the left, sample images from Shopping Street 2 dataset containing 6-DOF viewpoint change are shown. While, on the right, images from the widely used Gardens Point dataset and Stlucia dataset posing lateral viewpoint change are shown.

VPR techniques on the datasets proposed in [161]. The objective of this paper is to answer the question: *Can VPR state-of-the-art research be extended to resource-constrained aerial robotics and how can viewpoint change resulting from 6-DOF (degrees-of-freedom) platforms affect place matching performance?*

To explain the difference between a ground-based and aerial-based platform's viewpoint variation, the author shows in Fig. 5.1, a comparison between existing datasets and the datasets used in our work. The novel contributions of this chapter are as follows:

1. This chapter discusses and evaluates inter-platform VPR (particularly ground and aerial). This is important because the performance of a VPR technique immune to planar viewpoint changes cannot be generalized to an aerial platform.

2. The author presents the crucial metrics of processing power needs and memory commitment to be considered at the time of selecting a VPR technique against aerial robotics. These metrics directly effect the practicality of using any VPR technique in a resource-constrained, battery powered aerial robot.

It is important to note that although the datasets used in this evaluation particularly contain 3-dimensional challenging viewpoint variance, with illumination and temporal appearance change; it has been created using a hand-held rod to imitate vertical viewpoint variation. Both lateral and vertical position of the hand-held rod are continuously varied to simulate a

drone-mounted camera. Despite the contrived nature of dataset, baseline VPR techniques struggle (as shown later in sub-section 5.3.1) from such simulated 6-DOF viewpoint variation, "validating" the difficulty of this dataset.

The VPR techniques used for evaluation in this chapter are a subset of the methods discussed in Chapter 4 (and in [134]), and have shown promising results on different ground-based datasets. The comparison performed for aerial robotics in this chapter is kept fair by deploying all the techniques on a common platform. In addition to the matching performance of all techniques, the author derives the relations for processing power consumption which is an important factor of consideration for battery powered drones given limited flight-time [162]. Unlike ground-based platforms, aerial robots are also limited by the available physical memory for storing data, one reason being the increase in payload thus, faster battery drainage. Also, larger memory size translates to greater memory power consumption. The author therefore, gives the projected memory requirements of all techniques for storing each dataset (feature descriptors of reference images) as a complete map. While there is significant research into compact storage of robot maps and place selection as reviewed in [58], it is out of the scope of this work and thus, the author considers all reference image descriptors as nodes of the map.

5.2 Experimental Setup

The VPR techniques used in this chapter have already been introduced in Chapter 2 and their exact details have been kept similar to those of Chapter 4. Therefore, in this section the author only explains the datasets and the evaluation metrics used for this research.

5.2.1 Evaluation Datasets

The datasets used in this chapter are introduced by Maffra et al. in [161]. Essentially the authors performed three traverses of a shopping street in the center of Zurich city from different viewpoints and create two datasets. One of the three traverses serves as a constant reference in both the datasets, while the other two traverses act as query images. Ground-truth is provided for all three traversals in the form of timestamps. The details of these datasets are summarized in Table 6.1.

Since the three traversals were recorded with a Visual-Inertial sensor that stores images and timestamps as ROS (Robot Operating System) bag files, the author of this thesis wrote a

Table 5.1 Benchmark Place Recognition Datasets

Dataset	Traverse		Environment	Variation	
	Test	Reference		Viewpoint	Condition
Shopping Street 1	8577	7494	Urban	moderate	moderate
Shopping Street 2	4781	7494	Urban	strong	moderate

Fig. 5.2 Samples of images from Shopping Street 1 dataset are shown here. Top row consists of query images while the bottom row shows reference images. While this dataset contains illumination variation and dynamic objects, it does not have any extreme viewpoint variation. This makes Shopping Street 1 dataset a good reference in comparison to 6-DOF viewpoint change of Shopping Street 2 dataset (sub-section 5.2.1).

simple Python utility to extract images from a bag file with filenames as timestamps. This is provided here[1] for future ease-of-use of any datasets created using ROS-based platforms.

Shopping Street 1 Dataset

This dataset consists of the two traverses of shopping street captured with a hand-held setup as shown in Fig. 5.2. The undertaken traverses exhibit moderate viewpoint and appearance variation with adequate perceptual aliasing. While this dataset does not pose any significant 6-DOF viewpoint change as compared to existing VPR datasets, it serves as a good reference for the objective of this paper: observing the effect of extreme 6-DOF viewpoint change in comparison to moderate viewpoint changes. Therefore, the author evaluates the 8 state-of-the-art VPR techniques discussed in sub-section 4.2.1 on this dataset to give a qualitative and quantitative insight into their prowess under moderate viewpoint changes.

[1] https://github.com/MubarizZaffar/rosbagextraction/

Fig. 5.3 Samples of images from Shopping Street 2 dataset are shown here. Top row consists of query images taken using a rod-mounted camera while bottom row shows images taken by a handheld camera. Challenging viewpoint variation is depicted here, which is very similar to the variation experienced by a 6-DOF aerial robot.

Shopping Street 2 Dataset

The Shopping Street 2 dataset contains the interesting 6-DOF viewpoint change. This viewpoint change has been introduced by mounting the camera on a 4 meter long rod such that the motion of camera imitates the flying behavior of a drone. This dataset also contains significant illumination variation and temporal appearance change. The author shows some sample query and reference images in Fig. 5.3.

5.2.2 Evaluation Metrics

Matching Performance

In image-retrieval for VPR, area under the precision-recall curves (AUC) is a well-established evaluation metric, as widely reviewed in this thesis. Therefore, to maintain consistency in this chapter, the author only computes and reports AUC performance by utilising equation 2.3.

Processing Power Consumption

The power consumption of a CPU is directly related to the CPU utilisation of running processes as shown by authors in [163]. Over-time, this power consumption becomes a critical factor for battery powered aerial robots. Since, computationally intense processes

running for longer time-periods will quickly drain the battery, they lead to reduction of the single-charge flight-time of a drone. Therefore, the author builds upon the power consumption relations of [163] and derives the battery expense (Ampere-hours) for each of the 8 VPR techniques. The CPU power consumption is linked to CPU utilisation by the below equation 5.1.

$$P_c = P_i + (P_b - P_i) \times U \tag{5.1}$$

$$\begin{aligned} where;\ P_c &= Power\ consumption\ of\ CPU \\ P_i &= CPU\ power\ consumed\ in\ idle\ state \\ P_b &= CPU\ power\ consumed\ under\ full\ load \\ U &= CPU\ utilisation \end{aligned}$$

Given that the author uses the same computational platform i.e. Intel(R) Xeon(R) Gold 6134 CPU @ 3.20GHz for evaluating all 8 VPR techniques, P_i and P_b can be taken as constants while U is a variable parameter. Thus, by taking P_i as an offset a and $P_b - P_i$ as the slope s, equation 5.1 can further be modified as below.

$$P_c = a + (s \times U) \tag{5.2}$$

The CPU utilisation U can further be broken down into the CPU utilisation U_e for an image feature descriptor encoding and CPU utilisation U_m for feature descriptor matching. These two CPU utilisations correspond to the feature encoding time t_e for an input query image and query descriptor matching time t_m for M ($M = 7494$) reference images in the database. Since encoding an input query image and matching it with all the reference images in the database is the deployment application of VPR techniques, the power consumed P_q by such a process can be represented as;

$$P_q = P_e + P_m \tag{5.3}$$

$$P_e = a + (s \times U_e) \tag{5.4}$$

$$P_m = a + (s \times U_m) \tag{5.5}$$

Given that CPUs are powered from a constant voltage rail V (typically $V = 2.5$ volts), the ampere-hours consumed per query image Ah_q can be estimated from equation 5.6. Thus, the total Ah consumption Ah_t of each VPR technique for N query images and M reference images can be computed by using equation 5.7.

$$Ah_q = \frac{P_e \times t_e + P_m \times t_m}{V} \tag{5.6}$$

$$Ah_t = N \times Ah_q \tag{5.7}$$

Projected Memory Requirement

Although the ability to retrieve correct image matches is critical for a VPR technique, there is a trade-off between the amount of salient information that is encoded and the available on-board storage. Thus, although a VPR method can achieve excellent matching performance, its deploy-ability on an aerial platform depends on the memory footprint of its image descriptors. Therefore, for each of the 8 VPR techniques, the author provides a projected memory consumption for storing the descriptors of reference images corresponding to a complete environment traversal.

5.3 Results and Analysis

This section discusses the performance evaluation of all the employed VPR techniques. A separate subsection is allocated to each criterion including matching performance, computational power requirements and memory usage.

5.3.1 Matching Performance

For both the benchmark datasets, this sub-section outlines and compares the AUC under PR-curves of all the 8 VPR approaches. For a qualitative insight, the author has also displayed example scenarios where query images are successfully matched or mismatched by the employed VPR techniques.

Shopping Street 1 Dataset

For this dataset, Fig. 5.5 illustrates the PR-curves of all the employed approaches. The dataset contains mostly less-challenging planar (2-dimensional) viewpoint variation but has moderate illumination changes and occasionally observed dynamic objects, therefore, most of the techniques perform well. This shows the success of all the recently proposed VPR techniques provided moderate viewpoint variation in the dataset. Out of all the techniques and using AUC under PR-curves as an evaluation parameter, NetVLAD achieved state-of-the-art performance followed by AMOSNet/HybridNet, RMAC and CALC with very minimal differences. Examples of matches/mismatches are shown in Fig. 5.4.

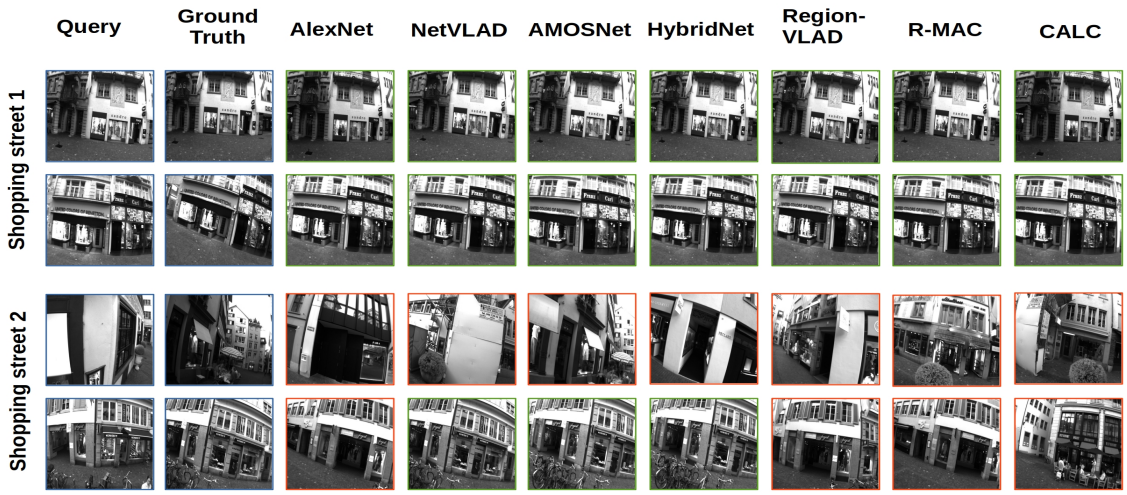

Fig. 5.4 Example images retrieved by all VPR techniques on both datasets are shown here. All techniques show excellent matching performance on Shopping Street 1 dataset, but struggle with 6-DOF viewpoint variation in Shopping Street 2 dataset.

Shopping Street 2 Dataset

The query and reference traverse exhibit strong viewpoint variation in this dataset (see Fig. 5.3). Fig. 5.6 reports the PR curves for this dataset; showing across-the-board decline in matching performance with NetVLAD still outperforming other VPR techniques.

The significant performance degradation for all other approaches (in comparison to NetVLAD) can be associated with the training of the CNN models. For instance, authors of NetVLAD [107] trained VGG-16 on an urban 250k place-centric Pittsburgh dataset exhibiting strong condition and viewpoint changes coupled with dynamic objects such as pedestrian, vehicles etc. Whereas, although HybridNet used strong condition-variant SPED dataset, the dataset intrinsically does not contain any viewpoint variation, thus, failing to perform on Shopping Street 2 dataset. Since, HybridNet is also the underlying model for Region-VLAD, where Region-VLAD does not explicitly tackle viewpoint variation; matching performance degrades under 6-DOF viewpoint change. Similarly for RMAC, VGG-16 was pre-trained on object-centric ImageNet dataset, therefore, it is not efficient in dealing with severe changes in viewpoint. Although, authors of CALC have trained their auto-encoder with viewpoint variant input images, the nature of variation is random planar projections which leads to the observed performance degradation for aerial place recognition.

In summary, a common observation across all baseline techniques is the decline of matching performance from lateral viewpoint variation of Shopping Street 1 dataset to 6-

Table 5.2 Computational Power Requirements

Computational Performance	**VPR Techniques** (Platform: Intel(R) Xeon(R) Gold 6134 CPU @ 3.20GHz with 32 cores, 64GB RAM)							
	AlexNet	**NetVLAD**	**AMOSNet**	**HybridNet**	**Cross-Region-BOW**	**R-MAC**	**Region-VLAD**	**CALC**
U_e	0.734	0.656	0.437	0.437	0.32	0.5	0.25	0.781
U_m	0.0312	0.036	0.03	0.03	0.1	0.371	0.031	0.0312
$t_e(sec)$	0.666	0.77	0.359	0.357	0.834	0.478	0.463	0.027
$t_m(sec)$	3.222	0.0374	0.614	0.584	1199.04	0.254	0.899	0.974
Ah_t	0.3128	0.2688	0.0931	0.0921	63.836	0.1768	0.0764	0.0272

DOF viewpoint variation of Shopping Street 2 dataset. However, the trend of this decline is different between NetVLAD and the remainder techniques, primarily due to the absence of viewpoint variation in the training datasets of latter. These observations outline the need and significance of large-scale, 6-DOF viewpoint variant datasets for training VPR techniques, especially for aerial robotics.

5.3.2 Processing Power Consumption

When we generally talk about aerial robotics or resource-constrained platforms, energy management is the key component for any on-board deployed application. Thus, while different VPR techniques have been proposed over the years, each achieving incremental matching performance improvement and immunity to challenging conditional variations; a thorough investigation of their practicality for VPR is presented in this sub-section. The author enlists in Table 6.1, the CPU utilisation U_e for feature encoding, CPU utilisation U_m for feature matching, feature encoding time t_e and feature matching time t_m. By taking $a = 0$, $s = 1$, $V = 2.5$ and $N = 4781$, we also enlist the total battery consumption Ah_t for all VPR techniques to give a comparative analysis. The units of times t_e and t_m were changed from seconds to hours for Ah_t computation. Since, CPU utilisation is a highly fluctuating variable therefore, we take its average over run-time of a process with a sampling rate of $0.01sec$. It can be clearly seen that Cross-Region-BOW is the most power-hungry VPR technique primarily due to its computationally intense feature matching methodology. On the other hand, CALC stands-out to be the most energy-efficient technique for VPR. Please note that the author does not explicitly optimize any of the VPR techniques for performance enhancement and the values of Ah_t in Table 6.1 will scale with the values of a and s.

5.3.3 Projected Memory Requirement

One of the well-researched areas in robotic navigation and mapping is the efficient storage and indexing of a robot map. This primarily involves selection of images that correspond to distinct places based on either time-interval [5], distance [60], distinctiveness [61] or memorability [2]. While many different techniques have been presented in this regard, the

underlying limitation is posed inherently by the memory footprint of a technique's image descriptor. Therefore, although place/image selection schemes can reduce the total number of images to be stored in a robot map, the size of the map will still scale with the size of a VPR technique's output descriptor. Thus, with abstraction to any place selection scheme employed, the author reports in Fig. 5.7, the projected memory consumption for all 8 VPR techniques, given that the descriptors of all reference images are to be stored. Please note that the size of reference database for both the evaluation datasets used in this work is same and hence information in Fig. 5.7 is equally applicable.

5.4 Summary

In this chapter, the author performed a thorough evaluation of visual place recognition state-of-the-art on two aerial place recognition datasets. It is shown that contemporary VPR techniques generally perform well on datasets containing moderate changes in viewpoint even under severe variations in illumination and conditions. However, the notable change of matching performance in between the two datasets (Shopping Street 1 and Shopping Street 2) reveals the extent of challenge posed by viewpoint variance; especially for 6-DOF (degrees-of-freedom) platforms like drones.

The author also presents the limitations of VPR techniques from computational and storage perspectives given the limited on-board resources and energy supply of an aerial robot. This evaluation is the first step into generalisability analysis of VPR techniques between different platforms and can be further extended upon proposal of more challenging aerial datasets in future. Generally, the motor power requirement for drones takes the huge chunk of battery storage rather than the computational platform, however our objective in this chapter was to see the local comparison between techniques for such power requirements.

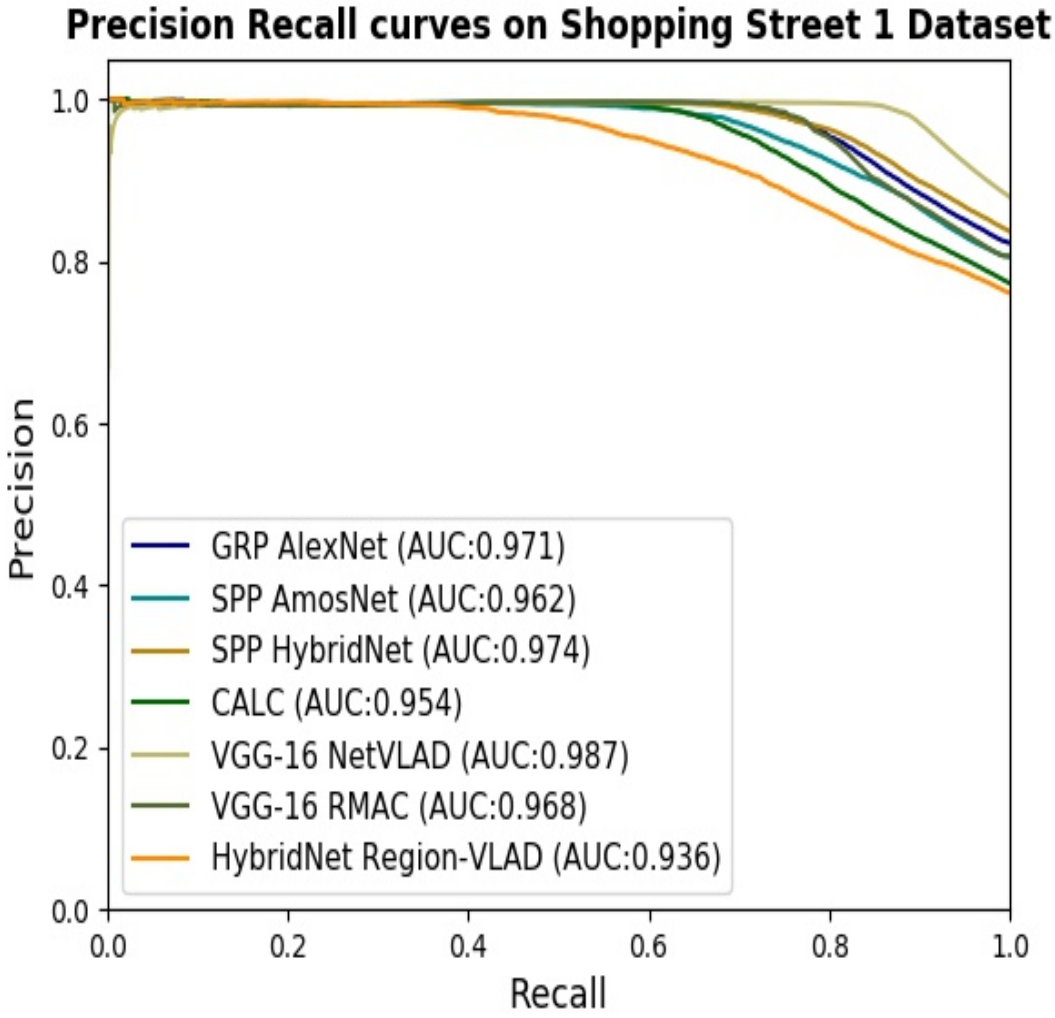

Fig. 5.5 AUC-PR curves of the employed VPR approaches on Shopping Street 1 Dataset are shown here. All VPR techniques achieve near-to-ideal matching performance on this dataset, advocating that the past few years of VPR research has been highly successful against planar viewpoint variations and conditional changes.

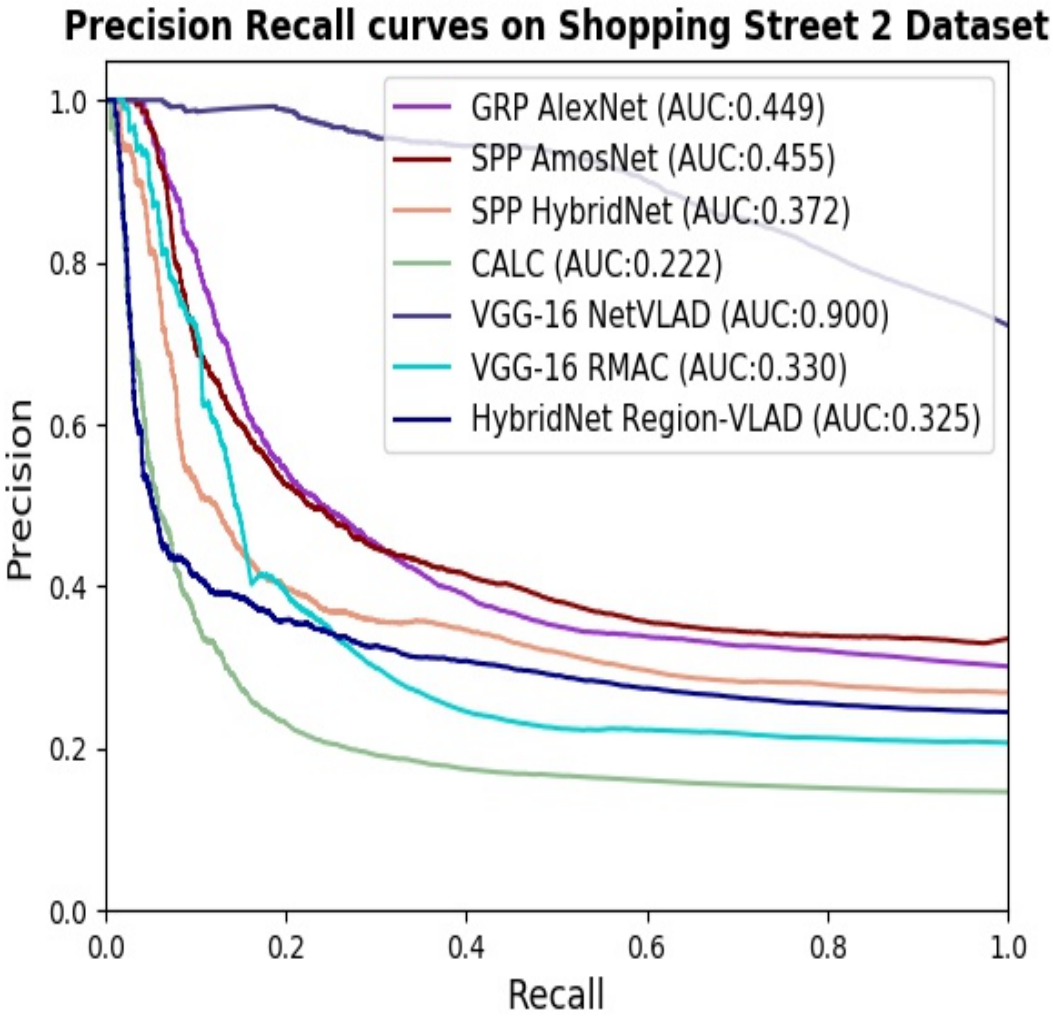

Fig. 5.6 AUC-PR curves of the employed VPR approaches on Shopping Street 2 Dataset are shown here. All VPR techniques clearly suffer from 6-DOF viewpoint variation in this dataset, with NetVLAD achieving state-of-the-art matching performance.

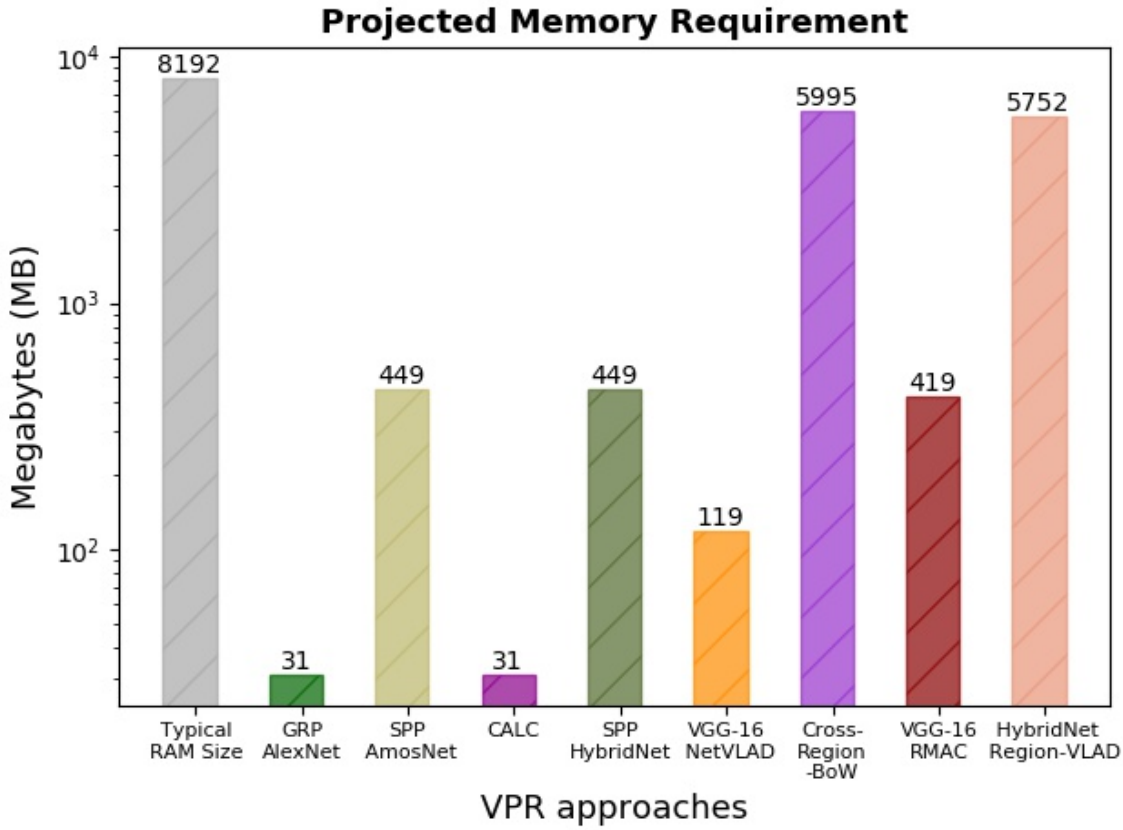

Fig. 5.7 Projected memory requirements for all the VPR approaches are shown here. The vertical axis is in logarithmic scale for clarity. The leftmost bar shows the typical RAM size of various development platforms. The reference features may not be loaded in RAM and could essentially be stored in an on-board SD Card which usually has similar storage capacity.

Chapter 6

CoHOG: A Light-weight, Compute-efficient and Training-free VPR Technique

The evaluations performed in Chapter 4 and Chapter 5 identified that the computational performance in the form of image retrieval time, training time and memory consumption are also important factors to be considered during the design of a VPR system. The existing VPR techniques either do not regard these requirements or the place matching performance of computationally-efficient techniques is significantly lower than the state-of-the-art. Therefore, in this chapter, the author presents a novel, compute-efficient and training-free approach based on Histogram-of-Oriented-Gradients (HOG) descriptor for achieving state-of-the-art performance-per-compute-unit in Visual Place Recognition (VPR). The inspiration for this approach (namely CoHOG) is based on the convolutional scanning and regions-based feature extraction employed by Convolutional Neural Networks (CNNs). By using image entropy to extract regions-of-interest (ROI) and regional-convolutional descriptor matching, this technique performs successful place recognition in changing environments. The author uses viewpoint- and appearance-variant public VPR datasets to report this matching performance, at lower RAM commitment, zero training requirements and 20 times lesser feature encoding time compared to state-of-the-art neural networks. The image retrieval time of CoHOG and the effect of CoHOG's parametric variation on its place matching performance and encoding time is also discussed.

6.1 Background

The Chapter 1 of this thesis introduced in detail the VPR problem and the various challenges within VPR. It is a well-defined, albeit a highly challenging one, since places change their appearance rapidly due to varying viewpoints and conditions. Other than the environmental variations, texture-less and low-informative scenes also pose difficulty to place matching. Fig. 1.4 has previously shown these different variations that a VPR system is expected to be robust against. However, in addition to these variations, a VPR systems is constrained by run-time memory, processing power and/or pre-deployment training needs.

Prior to the use of neural network based techniques, VPR research was primarily based on local and global handcrafted feature descriptors, as reviewed in Chapter 2 of this thesis. Local feature descriptors extract and describe keypoints (areas of interest) from an image, therefore they are primarily viewpoint invariant but suffer from illumination variation. Global feature descriptors, on the other hand, suffer from translational and/or rotational viewpoint change but they are moderately illumination invariant. Moving away from handcrafted feature descriptors, the application of Convolutional Neural Networks (CNNs) to VPR was first studied by Chen et al. [103]. Since then, different CNNs with and without architectural modifications have incrementally shown state-of-the-art VPR performance. However, CNNs (and Convolutional Auto-encoders as in [113]) require significant model training with their deployment accuracy directly linked to the size, inter-sample variance and nature of the training dataset. Training of VPR-specific CNNs requires large-scale labelled datasets of places from a multitude of environments, which is a practical limitation. Moreover, training of these CNNs requires dedicated Graphical Processing Units (GPUs) with training time usually ranging from a few days to a few weeks. One key limitation of neural network based techniques is their intense computational nature requiring significantly higher run-time memory and feature encoding time compared to handcrafted feature descriptors. Thus, while the success of these recent CNN-based techniques from the perspective of place matching is evident, their practical deployment in field is restricted. More specifically, such computational intensiveness raises concerns for deployability on resource-constrained platforms (including battery-powered aerial, micro-aerial and ground vehicles) as identified in [149] [164].

In this chapter, the author proposes a novel technique based on hand-crafted feature descriptors delivering state-of-the-art (or close to state-of-the-art) VPR performance with no training requirements compared to CNNs. The proposed technique has significantly lower feature encoding time and RAM commitment while delivering comparable place matching performance on challenging viewpoint- and conditionally-variant datasets. The inspiration for this approach is drawn from the following:

1. By design, CNNs are able to scan an entire image for a particular feature and irrespective of the location of that feature in an image, the same CNN filter (layer activations) will fire.

2. CNNs trained/fine-tuned for VPR have the ability to extract regions-of-interest (ROI) which are informative and distinct.

3. CNNs trained on condition-variant VPR datasets can internally learn representations of places/images which are immune to seasonal and illumination variations.

From the above list, both 1 and 2 contribute towards viewpoint invariance. This is further improved by manually introducing viewpoint variation in training datasets. Conditional invariance is predominantly the result of 3, not user-defined and essentially a black-box.

By deriving motivation from this behavior of CNNs, the proposed technique first computes the entropy map of an image and extracts information-rich regions from it. Each of these ROI are then locally described by dedicated HOG-descriptors. Secondly, the author uses convolutional matching of regional HOG-descriptors that provides viewpoint invariance. This regional-convolutional matching is based on standard matrix multiplication and is therefore compute-efficient. For illumination invariance, block normalization of HOG-descriptors is used, which shows acceptable performance on conditionally-variant datasets. The author's choice of HOG-descriptor is based on its reliable performance across illumination and seasonal variation as shown by McManus et al. [101], and its utility as an underlying feature descriptor for training a VPR-specific convolutional auto-encoder in [113]. The image retrieval scheme of CoHOG can be summarized by Fig. 6.1.

6.2 Methodology

This section presents the methodology adopted in this work that constitutes CoHOG. The proposed technique can be broken down into 7 primary blocks (as shown in Fig. 6.2) for end-to-end VPR image retrieval. The query image can be any incoming RGB camera frame which is converted to grayscale and resized to $W1 \times H1$. The robot map consists of precomputed HOG-descriptors of reference images. Please note that the author has used 'vanilla' HOG in this work, but the implementation computes HOG in the regional sense instead of the usual global fashion. A sub-section is dedicated to each of three crucial computational tasks of the proposed technique, namely HOG-descriptor computation, ROI extraction and regions-based convolutional matching.

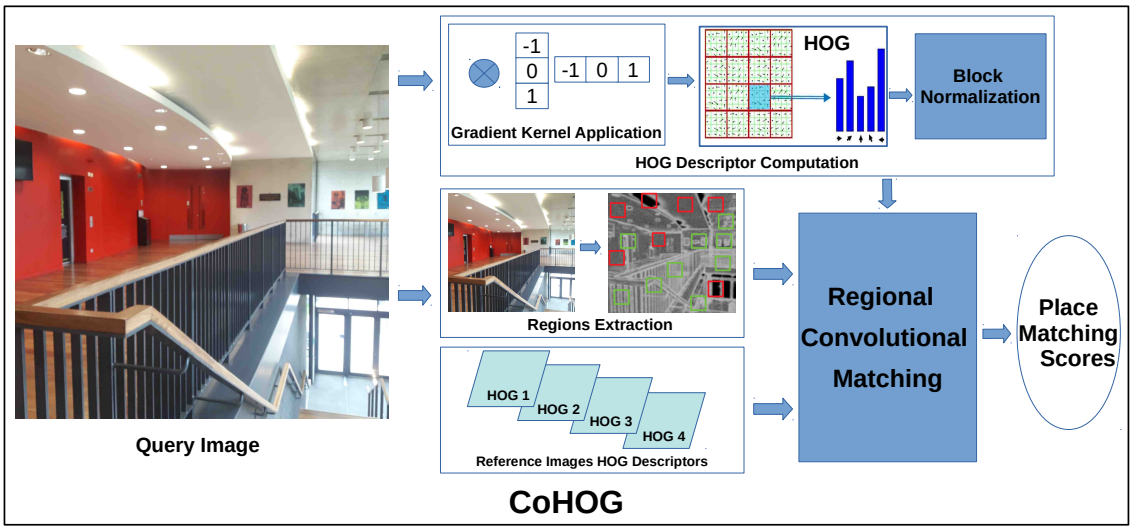

Fig. 6.1 The developed technique (CoHOG) is explained here. Each query image goes through ROI extraction and HOG computation, which are then fed to the convolutional matching block. This block outputs a similarity score against each reference image in the robot map. The green squares in region extraction block represent salient regions while the red squares are less-informative confusing regions.

6.2.1 ROI Extraction

Regions-of-interest based image matching has recently been the subject of significant VPR research [3] [110] [111]. In CoHOG, the author uses regions in an image that are information-rich. Firstly, the entropy map for each query image is computed using the following algorithm.

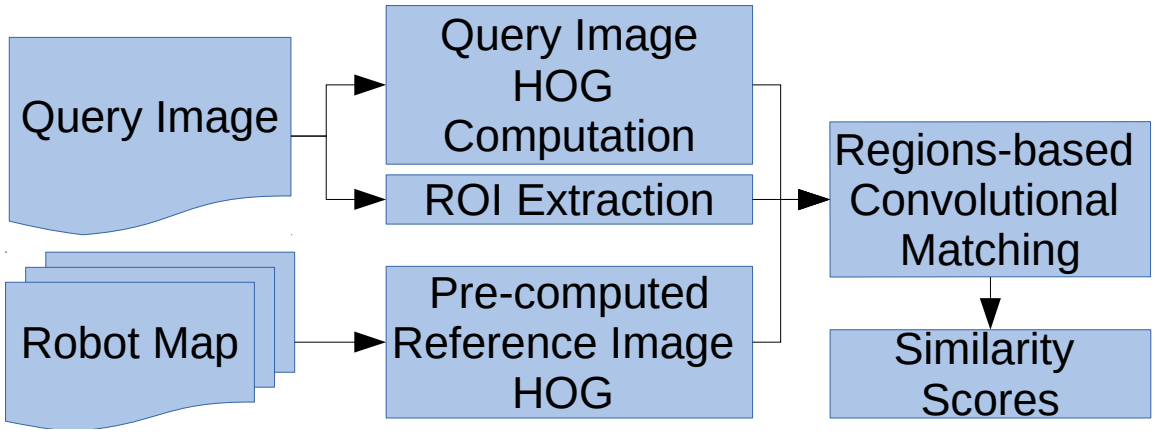

Fig. 6.2 The block-level overview of CoHOG is shown here.

Algorithm Computing Entropy Map

Entropy_Map = *Zeros_Matrix*$(W1, H1)$
Max_Entropy_Value = $log_2(256) = 8$
Create a Histogram of 256 *Pixel Intensity Bins*
Radius = *UserDefinedConstant*
for all *Pixels in Image* **do**
  Origin = *Pixel_Index*
  Local_Neighbourhood = *Circle*$(Origin, Radius)$
  Local_Neigh_List = *Append*$(Local\_Neighbourhood)$
end for
for all *Elements in Local_Neigh_List* **do**
  for all *Valid Pixels in Local_Neighbourhood* **do**
    if *Current_Pixel_Intensity lies in BinX* **then**
      Items_in_BinX = *Items_in_BinX* + 1
    end if
  end for
  Entropy_Map$(i, j) = log_2(No.\ of\ Filled\ Histogram\ Bins)$
  Clear all Histogram Bins
end for
Normalize Entropy_Map with Max_Entropy_Value

The entropy map has the same dimensions as the query image i.e. $W1 \times H1$ and example query images with entropy maps computed using this algorithm are shown in Fig. 6.3. The author now defines a region in an image as a $W2 \times H2$ image patch. Thus, a $W1 \times H1$

image with regions/patches of size $W2 \times H2$ contains N regions in total, whose goodness is represented by a matrix R;

$$Where; \quad N = (H1/H2) \times (W1/W2)$$

$$R = \begin{bmatrix} r_{11} & r_{12} & \dots & r_{1n} \\ r_{21} & r_{22} & \dots & r_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ r_{mn} & r_{mn} & \dots & r_{mn} \end{bmatrix}$$

$$Where; \quad m = H1/H2 \quad n = W1/W2$$
$$r_{ij} \in \mathbb{Z}_2 | \mathbb{Z}_2 = [0,1]$$
$$r_{ij} = 1 \,| Region = Good$$
$$r_{ij} = 0 \,| Region = Confusing$$

For evaluating the goodness r_{xy} of each region in R, the entropy map is represented as a matrix E. This entropy matrix has a size of $W1 \times H1$ with element values between $0-1$ (with 1 being the ideal value) and has the below shape.

$$E = \begin{bmatrix} e_{11} & e_{12} & \dots & e_{1W_1} \\ e_{21} & e_{22} & \dots & e_{2W_1} \\ \vdots & \vdots & \ddots & \vdots \\ e_{H_1W_1} & e_{H_1W_1} & \dots & e_{H_1W_1} \end{bmatrix}$$

$$where; \quad \{e_{ij} \in K \mid K \subseteq \mathbb{R} \wedge K = \{0, \dots, 1\}\}$$

The goodness r_{xy} of each region is calculated by thresholding the average entropy values of a $(W2 \times 2) \times (H2 \times 2)$ block size i.e. each block containing 4 regions (each region of size $W2 \times H2$), where all 4 regions have a common corner. Such a block-level evaluation provides consistency with HOG-descriptor computation, as shown later in sub-section 6.2.2. The stride of this block-level goodness evaluation is $Stride = W2 = H2$ and hence the total number of regional blocks for evaluation is $M = n-1 \times m-1$. All G regions which have an entropy score e_{xy} greater than or equal to the goodness threshold GT are selected for matching. Therefore, G is a variable depending on the scene being represented in an image and may vary from one query image to another. Selecting regions in this manner compared to

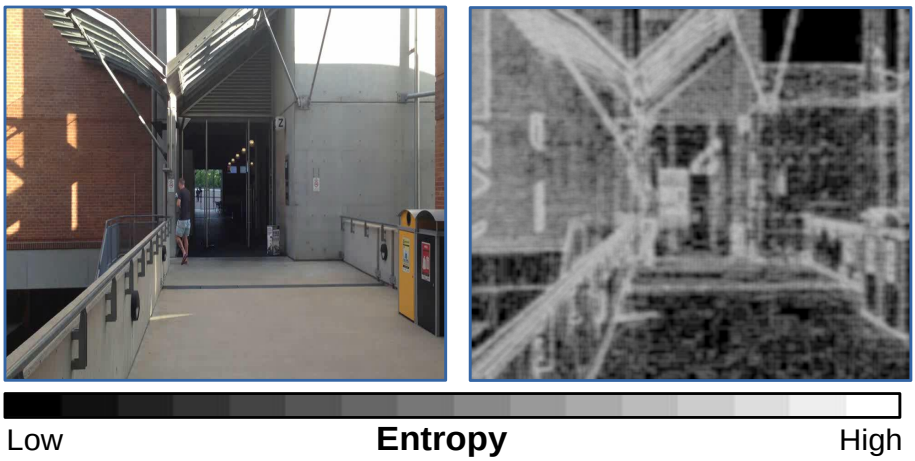

Fig. 6.3 Example of a query image [left] with its corresponding entropy map [right] is shown here. Texture-less walls and floors get filtered out as lower entropy areas which is consistent with the author's motivation to discard such regions.

the conventional Top-G (where G is a constant) regions selection provides more saliency and computational advantages. If an image has more confusing regions, only a few salient regions are selected. This helps in successfully matching low-textured images and is not possible with Top-G regions selection. Discarding confusing regions before regional convolutional matching also leads to lesser computational intensity. Fig. 6.4 shows examples of good regions extracted with varying GT.

6.2.2 HOG-descriptor Computation

Histogram-of-Oriented-Gradients (HOG) [165] [166] is a well-established handcrafted computer vision technique used originally for object detection. The end-to-end HOG-descriptor computation is quickly summarized as follows:

1. A gradient map is computed for an input grayscale image of size $W1 \times H1$.
2. A histogram-of-oriented-gradients (HOG) is created and computed for all N regions in the image, where every region has a size of $W2 \times H2$. Each regional-histogram has L bins, such that a bin is identified by a range of gradient angles assigned to it.
3. HOG computed previously is L2-normalised at a block level of size $(W2 \times 2) \times (H2 \times 2)$. This results in a descriptor of depth $4 \times L$ with the total number of block-level HOG-descriptors equal to M. Refer to sub-section 6.2.1, each ROI now has a

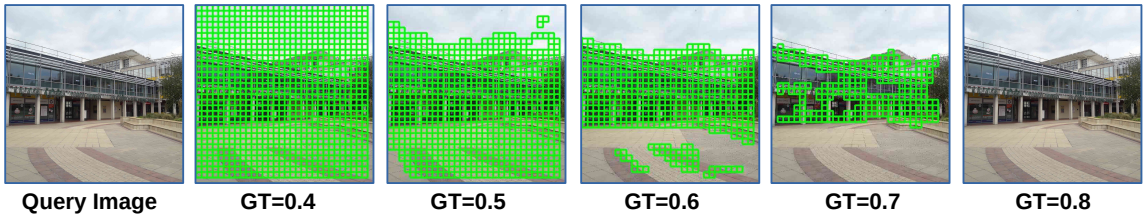

Fig. 6.4 ROI extracted by CoHOG are shown here with varying GT. Each good region is represented by a green colored square. Increasing GT reduces the number of regions selected by the proposed technique. A clear range exists between $GT = 0.5 - 0.7$, where confusing and low-informative regions coming from sky and texture-less walls/floors are filtered out, while maintaining a reasonable number of regions for subsequent regional convolutional matching.

corresponding HOG-descriptor of depth $4 \times L$ which is illumination invariant and can be easily indexed/retrieved.

6.2.3 Regions based Convolutional Matching

After HOG-descriptor computation, a query image is essentially converted into M regions with each region described by a vector of length $4 \times L$. Based on ROI evaluation, these M regions are reduced to G salient regions. This allows the author to shape the query image HOG-descriptor as a 2-dimensional matrix A with dimensions $[G, 4 \times L]$. The reference image is also composed of M regions with descriptors of depth $4 \times L$. Thus, the reference image is shaped as a matrix B with dimensions $[M, 4 \times L]$. Next, standard matrix multiplication is performed between A and B^T yielding a matrix C of dimensions $[G, M]$. Each row of matrix C represents a query image region and every column in C represents the cosine-matching scores against all reference image regions.

The author employs max-pooling across all rows of matrix C to find the best matched reference image candidate region for every query image region, which yields a vector D having length G. Finally, the author takes the arithmetic-mean of vector D giving us the similarity score of a query and reference image in the range of $0 - 1$. A query image is matched with all reference images, such that the reference image with the highest matching score is selected as the best match.

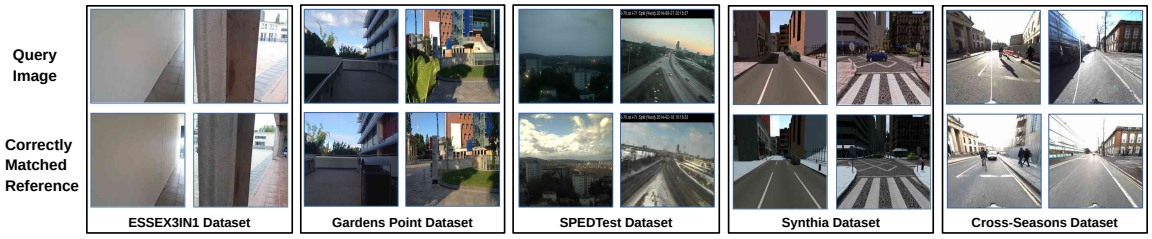

Fig. 6.5 Samples of correctly matched places by CoHOG on all 5 datasets are shown here. Given the viewpoint variations in ESSEX3IN1 [2] and Gardens Point datasets [1] datasets, CoHOG's regional-convolutional matching scheme can retrieve correct matches from the database. Even with the conditional variation in SPEDTest [110], Synthia [128] and Cross-Seasons datasets [129], the proposed technique is able to correctly match places. More samples of correctly and incorrectly matched places and the open-source technique are provided at https://github.com/MubarizZaffar/CoHOG_Results_RAL2019

6.3 Results and Analysis

This section first discusses the experimental setup used in this analysis including the VPR datasets, VPR techniques and evaluation metric used for assessing CoHOG's performance. The author then presents a detailed qualitative and quantitative comparison of CoHOG with state-of-the-art VPR techniques on the fronts of image matching, feature encoding time and run-time memory requirements. The image retrieval performance of CoHOG is also discussed and the effect on computational and matching performance by varying different parameters is shown to give the reader an insight into the selection of thresholds.

6.3.1 Experimental Setup

In order to evaluate CoHOG, the author has utilised 5 VPR datasets that represent all the challenges in VPR (as identified in Chapter 1). For viewpoint variation, the Gardens Point dataset [1] is used. Secondly, the ESSEX3IN1 dataset which was introduced in [2] and contains highly confusing and challenging images of places is used. Thirdly, the SPEDTest dataset is used which has been introduced in [103]. In this chapter, the author also employed the synthetically created Synthia [128] dataset, which consists of city-like traversal during Winter and Spring seasons. The number of query and reference images are 959 and 947, respectively. Finally, the author used the low-quality, highly dynamic and blurry Cross-Seasons dataset [129] consisting of 206 sunny query images and 202 dusk reference images. Results on this dataset present the failure-cases of CoHOG and identify important directions for future research.

For comparison with CoHOG, the author used all contemporary VPR techniques employed in Chapters 4 and 5. The implementation details, selected parameters and evaluation platform have all been kept similar to the setup of Chapter 4 for a fair comparison, except that AlexNet is used for the Region-VLAD approach instead of HybridNet. The author has also reported the performance for using Top-G (at G=200, 400 and 800) based regions selection with CoHOG. As the evaluation-metric, the author examined the place matching performance per compute unit of all VPR techniques. The extensive review performed by Lowry et al. in [8] identifies high precision to be a desirable characteristic of a VPR system due to the advent of false-positive prediction systems (as in [150] [167] [168]). On the other hand, authors in [113], [111], [149] and [134] have identified feature encoding time (t_e) to be a crucial computational metric. Therefore, by combining precision at 100% recall with encoding time per image, the author defines the Performance-per-Compute-Unit (PCU) as below.

$$PCU = Precision \times log(\frac{Max\ Feature\ Encoding\ Time}{Feature\ Encoding\ Time} + 9)$$

In the above equation, higher precision directly leads to higher PCU. However, for feature encoding time t_e, the author computes the logarithmic encoding time boost for a given VPR technique to provide a reasonable combination of precision and encoding time metrics. Thus, only exponential increase in encoding time for a highly precise VPR technique leads to increase in PCU. Maximum feature encoding time (t_{e_max}) belongs to the most computationally intensive VPR technique, which in this case is Cross-Region-BoW with the highest feature encoding time of 0.83 seconds. A scalar '9' is added to ensure that $PCU = Precision$ for the technique with $t_e = t_{e\_max}$, instead of $PCU = 0$, thus providing an interpretable scale.

6.3.2 Performance Evaluation

This section provides a detailed comparison of CoHOG with state-of-the-art VPR techniques on the frontiers of performance-per-compute-unit and run-time memory requirements. The reported performance is for $GT = 0.5$, $W1 = H1 = 512$, $W2 = H2 = 16$ and $L = 8$.

Place Matching Performance

This sub-section presents the PCU of CoHOG in comparison with other VPR techniques. While Fig. 6.6 shows the PCU of all techniques, the absolute values of precision at 100% recall and feature encoding time are listed in Table 6.1 for the reader's reference.

Table 6.1 Place Matching Precision, Feature Encoding Time and RAM Commitment

Performance Metric	**VPR Techniques** (Platform: Intel(R) Xeon(R) Gold 6134 CPU @ 3.20GHz with 32 cores, 64GB RAM, No GPU)												
	HOG	**AlexNet**	**AMOSNet**	**HybridNet**	**CALC**	**Cross-R-BOW**	**NetVLAD**	**R-VLAD**	**RMAC**	**Top-200**	**Top-400**	**Top-800**	**CoHOG**
Precision ESSEX3IN1	0.01	0.14	0.26	0.28	0.1	0.62	0.76	0.56	0.12	0.75	0.79	0.82	0.84
Precision Gardens	0.2	0.49	0.64	0.81	0.44	0.81	0.95	0.9	0.42	0.74	0.82	0.87	0.9
Precision SPEDTest	0.02	0.03	–	–	0.02	0.5	0.74	0.54	0.6	0.4	0.44	0.48	0.51
Precision Synthia	0.37	0.89	0.91	0.92	0.76	0.89	0.95	0.86	0.92	0.67	0.83	0.91	0.92
Precision CrossSeasons	0.5	0.85	0.93	0.96	0.67	0.9	0.97	0.87	0.83	0.68	0.75	0.43	0.65
Encoding Time(sec)	0.007	0.67	0.36	0.36	0.027	0.83	0.77	0.46	0.47	0.02	0.02	0.02	0.02
RAM Consumption(MBs)	0.02	47.04	4.22	4.22	2.3	0.58	1.21	47.04	0.58	0.06	0.06	0.06	0.06

CoHOG achieves state-of-the-art PCU on all the 5 datasets utilised in this chapter as shown in Fig. 6.6. The author also reports state-of-the-art precision on ESSEX3IN1 dataset and comparable precision on other datasets (except cross-seasons dataset), as listed in Table 6.1. The viewpoint variation in ESSEX3IN1 dataset is catered for by CoHOG's regional convolutional matching while confusing frames (and/or regions within) are handled by our entropy-based region extraction. This matching performance is qualitatively shown in Fig. 6.5. The author achieved close-to-ideal place matching precision on Gardens Point dataset and Fig. 6.5 shows samples of places correctly matched by the proposed technique despite the viewpoint variation. The nature of challenges handled in SPEDTest and Synthia datasets is also depicted in Fig. 6.5, where the author shows that under notable seasonal and illumination changes, CoHOG can still retrieve correct place matches. However, the cross-seasons dataset consisting of low-quality images with motion blur and significant dynamic objects identifies important limitations of the proposed gradient-based technique, that can intrinsically be handled by neural network-based techniques. Please note that the average number of regions employed by CoHOG are 730, 790, 780 and 750 on SPEDTest, Synthia, Gardens Point and ESSEX3IN1 datasets, respectively, but it still achieves better matching performance than Top-800, similar to our motivation in sub-section 6.2.1.

The precision-recall curves for CoHOG are presented in Fig. 6.7. In an environment-aware VPR system, conditional variations are predictable [169] and can either be avoided or the VPR system be switched accordingly. Thus, given the lower computational and zero training requirements, CoHOG presents the best overall utility for a computationally-efficient VPR system in changing environments.

Run-time Memory Requirements

Due to their intense computational requirements, neural network-based techniques have significantly higher run-time memory consumption which is an important factor for resource-constrained and battery-powered robotic platforms that are usually running multiple tasks simultaneously. The author reports the run-time memory consumption of all VPR techniques in Table 1, which shows that CoHOG is light-weight compared to the rest of VPR techniques.

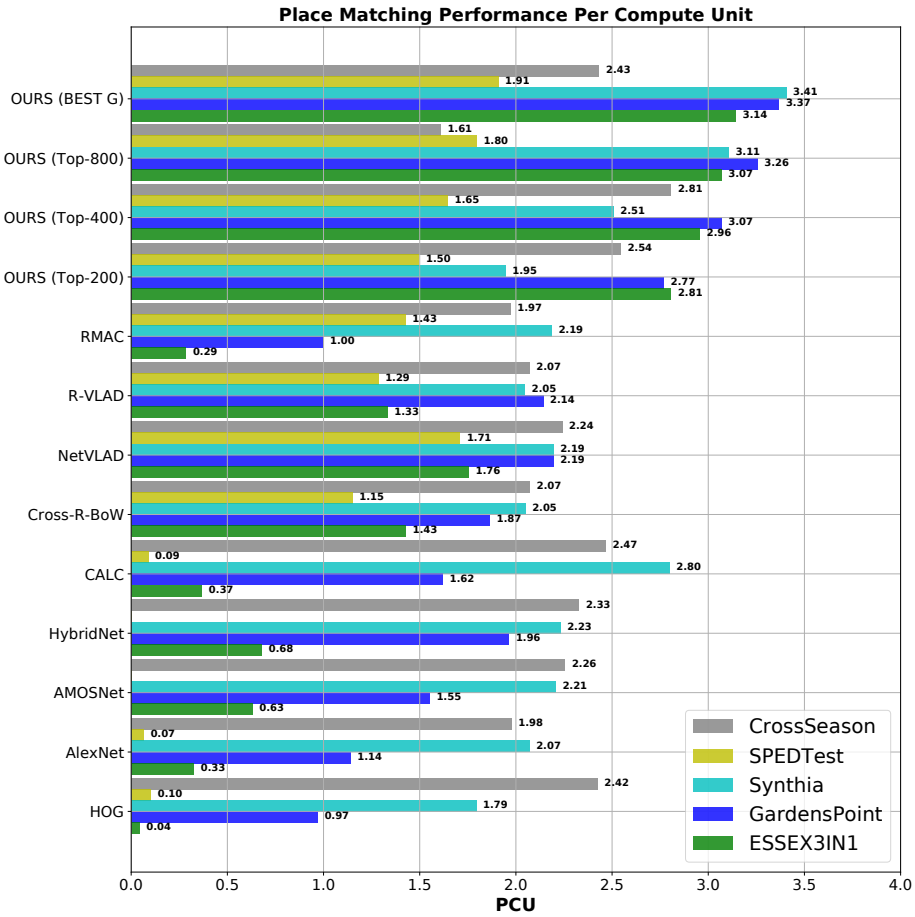

Fig. 6.6 The PCU of CoHOG is compared with all other VPR techniques. HybridNet and AMOSNet are trained on SPED dataset and thus not included for SPEDTest comparison.

This is because CoHOG intrinsically does not involve loading/deployment of any machine-learning models into RAM for feature extraction/description. The reported RAM commitment is only for encoding a single query image.

Descriptor Matching Time

The descriptor matching time (t_m) represents the time required to match the feature descriptors of 2 images and determines the retrieval performance of a VPR system. The image retrieval time (T) for any VPR system can be modelled as $T = t_e + O(Z) \times t_m$. Where, O(Z) represents the total number of prospective candidate matches and could be linear, logarithmic or other depending upon the employed neighbourhood selection mechanism (e.g., linear search, approximate nearest neighbour search etc.). The author further models t_m as $t_m = O(D) \times N1 \times N2$, where O(D) is the time required to match 2 descriptors of length D, N1 is the number

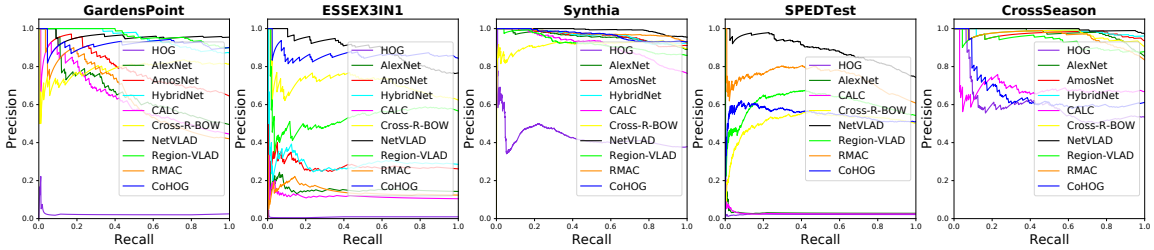

Fig. 6.7 The Precision-Recall curves for all 10 VPR techniques on the 5 datasets employed in this chapter are presented. (Vector Graphics: Zoom-In Supported)

of query image descriptors and N2 is the number of reference image descriptors. Theoretically, the value of t_m for CoHOG is $t_m = O(4 \times L) \times G \times M$. The values of t_e and t_m for the author's implementation of CoHOG are 0.02 sec and 0.2 msec, respectively, for the parameters specified in sub-section 6.3.2, such that the value of T will be $T = 0.02 + 0.0002 \times O(Z)$ sec.

Because it is computationally intractable to have a linear O(Z) for larger values of Z, different approaches exist to cater for this: 1) The total number of images in a map can be limited to a fixed value [58], 2) A spatial context can be introduced to search across images within a particular geographical radius [170], 3) A two-stage approach can be adopted to first extract possible candidate matches, followed by rigorous feature matching [120] [121], 4) Multi-processing systems can be employed to distribute the matching task across several processors. For further timing comparison between the techniques discussed in this work and understanding respective limitations, the author would refer the reader to Chapter 4 ([134]), provided the value of Z and the nature of O(Z) are known.

6.3.3 Parameter Sweep

This sub-section presents the effects of changing CoHOG's parameters. The parametric sweep is performed for GT, $W1$ and $W2$ on ESSEX3IN1 dataset. Each of the 3 parameters is first varied within a suitable range while keeping the other 2 constant, where the values of these constants are the same as used in sub-section 6.3.2. The author also shows the effect of varying $W1$ and $W2$ with a constant ratio.

The qualitative effect of variation in GT is already shown in Fig. 6.4 and the quantitative effect is reported in Fig. 6.8 (a). More salient regions and less confusing regions are selected with increasing GT, leading to improved matching performance. The quantitative contribution of GT to place matching performance is inherent to the places being represented in the dataset and may vary. While feature encoding-time is independent of GT, it depends on both $W1$ and $W2$, as reported in Fig. 6.8. Matching performance improves with increasing

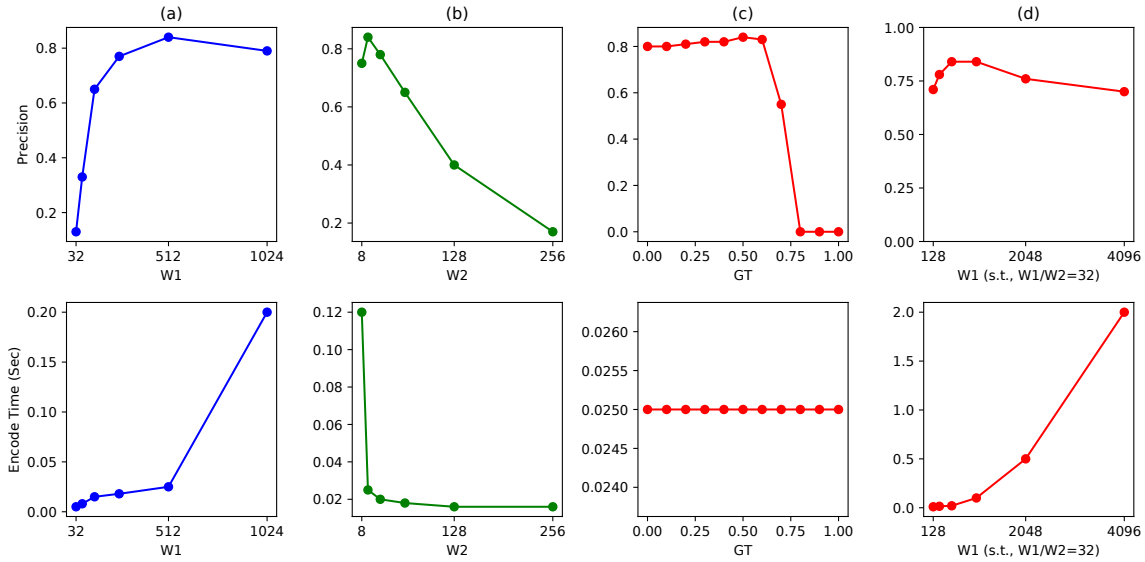

Fig. 6.8 The impact on CoHOG's performance by sweeping various thresholds within a suitable range is depicted here.

image-size [Fig. 6.8 (b)] as greater number of gradients now contribute to the regional HOG-descriptors, which also results in increased gradient bin-assignment time. Increasing the cell-size reduces viewpoint-invariance as lesser number of regions (N) are now available for regional-convolutional matching, thus reducing matching performance [Fig. 6.8 (c)]. The key take-away here is the critical ratio of $W1$ and $W2$ that determines the number of regions for entropy-evaluation and regional-convolutional matching. The area represented by each region in an image should be small enough to accommodate for viewpoint variation between adjacent regions and yet large enough for a suitable regional HOG-descriptor (i.e. each region should contain a reasonable number of intensity-change gradients). Fig. 6.8 (d) shows stable precision under a range of values for W1 and W2, given a constant value for $\frac{W1}{W2}$.

6.4 Summary

This chapter presented a light-weight, compute-efficient and training-free VPR technique (namely CoHOG) based on Histogram-of-Oriented-Gradients (HOG) descriptor that achieves state-of-the-art performance under computational constraints on standard VPR datasets. By evaluating on both viewpoint and appearance variant datasets, the utility of the proposed approach is discussed. The author shows highly precise place matching performance on view-

point variant datasets, while comparable precision is achieved on condition variant dataset. With zero training requirements, lower encoding time and lesser run-time memory footprint than neural networks, CoHOG promises better deployability in real-world applications. The technique presented in this work is agnostic in nature. Although, the author has used HOG as the regional descriptor but any feature descriptor can be plugged-in for robustness to viewpoint variations.

Chapter 7

Conclusions and Future Directions

Visual Place Recognition is a well-established, challenging and interesting research problem for both the computer vision and the robotics community. It represents the ability of a system to recognise a previously visited place using only visual information, under drastic viewpoint, illumination and/or seasonal changes. Other than the appearance variation challenges within VPR, computational requirements and storage needs are also crucial for real-world deployment of VPR in resource-constrained platforms.

The application domains of VPR are wide-spread and so are the researchers investigating it, which makes the VPR community diverse, multi-cultural, multi-talented and above all, exciting to be a part of. Moreover, the equipment required to undertake research in VPR is widely-accessible and several open-source datasets and VPR techniques exists for effective analysis and comparisons to quickly evaluate new research. This makes VPR open to a wide range of researchers coming from various backgrounds, which is generally not the case for most research domains that require state-of-the-art machinery and controlled environments. VPR is still far from saturation and several developments in hardware, newly proposed learning techniques, large scale datasets, inter-domain research combinations, better evaluation frameworks and incremental advances present a huge potential for any incoming VPR researcher.

This thesis is an amalgamation of various research gaps addressed within VPR. As such the author's research could be divided into 3 primary tracks: (a) Perceptual Aliasing within VPR and therefore predicting potential false-positives (Chapter 3), (b) Evaluation of VPR techniques, the metrics and the methodology for that (Chapter 4 and 5), (c) Incremental performance improvement by proposing novel VPR techniques (Chapter 6). Each of these chapters is a peer-reviewed publication and has been through a series of revisions. Because these chapters are mutually-independent, they identify parallel tracks for future research within VPR, which is the objective of this Chapter 7.

This chapter summarizes the presented research work in this thesis with potential future directions. Section 7.1 outlines the research and contributions presented in this thesis. Section 7.2 presents the potential future directions of research in the field of VPR.

7.1 Contributions Summary

This thesis presents the author's research performed within visual place recognition for satisfying the requirements of the degree of Masters by Dissertation. The major contributions of this thesis can be enlisted as follows:

1. Firstly, this thesis presents a detailed theoretical basis of the VPR research problem, its challenges, existing research works and the open-source datasets and techniques. The author also discusses, co-relates and distinguishes VPR from closely related topics of Visual-SLAM, Visual-Odometry, image matching and the correspondence problem.

2. The literature reviewed in this thesis is not limited only to VPR techniques, but instead research within semantic mapping, SLAM, hardware used within robotic vision and other relevant works that have potential for VPR have also been reviewed.

3. The author has presented in Chapter 3 a detailed case study advocating for the efficacy of predicting potential false-positives in VPR. This Chapter 3 presents heuristics that can effectively determine the salience of an image for VPR. An input image undergoes through an initial evaluation using 3 filters of memorability, staticity and entropy, prior to its use for VPR. The author has presented a detailed analysis of this filtering, including the individual contribution of each criterion, effect of sweeping each criterion, spatio-temporal filtering to avoid large physical gaps, computational and storage needs, benefits and limitations of the proposed approach.

4. In Chapter 4, an evaluation of 10 state-of-the-art VPR techniques is presented on 3 of the most challenging VPR datasets containing extreme viewpoint and illumination variations. This is the first time such an extensive evaluation has been performed for VPR techniques. The evaluation metrics include place matching performance using Precision-Recall curves, feature encoding time, descriptor matching time and memory footprint. Examples of images matched by some techniques, mismatched by all techniques and matched by all techniques are also shown.

5. Building upon the work in Chapter 4, the same evaluation is extended to aerial platforms in Chapter 5. In this evaluation, the author has utilised a recently proposed

Shopping Street dataset [161] containing 6-DOF viewpoint variation as seen by aerial robots. Other than the place matching performance using Precision-Recall curves, storage requirements, run-time memory consumption, the author has also presented a theoretical co-relation of image retrieval time of these 8 VPR techniques with electrical power requirements based on CPU consumption.

6. In Chapter 6, the author presents a novel, handcrafted, training-less VPR technique, namely CoHOG, that achieves state-of-the-art place matching performance per compute-unit and close to state-of-the-art place matching precision. This proposed technique is evaluated on 5 public VPR datasets and an extensive ablation study of the technique's parameters is performed.

7.2 Future Directions

One of the prime objectives of the author through this thesis is to leave a number of research ideas and open research questions for the VPR research community. These are not restricted only to the follow-up works and improvements to the contributions presented in this paper, but also to other research gaps that the author has identified during his work within VPR over the past 2 years. These are enlisted as follows:

1. The memorable maps framework is a very initial attempt at handling perceptual aliasing and predicting potential false-positives for VPR. It presents and proves the hypothesis that some images are confusing for VPR techniques, they will be mismatched and can be predicted. There is a significant room for improvement in the memorable maps framework (and generally within false-positives prediction for VPR) to achieve much better AUC boost and much lesser discarded true-positives than reported presently. Other salience evaluation criteria can be explored to achieve this.

2. The memorable maps framework coupled with different VPR techniques also enables the creation of a large-scale dataset containing 'good' and 'confusing' images for VPR state-of-the-art. Such a dataset should contain 4 classes of images: (a) confusing but matched, (b) good and matched, (c) confusing and mismatched, (d) good but mismatched. This could subsequently help in training an end-to-end neural network for classifying an image as good/bad for map-insertion.

3. The current thresholds in the memorable maps framework are constant values. Although their value is shown to change based on spatio-temporal filtering, this change of values is based only on avoiding large physical gaps and otherwise do not entail a

semantically-relevant change of thresholds. Ideally, in a continuously dynamic (e.g, city centre) environment, the staticity threshold should be relaxed and so should the memorability threshold in a continuously natural outdoor scenery. Moreover, the entropy threshold is not illumination-invariant, i.e, the entropy of the same place can be different if the images of this place are taken at drastically different times of the day. Therefore, if ET is taken as a constant and is not made illumination-conscious, the framework would end up discarding all images in a night traversal, which may not be desirable. This presents a significant room for improvement.

4. The existing version of the memorable maps framework is not compute-efficient. As reported, evaluating a single image based on all 3 criteria takes around 5 seconds. This presents a bottleneck for online map creation and presents significant room for improvement on computational fronts.

5. The evaluations performed in Chapter 4 are limited to small-scale datasets. This should be further extended to large-scale datasets, for example the complete Oxford RobotCar dataset [131]. The utility of other statistical evaluation metrics to VPR and the discrepancies of the currently used metrics needs to be explored, similar to the work of [135].

6. A key extension to the work in Chapter 5 would be the empirical analysis of the power consumption model proposed in a theoretical setting currently. While the existing work utilises a dataset simulating 6-DOF viewpoint variation, it would be useful to utilise or introduce new datasets created for VPR with real aerial platforms.

7. The work in Chapter 6 proposes that further investigation into the application of traditional handcrafted feature descriptors for VPR needs to be performed. Region extraction mechanism of CoHOG can be improved, sequential/temporal information can be integrated and the convolutional-regional matching mechanism can be improved with a more efficient matching scheme that exploits the geometric constraints on these regions.

8. Combining a range of different VPR techniques to complement the strengths and weaknesses (as identified in Chapters 4 and 5) of each other is also an interesting avenue. A hierarchical approach to this has recently been shown by authors in [115], but there is a need to perform this selection based on intelligent heuristics which should present better value and improved performance.

9. VPR as a domain has matured enough to look into its combination with other modules of robotics. In particular the combination of SLAM and VPR should be the logical next step to understand the value presented by state-of-the-art VPR systems to these SLAM systems. Another particularly interesting area is the combination of motion planning and coverage path planning advances with map creation and place recognition.

10. As reviewed in the literature review Chapter 2 of this thesis, SLAM enjoys specific techniques designed explicitly for the various range of sensors available today. VPR, especially as a stand-alone localisation/navigation system has only recently been investigated thoroughly and there is a significant room for further investigation based on sensors like event cameras, omni-directional cameras and RGB-D sensors.

References

[1] N. Sünderhauf, S. Shirazi, F. Dayoub, B. Upcroft, and M. Milford, "On the performance of convnet features for place recognition," in *IROS*. IEEE, 2015, pp. 4297–4304.

[2] M. Zaffar, S. Ehsan, M. Milford, and K. M. Maier, "Memorable maps: A framework for re-defining places in visual place recognition," *IEEE Transactions on Intelligent Transportation Systems*, 2020.

[3] Z. Chen, F. Maffra, I. Sa, and M. Chli, "Only look once, mining distinctive landmarks from convnet for visual place recognition," in *IROS)*. IEEE, 2017, pp. 9–16.

[4] S. Skrede, "Nordland dataset," https://bit.ly/2QVBOym, 2013.

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

[6] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," in *ECCV*. Springer, 2006, pp. 404–417.

[7] A. Oliva and A. Torralba, "Building the gist of a scene: The role of global image features in recognition," *Progress in brain research*, vol. 155, pp. 23–36, 2006.

[8] S. Lowry, N. Sünderhauf, P. Newman, J. J. Leonard, D. Cox, P. Corke, and M. J. Milford, "Visual place recognition: A survey," *IEEE Transactions on Robotics*, vol. 32, no. 1, pp. 1–19, 2015.

[9] A. J. Glover, W. P. Maddern, M. J. Milford, and G. F. Wyeth, "Fab-map+ ratslam: Appearance-based slam for multiple times of day," in *2010 IEEE international conference on robotics and automation*. IEEE, 2010, pp. 3507–3512.

[10] C. Cadena, L. Carlone, H. Carrillo, Y. Latif, D. Scaramuzza, J. Neira, I. Reid, and J. J. Leonard, "Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age," *IEEE T-RO*, vol. 32, no. 6, pp. 1309–1332, 2016.

[11] G. Tolias, Y. Avrithis, and H. Jégou, "To aggregate or not to aggregate: Selective match kernels for image search," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 1401–1408.

[12] A. Odo, S. McKenna, D. Flynn, and J. Vorstius, "Towards the automatic visual monitoring of electricity pylons from aerial images," in *15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications 2020*. SciTePress, 2020, pp. 566–573.

[13] D. P. Robertson and R. Cipolla, "An image-based system for urban navigation." in *Bmvc*, vol. 19, no. 51. Citeseer, 2004, p. 165.

[14] E. Johns and G.-Z. Yang, "From images to scenes: Compressing an image cluster into a single scene model for place recognition," in *2011 International Conference on Computer Vision*. IEEE, 2011, pp. 874–881.

[15] 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.

[16] S. Agarwal, Y. Furukawa, N. Snavely, I. Simon, B. Curless, S. M. Seitz, and R. Szeliski, "Building rome in a day," *Communications of the ACM*, vol. 54, no. 10, pp. 105–112, 2011.

[17] F. Fraundorfer, C. Engels, and D. Nistér, "Topological mapping, localization and navigation using image collections," in *2007 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2007, pp. 3872–3877.

[18] F. Zeng, A. Jacobson, D. Smith, N. Boswell, T. Peynot, and M. Milford, "Lookup: Vision-only real-time precise underground localisation for autonomous mining vehicles," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 1444–1450.

[19] A. Kendall, M. Grimes, and R. Cipolla, "Posenet: A convolutional network for real-time 6-dof camera relocalization," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2938–2946.

[20] M. Zaffar, S. Ehsan, M. Milford, D. Flynn, and K. McDonald-Maier, "Vpr-bench: An open-source visual place recognition evaluation framework with quantifiable viewpoint and appearance change," *arXiv preprint arXiv:2005.08135*, 2020.

[21] M. Zaffar, S. Ehsan, R. Stolkin, and K. M. Maier, "Sensors, slam and long-term autonomy: a review," in *2018 NASA/ESA Conference on Adaptive Hardware and Systems (AHS)*. IEEE, 2018, pp. 285–290.

[22] W. Burgard, M. Hebert, and M. Bennewitz, "World modeling," in *Springer handbook of robotics*. Springer, 2016, pp. 1135–1152.

[23] H. Zhang *et al.*, "Quantitative evaluation of feature extractors for visual slam," in *Fourth Canadian Conference on Computer and Robot Vision (CRV'07)*. IEEE, 2007, pp. 157–164.

[24] M. Montemerlo and S. Thrun, "Simultaneous localization and mapping with unknown data association using fastslam," in *2003 IEEE International Conference on Robotics and Automation (Cat. No. 03CH37422)*, vol. 2. IEEE, 2003, pp. 1985–1991.

[25] B. Williams, M. Cummins, J. Neira, P. Newman, I. Reid, and J. Tardós, "A comparison of loop closing techniques in monocular slam," *Robotics and Autonomous Systems*, vol. 57, no. 12, pp. 1188–1197, 2009.

[26] A. Elfes, "Sonar-based real-world mapping and navigation," *IEEE Journal on Robotics and Automation*, vol. 3, no. 3, pp. 249–265, 1987.

[27] C. Evers, A. H. Moore, and P. A. Naylor, "Acoustic simultaneous localization and mapping (a-slam) of a moving microphone array and its surrounding speakers," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 6–10.

[28] M. Kreković, I. Dokmanić, and M. Vetterli, "Echoslam: Simultaneous localization and mapping with acoustic echoes," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Ieee, 2016, pp. 11–15.

[29] J. Djugash, S. Singh, G. Kantor, and W. Zhang, "Range-only slam for robots operating cooperatively with sensor networks," in *Proceedings 2006 IEEE International Conference on Robotics and Automation, 2006. ICRA 2006.* IEEE, 2006, pp. 2078–2084.

[30] G. Grisettiyz, C. Stachniss, and W. Burgard, "Improving grid-based slam with rao-blackwellized particle filters by adaptive proposals and selective resampling," in *Proceedings of the 2005 IEEE international conference on robotics and automation*. IEEE, 2005, pp. 2432–2437.

[31] G. D. Tipaldi, M. Braun, and K. O. Arras, "Flirt: Interest regions for 2d range data with applications to robot navigation," in *Experimental Robotics*. Springer, 2014, pp. 695–710.

[32] S. Kohlbrecher, O. Von Stryk, J. Meyer, and U. Klingauf, "A flexible and scalable slam system with full 3d motion estimation," in *2011 IEEE International Symposium on Safety, Security, and Rescue Robotics*. IEEE, 2011, pp. 155–160.

[33] W. Hess, D. Kohler, H. Rapp, and D. Andor, "Real-time loop closure in 2d lidar slam," in *2016 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2016, pp. 1271–1278.

[34] A. J. Davison, I. D. Reid, N. D. Molton, and O. Stasse, "Monoslam: Real-time single camera slam," *IEEE transactions on pattern analysis and machine intelligence*, vol. 29, no. 6, pp. 1052–1067, 2007.

[35] E. Royer, M. Lhuillier, M. Dhome, and J.-M. Lavest, "Monocular vision for mobile robot localization and autonomous navigation," *International Journal of Computer Vision*, vol. 74, no. 3, pp. 237–260, 2007.

[36] K. Wang, Y. Liu, and L. Li, "A new algorithm for robot localization using monocular vision and inertia/odometry sensors," in *2012 IEEE International Conference on Robotics and Biomimetics (ROBIO)*. IEEE, 2012, pp. 735–740.

[37] T. Lemaire, C. Berger, I.-K. Jung, and S. Lacroix, "Vision-based slam: Stereo and monocular approaches," *International Journal of Computer Vision*, vol. 74, no. 3, pp. 343–364, 2007.

[38] J. Civera, D. Gálvez-López, L. Riazuelo, J. D. Tardós, and J. Montiel, "Towards semantic slam using a monocular camera," in *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2011, pp. 1277–1284.

[39] J.-H. Kim and M. J. Chung, "Slam with omni-directional stereo vision sensor," in *Proceedings 2003 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2003)(Cat. No. 03CH37453)*, vol. 1. IEEE, 2003, pp. 442–447.

[40] H. Tamimi, H. Andreasson, A. Treptow, T. Duckett, and A. Zell, "Localization of mobile robots with omnidirectional vision using particle filter and iterative sift," *Robotics and Autonomous Systems*, vol. 54, no. 9, pp. 758–765, 2006.

[41] L. Payá, A. Gil, and O. Reinoso, "A state-of-the-art review on mapping and localization of mobile robots using omnidirectional vision sensors," *Journal of Sensors*, vol. 2017, 2017.

[42] F. Endres, J. Hess, N. Engelhard, J. Sturm, D. Cremers, and W. Burgard, "An evaluation of the rgb-d slam system," in *2012 IEEE International Conference on Robotics and Automation*. IEEE, 2012, pp. 1691–1696.

[43] C. Kerl, J. Sturm, and D. Cremers, "Dense visual slam for rgb-d cameras," in *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2013, pp. 2100–2106.

[44] C. Kerl, J. Sturm, and D. Cremers, "Robust odometry estimation for rgb-d cameras," in *2013 IEEE international conference on robotics and automation*. IEEE, 2013, pp. 3748–3754.

[45] T. Whelan, M. Kaess, H. Johannsson, M. Fallon, J. J. Leonard, and J. McDonald, "Real-time large-scale dense rgb-d slam with volumetric fusion," *The International Journal of Robotics Research*, vol. 34, no. 4-5, pp. 598–626, 2015.

[46] A. S. Huang, A. Bachrach, P. Henry, M. Krainin, D. Maturana, D. Fox, and N. Roy, "Visual odometry and mapping for autonomous flight using an rgb-d camera," in *Robotics Research*. Springer, 2017, pp. 235–252.

[47] P. Elinas, R. Sim, and J. J. Little, "/spl sigma/slam: Stereo vision slam using the rao-blackwellised particle filter and a novel mixture proposal distribution," in *Proceedings 2006 IEEE International Conference on Robotics and Automation, 2006. ICRA 2006.* IEEE, 2006, pp. 1564–1570.

[48] L. M. Paz, P. Piniés, J. D. Tardós, and J. Neira, "Large-scale 6-dof slam with stereo-in-hand," *IEEE transactions on robotics*, vol. 24, no. 5, pp. 946–957, 2008.

[49] M. Tomono, "Robust 3d slam with a stereo camera based on an edge-point icp algorithm," in *2009 IEEE International Conference on Robotics and Automation*. IEEE, 2009, pp. 4306–4311.

[50] J. Engel, J. Stückler, and D. Cremers, "Large-scale direct slam with stereo cameras," in *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2015, pp. 1935–1942.

[51] G. Klein and D. Murray, "Parallel tracking and mapping for small ar workspaces," in *2007 6th IEEE and ACM international symposium on mixed and augmented reality*. IEEE, 2007, pp. 225–234.

[52] H. Rebecq, T. Horstschäfer, G. Gallego, and D. Scaramuzza, "Evo: A geometric approach to event-based 6-dof parallel tracking and mapping in real time," *IEEE Robotics and Automation Letters*, vol. 2, no. 2, pp. 593–600, 2016.

[53] A. R. Vidal, H. Rebecq, T. Horstschaefer, and D. Scaramuzza, "Ultimate slam? combining events, images, and imu for robust visual slam in hdr and high-speed scenarios," *IEEE Robotics and Automation Letters*, vol. 3, no. 2, pp. 994–1001, 2018.

[54] E. Mueggler, G. Gallego, H. Rebecq, and D. Scaramuzza, "Continuous-time visual-inertial odometry for event cameras," *IEEE Transactions on Robotics*, vol. 34, no. 6, pp. 1425–1440, 2018.

[55] J. Fuentes-Pacheco, J. Ruiz-Ascencio, and J. M. Rendón-Mancha, "Visual simultaneous localization and mapping: a survey," *Artificial intelligence review*, vol. 43, no. 1, pp. 55–81, 2015.

[56] T. Taketomi, H. Uchiyama, and S. Ikeda, "Visual slam algorithms: a survey from 2010 to 2016," *IPSJ Transactions on Computer Vision and Applications*, vol. 9, no. 1, p. 16, 2017.

[57] Y. Jin, D. Mishkin, A. Mishchuk, J. Matas, P. Fua, K. M. Yi, and E. Trulls, "Image matching across wide baselines: From paper to practice," *arXiv preprint arXiv:2003.01587*, 2020.

[58] I. Kostavelis and A. Gasteratos, "Semantic mapping for mobile robotics tasks: A survey," *RAS*, vol. 66, pp. 86–103, 2015.

[59] E. Pepperell, P. Corke, and M. Milford, "Towards persistent visual navigation using smart," in *Proceedings of Australasian Conference on Robotics and Automation*. ARAA, 2013.

[60] S. Garg and M. Milford, "Straightening sequence-search for appearance-invariant place recognition using robust motion estimation," *ACRA, ARAA*, 2017.

[61] A. Chapoulie, P. Rives, and D. Filliat, "Topological segmentation of indoors/outdoors sequences of spherical views," in *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2012, pp. 4288–4295.

[62] H. Korrapati, J. Courbon, Y. Mezouar, and P. Martinet, "Image sequence partitioning for outdoor mapping," in *2012 IEEE International Conference on Robotics and Automation*. IEEE, 2012, pp. 1650–1655.

[63] R. Paul and P. Newman, "Self help: Seeking out perplexing images for ever improving navigation," in *2011 IEEE International Conference on Robotics and Automation*. IEEE, 2011, pp. 445–451.

[64] A. Ranganathan and F. Dellaert, "Bayesian surprise and landmark detection," in *2009 IEEE International Conference on Robotics and Automation*. IEEE, 2009, pp. 2017–2023.

[65] Y. Girdhar, P. Giguere, and G. Dudek, "Autonomous adaptive underwater exploration using online topic modeling," in *Experimental Robotics*. Springer, 2013, pp. 789–802.

[66] L. Murphy and G. Sibley, "Incremental unsupervised topological place discovery," in *2014 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2014, pp. 1312–1318.

[67] Y. Girdhar and G. Dudek, "Online navigation summaries," in *2010 IEEE International Conference on Robotics and Automation*. IEEE, 2010, pp. 5035–5040.

[68] R. Paul, D. Feldman, D. Rus, and P. Newman, "Visual precis generation using coresets," in *2014 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2014, pp. 1304–1311.

[69] H. Karaoguz and H. I. Bozma, "Reliable topological place detection in bubble space," in *2014 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2014, pp. 697–702.

[70] M. Demir and H. I. Bozma, "Automated place detection based on coherent segments," in *2018 IEEE 12th International Conference on Semantic Computing (ICSC)*. IEEE, 2018, pp. 71–76.

[71] E. A. Topp and H. I. Christensen, "Detecting structural ambiguities and transitions during a guided tour," in *2008 IEEE International Conference on Robotics and Automation*. IEEE, 2008, pp. 2564–2570.

[72] A. Ranganathan, "Detecting and labeling places using runtime change-point detection and place labeling classifiers," Oct. 15 2013, uS Patent 8,559,717.

[73] E. Pepperell, P. I. Corke, and M. J. Milford, "All-environment visual place recognition with smart," in *2014 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2014, pp. 1612–1618.

[74] M. Milford, "Vision-based place recognition: how low can you go?" *The International Journal of Robotics Research*, vol. 32, no. 7, pp. 766–789, 2013.

[75] B. Talbot, S. Garg, and M. Milford, "Openseqslam2. 0: An open source toolbox for visual place recognition under changing conditions," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 7758–7765.

[76] M. Milford and A. George, "Featureless visual processing for slam in changing outdoor environments," in *Field and Service Robotics*. Springer, 2014, pp. 569–583.

[77] S. Garg, B. Harwood, G. Anand, and M. Milford, "Delta descriptors: Change-based place representation for robust visual localization," *arXiv preprint arXiv:2006.05700*, 2020.

[78] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *IJCV, Springer*, vol. 60, no. 2, pp. 91–110, 2004.

[79] S. Se, D. Lowe, and J. Little, "Mobile robot localization and mapping with uncertainty using scale-invariant visual landmarks," *IJRR*, vol. 21, no. 8, pp. 735–758, 2002.

[80] H. Andreasson and T. Duckett, "Topological localization for mobile robots using omni-directional vision and local features," *IFAC Proceedings Volumes*, vol. 37, no. 8, pp. 36–41, 2004.

[81] E. Stumm, C. Mei, and S. Lacroix, "Probabilistic place recognition with covisibility maps," in *IROS*. IEEE, 2013, pp. 4158–4163.

[82] J. Košecká, F. Li, and X. Yang, "Global localization and relative positioning based on scale-invariant keypoints," *Robotics and Autonomous Systems*, vol. 52, no. 1, pp. 27–38, 2005.

[83] A. C. Murillo, J. J. Guerrero, and C. Sagues, "Surf features for efficient robot localization with omnidirectional images," in *Proceedings of IEEE ICRA*, 2007, pp. 3901–3907.

[84] M. Cummins and P. Newman, "Appearance-only slam at large scale with fab-map 2.0," *IJRR*, vol. 30, no. 9, pp. 1100–1123, 2011.

[85] A. Glover, W. Maddern, M. Warren, S. Reid, M. Milford, and G. Wyeth, "Openfabmap: An open source toolbox for appearance-based loop closure detection," in *2012 IEEE International Conference on Robotics and Automation*. IEEE, 2012, pp. 4730–4735.

[86] W. Maddern, M. Milford, and G. Wyeth, "Cat-slam: probabilistic localisation and mapping using a continuous appearance-based trajectory," *IJRR*, vol. 31, no. 4, pp. 429–451, 2012.

[87] M. Agrawal, K. Konolige, and M. R. Blas, "Censure: Center surround extremas for realtime feature detection and matching," in *European Conference on Computer Vision*. Springer, 2008, pp. 102–115.

[88] K. Konolige and M. Agrawal, "Frameslam: From bundle adjustment to real-time visual mapping," *IEEE Transactions on Robotics*, vol. 24, no. 5, pp. 1066–1077, 2008.

[89] E. Rosten and T. Drummond, "Machine learning for high-speed corner detection," in *ECCV*. Springer, 2006, pp. 430–443.

[90] C. Mei, G. Sibley, M. Cummins, P. Newman, and I. Reid, "A constant-time efficient stereo slam system," in *Proceedings of the British machine vision conference*, vol. 1, no. 2009. BMVA Press, 2009.

[91] J. Sivic and A. Zisserman, "Video google: A text retrieval approach to object matching in videos," in *null*. IEEE, 2003, p. 1470.

[92] A. Angeli, S. Doncieux, J.-A. Meyer, and D. Filliat, "Incremental vision-based topological slam," in *IROS*. Ieee, 2008, pp. 1031–1036.

[93] K. L. Ho and P. Newman, "Detecting loop closure with scene sequences," *IJCV*, vol. 74, no. 3, pp. 261–286, 2007.

[94] J. Wang, H. Zha, and R. Cipolla, "Combining interest points and edges for content-based image retrieval," in *IEEE International Conference on Image Processing 2005*, vol. 3. IEEE, 2005, pp. III–1256.

[95] D. Filliat, "A visual bag of words method for interactive qualitative localization and mapping," in *ICRA*. IEEE, 2007, pp. 3921–3926.

[96] A. C. Murillo and J. Kosecka, "Experiments in place recognition using gist panoramas," in *ICCV Workshops*. IEEE, 2009, pp. 2196–2203.

[97] G. Singh and J. Kosecka, "Visual loop closing using gist descriptors in manhattan world," in *ICRA Omnidirectional Vision Workshop*, 2010, pp. 4042–4047.

[98] M. Calonder, V. Lepetit, M. Ozuysal, T. Trzcinski, C. Strecha, and P. Fua, "Brief: Computing a local binary descriptor very fast," *IEEE T-PAMI*, vol. 34, no. 7, pp. 1281–1298, 2011.

[99] N. Sünderhauf and P. Protzel, "Brief-gist-closing the loop by simple means," in *IROS*. IEEE, 2011, pp. 1234–1241.

[100] H. Badino, D. Huber, and T. Kanade, "Real-time topometric localization," in *ICRA*. IEEE, 2012, pp. 1635–1642.

[101] C. McManus, B. Upcroft, and P. Newmann, "Scene signatures: Localised and point-less features for localisation," *Robotics, Science and Systems Conference*, 2014.

[102] Z. Chen *et al.*, "Deep learning features at scale for visual place recognition," in *ICRA*. IEEE, 2017, pp. 3223–3230.

[103] Z. Chen, O. Lam, A. Jacobson, and M. Milford, "Convolutional neural network-based place recognition," *preprint arXiv:1411.1509*, 2014.

[104] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "Overfeat: Integrated recognition, localization and detection using convolutional networks," in *2nd International Conference on Learning Representations, ICLR 2014*, 2014.

[105] L. Liu, C. Shen, and A. van den Hengel, "The treasure beneath convolutional layers: Cross-convolutional-layer pooling for image classification," in *CVPR*, 2015, pp. 4749–4757.

[106] A. Babenko and V. Lempitsky, "Aggregating deep convolutional features for image retrieval," *arXiv preprint arXiv:1510.07493 ICCV*, 2015.

[107] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "Netvlad: Cnn architecture for weakly supervised place recognition," in *CVPR*, 2016, pp. 5297–5307.

[108] H. Jégou, M. Douze, C. Schmid, and P. Pérez, "Aggregating local descriptors into a compact image representation," in *CVPR*. IEEE Computer Society, 2010, pp. 3304–3311.

[109] G. Tolias, R. Sicre, and H. Jégou, "Particular object retrieval with integral max-pooling of cnn activations," *arXiv:1511.05879, ICLR*, 2016.

[110] 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.

[111] 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*, 2019.

[112] J. Mao, X. Hu, X. He, L. Zhang, L. Wu, and M. J. Milford, "Learning to fuse multiscale features for visual place recognition," *IEEE Access*, vol. 7, pp. 5723–5735, 2018.

[113] N. Merrill and G. Huang, "Lightweight unsupervised deep loop closure," *arXiv preprint arXiv:1805.07703, Robotics Science and Systems Conference*, 2018.

[114] M. Chancán, L. Hernandez-Nunez, A. Narendra, A. B. Barron, and M. Milford, "A hybrid compact neural architecture for visual place recognition," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 993–1000, 2020.

[115] S. Hausler and M. Milford, "Hierarchical multi-process fusion for visual place recognition," *arXiv preprint arXiv:2002.03895*, 2020.

[116] F. Zeng, A. Jacobson, D. Smith, N. Boswell, T. Peynot, and M. Milford, "I2-s2: Intra-image-seqslam for more accurate vision-based localisation in underground mines," 2018.

[117] F. Zeng, A. Jacobson, D. Smith, N. Boswell, T. Peynot, and M. Milford, "Timtam: Tunnel-image texturally accorded mosaic for location refinement of underground vehicles with a single camera," *IEEE Robotics and Automation Letters*, vol. 4, no. 4, pp. 4362–4369, 2019.

[118] D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superpoint: Self-supervised interest point detection and description," in *CVPR Workshops*, 2018, pp. 224–236.

[119] M. Dusmanu *et al.*, "D2-net: A trainable cnn for joint description and detection of local features," in *CVPR*, 2019, pp. 8092–8101.

[120] L. G. Camara, C. Gäbert, and L. Preucil, "Highly robust visual place recognition through spatial matching of cnn features," *ResearchGate Preprint*, 2019.

[121] P.-E. Sarlin, C. Cadena, R. Siegwart, and M. Dymczyk, "From coarse to fine: Robust hierarchical localization at large scale," in *CVPR*, 2019, pp. 12 716–12 725.

[122] E. Stenborg, C. Toft, and L. Hammarstrand, "Long-term visual localization using semantically segmented images," in *2018 IEEE ICRA*, 2018, pp. 6484–6490.

[123] J. L. Schönberger, M. Pollefeys, A. Geiger, and T. Sattler, "Semantic visual localization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6896–6906.

[124] T. Naseer, G. L. Oliveira, T. Brox, and W. Burgard, "Semantics-aware visual localization under challenging perceptual conditions," in *2017 IEEE ICRA*, 2017, pp. 2614–2620.

[125] Y. Hou, H. Zhang, and S. Zhou, "Evaluation of object proposals and convnet features for landmark-based visual place recognition," *Journal of Intelligent & Robotic Systems*, vol. 92, no. 3-4, pp. 505–520, 2018.

[126] A. Torii, J. Sivic, T. Pajdla, and M. Okutomi, "Visual place recognition with repetitive structures," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 883–890.

[127] A. Torii *et al.*, "24/7 place recognition by view synthesis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1808–1817.

[128] G. Ros, L. Sellart, J. Materzynska, D. Vazquez, and A. M. Lopez, "The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3234–3243.

[129] M. Larsson, E. Stenborg, L. Hammarstrand, M. Pollefeys, T. Sattler, and F. Kahl, "A cross-season correspondence dataset for robust semantic segmentation," in *CVPR*, 2019, pp. 9532–9542.

[130] H. Badino, D. Huber, and T. Kanade, "Visual topometric localization," in *2011 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2011, pp. 794–799.

[131] W. Maddern, G. Pascoe, C. Linegar, and P. Newman, "1 year, 1000 km: The oxford robotcar dataset," *The International Journal of Robotics Research*, vol. 36, no. 1, pp. 3–15, 2017.

[132] R. Sahdev and J. K. Tsotsos, "Indoor place recognition system for localization of mobile robots," in *2016 13th Conference on Computer and Robot Vision (CRV)*. IEEE, 2016, pp. 53–60.

[133] J. Mount and M. Milford, "2d visual place recognition for domestic service robots at night," in *2016 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2016, pp. 4822–4829.

[134] 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, IEEE ICRA Workshop on Database Generation and Benchmarking*, 2019.

[135] B. Ferrarini, M. Waheed, S. Waheed, S. Ehsan, M. J. Milford, and K. D. McDonald-Maier, "Exploring performance bounds of visual place recognition using extended precision," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 1688–1695, 2020.

[136] M. Warren, D. McKinnon, H. He, and B. Upcroft, "Unaided stereo vision based pose estimation," 2010.

[137] A. Khosla, A. S. Raju, A. Torralba, and A. Oliva, "Understanding and predicting image memorability at a large scale," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 2390–2398.

[138] W. Hartmann, M. Havlena, and K. Schindler, "Predicting matchability," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 9–16.

[139] M. Dymczyk, E. Stumm, J. Nieto, R. Siegwart, and I. Gilitschenski, "Will it last? learning stable features for long-term visual localization," in *2016 Fourth International Conference on 3D Vision (3DV)*. IEEE, 2016, pp. 572–581.

[140] T. F. Brady, T. Konkle, G. A. Alvarez, and A. Oliva, "Visual long-term memory has a massive storage capacity for object details," *Proceedings of the National Academy of Sciences*, vol. 105, no. 38, pp. 14 325–14 329, 2008.

[141] T. Konkle, T. F. Brady, G. A. Alvarez, and A. Oliva, "Scene memory is more detailed than you think: The role of categories in visual long-term memory," *Psychological Science*, vol. 21, no. 11, pp. 1551–1556, 2010.

[142] A. Mousavian, J. Košecká, and J.-M. Lien, "Semantically guided location recognition for outdoors scenes," in *2015 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2015, pp. 4882–4889.

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

[144] P. Zhang, J. Wang, A. Farhadi, M. Hebert, and D. Parikh, "Predicting failures of vision systems," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 3566–3573.

[145] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning deep features for scene recognition using places database," in *Advances in neural information processing systems*, 2014, pp. 487–495.

[146] J. Redmon and A. Farhadi, "Yolo9000: better, faster, stronger," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7263–7271.

[147] S. Van der Walt *et al.*, "scikit-image: image processing in python," *PeerJ*, vol. 2, p. e453, 2014.

[148] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[149] 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 ICRA 2019 Workshop on Aerial Robotics*, 2019.

[150] E. Olson and P. Agarwal, "Inference on networks of mixtures for robust robot mapping," *IJRR*, vol. 32, no. 7, pp. 826–840, 2013.

[151] T. Naseer, L. Spinello, W. Burgard, and C. Stachniss, "Robust visual robot localization across seasons using network flows," in *Twenty-Eighth AAAI Conference on Artificial Intelligence*, 2014.

[152] C. Valgren and A. J. Lilienthal, "Sift, surf & seasons: Appearance-based long-term localization in outdoor environments," *RAS*, vol. 58, no. 2, pp. 149–156, 2010.

[153] A. Pronobis, B. Caputo, P. Jensfelt, and H. I. Christensen, "A discriminative approach to robust visual place recognition," in *IROS*. IEEE, 2006, pp. 3829–3836.

[154] S. Garg, N. Suenderhauf, and M. Milford, "Lost? appearance-invariant place recognition for opposite viewpoints using visual semantics," `arXiv:1804.05526 [cs.RO]`, 2018.

[155] A. Ranganathan, S. Matsumoto, and D. Ilstrup, "Towards illumination invariance for visual localization," in *ICRA*. IEEE, 2013, pp. 3791–3798.

[156] C.-C. Wang and et al., "Simultaneous localization, mapping and moving object tracking," *IJRR*, vol. 26, no. 9, pp. 889–916, 2007.

[157] T. Cieslewski, S. Choudhary, and D. Scaramuzza, "Data-efficient decentralized visual slam," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 2466–2473.

[158] "Region-bow," https://github.com/scutzetao/IROS2017_OnlyLookOnce.

[159] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," `arXiv:1409.1556 [cs.CV]`, 2014.

[160] "Rmac," https://github.com/gtolias/rmac.

[161] F. Maffra, Z. Chen, and M. Chli, "Viewpoint-tolerant place recognition combining 2d and 3d information for uav navigation," in *ICRA*, 2018.

[162] N. Bezzo and et al., "Online planning for energy-efficient and disturbance-aware uav operations," in *IROS*. IEEE, 2016, pp. 5027–5033.

[163] X. Fan, W.-D. Weber, and L. A. Barroso, "Power provisioning for a warehouse-sized computer," in *ACM SIGARCH computer architecture news*, vol. 35, no. 2. ACM, 2007, pp. 13–23.

[164] F. Maffra, L. Teixeira, Z. Chen, and M. Chli, "Real-time wide-baseline place recognition using depth completion," *IEEE Robotics and Automation Letters*, vol. 4, no. 2, pp. 1525–1532, 2019.

[165] W. T. Freeman and M. Roth, "Orientation histograms for hand gesture recognition," in *International workshop on automatic face and gesture recognition*, vol. 12, 1995, pp. 296–301.

[166] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *CVPR*, vol. 1. IEEE, 2005, pp. 886–893.

[167] N. Sünderhauf and P. Protzel, "Switchable constraints for robust pose graph slam," in *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2012, pp. 1879–1884.

[168] Y. Latif, C. Cadena, and J. Neira, "Robust loop closing over time for pose graph slam," *The International Journal of Robotics Research*, vol. 32, no. 14, pp. 1611–1626, 2013.

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

[170] J. Surber, L. Teixeira, and M. Chli, "Robust visual-inertial localization with weak gps priors for repetitive uav flights," in *2017 IEEE ICRA*, 2017, pp. 6300–6306.