COMPLEMENTARITY-BASED VISUAL PLACE RECOGNITION IN CHANGING ENVIRONMENTS

Maria Waheed

A thesis submitted for the degree of

Doctor of Philosophy in Computer Science

School of Computer Science and Electronic Engineering

University of Essex

September 2024

Abstract

Localization has long been a key research topic in computer science due to its applications in autonomous vehicles, surveillance, security, and indoor positioning systems. Visual Place Recognition (VPR), which detects previously visited locations through visual data, is a significant area within localization. Solving VPR is complex due to environmental and viewpoint variations. Despite numerous high-performing algorithms, no universal technique can address all variations with complete accuracy; each has its strengths and weaknesses for specific variations. This thesis introduces a novel element: the concept of complementarity among VPR techniques. It defines complementarity in the context of VPR algorithms and establishes the existence and degree of complementarity among existing methods. This revolutionary approach contributes significantly to developing more efficient VPR ensemble setups. It uses complementarity as a guideline to combine highly complementary techniques and avoid redundant pairings. This research introduces SwitchHit, a probabilistic, complementarity-based switching system that dynamically selects the most suitable VPR technique based on complementarity. Unlike traditional methods that run multiple techniques simultaneously, SwitchHit intelligently switches to a better technique when necessary. This feature distinguishes SwitchHit from other setups and significantly enhances VPR performance in terms of accuracy. The thesis also explores SwitchFuse, a hybrid model that combines the strengths of switching and fusion strategies for improved VPR accuracy, outperforming other similar setups, including SwitchHit and existing multi-fusion systems. The findings highlight the importance of innovative approaches, informed by complementarity analysis, to develop robust and efficient VPR systems. Additionally, it examines the utility of universal voting schemes within ensemble setups, demonstrating their potential to refine VPR accuracy further.

This work lays a foundation for future research in leveraging complementarity and ensemble methods in autonomous navigation and beyond.

Declaration

I hereby affirm that, except where explicitly referenced to the work of others, the contents of this thesis are original and have not been submitted, either wholly or in part, for any other degree or qualification at this or any other institution. This thesis represents my own efforts and contains no material that is the product of collaboration, except as delineated in the text and the Acknowledgements.

Maria Waheed

April 2024

To the loving memory of my late grandfather; and to my cherished parents, beloved husband, and dear sisters.

Acknowledgements

With heartfelt appreciation, I acknowledge the instrumental roles played by Prof. Klaus McDonald-Maier and Dr. Shoaib Ehsan in supervising my PhD. Prof. Klaus opened doors to opportunities that shaped my academic and professional paths. Dr. Shoaib Ehsan, a beacon of unwavering encouragement, consistently believed in my ability to overcome challenges. His mentorship has been pivotal in my development. Thanks to Prof. Michael Milford for his invaluable insights, enriching my research and broadening my perspective. Special thanks to my colleagues and friends, especially Bruno Ferrarini, for his indispensable support and guidance. Thanks also to Bruno Arcanjo and Oliver Grainge for their support and camaraderie. My family has been my consistent backbone throughout this journey. To my sister and brother, Sania and Irtaza, for their ongoing assistance, care, and encouragement, and to my beloved parents, Dr. Waheed and Faiza, and my youngest sister, Sahar, for their boundless love, backing, and support. I am equally thankful to my grandparents and my Ammi, Saima, who have enveloped me with love, warmth and kindness throughout this journey. And finally, a special tribute to my loving husband, Afseh, whose unwavering companionship and support during the challenging thesis writing phase have been my anchor. Thank you all for being an integral part of this significant chapter in my life.

Contents

List of Figures

List of Tables

Abbreviations

AUC	Area Under Curve
CNN	Convolutional Neural Network
DOF	Degree of Freedom
EP	Extended Precision
FN	False Negative
FP	False Positive
GPS	Global Positioning System
GPU	Graphical Processing Unit
PN	Predicted Negative
PP	Predicted Positive
PR-Curve	Precision-Recall Curve
TN	True Negative
TP	True Positive
UAV	Unmanned aerial vehicle
VPR	Visual Place Recognition

Chapter 1

Introduction

1.1 Visual Place Recognition (VPR)

Visual Place Recognition (VPR) is a critical component of modern computer vision and robotics, designed to enable autonomous systems, such as robots and autonomous vehicles, to recognise and distinguish locations based on visual inputs. Utilizing images captured by onboard cameras or sensors, VPR helps these systems navigate and make sense of their surroundings by identifying whether a current view matches a previously recorded location as illustrated in Figure 1.1. This technology is crucial for maintaining accurate navigation especially in environments where GPS signals are unreliable. VPR must effectively handle various visual changes in the environment, including alterations in lighting, weather conditions, and seasonal variations, ensuring consistent and reliable recognition performance across different conditions [14–24]. VPR holds a critical role in numerous applications, including autonomous navigation (for both ground and aerial robots), augmented reality, and long-term localization in dynamic environments. It is

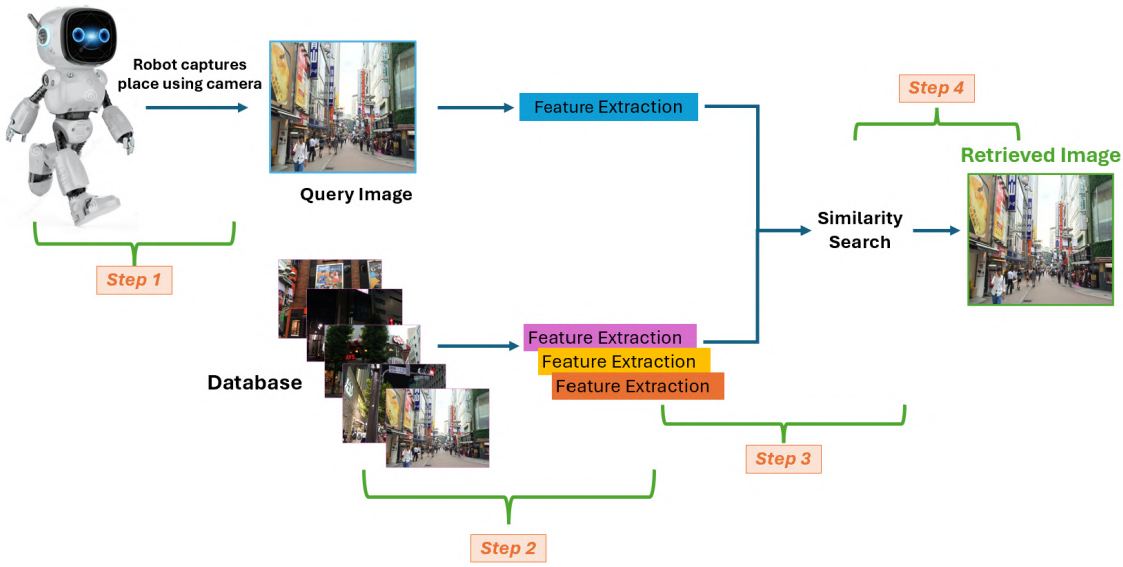

Fig. 1.1. An Overview of the Basic Visual Place Recognition task steps based on image retrieval. Images taken from [1]

essential for the development of robust autonomous systems that can operate in unstructured environments without relying heavily on GPS, which can be inaccurate in urban canyons or indoors, or entirely unavailable [25].

One of the key contributions of VPR to computer vision and robotics is its ability to deal with the vast amount of variability in real-world environments. By enabling machines to recognise places under different conditions, VPR supports the creation of more flexible and adaptable navigation systems. This capability is particularly crucial for tasks such as route planning, re-localization after getting lost, and understanding environmental changes over time [26]

1.2 Key Problems in Solving VPR

Solving VPR tasks involves overcoming a range of complex challenges, primarily due to the dynamic and unpredictable nature of real-world environments. The following sections describe the primary and most frequent environmental challenges in VPR applications.

Fig. 1.2. A place from Nordland dataset in winter and summer. Images taken from [2].

1.2.1 Environmental Changes

Environmental changes, including variations in lighting (day to night, shadows), weather (rain, fog, snow), and seasons, significantly alter the appearance of places such as the examples in Figure 1.2. To address these challenges, VPR systems must develop adaptable and robust recognition algorithms. Techniques like domain adaptation and the use of invariant features have been explored to mitigate these effects [27]. Domain adaptation involves aligning the feature distributions from different environmental conditions, allowing models to perform consistently across various domains without extensive retraining. Meanwhile, the use of invariant features ensures that the extracted features remain stable despite changes in lighting, weather, and viewpoint, thus maintaining re-

liable recognition performance across diverse conditions.

Fig. 1.3. Examples of images with dynamic objects. Images taken from [3,4].

1.2.2 Dynamic Objects

The presence of dynamic objects, such as vehicles, pedestrians, and animals, can obstruct important features of the environment such as those presented in Figure 1.3, leading to recognition failures. VPR systems must either be capable of ignoring these transient obstructions or incorporating them into their recognition process in a way that does not compromise the accuracy of place recognition [28].

Fig. 1.4. Examples of images appearing different due to change in viewpoint. Images taken from [5]

1.2.3 Viewpoint and Scale Variability

Changes in the viewpoint and scale at which a scene is observed can dramatically affect its visual appearance as presented in Figure 1.4 and Figure 1.5, posing significant challenges for VPR systems. Algorithms must be capable of recognizing a place regardless of the observer's position, orientation, or distance. This requires sophisticated geometric transformations and scale-invariant features to ensure reliable place recognition across different viewpoints [29].

Fig. 1.5. Examples of images demonstrating the different types of viewpoint variations. Images taken from [5, 6]

1.2.4 Perceptual Aliasing

Perceptual aliasing occurs when distinct locations appear visually similar, leading to incorrect place recognition such as the examples in Figure 1.6. This issue is prevalent across all types of environments, as visually similar features can confuse VPR systems. Addressing perceptual aliasing requires sophisticated feature extraction and matching strategies that can discern subtle differences between visually similar places [30]. Research has shown that advanced techniques, such as context-aware methods and geometric verification, are effective in mitigating the effects of perceptual aliasing in diverse settings [14].

Fig. 1.6. Examples demonstrating perceptual aliasing; Different places places generating a similar visual making them look the same. Images taken from [5]

1.2.5 Summary of Key Problems in VPR

In conclusion, the field of VPR is constantly evolving and leveraging advancements to develop more sophisticated and resilient algorithms to tackle these challenges. The goal is to achieve a level of visual place recognition that can closely mimic or even surpass human capabilities in navigating and understanding complex environments. One such effort led to the exploration of utilizing multiple algorithms or VPR techniques to enhance the accuracy and computational performance of VPR systems [31]. This approach was inspired by the use of multiple sensor in robotics, which provide a more accurate and comprehensive understanding of the environment [32]. For example, a combination

of cameras, LiDAR, and ultrasonic sensors provides more comprehensive information about the surroundings including visual, depth, and proximity data [33]. Additionally, the multiple-sensor approach is also utilised to tackle cases where one sensor fails or malfunctions, the robot still has a backup to rely on to keep going. Similarly, relying on more than one VPR algorithm boosts the chances of obtaining accurate input and improves the overall effectiveness of solving the VPR task. This thesis focuses on several aspects encountered in the journey of designing VPR systems that utilise multiple algorithms efficiently. Some notable work has been presented in [31], [34] which have played the role of a stepping stone for the research conducted. The experiments conducted and results collected aim to provide knowledge for designing better ensemble VPR (Chapter 3) and further continue to test this knowledge by designing different systems (SwitchHit & Switch-Fuse) demonstrating the utility of such ensemble methods as presented in Chapter 4 and 5. Lastly, this thesis explores the component of Voting within ensemble VPR systems, which is underexplored and understudied, highlighting the potential benefits and implications of universal voting methods.

1.3 Research Questions

This section introduces the main research questions that the thesis will address:

- How can complementarity among different Visual Place Recognition (VPR) techniques be effectively measured and quantified?
- What methods can be developed to dynamically select or combine VPR techniques based on their assessed complementarity to improve recognition performance?

- How do these innovative systems, designed around the concept of complementarity, perform under varied and challenging real-world conditions?
- How do different universal voting schemes influence the performance of ensemble VPR systems in terms of accuracy and reliability across diverse environments?

1.4 Thesis Contributions

This dissertation establishes the concept of complementarity in Visual Place Recognition systems, proposing a novel methodology to assess and harness this attribute. It details the development of innovative systems that use intelligent technique selection or fusion based on complementarity, demonstrably enhancing recognition accuracy and illustrating the methodology's effectiveness in practical applications. The details of each contribution are outlined below:

- The introduction of the concept of Complementarity followed by an exploration and systematic study of the existence of complementarity among the different state-of-the-art VPR techniques. Further, designing a framework utilizing a McNemar's test-like approach to determine the levels of complementarity between VPR technique pairs. Providing insightful information for future endeavours to designing ensemble VPR systems on the basis of complementarity.
- Designing and presenting the SwitchHit System which operates by carefully by utilizing the information discovered on complementarity in the previously mentioned contribution and follows a probabilistic model to allow for dynamic switching between the available complementary techniques, so as to avoid the use of

brute force and rather perform an efficient selection of the optimal method hence significantly improving the VPR performance in terms of accuracy.

- The third contribution presented is a hybrid system design entitled "Switch-Fuse" which tackles shortcomings of both the SwitchHit system as well as other existing Multi-Fusion systems such as presented in [31] [34]. It is an inventive approach combining both the robustness of switching VPR techniques based on complementarity and The impact of integrating the carefully selected techniques to significantly improve performance. SwitchFuse holds a structure superior to the basic fusion methods as, instead of simply fusing all or any random techniques, it is structured to first switch and then select the best possible VPR techniques for fusion, according to the query image, which together as a hybrid model substantially improve performance on all major VPR data sets.

- Lastly, the introduction, exploration and utilization of universal voting schemes to improve VPR accuracy in ensemble VPR set ups. Voting which is a common aspect to almost all types of ensemble VPR setups remained previously under-researched for VPR, instead followed the common basic practise of voting without having tested the several other available voting schemes to evaluate their differences. The idea for this stems from an observation that different voting methods result in different outcomes for the exact same type of data tested hence the use of any voting scheme should not be a trivial or random task. We test this observation for VPR systems, illustrating that it stands true and then propose the best, worst and satisfactory voting methods that can be employed.

1.5 Thesis Structure

This thesis is organised into seven chapters overall, as follows;

- Chapter 2 presents a detailed literature review on Visual Place Recognition, its complications, core methodologies and the recent introduction of innovative approaches i.e. different types of ensemble set ups to solve the VPR problem. Additionally, a description on the types of datasets widely employed and different metrics used to evaluate the techniques and their performance is also provided.

- Chapter 3 introduces the concept of complementarity among different VPR techniques and its significance. With the use of a framework, the chapter further presents how the complementarity is determined and the level of complementarity measured and presented utilizing multiple evaluation metrics. The results for complementarity are presented for 10 major VPR datasets utilzing 8 state-of-the-art VPR technqiues.

- Chapter 4 presents the SwitchHit system in detail, its inspiration and design. The system is based on a probabilistic model to allow for dynamic switching between multiple VPR methods to ensure a switch to the optimal technique given the query image. This is achieved utilizing a Bayes' theorem inspired framework that updates the matching probability of a system for the given query image based on prior information and likelihood of matching correctly. The results for the system tested are presented for six widely employed VPR datasets with a combination of state-of-the-art VPR techniques utilised within the SwitchHit system, selected on the basis of the complementarity knowledge of these techniques.

- Chapter 5 presents the Switch-Fuse system in details, the shortcoming it addresses of the other existing ensemble VPR systems. The chapter discusses in detail the design and structure of the hybrid model of Switch-Fuse covering the methodology and all steps involved starting with a tripartite model with each component consisting of VPR techniques dedicated to different types of variation that can be encountered. The query image is input to all three units of the system, where the probability of a correct match is calculated by the primary technique in each unit, and switching is conducted to select an alternate technique where required. Finally, a technique is selected by each unit, each of which then undergoes fusion where the normalized distance vectors are added ensure a significant enhancement in performance. The results are evaluated using different evaluation metrics for six datasets while the results for Switch-Fuse are compared with each technique, SwitchHit, and multi-process fusion systems.

- Chapter 6 presents the use of universal voting schemes for VPR and it opens a discussion to the observation that different voting schemes result it different results for the same data hence the choice of a voting method must be a well curated decision. This chapter analyses several universal voting schemes to determine if the observed principles apply to VPR tasks involving voting, similar to other research fields. Furthermore, it aims to maximize the place detection accuracy of a VPR ensemble setup and identify the optimal voting schemes for selection. The experiments in the chapter present five different widely used voting methods from other research fields and applies them to a standard ensemble VPR system to present how the produced results vary, the significance of this difference in performance and suggest what voting schemes are then better than others given the type of VPR

task to be performed.

1.6 List of Publications

Following is the list of publications made during the course of this PhD:

Chapter 3:

- Waheed, M., Milford, M., McDonald-Maier, K., & Ehsan, S. (2021). Improving visual place recognition performance by maximising complementarity. *IEEE Robotics and Automation Letters*, 6(3), 5976-5983.

Chapter 4:

- Waheed, M., Milford, M., McDonald-Maier, K., & Ehsan, S. (2022, October). Switchhit: A probabilistic, complementarity-based switching system for improved visual place recognition in changing environments. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (pp. 7833-7840). IEEE.

Chapter 5:

- Waheed, M., Waheed, S., Milford, M., McDonald-Maier, K., & Ehsan, S. (2023, October). A Complementarity-Based Switch-Fuse System for Improved Visual Place Recognition. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (pp. 5195-5202). IEEE.

Chapter 6:

- Waheed, M., Milford, M., Zhai, X., McDonald-Maier, K., & Ehsan, S. (2023). An Evaluation and Ranking of Different Voting Schemes for Improved Visual Place Recognition. In *2023 ICRA Workshop on Active Methods in Autonomous Navigation*.

Chapter 2

Literature Review

Visual Place Recognition (VPR) is a fundamental task in the field of computer vision and robotics, aiming to enable autonomous systems to recognise previously visited places based on visual cues, as illustrated in Figure 2.1. The core challenge of VPR lies in the system's ability to identify a location despite significant changes in appearance due to variations in lighting, weather conditions, seasonal changes, and dynamic elements within the environment. The primary challenges in VPR stem from the variability and complexity of real-world environments caused by appearance changes, viewpoint variations, dynamic objects, and perceptual aliasing, as discussed in detail in Chapter 1.

This chapter further dives into the past, present and future of VPR by presenting a detailed survey on the different core methodologies that are part of solving the VPR problem. With years of research been done on VPR there are two major categories the core methodologies can be divided into, with a few other innovative approaches that have been introduced in recent times. However, in the broadest sense the two categories are the traditional approaches referring to feature based methods and secondly modern

approaches primarily referring to deep learning methods. A detailed description of each category is presented ahead to further understand what are the difference, principles and existing challenges.

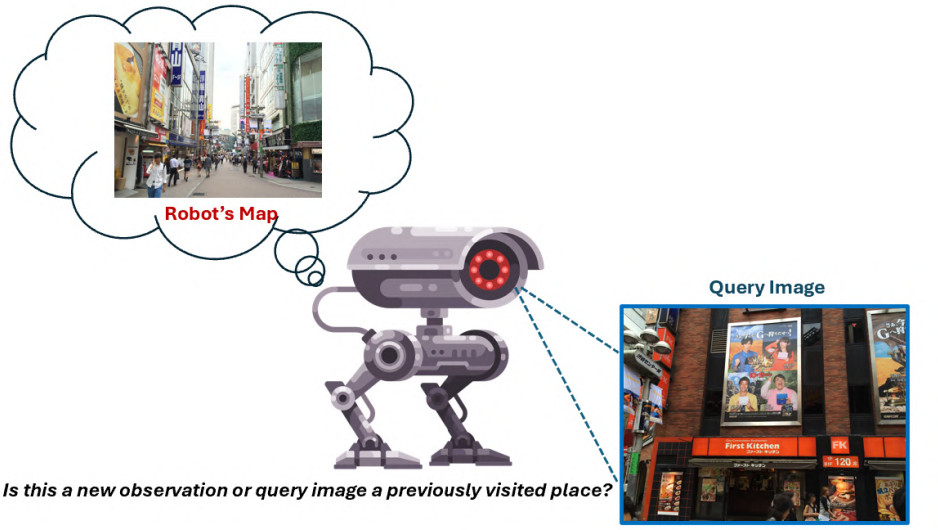

Fig. 2.1. Visual place recognition success: autonomous systems identifying places amid environmental changes using visual cues. Images taken from [1]

2.1 Pre-deep-learning era in VPR

The Pre-deep-learning era in VPR is just another way of referring to feature based approaches due to their vast history and use for solving the VPR problem. Feature based approaches have garnered significant attention for their robustness and efficacy in handling the complex challenges associated with place recognition. The analysis delves into the essence of feature-based methods, underscoring their foundational principles, categorizations, challenges, and advancements.

Beginning from the foundational principles of feature-based approaches for VPR, they rely on the extraction and matching of distinctive visual features from images. These

features, both local and global, serve as the fundamental building blocks for recognizing places under varying environmental conditions. The principle behind these methods is to identify and describe unique aspects of an image, such as edges, corners, textures, or entire scene layouts, which can be invariant to changes in scale, rotation, illumination, and viewpoint. They way these operate is categorized into the following types explained below.

2.1.1 Local Descriptors

Local descriptors pertain to specific interest points within an image, such as keypoints detected by algorithms like Scale-Invariant Feature Transform (SIFT) [29] which extract distinctive invariant features from images that are invariant to image scale and rotation, and partially invariant to change in illumination and 3D camera viewpoint. SIFT and it is successors have enabled a wide range of applications and advanced vision systems, including automatic panorama creation [35,36], object recognition [37], large-scale image retrieval [38], video object retrieval [39], place recognition [40], categorization [41–44], robot localization [45], shot location [46], texture [47,48] and gesture recognition [49] showcasing the versatility of local invariant features. Building upon this foundation, alternative approaches like Speeded Up Robust Features (SURF) [50] offer quicker feature extraction while maintaining performance, enhancing the practicality of local features for various applications. SURF, along with Oriented FAST and Rotated BRIEF (ORB) [51], which combines efficient keypoint detection and a robust binary descriptor, highlight the continuous evolution of feature extraction techniques, ensuring their relevance and utility in real-time VPR tasks.

2.1.2 Global Descriptors

Contrarily, global descriptors encapsulate the overall characteristics of an image, offering a holistic view of the scene. Hence global descriptors capture the overall characteristics of an image, offering an alternative to the detailed but computationally expensive local descriptors. Firstly, Global Features from Accelerated Segment Test (GIST) [52] descriptors summarize the spatial layout of the image, capturing the dominant spatial structures without focusing on individual objects or features. GIST descriptors are efficient to compute and have been used in VPR to quickly filter candidate locations before more detailed analysis and is used for matching place images in [53–56]. Next, Histogram of Oriented Gradients (HOG) [57] descriptors capture edge or gradient structures that are invariant to geometric and photo-metric transformations, except for object orientation. Though originally designed for human detection, HOG has found applications in VPR due to it is ability to represent the structural essence of a scene as used in [58]. Hence techniques like GIST & HOG generate global descriptors that summarize the spatial layout or dominant gradients of an image, facilitating coarse matching and initial filtering in VPR systems.

It is also important to know that the process of feature extraction involves detecting points of interest and computing their descriptors, which are then used to represent and match images. However, the matching phase, especially for local features, often employs algorithms like Approximate Nearest Neighbors (ANN) for efficient comparison. For global features, similarity metrics such as cosine similarity or Euclidean distance are commonly used to gauge the resemblance between scene representations.

2.1.3 Descriptor Aggregation Methods

Another important notable approach to mention are the Descriptor Aggregation Methods. These methods, in order to represent a place effectively, extract features that are often aggregated into a compact representation, using methods like Bag-of-Words (BoW) [41] and Vector of Locally Aggregated Descriptors (VLAD). BoW is actually inspired by text retrieval, and it is models represent an image as a histogram of visual word occurrences, where visual words correspond to cluster centres in the feature space. BoW models, despite their simplicity, have shown effectiveness in VPR, especially when combined with powerful feature detectors and descriptors. Secondly, Vector of Locally Aggregated Descriptors (VLAD) [59] is then an extension of the BoW model that aggregates feature descriptors themselves, rather than their occurrences. It captures the distribution of features around cluster centres, improving place recognition accuracy. Moreover, the only differences are the pre-trained network, PlaceNet [92] instead of VGG-M [93], and the post-processing phase using VLAD instead of BoW.

Indeed feature-based methods have consistently been used throughout for solving VPR and have evolved over time to ensure better efficiency, however despite their strength several challenges remain when utilizing these approaches including high computational costs, sensitivity to environmental changes, and the issue of perceptual aliasing where different places appear indistinguishably similar. To then address these challenges, recent advancements have focused on enhancing feature descriptors, optimizing matching algorithms, and integrating machine learning techniques for adaptive feature selection and matching.

2.2 Deep Learning in VPR

Deep learning has transformed Visual Place Recognition (VPR) by offering robust techniques to address challenges like changing viewpoints, lighting, and seasonal shifts. These methods, ranging from supervised to unsupervised and semi-supervised approaches, have been pivotal in tackling real-world complexities.

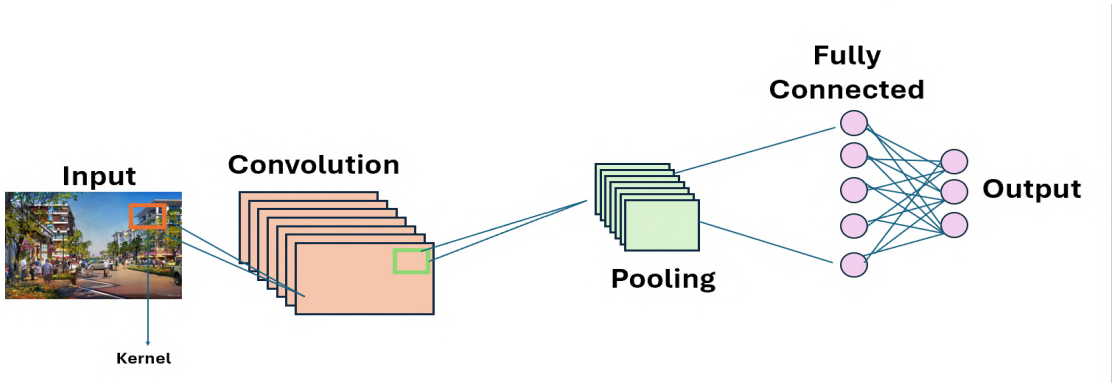

Fig. 2.2. Basic CNN architecture: Layered convolutional filters combining outputs to form final descriptor

Supervised deep learning techniques have seen success by using large labeled datasets to learn intricate visual features. A common strategy involves pre-trained frameworks. Early models like OxfordNet, GoogLeNet, AlexNet, and VGG16 were initially trained on large datasets for image classification. One approach utilized OxfordNet and GoogLeNet to extract VLAD descriptors, creating compact image descriptors [60]. As shown in Figure 2.2, CNN architectures typically use layered convolutional filters to extract features, which are then combined to form a final descriptor. Further advancements used AlexNet and Places205 for feature extraction, enhancing robustness to environmental changes [8]. In landmark-based methods, Edge Boxes were utilized for landmark detection, followed by AlexNet for feature extraction to isolate stable elements in dynamic scenes [61]. Another method used BING and AlexNet's pooling layer for efficient

descriptor extraction [62].

Transfer learning has been crucial in enhancing pre-trained frameworks by allowing models to adapt to VPR's specific challenges. By leveraging models like AlexNet, GoogLeNet, and VGG16, initially trained on datasets like ImageNet, transfer learning enables effective reuse of visual patterns. In one approach, AlexNet (pre-trained on Places205) was fine-tuned to handle environmental changes [8]. Another method adapted AlexNet365 from the Places365 dataset to extract region-based features for VPR [63]. The IVPR-SSADTL model further utilizes deep transfer learning with a pre-trained MixNet for VPR tasks, improving generalization across environments [64].

Region-based approaches have also been key. One method leveraged VGG16 to extract salient regions, focusing on areas providing strong visual cues [65]. In another approach, Region-VLAD used AlexNet365 to aggregate these regional features into robust descriptors, handling diverse conditions effectively [63].

End-to-end frameworks like NetVLAD have set a standard in VPR by combining VGG/AlexNet with a NetVLAD layer, using triplet loss to aggregate local features into global descriptors [1]. This method allows networks to learn compact yet distinctive place descriptors. The concept was extended to 3D LiDAR data, where PointNet was used to capture spatial structures in LiDAR point clouds [66].

Transformer-based architectures have emerged, further advancing VPR. PLACEFORMER employs a transformer model to capture global and local features, directly addressing illumination and viewpoint changes [67]. R2FORMER refines region-level features using a transformer network [68], while TRANsVPR combines CNNs and transformers for learning global and local descriptors [69]. MixVPR offers holistic feature aggregation using pre-trained backbones to achieve high-performance global descriptors [70].

CosPlace rethinks VPR as a classification problem, partitioning large geographical areas into cells and training the model to recognize these cells as distinct classes [71]. This allows CosPlace to learn discriminative descriptors without requiring extensive negative sample mining, achieving state-of-the-art results with reduced memory usage, making it scalable for large-scale VPR tasks.

While supervised learning dominates VPR, unsupervised deep learning techniques provide alternatives when labeled data is scarce. GAN-based methods, such as using Generative Adversarial Networks (GANs) for domain translation, transform images between conditions like seasons to learn invariant features [72]. Another approach expands this concept to 3D LiDAR data, using GANs to learn stable features across diverse LiDAR conditions [73]. In addition, autoencoder-based methods have been used to map images into a HOG descriptor space for loop closure [74].

Semi-supervised learning techniques bridge the gap between supervised and unsupervised methods, using both labeled and unlabeled data. One framework combined weakly supervised and unsupervised learning to derive domain-invariant features, employing an attention-aware VLAD module [75]. Another approach used a self-supervised method to disentangle place-related features, while SegMap, an autoencoder-like network, was presented for learning 3D point cloud segments [76].

Beyond these primary categories, parallel frameworks like SRALNet and ROMS integrate features such as semantic priors and multimodal data to enhance robustness in dynamic environments. Hierarchical frameworks, including a multi-process fusion technique and X-Lost, combine features from multiple sources to filter loop candidates, improving recognition accuracy [34, 77].

In summary, the diverse deep learning techniques in VPR, from supervised approaches

like PLACEFORMER and CosPlace to unsupervised methods like GANs and autoencoders, and further to semi-supervised and complex frameworks like SRALNet, illustrate the rapid evolution of the field, offering increasingly sophisticated solutions. More detailed discussions on deep learning in VPR can be found in survey papers such as [21, 23, 78].

2.3 Sequence-based Methods

Sequence-based methods leverage temporal information by considering sequences of images instead of individual frames. This approach exploits the continuity and dynamics within a traversal, providing additional cues for place recognition. By analyzing sequences, these methods can mitigate the effects of transient occlusions and significant appearance changes, enhancing the robustness of place recognition. SeqSLAM [26], for instance, introduced a novel approach to visual navigation under changing conditions by matching coherent sequences of images rather than individual frames. This method successfully demonstrated robust place recognition across extreme environmental changes, such as transitions from day to night and changes in seasons.

However, the effectiveness of sequence-based methods can be diminished by the computational complexity associated with processing and matching image sequences, especially for long traversals. To tackle these shortcomings, new sequence-based methods are persistently being introduced. For example, SeqVLAD [79] proposes a sophisticated approach that categorizes and benchmarks various techniques for integrating information across individual images to enhance recognition accuracy. It explores the potential use of transformers instead of conventional CNNs and introduces SeqVLAD, an ad-hoc sequence-level aggregator.

Similarly, SeqNet [80] is another approach that introduces an innovative framework that emphasizes the use of sequential image data by developing a hybrid system that generates initial match hypotheses through learned sequential descriptors, effectively capturing temporal dynamics and enabling robust place recognition even in challenging environments. Adding to the advancement of sequence-based methods is MATC-Net [81], which introduces a multi-scale asymmetric temporal convolution network. MATC-Net provides a compact sequence representation that integrates temporal sequence information effectively, optimizing loop closure detection (LCD) tasks in hierarchical visual place recognition (VPR) frameworks. By generating sequential and global features, MATC-Net demonstrates improved performance across different challenging datasets, enhancing the robustness and efficiency of VPR systems.

Lastly and most recent is the Sequence Descriptor approach introduced in [82] that proposes a method that effectively captures and utilizes both spatial and temporal information from sequences of images, aiming to overcome challenges such as environmental changes and perceptual aliasing that often hinder VPR systems. These are just some examples of endeavors focusing on perfecting sequence-based methods. There is more literature being introduced every year, such as [83–86], indicating that sequence-based methods are continually evolving and hold potential for further improvements and solutions.

2.4 Fusion and Complementarity-Based Methods

Fusion-based techniques play a crucial role in enhancing the performance of visual place recognition (VPR) systems by integrating multiple complementary methods. These tech-

niques aim to leverage the strengths of different approaches, providing more flexible, reliable, and accurate place recognition capabilities across various scenarios and environments.

2.4.1 Fusion-Based Methods

With a wide variety of fusion-based VPR methods that can be designed, the common denominator is their design to overcome the limitations of relying on a single type of method or data source. This approach aims to provide more robust and accurate place recognition capabilities. Notable examples of such fusion methods are presented in works like [31] and [34], which discuss and demonstrate the fusion of multiple VPR techniques to produce enhanced VPR accuracy. Inspired by multi-modal fusion practices that integrate data from sensors like visual, depth, LIDAR, and GPS, Probabilistic Robotics [87] is an example of work that discusses how different sensors compensate for each other's weaknesses. For example, LIDAR can complement visual sensors in poor lighting. By combining data from these sources, robots achieve more accurate and robust performance. It emphasizes how fusing diverse sensor inputs strengthens decision-making and localization, embodying the concept implicitly. Similarly, in the context of SLAM (Simultaneous Localization and Mapping), [88] discusses how integrating multiple sensors, such as visual sensors and LIDAR, helps compensate for the limitations of individual sensors, resulting in improved robustness and accuracy across different environments. This approach reflects the implicit use of complementarity, where multiple sensor inputs enhance the system's overall performance. Additionally, [89] demonstrates how the inception architecture captures complementary features at multiple scales by using filters of varying sizes. This fusion of diverse features allows for better generalization across

different image types, further illustrating the power of complementarity in deep learning.

subsectionComplementarity in Computer Vision, Robotics and VPR

While multi-sensor fusion across various fields has demonstrated the importance of integrating complementary inputs, the same principle is crucial when considering the fusion of techniques within VPR systems. An extremely important component to consider, which up until now has not been fully studied, is the consideration of complementarity among these fused VPR techniques. Joining two highly performing VPR techniques with the assumption that both will compensate for each other is misguided, as the high performance of both methods can be redundant, making the effort fruitless. This raises the question of complementarity among VPR techniques—whether it exists, how it can be measured, and its significance. These questions are researched, answered, and discussed in the next chapter.

Complementarity is not only crucial in VPR but has been effectively leveraged in other areas of feature detection. For instance, [90] introduced the concept of mutual coverage, which evaluates how combining feature detectors enhances spatial feature coverage across an image. By measuring how well-combined interest points cover the image, the complementarity of detectors can be assessed. The study demonstrates how different detectors complement each other in vision tasks by improving spatial feature distribution, showcasing the benefits of combining complementary techniques. Similarly, [91] explores how detectors like the Scale-invariant Feature Operator (SFOP) perform well when combined with others, particularly under challenging conditions like JPEG compression and light reduction. This study emphasizes that complementarity can enhance detector performance by compensating for weaknesses in one method with strengths in another. For instance, SFOP performs better on simpler scenes, while other detectors

excel in more complex environments, showcasing complementarity in action. In another study, [92] investigated how combining local feature detectors improves overall robustness by leveraging the strengths of different detectors. Detectors often have varying strengths under different conditions (e.g., lighting or scale changes), and using complementary detectors together helps overcome the limitations of individual methods. This combination approach enhances feature detection, especially in challenging or dynamic environments. Further research by [93] extended this concept by using mutual coverage to quantify complementarity between feature detectors. The authors demonstrated that detectors identifying features in different areas of an image provide complementary advantages, significantly improving detection performance. This metric was shown to predict the effectiveness of detector combinations, especially in real-time applications, where complementary techniques are essential for robust feature detection. Moreover, the work by [94] demonstrates the effectiveness of predicting and integrating maximally complementary techniques to boost the performance of baseline methods. Although this work was published subsequent to the research conducted in this thesis, it underscores the relevance and potential for significant improvements in visual place recognition by harnessing complementary strengths.

These examples highlight the broad applicability and benefits of complementarity-based approaches in enhancing performance and robustness across various domains. By leveraging complementary strengths, systems can overcome individual limitations, leading to significant improvements in performance, as demonstrated in both VPR and related fields.

2.5 Evaluation Metrics Utilised For VPR

Evaluation metrics for Visual Place Recognition (VPR) are crucial for assessing the performance of VPR systems across various conditions and challenges. These metrics help in understanding the effectiveness, reliability, and limitations of different VPR approaches. However, there are some common practises that differ between the robotics and computer vision communities. For example robotics mostly focuses on high precision hence usually requires a single correct match for localisation estimates and therefore more commonly employs evaluation metrics such as Precision Recall Curves, AUC for PR Curves and F1-Score. On the other hand the computer vision community for the most part uses Recall@N and mean-Average Precision (mAP). This section discusses evaluation metrics predominantly used for both, computer vision and robotics tasks. The following are some commonly used evaluation metrics in VPR:

2.5.1 Precision and Recall Curves

Precision-Recall (PR) curves are used for evaluating the performance of Visual Place Recognition (VPR) systems, especially in scenarios where there is a significant imbalance between the classes of interest, typically, the number of non-matching places far exceeds the number of matching places. PR curves plot the precision (the ratio of true positive outcomes to the total predicted positives) against the recall (the ratio of true positive outcomes to the actual total positives) at various threshold settings as illustrated in Figure 2.3. For VPR, precision measures how many of the identified places correctly match the query place, while recall measures how many of the actual matching places the system can identify

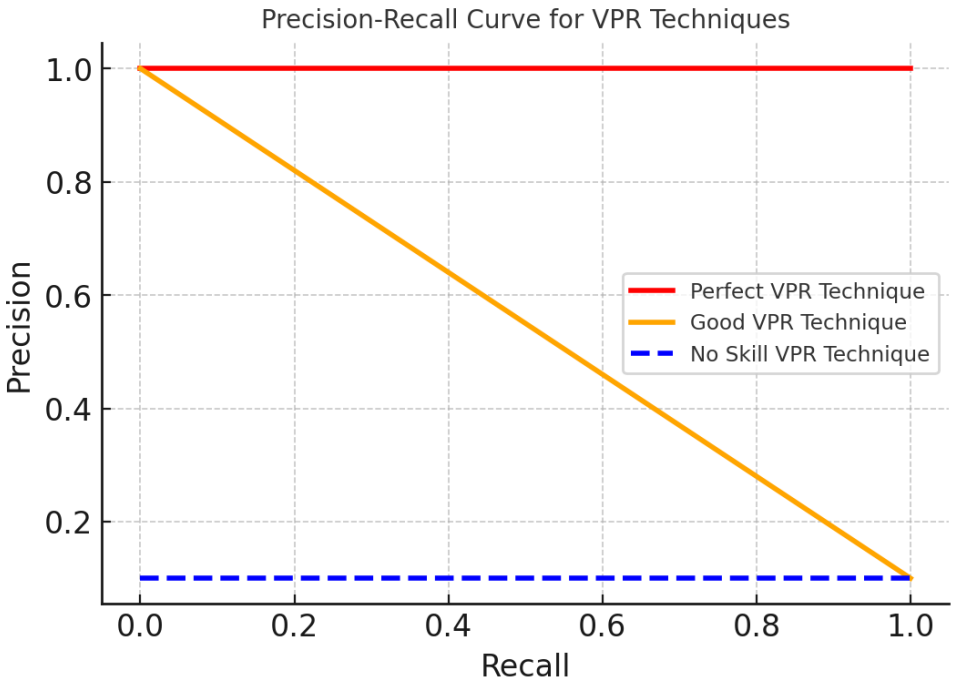

Fig. 2.3. A generic illustration explaining how to interpret a PR-Curve for any VPR model.

Precision quantifies the number of correct positive predictions made. It is crucial in scenarios where the cost of a false positive is high. The equation for precision is given by:

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

Recall measures the ability of a model to identify all relevant instances within a dataset. High recall is essential in situations where missing a positive instance has a significant penalty. The recall equation is:

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

In VPR, precision could relate to how accurately a system identifies images of the same location, while recall would measure the system's ability to retrieve all instances of

a given place.

There are several reasons for why PR curves are used so widely for the evaluation of VPR performance. Some of the reasons are as follows;

- Balancing Precision and Recall: In VPR, it is crucial to maintain a balance between identifying all relevant places (high recall) and ensuring that the identified places are correct (high precision). PR curves provide a visual representation of this trade-off, helping researchers to choose an optimal threshold that balances these metrics according to their specific application needs.

- Handling Class Imbalance: PR curves are particularly useful in VPR due to the common issue of class imbalance (far more negative samples of non-matching places than positive samples of matching places). They offer a more informative measure of performance in such conditions than metrics like accuracy.

- Model Comparison with AUC: By comparing the area under the PR curves (AUC-PR) of different models, researchers can evaluate and choose the best-performing model for their specific VPR application. A higher AUC-PR indicates better overall performance in terms of precision and recall balance.

In conclusion, PR curves play a crucial role in the evaluation and development of VPR systems, especially in addressing the challenges posed by class imbalance. Their use enables a deeper understanding of the trade-offs between precision and recall, guiding the optimization and selection of VPR systems for practical applications.

2.5.2 Area-Under Curve (AUC) of PR-Curves

The Area Under the Curve (AUC) of Precision-Recall (PR) curves is a critical evaluation metric for Visual Place Recognition (VPR) [95], especially when dealing with imbalanced datasets and it is often used as a performance metric for VPR [6,63,74,96]. Additionally, AUC is a suitable criterion for applications requiring high Precision and Recall. The PR curve illustrates the trade-off between precision (the proportion of true positive results in all positive predictions) and recall (the proportion of true positive results in all actual positives) at various threshold levels as explained in equations 2.1 and 2.2.

In VPR, where the negative (non-place matches) vastly outnumber the positive (correct place matches), the AUC of PR curves becomes particularly informative. This metric is more sensitive to changes in the number of false positives among the minority class (the places of interest), making it a preferred choice for evaluating the performance of VPR systems in scenarios with a significant class imbalance.

Area-Under Curve (AUC) is the area underlying a Precision-Recall (PR) Curve [95], commonly used as a performance metric for VPR [6, 63, 74, 96]. AUC-PR is a suitable criterion for applications requiring high Precision and Recall, especially in imbalanced datasets where PR curves are more informative than ROC curves. The AUC-PR measures the area under the PR curve, where a higher AUC represents better VPR performance, with an AUC of 1 indicating optimal precision and recall balance.

2.5.3 F1 Score

The F1 score is a crucial evaluation metric for Visual Place Recognition (VPR), especially significant in scenarios where both the precision of place identifications and the

recall of the system (it is ability to identify all relevant places) are equally important. This balanced approach makes the F1 score an invaluable tool for assessing the overall performance of VPR systems. For VPR, a high F1 Score would indicate both low false positives and low false negatives, ideal for reliable place recognition. The F1 Score is calculated as:

$$F1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \tag{2.3}$$

F1 score provides a more comprehensive understanding of a system's performance than using either precision or recall alone. This is particularly important in real-world applications of VPR, such as autonomous navigation, where failing to recognize a place correctly (low recall) or misidentifying a place (low precision) can have serious consequences. The F1 score helps in identifying VPR systems that are not only accurate but also reliable across various conditions and challenges.

2.5.4 Mean Average Precision (mAP)

Mean Average Precision (mAP) is a another widely used metric for Visual Place Recognition (VPR) and other areas of computer vision, particularly for tasks involving image retrieval and object detection. mAP provides a single-figure measure of quality across recall levels, encapsulating both the precision and recall of a system into a comprehensive indicator of performance.

The concept of Average Precision (AP) originates from the area of information retrieval and is used to evaluate the quality of results returned by a search system. For a single query, AP is calculated by taking the mean of the precision scores at each rank

where a relevant document is found, up to a certain cut-off rank, considering all relevant documents. However, in the context of VPR, where the task often involves identifying images of the same place from a database, mAP averages the AP scores across all queries. This averaging process allows mAP to summarize the performance of the VPR system across it is entire dataset, providing a holistic view of it is effectiveness.

2.5.5 Extended Precision (*EP*)

Extended Precision (*EP*) is a metric introduced to evaluate Visual Place Recognition (VPR) techniques in robotics, specifically designed to improve upon traditional evaluation methods by offering a more nuanced view of a system's performance [97]. The *EP* metric addresses the limitations of existing metrics by combining two key performance indicators: the Precision at minimum Recall (PR0) and the Recall value at which Precision drops from 100% (RP100).

The formula for Extended Precision (*EP*) combines the Precision at minimum Recall (*PR0*) and the Recall value at which Precision drops from 100% (*RP100*) into a single scalar value:

$$EP = \frac{PR0 + RP100}{2} \tag{2.4}$$

PR0 is the Precision at the minimum Recall value, which is determined by the number of False Positives (FP) before the first True Positive (TP). *RP100* represents the highest value of Recall that can be achieved without any FP. If the PR0 is less than 1, RP100 is set to 0, and EP depends only on the Precision at minimum Recall.

The *EP* metric provides a comprehensive measure of a VPR method's performance,

considering both the accuracy of initial matches and the robustness against incorrect matches as more results are considered. Secondly, by encapsulating performance into a single scalar value, EP facilitates straightforward comparisons between different VPR techniques, helping identify which methods are more suitable for particular environments or tasks. This is crucial for the development of more reliable and effective localization systems in robotics, as it allows for the assessment of a method's upper and lower performance bounds and the identification of statistically significant performance differences.

2.5.6 Summary of Evaluation Metrics in VPR

In conclusion, evaluation metrics are crucial for VPR systems, offering insights into their performance, reliability, and computational efficiency. Metrics like Precision, Recall, F1 Score, and mAP highlight the balance between accuracy and the practical utility of VPR systems. Beyond performance, considerations of computational demand and adaptability to environmental changes are vital as demonstrated in [98–100]. Ultimately, these metrics guide the ongoing development and refinement of VPR technologies, ensuring they meet the diverse needs of real-world applications.

2.6 Datasets for Visual Place Recognition

Datasets used for testing Visual Place Recognition (VPR) systems are crucial for evaluating the effectiveness and robustness of these systems across various environmental conditions, lighting changes, and structural variations. These datasets are collected and curated to represent a wide range of real-world scenarios that autonomous robots might

encounter. There are several example of such scenarios including urban and street views, indoor setting, seasonal/weather variation or even synthetic datasets as shown in Figure 2.4 and 2.5. Accordingly these datasets can be collected using relevant means such as mounted cameras, hand-held devices, repeated traversals, or again even synthetically generated using computer graphics and simulation environment.

2.6.1 GardensPoint

Captured on the Queensland University of Technology campus, GardensPoint includes sequences from day and night with significant viewpoint and appearance changes [3, 8]. This dataset is crucial for testing the robustness of VPR algorithms under varying lighting conditions and daily changes. It helps in evaluating how well VPR systems can adapt to changes in illumination and appearance, which are common in urban environments.

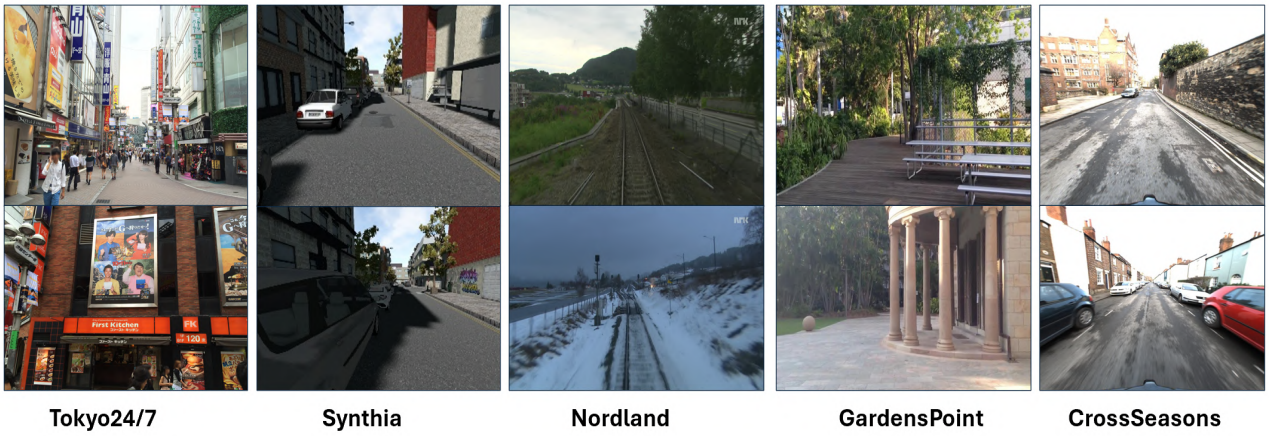

Fig. 2.4. A collection of images from widely employed VPR datasets. Images taken from [1], [7], [2], [3, 8], [4].

2.6.2 Tokyo 24/7

Tokyo 24/7 is designed to test VPR systems across different times of the day, including day, dusk, and night scenes in urban Tokyo [101, 102]. This dataset emphasizes the challenge of recognizing places under extreme lighting variations and in dense urban environments, which are critical for autonomous driving and navigation applications in metropolitan areas. The dataset's variety in time-of-day captures helps in understanding the performance of VPR systems under diverse lighting conditions.

2.6.3 Essex3IN1

Essex3IN1 comprises images from three distinct indoor environments captured under three different lighting conditions [5]. This dataset is particularly useful for evaluating VPR performance in controlled indoor settings, such as offices, homes, and public buildings. It allows researchers to analyse how VPR systems handle indoor lighting variations, which can significantly affect recognition accuracy in real-world indoor applications.

2.6.4 SPEDTest

The SPEDTest dataset, derived from the larger SPED (Semantic Place Description) dataset, focuses on testing VPR robustness against changes in scene content due to the presence of different objects and variations in scene layout [9]. This dataset is essential for evaluating the adaptability of VPR systems to dynamic environments where the arrangement and presence of objects can change frequently, such as in shopping malls or busy public spaces.

Fig. 2.5. A collection of images from widely employed VPR datasets. Images taken from [9], [10–13].

2.6.5 Cross-Seasons

Cross-Seasons contains images of the same locations captured in different seasons, presenting challenges related to changes in vegetation, weather, and lighting conditions [4]. This dataset is pivotal for assessing the capability of VPR systems to handle seasonal variations, which is important for applications involving long-term outdoor navigation where environments change significantly across seasons.

2.6.6 SYNTHIA

The SYNTHIA dataset consists of synthetic images generated from a virtual city environment [7] . It offers variations in weather, seasons, and lighting conditions, providing a controlled yet challenging environment for testing VPR systems. This dataset is particularly valuable for benchmarking VPR algorithms in a virtual setting before real-world deployment, as it allows for extensive testing under a wide range of controlled conditions.

2.6.7 Nordland

Captured from a train journey through Norway, the Nordland dataset showcases changes in scenery across four seasons [2, 103]. It is widely used for testing the ability of VPR systems to handle extreme seasonal variations, making it a standard for evaluating performance under significant environmental changes. This dataset's unique longitudinal aspect makes it particularly useful for long-term place recognition research.

2.6.8 Corridor

The Corridor dataset involves indoor environments with repetitive structures, such as hallways and corridors, posing challenges related to perceptual aliasing [11]. This dataset is critical for testing VPR systems in environments where different parts of the scene look remarkably similar, such as in office buildings and universities, where the challenge is to distinguish between visually similar locations.

2.6.9 17-PLACES

This dataset consists of 17 different places developed at two locations with changing illumination conditions throughout the day and night [13]. It includes images of various indoor settings like hallways, bedrooms, and laboratories. This dataset is useful for evaluating VPR systems in diverse indoor environments, providing a comprehensive testbed for systems designed for real-time applications.

2.6.10 Living-room

The Living Room dataset is composed of high-quality images taken by a robot during a home exploration task, capturing images from a near-floor viewpoint [12]. This dataset provides a unique perspective that differs from typical human-eye-level captures, making it useful for testing VPR systems in domestic robotics applications where the viewpoint is closer to the ground.

2.6.11 Datasets Conclusion

The datasets detailed in this section represent a selection of the diverse VPR resources leveraged throughout this thesis for experimentation. They showcase a range of environments, conditions, and challenges in visual place recognition. However, it is important to acknowledge that these examples are just a subset of the broader array of datasets available in the field such as [104–119], and each offers unique properties for testing and researching VPR technologies

2.7 Summary

This chapter not only explores the evolution, methodologies, and challenges of Visual Place Recognition (VPR) but also presents a comprehensive backdrop against which the current research is positioned. Through a detailed review of traditional and modern VPR approaches, significant strides in the field are underscored, while highlighting gaps that still limit VPR systems' performance in dynamic and complex environments.

A critical gap identified is the lack of complementarity assessment among fused VPR

techniques. The concept of complementarity is pivotal in developing robust and efficient fusion methods, a key point this thesis tries to showcase. Complementarity ensures that combined techniques enhance each other's strengths while compensating for weaknesses, chapter 3 explores this idea in detail and illustrates it through experiments. However, many existing fusion-based techniques lack a focus on well-studied complementarity, thereby missing opportunities for enhanced performance. This section critically analyses related works, highlighting their shortcomings and the research gaps they leave open, which this thesis aims to address. The study on multi-process fusion [31] discusses combining multiple techniques to improve visual place recognition. While insightful, it assumes complementarity without exactly verifying it. The static fusion process does not adapt to changing conditions, potentially leading to situations where techniques are no longer complementary. Additionally, it lacks empirical evidence showing complementarity in various scenarios, resulting in possible redundancy, meaning that multiple techniques might overlap in their functionality, leading to overlapping efforts and missed opportunities to harness unique strengths and boost performance. Hierarchical multi-process fusion [34] presents a novel approach by combining techniques through a hierarchical structure. Despite improvements over the multi-process fusion approach, it lacks complementarity focus. The hierarchical fusion does not adapt dynamically, and there is no detailed analysis of how techniques complement each other in real-time, potentially again leading to redundancy rather than enhancement. Lastly, the exploration of hyperdimensional computing [120] leverages high-dimensional vectors for robust data fusion. However, it remains largely theoretical with limited practical applications demonstrating real-time complementarity. There is a lack of empirical validation showing how hyperdimensional techniques complement traditional methods. The complexity

of implementation also makes it difficult to evaluate complementarity, leaving a gap in practical guidelines. Despite the potential improvements these methods offer, their lack of emphasis on well-studied complementarity represents a significant research gap. Adaptive fusion strategies that dynamically assess and leverage complementary strengths are needed for optimal performance across diverse environments. This thesis addresses these gaps by emphasizing on proposing a framework to study complementarity among different VPR methods. The proposed framwork is also used as the basis of an adaptive fusion/switching strategy dynamically selects and weights techniques based on their complementary strengths, ensuring performance enchancement. Comprehensive evaluations using diverse datasets and real-world experiments validate these methods, ensuring robustness across different conditions. Comparative analyses with the different VPR techniques highlight the advantages and potential drawbacks of the proposed approach later in the thesis. By addressing the shortcomings of existing fusion methods through complementarity, this thesis demonstrates significant performance improvements.

Chapter 3

Proposed Complementarity Framework[1]

Chapter 1 discussed the importance of the need for new and innovative approaches to solve the VPR problem and how developing new methods from scratch is not the only solution. A notable and promising method discussed previously is the multi-process fusion methodology which is a type of ensemble VPR set up. Chapter 1 however goes on further to identify a major shortcoming in such existing systems which is the assumption that the VPR methods fused together complement each other. This chapter puts forth a solution to address this by presenting a well-defined criterion for selecting and combining different VPR methods from a wide range of available options. This is achieved by the introduction of the concept of complementarity among the VPR methods and the framework presented that systematically explores complementarity identifies combinations which can result in better performance. The framework acts as a sanity check to find the complementarity between two techniques by utilising a McNemar's test-like approach, explained ahead in this chapter. It further allows for an estimation of upper and

[1]The work is published in IEEE RA-L (IEEE Robotics and Automation Letters) 2021: vol. 6, no. 3, pp. 5976-5983, July 2021, doi: 10.1109/LRA.2021.3088779.

lower complementarity bounds for the VPR techniques to be combined as illustrated in Figure 3.1, along with an estimate of maximum VPR performance that may be achieved. Based on this framework, results are presented for eight different VPR methods on ten widely-used VPR datasets showing the potential of different combinations of techniques for achieving better performance.

3.1 The Need for Complementartity

Chapter 1 presents VPR as a fundamental yet challenging task that still remains an open problem in the field of robotics and computer vision research. It has been the subject of significant advancements in recent times, introducing several new types of innovative approaches to perform this task. More specifically the recent approach that has drawn attention is the idea of an ensemble VPR system, for example those that ones introduced in [31], [34].

This new approach combines several image processing methods and negates the requirement for multiple sensors to improve VPR performance in terms of accuracy. The concept comes from the empirical data which suggests that some VPR methods are more suitable for certain types of environments and scenarios than others [121]. Hence, utilising multiple VPR techniques simultaneously may compensate for each other's weaknesses. Although these systems mentioned above exhibit promising results, they do not provide a well-defined criterion for selection of VPR techniques based on complementarity out of the available options. Supposing that the fused VPR methods will complement each other in all cases is not a valid assumption and may have detrimental effect on performance and computation. For example, if the VPR techniques that are combined are

redundant, they will not achieve higher performance and will only add to the computational cost which may not be suitable for resource-constrained systems. Hence, complementarity information is vital and can enable a multi-process fusion based system to make an informed decision regarding selection of VPR techniques from available options.

This chapter presents the idea that complementarity of VPR methods has not been studied systematically so far. It bridges this gap and puts forth a framework that can be used as a sanity check for the selection of complementary pairs of VPR techniques for multi-process fusion systems. This framework is based on a McNemar's test-like approach [122], [123] that categorizes each VPR outcome from a technique as either success or failure (considering ground truth information). The framework allows estimation of upper and lower complementarity bounds for the VPR techniques to be combined, along with an estimate of maximum VPR accuracy that may be achieved. This framework is then employed for eight VPR methods to identify highly complementary pairs on widely used VPR data sets.

The reminder of this chapter is organised as follows. Section 3.2 provides an introduction to Complementarity framework. Section 3.3 includes the experimental set up and design. The results and insights produced after the experimentation are presented and discussed in Section 3.4. Finally, a summary of this chapter is given in Section 3.5

3.2 Proposed Complementarity Framework

This section presents the framework for computing complementarity, for establishing the upper and lower complementarity bounds, and for estimating the maximum achievable VPR performance in terms of accuracy by a muti-process fusion system. This framework

may be employed on an arbitrary number of VPR methods to determine the improved selection from among the pool of techniques available. It may also be utilised as a sanity check on whether the VPR techniques that a multi-process fusion system has assembled for integration are likely to improve accuracy. The framework employs a McNemar's test like approach to perform a case-by-case analysis of each VPR technique to compute the complementarity of the given technique with other available methods. Precision-recall curves, F-scores and accuracy percentage are usually utilised as performance metrics for VPR methods. Although viable for some applications / scenarios, these performance metrics do not provide the specific information that tells where exactly does a VPR method succeed or fail, and does not show the whole picture. For example, two VPR methods compared over a dataset of 100 images using these performance metrics may appear to have same performance if they both are able to match 70 images (out of 100). However, it is highly likely that the set of 70 images successfully matched by the first VPR method is not the same set that is also correctly matched by the second VPR technique. This neglected piece of information is critical for determining complementarity of different VPR methods, and is vital knowledge to have specifically when dealing with multi-process fusion systems.

McNemar's test is a form of chi-squared test with one degree of freedom that evaluates the performance of two algorithms based on their outcomes on a case-by-case basis over the same dataset. For utilizing McNemar's test, a criterion is needed to determine whether a test case results in success or failure. The proposed framework is loosely inspired by the McNemar's test as a pairwise analysis is performed on VPR methods on a case-by-case basis over the same dataset. The two VPR methods in question would produce results in the form of correct or incorrect matches verified using ground truth.

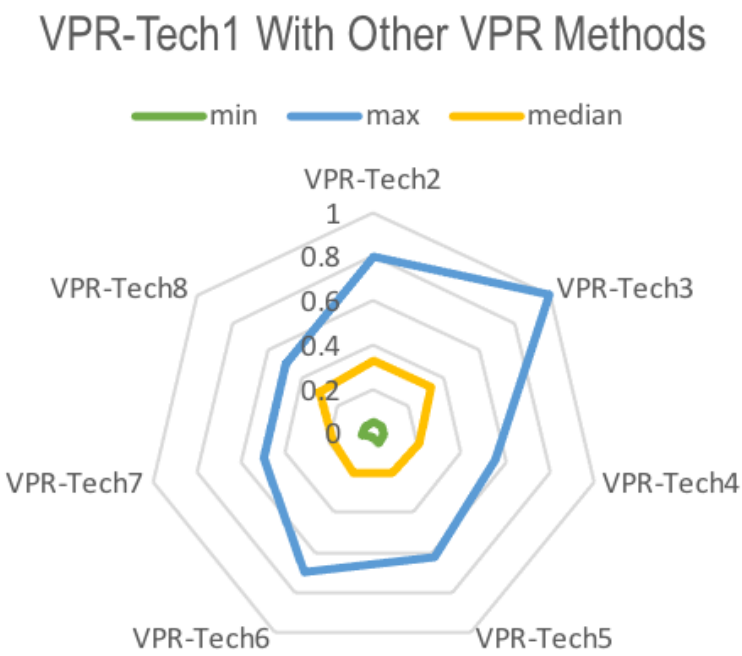

Fig. 3.1. Sample output of the complementarity framework: Primary VPR-Tech1 is combined with secondary methods. The lines green (minimum), blue (maximum), and yellow (median) complementarity bounds.

This data may then be divided into four possible cases as shown in Figure 3.2: first being the number of images where both algorithms are able to match the images correctly, second where the first algorithm matched correctly while the second produced an incorrect match, then vice versa and finally where both algorithms failed and produced incorrect matches. For computing complementarity, the prime focus remains on case two and three as these hold the number of images where the two algorithms perform differently and can help boost each other's performance. The equations presented in this chapter, are all developed as part of this research to provide a robust framework.

3.2.1 Computing complementarity.

Let *A* be the primary VPR technique. Let *B* be a VPR method that may be combined with *A* in a multi-process fusion system to enhance VPR accuracy over an image dataset *D*. VPR performance in terms of accuracy is defined as the ratio of number of images of *D* that are correctly matched (verified by groundtruth) to the total number of images of *D*. The complementarity is calculated by the following equation, which is an original formulation in this work:

$$CBA = \frac{T}{M} \tag{3.1}$$

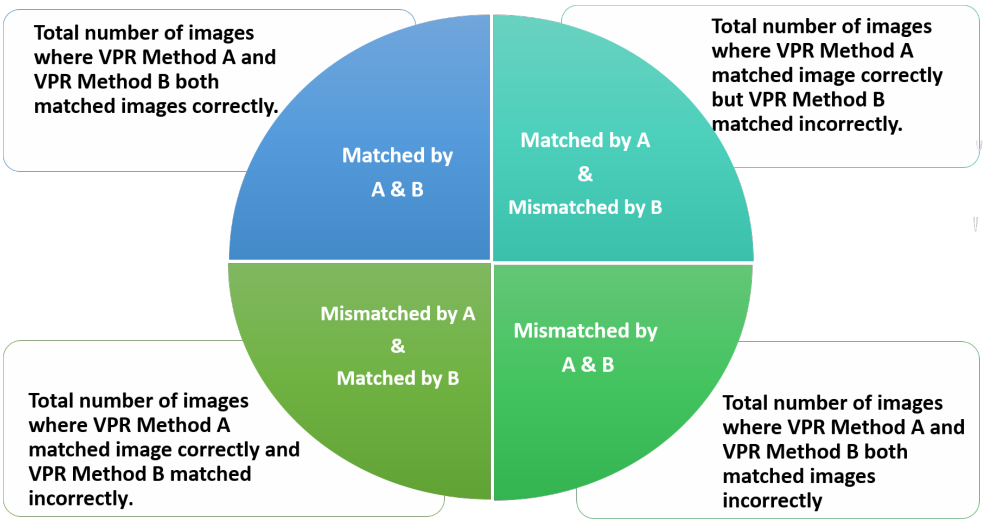

Fig. 3.2. Possible outcomes of pairwise analysis of VPR methods on a case-by-case basis over the same dataset.

Where *CBA* is the complementarity of *B* with *A*; *T* is the number of images of *D* which are incorrectly matched by *A* but correctly matched by *B* when the two methods are run; *M* is the number of images of *D* that are incorrectly matched by *A* when run. A large value of *CBA* implies that *B* complements *A* well on dataset *D* and will result in a

potential increase in VPR accuracy. On the other hand, a small value of *CBA* means that *B* does not complement *A* well. In other words, *A* and *B* are redundant, and combining *A* with *B* will increase computational cost without any substantial increase in VPR accuracy.

3.2.2 Establishing complementarity bounds.

It is interesting to further explore the upper and lower extremities of complementarity of *B* with *A*. Let *K* be the set of *n* individual datasets on which *A* and *B* are run.

$$K = \{D_1, D_2, D_3, \ldots . D_n\} \tag{3.2}$$

Let *J* be the set of complementarity scores (*B* with *A*) computed over *n* dataset in *K*.

$$J = \{CBA_1, CBA_2, CBA_3 \ldots CBA_n\} \tag{3.3}$$

The upper complementarity bound is then established as

$$U = max\{CBA_1, CBA_2, CBA_3 \ldots CBA_n\} \tag{3.4}$$

The lower complementarity bound is estimated as

$$L = min\{CBA_1, CBA_2, CBA_3 \ldots CBA_n\} \tag{3.5}$$

The median of complementarity of *B* with *A* is computed as

$$Q = median\{CBA_1, CBA_2, CBA_3 \ldots CBA_n\} \tag{3.6}$$

3.2.3 Explanation of Complementarity Scores

In this section, we provide an explanation of the complementarity scores used in this thesis, including their mathematical foundation, interpretation, and handling of undefined values. The complementarity score between two VPR techniques is a measure of how well the techniques compensate for each other's weaknesses. Mathematically, the complementarity score C is defined as:

$$C(A, B) = \frac{N_{A \cap B}}{N_A + N_B - N_{A \cap B}}$$

where:

- N_A is the number of proposed matches by technique A,
- N_B is the number of proposed matches by technique B,
- $N_{A \cap B}$ is the number of common proposed matches by both techniques A and B.

3.2.4 Motivation and Interpretation of Complementarity Scores

The motivation for using complementarity scores is to quantify how well two VPR techniques compensate for each other's weaknesses, guiding the selection of techniques that improve overall VPR performance. A complementarity score ranges from 0 to 1, with specific interpretations for these boundary values:

- **Complementarity Score of 0**: Indicates total redundancy, where the techniques make the same proposed matches, correct or incorrect. This means that combining these techniques would add no value, as they do not complement each other and would only increase computational cost without improving VPR performance.

- **Complementarity Score of 1**: Signifies perfect complementarity, where each technique makes entirely different proposed matches. This maximizes the overall chance of correct matches, as one technique's failures are fully compensated by the other, leading to significant improvements in VPR performance.

3.2.5 Handling Undefined Values

If the denominator in the formula becomes zero ($N_A + N_B - N_{A \cap B} = 0$), the complementarity score C is undefined. This occurs if $N_A = 0$ and $N_B = 0$, meaning both techniques fail to make any proposed matches. Practically, this is handled by:

- Assigning a complementarity score of 0, as neither technique contributes to recognition.
- Excluding such pairs from the analysis to prevent distortion of results.

3.2.6 Ensuring Valid Scores

By the formula design, the complementarity score cannot exceed 1, as the numerator represents the subset of proposed matches common to both techniques, and the denominator represents the union of proposed matches by both techniques. The complementarity score is a robust metric for evaluating the synergistic potential of different VPR techniques. A score of 0 indicates redundancy, while a score of 1 indicates complete complementarity, providing a clear measure for enhancing VPR performance in terms of accuracy through technique combinations. Handling undefined values by default assignment or exclusion ensures the robustness and reliability of the analysis.

3.2.7 Estimating maximum achievable performance.

It is beneficial to estimate the maximum achievable VPR performance (in terms of accuracy) of a multi-process fusion system over a dataset at an early stage. This is estimated as follows:

$$MAPE = \frac{(T + W + X)}{Y} \tag{3.7}$$

Where *MAPE* is the maximum achievable VPR performance estimate for the fusion system over a dataset *D*; *T* is the number of images of *D* which are incorrectly matched by *A* but correctly matched by *B* when the two methods are run; *W* is the number of images of *D* which are correctly matched by *A* but incorrectly matched by *B* when the two methods are run; ; *X* is the number of images of *D* which are correctly matched by both *A* and *B* when the two methods are run; *Y* is the total number of images of *D*.

3.3 Complementarity Experimental Setup

This section demonstrates the use of the proposed complementarity framework by comparing several VPR techniques in a pairwise manner, tested on multiple widely used VPR datasets [124] as listed in Table 3.1 that are used for the experiments, namely GardensPoint [3], 24/7 Query [101], Essex3in1 [5], SPEDTest [9], Cross-Seasons [4], Synthia [7], Corridor [11], 17-Places, Living room [12], and Nordland [2].

The implementation details of the eight VPR techniques that are utilised in the experiments are given below [125].

AlexNet: The use of AlexNet for VPR was studied by [8], who suggested that *conv3* is the most robust to conditional variations. The *conv3* layer refers to the third convolutional layer of the AlexNet architecture, which has been shown to effectively capture

TABLE 3.1: VPR-BENCH DATASETS USED FOR DETERMINING COMPLEMENTARITY

Dataset	Environment	Query Images	Ref Images	Viewpoint-Variation	Conditional-Variation
GardensPoint	University Campus	200	200	Lateral	Day-Night
24/7 Query	Outdoor	375	750	6-DOF	Day-Night
ESSEX3IN1	University Campus	210	210	6-DOF	Illumination
SPEDTest	Outdoor	607	607	None	Seasonal and Weather
Cross-Seasons	City-Like	191	191	Lateral	Dawn-Dusk
Synthia	City-like(Synthetic)	947	947	Lateral	Seasonal
Nordland	Train Journey	1622	1622	None	Seasonal
Corridor	Indoor	111	111	Lateral	None
17-Places	Indoor	406	406	Lateral	Day-Night
Living-room	Indoor	32	32	Lateral	Day-Night

mid-level features that are less sensitive to changes in environmental conditions. Gaussian random projections are used to encode the activation-maps from *conv3* into feature descriptors. The implementation of AlexNet for this purpose is similar to the one employed by [74].

NetVLAD: The original implementation of NetVLAD was in MATLAB, as released by [1]. The Python part of this code was open-sourced by [126]. The model selected for evaluation is *VGG-16*, which is a deep convolutional neural network architecture known for its 16 layers, trained in an end-to-end manner on Pittsburgh 30K dataset [1] with a dictionary size of 64 while performing whitening on the final descriptors.

AMOSNet: This technique was proposed by [127], where a CNN was trained from scratch on the SPED dataset. The authors presented results from different convolutional layers by implementing spatial pyramidal pooling on the respective layers. While the original implementation is not fully open-sourced, the trained model weights are shared by authors.

HybridNet: While AMOSNet was trained from scratch, [127] took inspiration from transfer learning for HybridNet and re-trained the weights initialised from the top-5 convolutional layers of CaffeNet [128] on SPED dataset. This work implements HybridNet

using *conv5* of the shared HybridNet model [129].

RegionVLAD: This technique is introduced and open-sourced by [63]. This work uses AlexNet (trained on the Places365 dataset) as the underlying CNN. The total number of regions of interest is set to 400, and uses *conv3* for feature extraction. The dictionary size is set to 256 visual words for VLAD retrieval. Cosine similarity is subsequently used for matching descriptors of query and reference images.

CALC: The use of convolutional auto-encoders for VPR was proposed by [74], where an auto-encoder network was trained in an unsupervised manner to re-create similar *Histogram of Oriented Gradients (HOG)* descriptors for viewpoint variant (cropped) images of the same place. The model parameters use are from 100,000 training iterations. Cosine-matching is used for descriptor comparison.

Histogram of Oriented Gradients (HOG): HOG is one of the most widely used *hand-crafted feature descriptors* [58]. This work uses a cell size of 16 $\times$ 16 and a block size of 32 $\times$ 32 for an image size of 512$\times$512 for the implementation. The total number of histogram bins is set to 9. Cosine-matching between HOG descriptors of various images is used to find the best place match.

CoHOG: This technique uses image entropy for region-of-interest extraction. The regions are subsequently described by dedicated HOG descriptors, and these regional descriptors are convolutionally matched to achieve lateral viewpoint-invariance. It is an open-source technique and uses an image size of 512 $\times$ 512, cell size of 16 $\times$ 16, bin-size of 8, and an entropy-threshold (ET) of 0.4. CoHOG also uses cosine-matching for descriptor comparison [130].

3.4 Complementarity Results and Analysis

This section presents the results generated by utilizing the proposed framework over a set of eight VPR techniques on various standard VPR datasets. The figures ahead illustrate the complementarity scores of different VPR methods with each other across these datasets, allowing for a visual analysis of how different pairs of VPR methods exhibit varied complementarity levels.

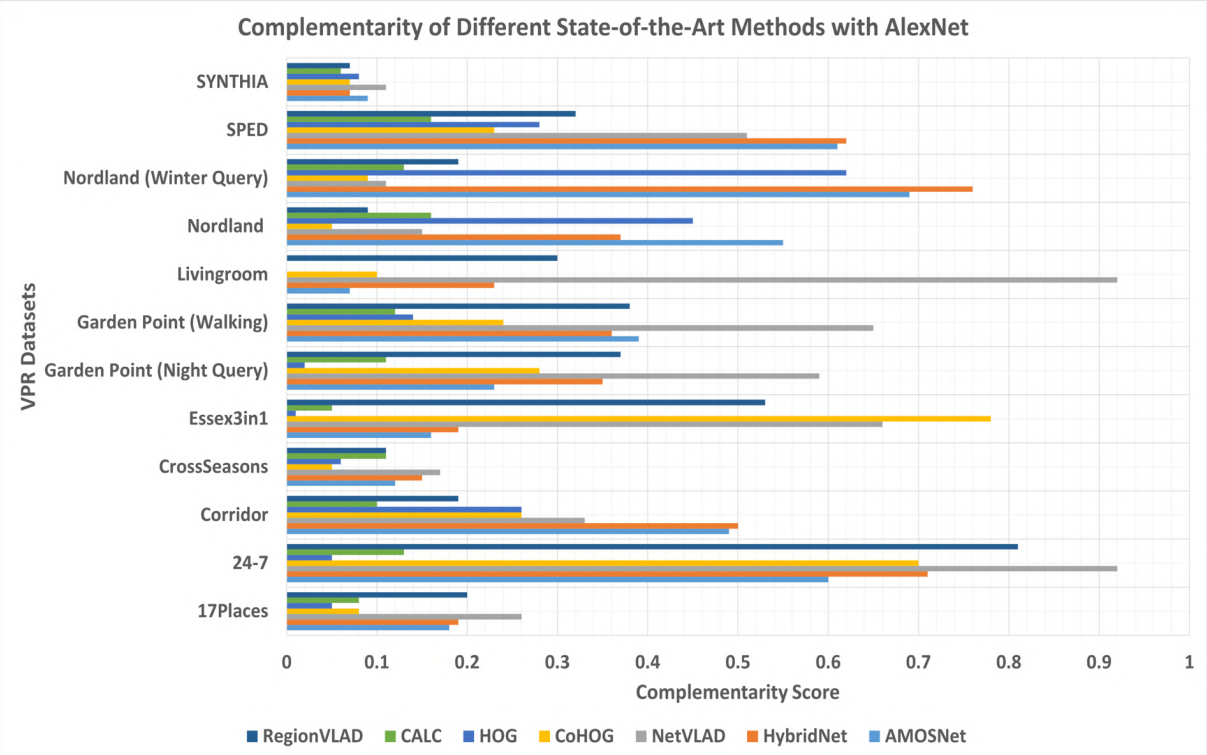

Fig. 3.3. Complementarity of VPR methods with AlexNet on Multiple VPR datasets.

One of the most intriguing findings is how certain VPR techniques, which perform poorly on their own, show remarkably high complementarity scores when paired with others. This phenomenon occurs when the errors made by one technique are systematically different from the errors made by another, allowing them to effectively cover

each other's weaknesses. Figure 3.3 for example, demonstrates high complementarity of AlexNet with NetVLAD, HybridNet, and RegionVLAD on several datasets. For example, NetVLAD achieves complementarity scores of 0.9, 0.65, 0.65, and 0.9 on the 24-7, Essex3in1, GardenPoint, and Livingroom datasets, respectively. Similarly, HybridNet shows strong performance on the 24-7, Corridor, Nordland, and SPED datasets, while RegionVLAD achieves the highest scores on the 24-7 and Essex3in1 datasets.

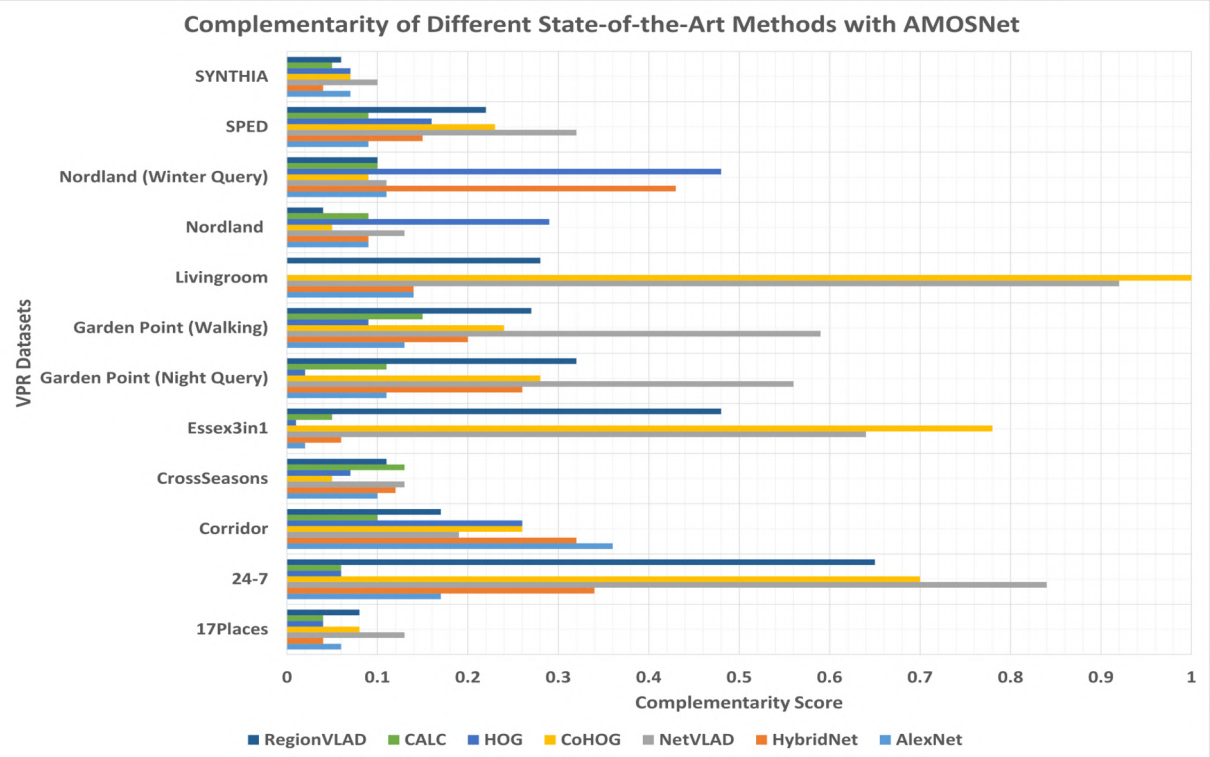

Fig. 3.4. Complementarity of VPR methods with AMOSNet on Multiple VPR datasets.

In another interesting example, CoHoG and NetVLAD are the only methods that complement AMOSNet well as illustrated in Figure 3.4. For instance, AMOSNet and CoHoG achieve high complementarity scores of 0.7, 0.8, and 1 on the 24-7, Essex3in1, and Livingroom datasets, respectively. The combination of AMOSNet and NetVLAD reaches scores of 0.85, 0.65, and 0.9 on the same datasets. This suggests that combining AMOSNet with

CoHoG or NetVLAD can lead to a robust VPR system, while CALC consistently scores low, indicating poor complementarity with AMOSNet.

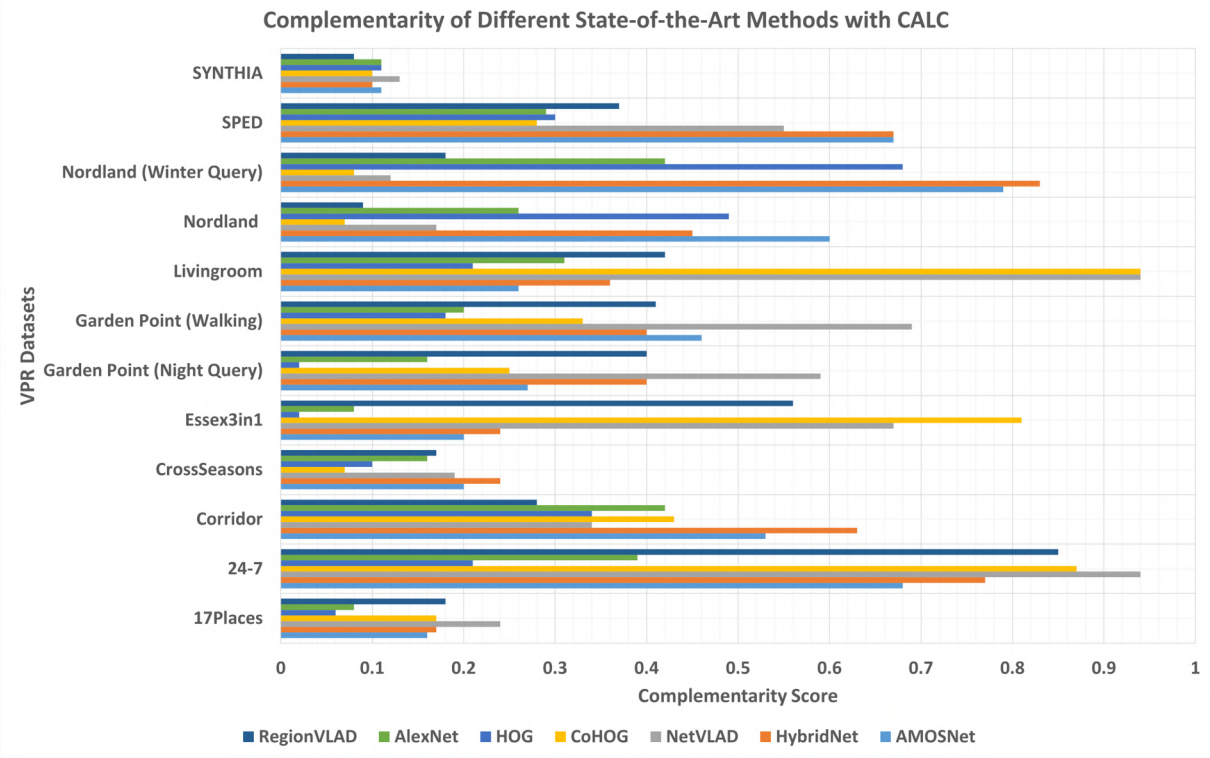

Fig. 3.5. Complementarity of VPR methods with CALC on Multiple VPR datasets.

When considering CALC as the primary VPR technique as shown in Figure 3.5, CoHoG and NetVLAD emerge as suitable partners. CoHoG exhibits high complementarity scores on the 24-7, Essex3in1, Livingroom, and SPED datasets, while NetVLAD matches CALC well on the 24-7, Essex3in1, GardenPoint, and Livingroom datasets. Despite CALC's generally lower standalone performance, its combination with these techniques significantly boosts overall performance. This highlights the importance of evaluating complementarity independently of individual performance metrics.

The complementarity levels are also presented in the form of radar charts (Figure 3.6), representing the lower and upper bounds of complementarity of each VPR tech-

Fig. 3.6. Max (upper bound), Min (lower bound), and Median complementarity of VPR methods with: AlexNet, AMOSNet, CALC, CoHoG, HoG, HybridNet, NetVLAD, RegionVLAD.

nique with all other methods. These charts provide a holistic view of how much the complementarity levels vary among different techniques. For example, combinations with AlexNet show the largest upper bounds with NetVLAD, RegionVLAD, and HybridNet, while CALC has the smallest bounds. AMOSNet combinations exhibit the highest upper bounds with NetVLAD and CoHoG, whereas HybridNet and CALC have the smallest bounds.

The detailed numerical results of the remaining figures for different datasets, including the exact complementarity scores for each technique pair across all datasets recorded

TABLE 3.2: MAXIMUM ACHIEVABLE PERFORMANCE ESTIMATE FOR DIFFERENT COMBINATIONS OF VPR METHODS ON STANDARD DATASETS

VPR Combinations	17Places	24-7	Corridor	CrossSeasons	Essex3in1	Garden Point	Livingroom	Nordland	SPED	SYNTHIA
AlexNet + AMOSNet	43.1	87.2	73.8	32.9	28.0	54.0	62.5	83.6	81.3	32.2
AlexNet + CALC	36.2	72.0	54.0	31.9	19.0	34.0	59.3	53.8	59.6	30.62
AlexNet + CoHoG	44.0	95.7	71.17	28.7	82.8	48.5	96.8	50.8	63.4	32.7
AlexNet + HoG	33.7	69.3	62.1	28.2	15.7	36.0	59.3	79.9	65.2	32.1
AlexNet + HybridNet	43.3	90.6	74.7	35.0	30.1	52.0	68.7	87.4	81.7	31.3
AlexNet + NetVLAD	**48.2**	97.6	65.7	**36.1**	70.9	65.5	96.8	52.7	76.2	**34.2**
AlexNet + RegionVLAD	44.5	94.1	58.5	31.9	60	53.5	71.8	57.1	67.2	31.2
AMOSNet + CALC	41.8	85.6	63.0	35.0	30.0	55.5	56.2	83.4	81.3	30.4
AMOSNet + CoHoG	44.0	95.4	69.3	29.3	84.2	60.5	**100**	83.1	84.1	32.5
AMOSNet + HoG	42.11	85.6	69.3	30.8	27.1	52.5	56.2	90.3	82.8	32.5
AMOSNet + HybirdNet	42.11	89.8	72.0	34.5	30.9	57.9	62.5	89.5	82.5	29.8
AMOSNet + NetVLAD	47.2	97.6	66.6	35.0	73.8	75.5	96.8	83.6	**86.1**	34.1
AMOSNet + RegionVLAD	44.5	94.6	65.7	33.5	61.9	62.0	68.7	83.4	84.0	31.2
CALC + CoHoG	42.3	94.3	54.9	25.1	83.3	45.5	96.8	26.6	59.1	29.6
CALC +HoG	34.4	63.4	47.7	27.2	13.3	33.0	53.1	75.0	60.1	30.6
CALC + HybridNet	42.3	94.3	54.9	25.1	83.3	45.5	96.8	26.6	59.1	29.6
CALC + NetVLAD	47.0	97.3	47.7	34.5	71.4	63.5	96.8	30.0	74.4	32.1
CALC + RegionVLAD	43.3	93.0	43.2	32.9	61.4	51.5	65.6	35.1	64.4	28.4
CoHoG + HoG	42.8	94.6	63.9	19.3	82.3	47.5	96.8	73.5	62.9	31.2
CoHoG + HybridNet	45.0	95.4	**77.4**	31.9	84.7	58.5	**100**	86.2	83.1	31.4
CoHoG + NetVLAD	45.8	97.6	61.2	27.2	**88.5**	74.5	96.8	22.3	74.6	32.9
CoHoG + RegionVLAD	44.0	97.0	59.4	25.6	86.1	59.5	96.8	28.4	64.0	31.5
HoG + HybridNet	42.6	90.1	75.6	33.5	29.5	50.5	62.5	**91.4**	82.5	31.3
HoG + NetVLAD	45.8	97.6	61.2	27.2	**88.5**	74.5	96.8	22.3	74.6	32.5
HoG + RegionVLAD	43.3	93.8	54.9	26.1	59.0	50.0	65.6	74.3	68.6	32.5
HybridNet + NetVLAD	45.0	94.9	72.9	**36.1**	61.4	61.0	75.0	86.4	83.0	29.7
HybridNet + RegionVLAD	45.0	94.9	72.9	**36.1**	61.4	61.0	75.0	86.4	83.0	29.7
NetVLAD + RegionVLAD	46.7	**98.4**	51.3	35.6	79.5	**77.0**	96.8	31.3	77.5	33.1

in Table 3.2, are provided in Appendix A. These results highlight the significant improvements that can be achieved by maximizing complementarity, offering insights into the improved selection of VPR techniques for ensemble setups.

3.5 Complementarity Summary

The well-defined Complementarity framework presented in this chapter is essential for determining the viability of combining different VPR methods for a multi-process fusion system. The complementarity information computed through the proposed framework helps to select the best possible combination of VPR techniques to ensure performance improvement in fused systems. The framework is based on a McNemar's test-like approach [122], [123] that categorizes each VPR outcome from a technique as either

success or failure (considering ground truth information). It allows the estimation of upper and lower complementarity bounds for the VPR techniques to be combined, along with an estimate of the maximum VPR performance (in terms of accuracy) that may be achieved. This chapter provides a significant contribution to the development of ensemble VPR methods proposed within this thesis. For exmaple one of the key insights from our analysis is the identification of VPR techniques that are particularly well-suited for specific environmental conditions. For instance, NetVLAD consistently shows high complementarity scores across different datasets, making it highly effective for day-night scenarios. Similarly, CoHoG and AMOSNet demonstrate high complementarity in scenarios with significant seasonal variations, suggesting their suitability for environments where appearance changes drastically with seasons. HybridNet and RegionVLAD exhibit strong performance under varying illumination conditions, making them ideal candidates for scenarios with fluctuating lighting. These insights into the suitability of specific VPR techniques for different environmental conditions informed the design of the Switch-Fuse System described in Chapter 5 ahead. By selecting techniques based on their demonstrated strengths in particular scenarios, the system dynamically switches and fuses the most appropriate techniques to enhance overall matching accuracy.

The remainder of this thesis utilises the insights and results produced in this chapter which further reiterates the importance of the findings presented in this chapter. To this end, Section 3.1 explained the need for complementarity and its importance that had, to this author's knowledge, had been overlooked to date. Section 3.2 introduced the Complementarity framework and its design. While Section 3.3 is dedicated for a demonstration of the framework utilizing a large experimental setup. This explains, in detail, where and how the framework is employed and produces interesting cases for

future reference and deeper analysis along with critical insights that are the basis of later chapters presented ahead.

Chapter 4

SwitchHit[1]

Chapter 3 introduced the concept and importance of complementarity for different ensemble VPR set ups and provided insights to evaluate level of complementarity among different VPR techniques. Based on the findings presented in Chapter 3 , this chapter is dedicated to the research conducted for designing a complementarity-based Switching system. SwitchHit, unlike other existing multi-fusion systems for VPR, does not simply run all techniques at once to enhance performance, rather predicts the probability of correct match for an incoming query image and dynamically switches to another complementary technique, as required.

Further motivation for SwitchHit is driven by the lack of a universal VPR technique that can work in all types of environments, on a variety of robotic platforms, and under a wide range of viewpoint and appearance changes. Nonetheless, recent work has shown the potential of combining different VPR methods intelligently for some specific VPR datasets to achieve better performance. This, however, requires ground truth informa-

[1]The work presented in this chapter has been accepted and presented at IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Kyoto, Japan, 2022.

tion (correct matches) which is not available when a robot is deployed in a real-world scenario. Moreover, running multiple VPR techniques in parallel may be prohibitive for resource-constrained embedded platforms. SwitchHit, on the other hand, is built on a probabilistic model and knowledge of complementarity of different VPR techniques and operates by switching to the best available VPR techniques. This innovative use of multiple VPR techniques allow SwitchHit to be more efficient and robust than other combined VPR approaches employing brute force and running multiple VPR techniques at once. Thus, it is more suitable for resource constrained embedded systems and achieving an overall superior performance from what any individual VPR method in the system could have achieved running independently. Figure 4.1 depicts an example of how the **SwitchHit** system ***selects*** and ***switches*** between VPR methods for each query image to ensure the selection of the best VPR technique for a given query, to maximize performance. The system does this by predicting the probability of each technique correctly matching the query image and switching from a technique with low chances of correctly matching to a technique with higher chances of correctly matching the query image. The top of the figure displays the fluctuating pattern of switches between different VPR techniques. The bottom presents the total percentage of each VPR technique selected using SwitchHit and the final results such a system produces which clearly indicate a surge in the total number of correctly matched images for the chosen data set.

4.1 Shortcomings of Existing Ensemble VPR Setups

Instead of another attempt to develop a new VPR technique from scratch, a well-received and intuitive solution was put forward in [31], [34] that introduced the concept of multi-

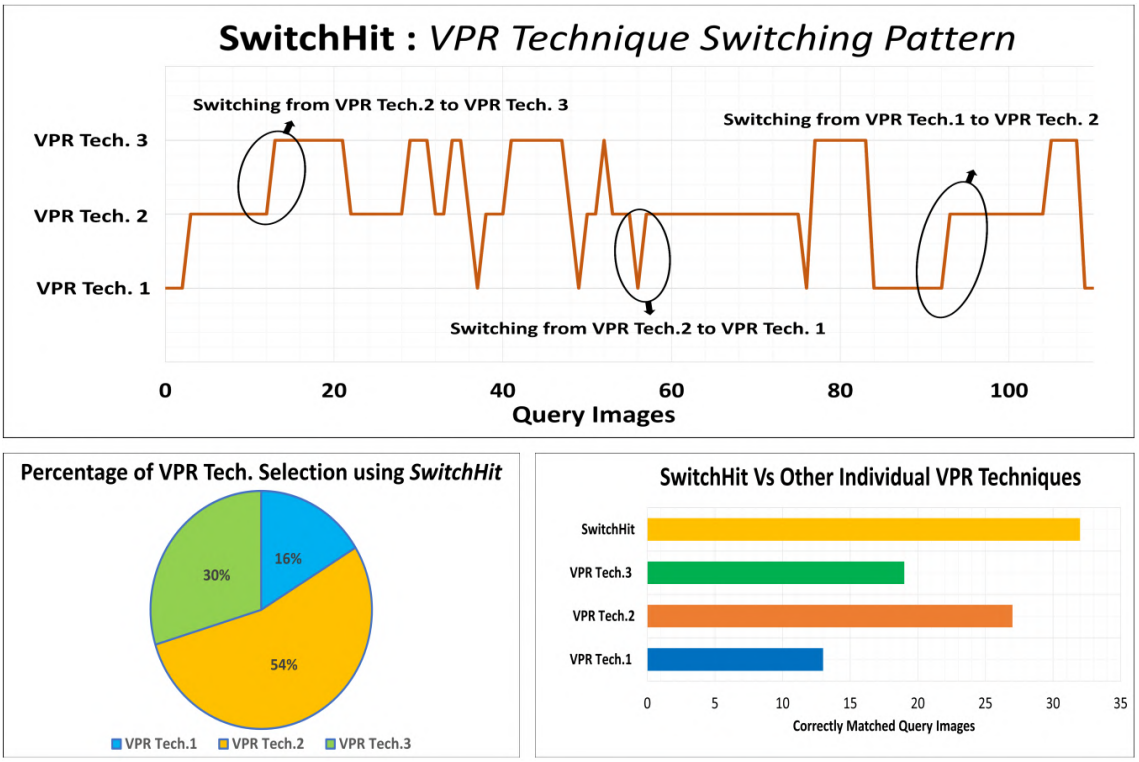

Fig. 4.1. SwitchHit: Dynamically optimizing VPR with dynamic VPR technique selection and switch

process fusion between different VPR techniques. They take inspiration from the practise of fusing multiple sensors to improve place recognition performance as that has been the focus of several research works [8], [74], [1]. However, multi-sensor approaches help boost performance, they do carry certain disadvantages, such as expensive and bulky sensors, and potentially significant increase in computation. To overcome these shortcomings, the concept of fusing multiple VPR techniques gained popularity. The authors of [127] combined multiple image processing methods into a merged feature vector using a convex optimization approach to decide the best match from the sequence of images generated. The effort did generate some promising results over multiple datasets but had limited overall performance due to the absence of sequential information. Similarly, a multi-process fusion system was introduced in [31] which combined multiple VPR meth-

ods using a Hidden Markov Model (HMM) to identify the optimal estimated location over a sequence of images. In continuation of this, a three-tier hierarchical multi-process fusion system was then presented in [34] which was customizable and may be extended to any arbitrary number of tiers. A different place recognition method is used in each tier to compare the query image with the provided sequence of images. The existing research conducted for advancing VPR via such approaches presented promising results and undoubtedly held even more potential.

However, two major and common shortcomings reoccurring throughout were the absence of complemenentarity and the use of brute force to run multiple techniques simultaneously. Firstly, instead of a carefully curated group of complementary VPR methods, a selection of somewhat random VPR techniques were chosen to be fused together, often on the basis of each method's individual performance not considering the redundancy that two otherwise high-preforming techniques might have with each other. The work presented in Chapter 3 presents insights to tackle this issue however the utility of the knowledge presented in Chapter 3 was yet to be tested. Secondly, employing multiple VPR techniques simultaneously to fuse their results to enhance performance is not always an efficient approach especially for resourc,e constrained environments.

SwitchHit attempts to tackle both these issues by basing its tested group of VPR techniques on careful selection by considering the information provided on complementarity and follows a probabilistic model to allow for dynamic switching between the available complementary techniques as to avoid the use of brute force.

4.2 SwitchHit Methodology

This section discusses the probabilistic complementarity-based switching system that estimates the probability of the primary VPR technique correctly matching the query image.

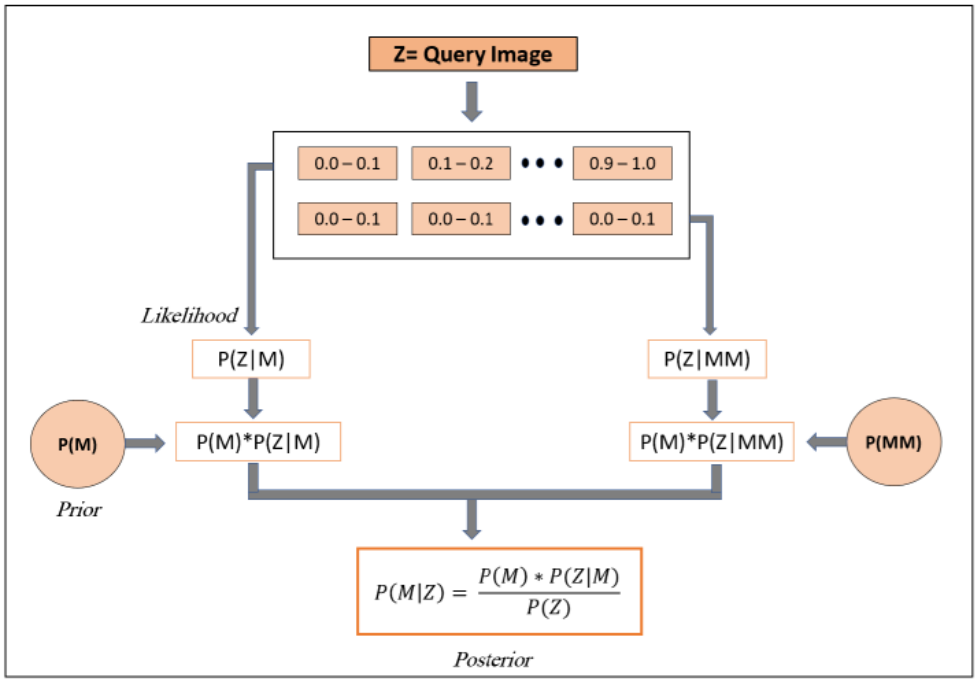

Fig. 4.2. Bayes' Theorem inspired framework: Updating VPR matching probabilities using priors and event likelihoods"

While if the probability of match is lower than the set threshold the system looks for the best alternative technique by calculating and selecting the technique with highest complementarity to the primary VPR technique.

The framework is based on the idea of Bayes inference which is a method of statistical inference using Bayes' theorem to update the probability for a hypothesis once more evidence is provided. The framework employs the basis of this statistical approach in terms that the hypothesis is that the incoming query image is correctly matched. The evidence is the matching score the VPR technique computes for every query image. Training

the system on several data sets provide us with the *prior* probability of a correct match the VPR techniques have overall and the *likelihood* of correctly matching the query image given a certain matching score range, which is also computed during training. The prior information is used to estimate the *posterior* probability of matching correctly given the input query image matching score. This will help the system to avoid running even after several incorrect matches and help regulate performance although with access to ground truth at the training stage. Furthermore, this decision is then the guiding factor to computing the complementarity of the primary VPR technique with other available VPR techniques to allow a dynamic switch to a better alternative technique for the query image.

The system runs by performing six major steps that are explained in detail ahead and also illustrated in Figure 4.2 which depicts the Bayes' theorem inspired framework that updates the matching probability of a system for the given query image based on prior information and likelihood of matching. Where $P(M_{match})$ and $P(M_{mismatch})$ is probability of match and mismatch respectively. $P(M_{match})$ and $P(Z|M_{mismatch})$ is probability of event *Z* occurring given its match or mismatch respectively while figure 4.3 explains the second component of the framework that is the selection and switching to a VPR Technique with the highest complementarity if the probability of match *(Posterior)* is below the threshold value. The system begins by training for the data sets mentioned in the Table 4.1 to gather the prior and likelihood values to determine later whether a switching step is required and finally, if needed, switches to a VPR technique that is the best alternative for the given the query image. Below each step performed is explained in detail along with their mathematical representation.

A. Computing Probability of Total System Match and Mismatch *(Prior)*. These are

the equations proposed in this thesis to compute the probability of correct match that a VPR technique has overall for given data set. Where $P(M_{match})$ is the probability of total correct matches which is calculated by the total number of correct matches in the data set divided by the total number of images in the given data set. This is vice versa for $P(M_{mismatch})$ which is the probability of total incorrect matches for the dataset.

$$P(M_{\text{match}}) = \frac{Total\ No.\ of\ matches\ in\ Dataset}{Total\ No.\ of\ Images\ in\ Dataset} \tag{4.1}$$

$$P(M_{\text{mismatch}}) = \frac{Total\ No.\ of\ Mismatches\ in\ Dataset}{Total\ No.\ of\ Images\ in\ Dataset} \tag{4.2}$$

B. Computing Probability of Any Score Event given its match or mismatch *(Likelihood)*. These equations compute the probability of any score event/range occurring given that it is correctly or incorrectly matched by the VPR technique. $P(Z|M_{match})$ is the probability of each score range given that it is correctly matched by a technique. This is calculated by a solving the fraction between number of correct matches given a certain score range and the total number of images or entries occurring in the given score range. This is vice versa for $P(Z|M_{mismatch})$ which is the probability of each score range given that it is incorrectly matched by a given technique. These equations are used for each score range considered in this experimentation beginning from 0 and ending at 1 with an interval of 0.1 between each range.

$$P(Z|M_{\text{match}}) = \frac{W}{X} \tag{4.3}$$

Where $P(Z|M_{match})$ is the probability of each score range given that it correctly matched by a technique and W is a number of matches within the given score range and X is the

total number of images within given score range.

$$P(Z|M_{\text{mismatch}}) = \frac{Y}{X} \tag{4.4}$$

Where $P(Z|M_{mismatch})$ is the probability of each score range given that it's not correctly matched by a technique, Y is the number of mismatches within the given score range and X is the total number of images within given score range.

C. Computing Probability that Query Image is Matched Given Input Score Event *(Posterior)*. This equation computes the posterior probability of the VPR technique correctly matching the image given the input query matching score generated. Where $P(M_{match})$ is the probability of match by the primary technique overall which is the prior in the framework. $P(Z|M_{match})$ is the likelihood for the VPR technique given it will correctly match for a certain score event. This produces an updated but non-normalized probability distribution between the matching and mismatching. Finally, $P(Z)$ is the marginalization in the equation and is the summation of both updated non-normalized distribution of match and mismatch. In other words $P(Z)$ is the summation of $P(Z|M_{match})*P(M_{match})$ and $P(Z|M_{mismatch})*P(M_{mismatch})$.

$$P(M_{\text{match}}|Z) = \frac{P(M_{\text{match}}) * P(Z|M_{\text{match}})}{P(Z)} \tag{4.5}$$

D. Determining VPR Technique for Switching. The posterior probability calculation allows us to predict the level of certainty or confidence with which the technique will correctly match the query image. When this value of probability is lower than the accepted value (0.5) the system attempts to switch to another technique complementary to the current primary technique. The system calculates the probability of complementarity that the primary technique has to the other available VPR techniques.

Once the technique with the highest complementarity is determined, the system switches toward this technique and determines the new posterior probability of matching the query image.

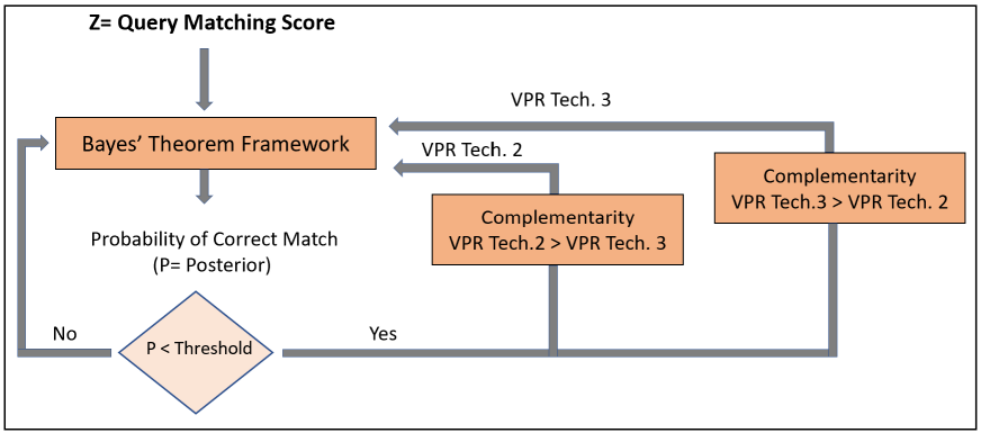

Fig. 4.3. Select and Switch: Adapting to the Most Complementary VPR Technique Below Threshold Probabilities

E. Calculating Probabilities of Complementarity This equation computes the complementarity for the given query image that the primary technique has to the other available VPR methods in the system. We define the terms score event and event Z to clarify their usage in the calculations. A score event refers to a specific range within the total score range of 0 to 1, divided into equal intervals. For example, dividing the range into intervals of 0.1 results in the following score events: [0.0-0.10], [0.10-0.20], ..., [0.90-1.0]. An event Z corresponds to one of these score event ranges. For instance, event Z as [0.30-0.40] represents scores within that interval. These divisions help analyse score distributions and occurrences. Refer to Figure 4.2 for a visual representation of the score range divisions and corresponding score events. Where $P(Z_{\text{Q}} \mid M_{\text{match by A}})$ and $P(Z_{\text{Q}} \mid M_{\text{mismatch by A}})$ are the probabilities of the certain score event for the query image given it is matched or mismatched by technique A. Similarly, $P(Z_{\text{Q}} \mid M_{\text{match by B}})$

and $P(Z_{\text{Q}} \mid M_{\text{mismatch by B}})$ are the probabilities of the certain score event for the query image given it is matched or mismatched by technique *B*.

$$P(CAB) = \frac{P(Z_Q|M_{\text{match by A}}) * P(Z_Q|M_{\text{match by B}})}{P(Z_Q|M_{\text{mismatch by A}}) * P(Z_Q|M_{\text{mismatch by B}})} \tag{4.6}$$

The equation computes the complementarity of *A* with *B* (CAB). This is the complementarity the two techniques have to each other given a certain matching score range, i.e., query image matching score range.

F. The Dynamic Switch. Once a successful loop of switching has taken place from the primary to the selected secondary VPR technique, the same Bayes inference inspired framework is implemented to predict the posterior probability of correct match for the new temporary primary technique. If the probability of correct match produced is above the predetermined threshold, the reference image matched by this technique is considered the final result i.e. the correct match. If however, this too fails to produce a satisfactory probability for correct match, the system switches again to the next best option to observe its results. Given that the probability for match by the third technique is satisfactory, the reference image it matches the query image to will be considered the final result. If not, then in the worst case the system selects the technique with the best probability among the group and considers the result produced by this technique. This, ensures that it exhausts all possible options the system could undergo to correctly match an image and improving overall performance of the system by producing better results than any individual VPR technique from the system could have produced independently.

TABLE 4.1: COMBINATIONS OF VPR TECHNQIUES TESTED ON EACH DATASET FOR SWITCHHIT

VPR Datasets	VPR Technique Combinations		
Corridor	CALC, HoG, NetVLAD	CoHoG, HybridNet, CALC	NetVLAD, AMOSNet, CoHoG
Livingroom	AMOSNet, CoHoG, NetVLAD	AlexNet, NetVLAD, RegionVLAD	CALC, CoHoG, AlexNet
ESSEX3IN1	CALC, CoHoG, HybridNet	CoHoG, NetVLAD, HoG	AlexNet, NetVLAD, RegionVLAD
GardenPoint	NetVLAD, RegionVLAD, CoHoG	AlexNet, NetVLAD, RegionVLAD	CALC, AMOSNet, NetVLAD
Cross-Seasons	AlexNet, NetVLAD, HybridNet	CoHoG, HoG, NetVLAD	CoHoG, HoG, AlexNet
SYNTHIA	CALC, HybridNet, CoHoG	RegionVLAD, NetVLAD, AlexNet	AlexNet, NetVLAD, CoHoG

4.3 SwitchHit Experimental Setup

This section discusses the choice of VPR techniques selected to be employed within the SwitchHit system along with a wide variety of the VPR datasets used to test the performance of SwitchHit. The experimental set up to test SwitchHit was designed to ensure that the maximum number of state-of-the-art VPR techniques are tested along with ensuring the testing is performed on datasets that consist of all major types of variations that can be possibly be encountered. Table 4.1 lists the several combinations of VPR techniques, covering multiple majorly employed methods, that were selected beforehand to test SWitchHit. This again was not a random selection but rather a carefully curated group of combinations based on the knowledge of complementarity provided in Chapter 3. To ensure a comprehensive evaluation, k-fold cross-validation with a fold size of 4 was employed. This approach allowed testing on the entire dataset, ensuring that the final performance reflects the system's behavior across diverse scenarios. Additionally, the deployment in unseen environments and the reliance on a relatively large training dataset is, however, a limitation. Future work can explore strategies like transfer learning to improve adaptability in new environments without requiring extensive training data, further enhancing the system's practicality in diverse real-world settings.

4.4 SwitchHit Results and Analysis

The results gathered and collected for evaluating the SwitchHit performance are presented to show case the overall performance improvement over various datasets. Moreover, to also depict how SwitchHit, merely by making intelligent switches to more optimal techniques, is able to produce improved results that even the highest performing individual VPR technique can achieve as a stand-alone choice. SwitchHit, unlike other ensemble VPR set ups does not employ brute force rather applies a more structured approach to only run one selected method at any given time. The performance is evaluated in terms of total number of correctly matched images for a dataset as well as the PR-curves generated to observe SwitchHit accuracy by contrast to the stand-alone VPR methods.

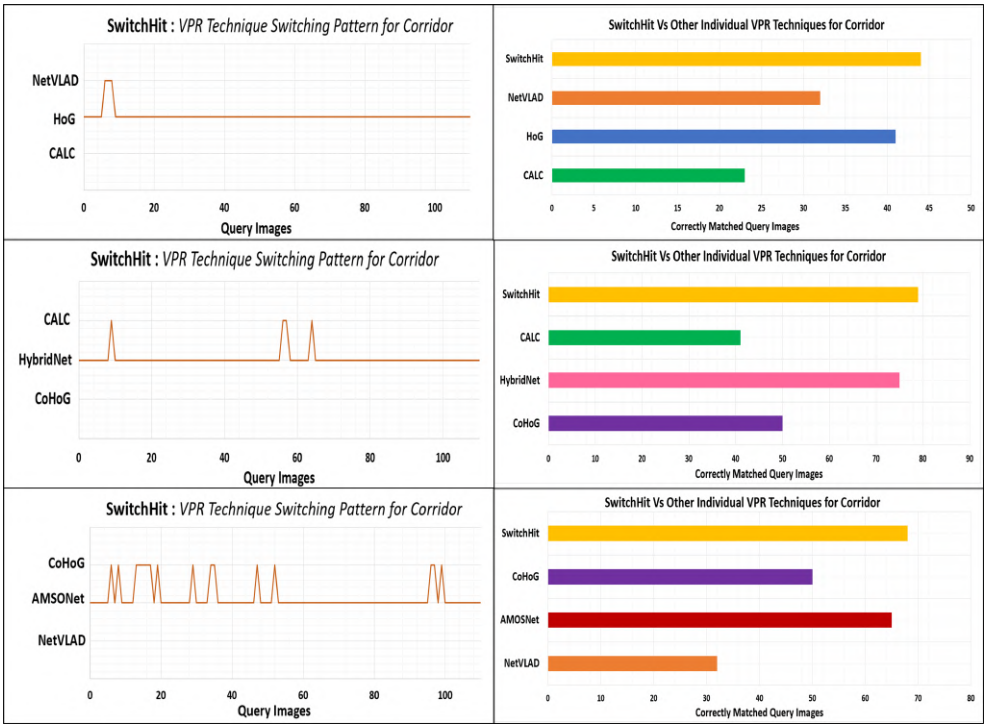

Fig. 4.4. Switching patterns and total Number of correct matches for Corridor dataset.

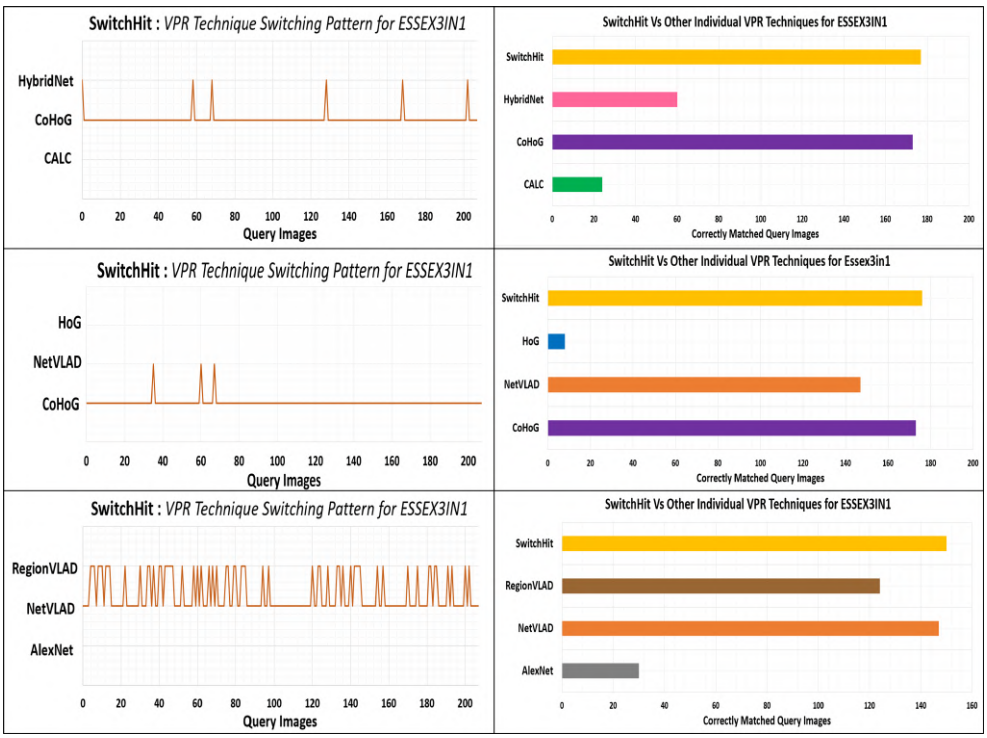

Fig. 4.5. Switching patterns and total Number of correct matches for ESSEX3IN1 dataset.

Total Correct Matches For a VPR Dataset: Measuring the total number of correct matches vs. the number of incorrect matches for all the tested images in a dataset is the most basic method to simply visualize the difference two VPR methods have in terms of performance, for a sample case see the example provided in Figure 4.1. SwitchHit and its performance is presented in this simplistic but informative manner to show how it outperforms other VPR methods including the ones SwitchHit itself employs. The results for this are presented in terms of bar charts for representation. Figures 4.4 to 4.10 illustrate these results. Additionally, x-axis ranges differ due to the large differences in dataset sizes, ensuring clear and accurate representation of each dataset's performance metrics.

The Switching Patterns: These results are unique and specific for a setup like SwitchHit and hence an interesting manner is proposed for representing exactly what is happen-

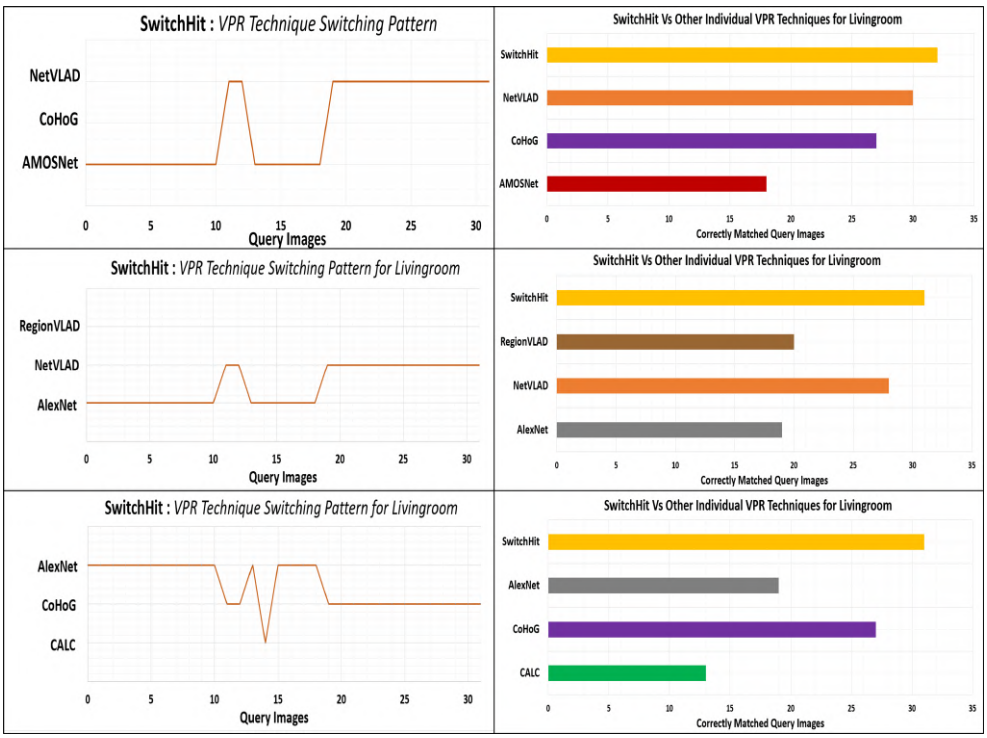

Fig. 4.6. Switching patterns and total Number of correct matches for Livingroom dataset.

ing within a SwitchHit system as it goes through various query images or places and tries to correctly match the image/place. For cases when the system predicts a low probability score of correct match it switches to either one of two options the available choosing the one with the higher complementarity to the currently running VPR technique. This allows for the highest chance that the next chosen method will be able to correctly match the query image. This forms various patterns depicting the switches that were made with a example case provided in Figure 4.1

Precision-Recall and AUC: True-Positives (TP) are the places that a VPR matched correctly, False-Positives (FP) are those erroneously matched, and False-Negatives (FN) are real positive matching places discarded by the VPR technique. For reference, chapter 2 discusses in detail the use of PR-curves and AUC; their significance and methodology is further detail.

4.5 Results and Performance of SwitchHit on Various VPR Datasets

Figures 4.4 to 4.8 present the results in a unique manner, depicting the switching pattern of SwitchHit for various datasets along with the increase in performance in terms of correctly matched images. The results for the Corridor dataset in Figure 4.4 show that all three combinations tested present varied switching patterns, with each combination correctly matching an average of three to four more images than any individual VPR technique. For instance, the combination of CALC, HoG, and NetVLAD demonstrates significant performance improvement.

Similarly, Figure 4.5 illustrates the results for the ESSEX3IN1 dataset where the combination of CALC, CoHoG, and HybridNet outperforms the best standalone VPR technique. SwitchHit correctly matches four to five more images than CoHoG, which has the highest individual performance. Another notable example is the combination of AlexNet, NetVLAD, and RegionVLAD, where SwitchHit mostly shifts between NetVLAD and RegionVLAD, matching three more images correctly than the best individual technique.

The Livingroom dataset results in Figure 4.6 reveal that SwitchHit improves performance by two images while switching between AMOSNet and NetVLAD. The combination of AlexNet, NetVLAD, and RegionVLAD exceeds NetVLAD's performance by matching three more images correctly. Additionally, the combination of CALC, CoHoG, and AlexNet, which are not the best VPR techniques for this dataset, improves performance by four images and matches NetVLAD's performance.

Lastly, Figure 4.7 shows the results for the GardensPoint dataset, where SwitchHit makes successful switches between NetVLAD and RegionVLAD, correctly matching more

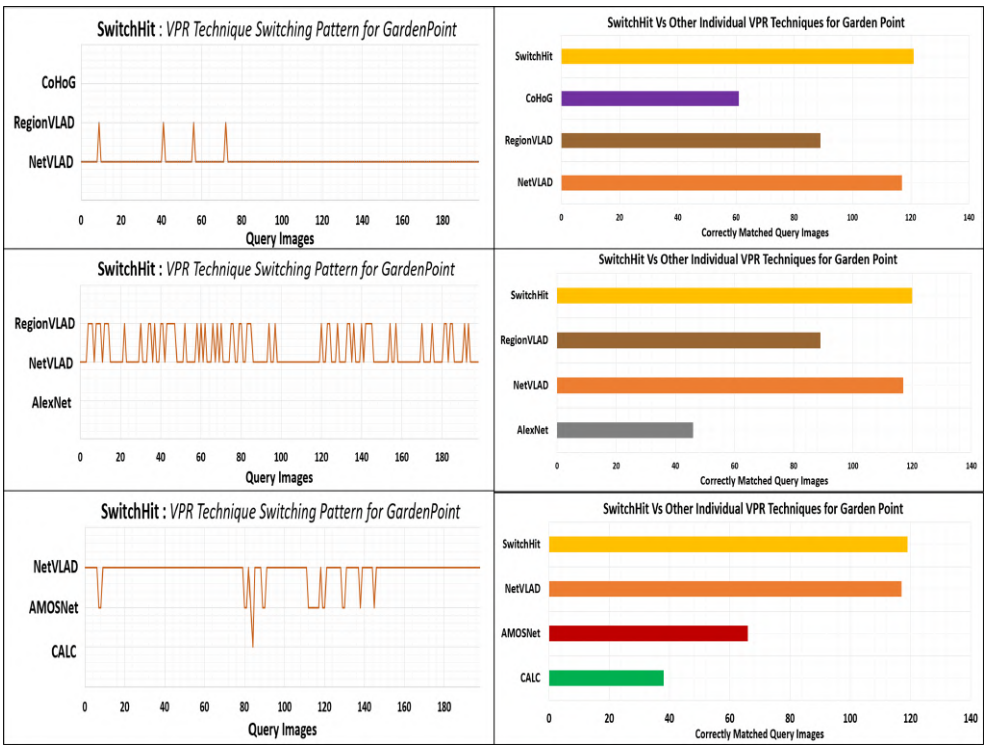

Fig. 4.7. Switching patterns and total Number of correct matches for GardensPoint dataset.

images than the highest-performing individual VPR technique.

The detailed numerical results and switching patterns of the all other datasets tested are provided in Appendix B. These results emphasize the substantial performance improvements achieved by SwitchHit through switching, demonstrating its effectiveness in enhancing VPR accuracy beyond the capabilities of individual techniques.

4.6 SwitchHit Summary

This chapter presents SwitchHit, an innovative and intelligent switching system for VPR techniques based on complementarity. The experiments conducted to test the system show how it selects and employs the best of each VPR technique, even those that are relatively low performing overall. It is built on the empirical data presented in Chapter

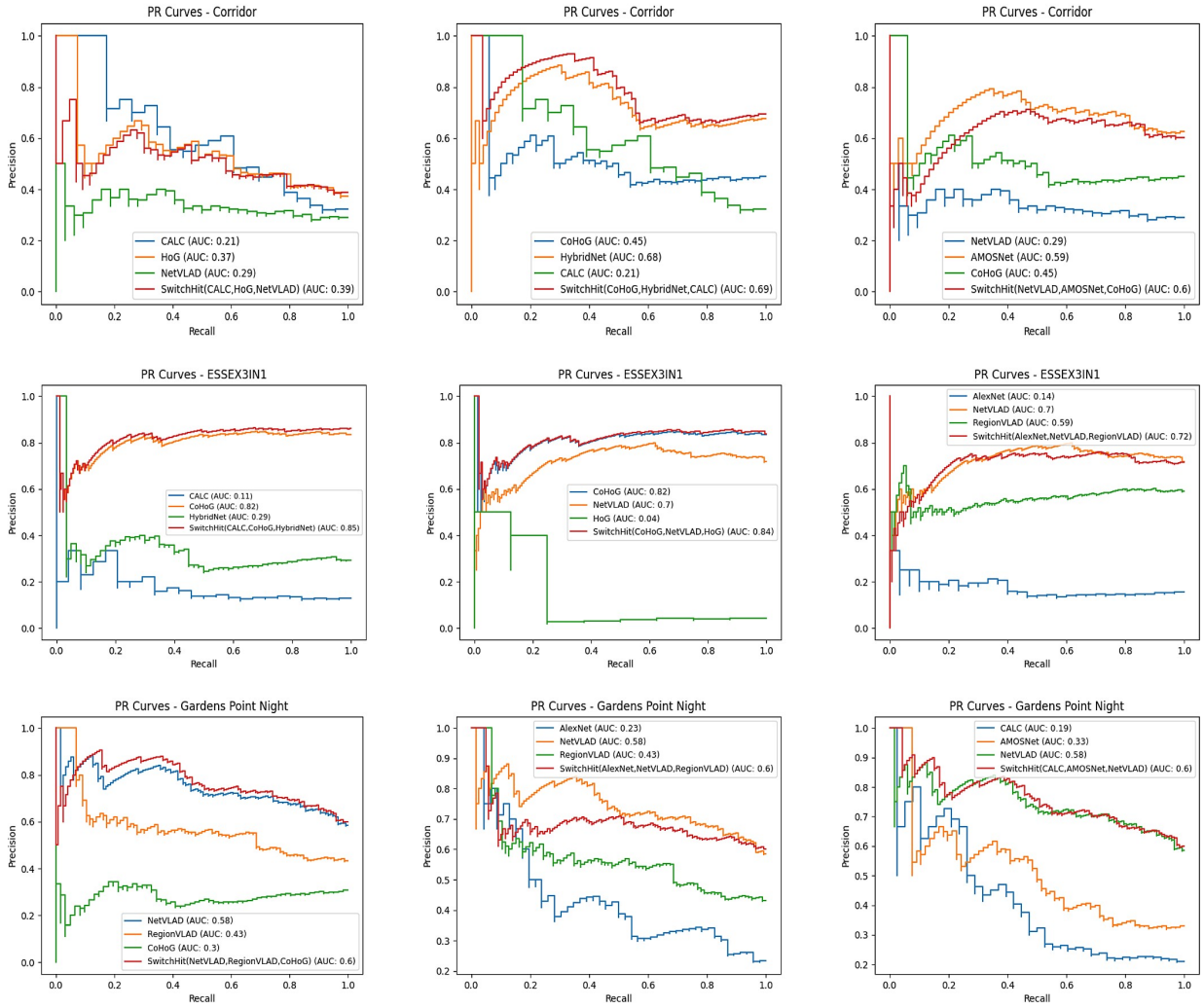

Fig. 4.8. PR curves for Corridor, ESSEX3IN1 and GardensPoint datasets illustrating SwitchHit performance in comparison to all other individual VPR techniques for each data set.

3, which demonstrates that for combining two techniques via fusion or switching, the selection of two high-performing VPR methods is not always the most efficient approach; rather, knowledge of their complementarity is essential to ensure maximum performance improvement and avoid redundancy. This requirement can be fulfilled by two high-performing techniques, two low-performing techniques, or one of each. Consequently, the observed cases in results where an otherwise low-performance VPR technique in-

cluded in SwitchHit significantly boosts performance. The insights and results further reiterate that there are several more ideas to be generated to solve the VPR problem informatively even just using the existing VPR techniques available. Although SwitchHit aims to optimize performance through intelligent switching, it is important to note that the computational and storage requirements vary based on the complexity of the techniques employed. While improvements in computation and storage were not the main focus, avoiding brute force methods provides a theoretical advantage. However, the system's dependency on large datasets to effectively learn the complementarity between different VPR techniques can be a challenge in data-scarce environments. Despite this, SwitchHit offers an interesting approach to VPR, enhancing recognition accuracy and robustness by leveraging the strengths of diverse techniques.

Chapter 5

SwitchFuse[1]

Chapter 4 presents the SwitchHit system, designed to allow efficient Switching among the available VPR methods to boost VPR performance in terms of accuracy. The system was presented as a solution to the approach of using brute force in ensemble VPR methods. However, although SwitchHit is substantial step towards an innovative approach for solving the VPR problem it is not without its limitations. The multi-fusion work presented in [31], [34] which utilised a fusion methodology that runs multiple techniques to then merge the results for each. SwitchHit, on the other, is based on switching between the VPR techniques at the required time to select the VPR method with the highest chance of correctly matching the query image. The limitations however that exist for both these approaches are non-negligible and hence require an interesting solution. First being that the multi-fusion approach, including methods described in [27] and [28], is not the most sophisticated due to its use of brute force, while switching, as seen in the SwitchHit architecture, is restricted to the availability of a suitable VPR technique and its correct

[1]The work presented in this chapter has been accepted and presented at the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS2023),Detroit, Michigan, USA

identification to have a better chance of recognizing the location correctly.

This chapter presents a solution that addresses both limitations by combining two otherwise separate approaches, bringing together their strengths while minimizing their weaknesses. The system presented, SwitchFuse, effectively integrates the robustness of switching VPR techniques based on complementarity with the effectiveness of fusing carefully selected techniques to significantly boost performance. Unlike basic fusion methods that simply combine random techniques, SwitchFuse is designed to first switch and then select the most suitable VPR techniques for fusion based on the query image. By uniting these two processes—switching and fusing—into a hybrid model, SwitchFuse achieves substantial performance improvements across major VPR datasets, as demonstrated through PR curves.

For reference the remaining chapter is organised as follows. Section 5.1 presents the path from of Switch or Fuse to SwitchFuse, Section 5.2 introduces SwitchFuse as a system, design and structure. Section 5.3 presents the experiments conducted to test the SwitchFuse system to evaluate its performance improvement followed by Section 5.4 that discusses and evaluates the results produced via different performance metrics. Lastly, Section 5.5 is dedicated to a conclusive discussion on what SwitchFuse is able to achieve and whether its a system leading towards more intelligent and sophisticated system ideas for VPR improvement.

5.1 Switch or Fuse to SwitchFuse

There are many VPR techniques available today, some with excellent performance. However, there is still no universal VPR system that is robust to all environmental variations.

Leaving us a with a pool of strong VPR techniques but each with its own set of advantages and disadvantages. As discussed in Chapters 3 and 4 the research endeavours have shifted their focus from developing an entirely new VPR technique to utilizing the existing techniques to achieve their maximum potential. Very interesting work in this regards is the concept of multi-fusion systems that introduces the concept of multi-process fusion between different VPR techniques presented by [31], [34]. Chapter 3 provides substantial knowledge for further designing such ensemble VPR systems and provides a guideline to follow to achieve higher efficiency. With this new information available, which is the complementarity of the VPR methods, it opens gates for many more interesting paths to take, SwitchHit (Chapter 4), is an example of such a case. This means for a scenario where an ensemble VPR approach appear to be beneficial due to the lack of a universal method, a choice of whether a switching or fusing ensemble VPR set up needs to be made. Both these approaches with their advantages carry some shortcomings and having to make a choice means baggage of the limitations of either,

SwitchFuse, is based on observing that both fusion and switching methods have the potential to improve performance. Switching takes precedence over fusion in some cases as it merely shifts to the better-performing algorithm and ends up selecting just one rather than running all or multiple techniques at once to combine their results. But simultaneously switching does fail at times when all available techniques have a probability of a proposed match that is below threshold, a major shortcoming faced by SwitchHit. In such a case SwitchHit selects the technique with the highest probability but not necessarily the best option. However, introducing fusion to the scenario mitigates this problem by enhancing the ability to fuse a few selected VPR techniques with a high probability of matching correctly.

5.2 Methodology

The SwitchFuse system has a tripartite model containing three units designed to incorporate a variety of VPR techniques, categorized on the basis of their performance for different types of variations. However, the system is not restricted to a variation categorization and can be employed for any other type of classification of different VPR techniques. The switching component allows for the selection of the best suited techniques for fusion to ensure improved performance. Such an approach saves us from having to use brute force and fuse all or some random techniques together, rather it intelligently selects from the pool of techniques in different units using a Bayes theorem inspired framework and fuses the final selected and most suitable techniques only.

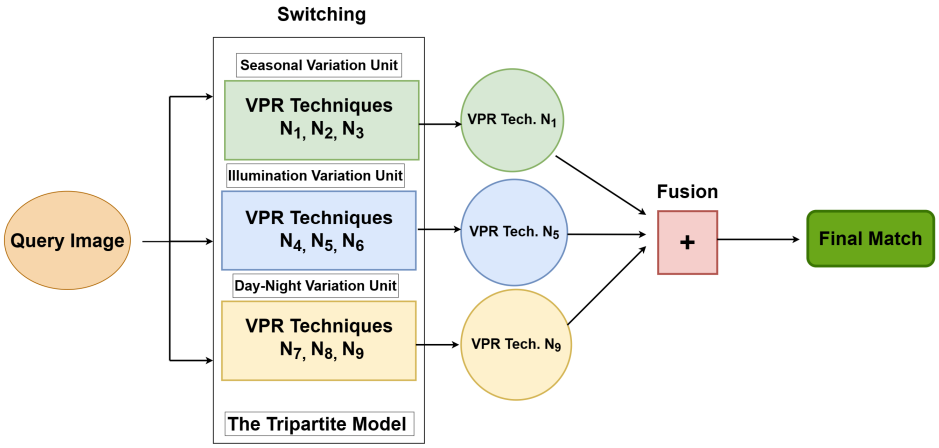

Fig. 5.1. The SwitchFuse System, a tripartite model, selects and fuses the best VPR techniques to enhance match accuracy.

Figure 5.1 illustrates the SwitchFuse System, a tripartite model consisting of three well-curated tiers based on complementarity and environmental variations. Each unit of the model determines one VPR technique with the highest probability of correctly matching the query image. These determined VPR techniques then undergo fusion, in-

volving their feature vectors to calculate combined similarity vectors for determining the proposed match, thus improving performance. The three units are categorized based on their performance on various major types of environmental variations, as illustrated in Figure 5.1. The query image is input to all three units of the system, where the probability of proposed is calculated by the primary technique in each unit. Switching is conducted to select an alternate technique when required.

As illustrated in Figure 5.2, each unit of the tripartite model selects the best VPR technique available by calculating the probability of proposed match and determining the VPR technique with the highest probability as the final selected VPR Technique. Finally, a single technique is selected by each unit, so all three units select one VPR technique, each of which has the highest likelihood of correctly matching the query image. These selected techniques then undergo fusion, where the next step is to add the normalized distance vectors produced by each of these techniques, as displayed in Figure 5.3, ensuring a significant enhancement in performance. To provide a clearer understanding, the following pseudo-code outlines the algorithm used in the tripartite model:

Algorithm 1 Tripartite Model Algorithm for VPR Technique Selection

Input: Query Image Q, VPR Techniques $N = \{N_1, N_2, ..., N_n\}$
Output: Selected VPR Techniques $\{N_1, N_2, N_3\}$
for each Unit U_i in {Unit1, Unit2, Unit3} **do**
    Set $n = 1$
    while $n \leq N$ **do**
        Calculate P_n (Posterior probability for VPR technique N_n given Q)
        if $P_n > 0.5$ **then**
            Select N_n as the best VPR technique for U_i
            break
        else if $n < N$ **then**
            $n = n + 1$
        else
            Select $N_{(P)}$ (VPR technique with the highest posterior probability)
        end if
    end while
end for
Fuse the selected VPR Techniques $\{N_1, N_2, N_3\}$ to determine the final match

The system can be primarily divided into two main steps starting from performing switching for each unit of the model and then the selected VPR techniques undergo fusion to determine the proposed match for the query image.

5.2.1 Switching

A. Input Query Image to Each Component of the Tripartite Model: The first step is the query image provided as an input to each of the three parts of the system. The query image then undergoes several steps and calculations for the final selection of a best suited VPR technique from each component. Let*X* be the set of query images in a data set and *Z* be the set of components of the tripartite model performing switching.

$$X = \{Q_1, Q_2, Q_3, \ldots . Q_n\} \tag{5.1}$$

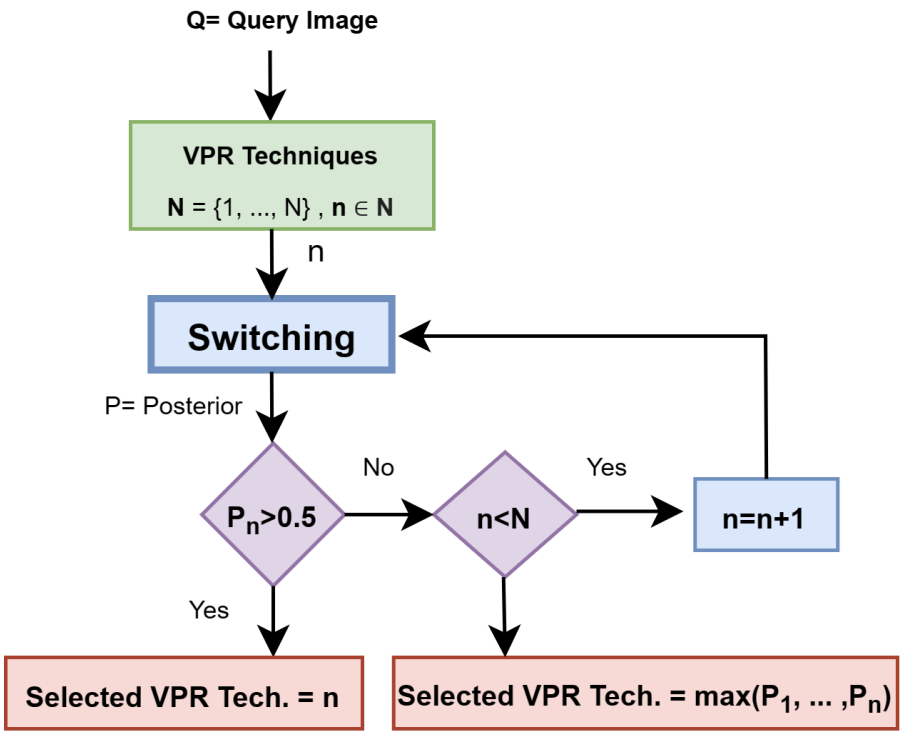

Fig. 5.2. One unit of the tripartite model calculates and selects the VPR technique with the highest match probability.

$$Z = \{A, B, C\} \tag{5.2}$$

B. Applying Switching to Each Unit to Determine A Single-Suited VPR Technique: This step is divided into further sub-steps that help determine a single technique as an output by each unit of the system. All of these sub-steps are performed individually for each component.

Computing Probability that Query Image will be Correctly Matched (Posterior): This equation, based on Bayes' theorem, computes the posterior probability of the VPR technique correctly matching the image given the input query matching score. Here,

$P(M_{\text{match}})$ represents the probability of a match by the primary technique overall, which is the prior probability of match. This prior probability can be estimated from the training data by calculating the ratio of correctly matched instances to the total number of instances. $P(Z|M_{\text{match}})$ is the likelihood that the VPR technique will correctly match given a certain matching score. This likelihood can be derived from the performance characteristics of the VPR technique observed during validation. This produces an updated but non-normalized probability distribution between the matching and mismatching. Finally, *P(Z)* which is the marginalization in the equation is the summation of both updated non-normalized distribution of match and mismatch i.e.. *P(Z)* is the summation of $P(Z|M)*P(M_{match})$ and $P(Z|M_{mismatch})*P(M_{mismatch})$. Finally, *P(M|Z)* which is the calculated posterior, is used to determine the probability of correctly matching the given image.

$$P(M_{\text{match}}|Z) = \frac{P(M_{\text{match}}) * P(Z|M_{\text{match}})}{P(Z)} \tag{5.3}$$

Determining VPR Technique for Switching: The posterior probability calculation allows us to predict the level of certainty or confidence with which the technique will correctly match the query image. While in case this value of probability is lower than the accepted value (0.5) the system attempts to switch to another technique complementary to the current primary technique. The system calculates the probability of complementarity that the primary technique has to the other available VPR techniques. Once the technique with the highest complementarity is determined the system switches towards this technique and determines the new posterior probability of matching the query image.

Calculating Probabilities of Complementarity: This equation computes the comple-

mentarity for the given query image that the primary technique has to the other available VPR methods in the system. Where $P(Z_Q|M_{match\ by\ A})$ and $P(Z_Q|M_{mismatch\ by\ A})$ is the probability of the certain score for query image given that it is matched or mismatched by technique *A*. Similarly $P(Z_Q|M_{match\ by\ B})$ and is the probability of the certain score event for query image given its matched or mismatched by technique *B*. The equation computes the complementarity of *A* with *B (CAB)* i.e. the complementarity the two techniques, have to each other given a certain matching score i.e. query image matching score. Finally, the system switches to the technique with the higher *P(CAB)*.

$$P(CAB) = \frac{P(Z_Q|M_A) * P(Z_Q|M_B)}{P(Z_Q|M_{\text{mismatch}\,A}) * P(Z_Q|M_{\text{mismatch}\,B})} \tag{5.4}$$

The Dynamic Switching: The posteriors for each technique are constantly checked against the threshold probability of above 0.5 to proceed. If the probability of match is below threshold the system will switch to another technique and perform the same steps to determine the probability of match until a suitable technique with satisfactory probability of match is found. In case no such technique can be found the system selects whichever technique has the highest probability of match. This selected technique is considered the output or prime selected technique of a single unit.

The Chosen VPR Techniques: The output is a set of three selected VPR techniques, one selected by each unit from the tripartite model. Let *N* be the set of selected techniques for any given query image where $n \in N$ and *n* represents an individual selected VPR technique.

$$N = \{n_1, n_2, n_3\} \tag{5.5}$$

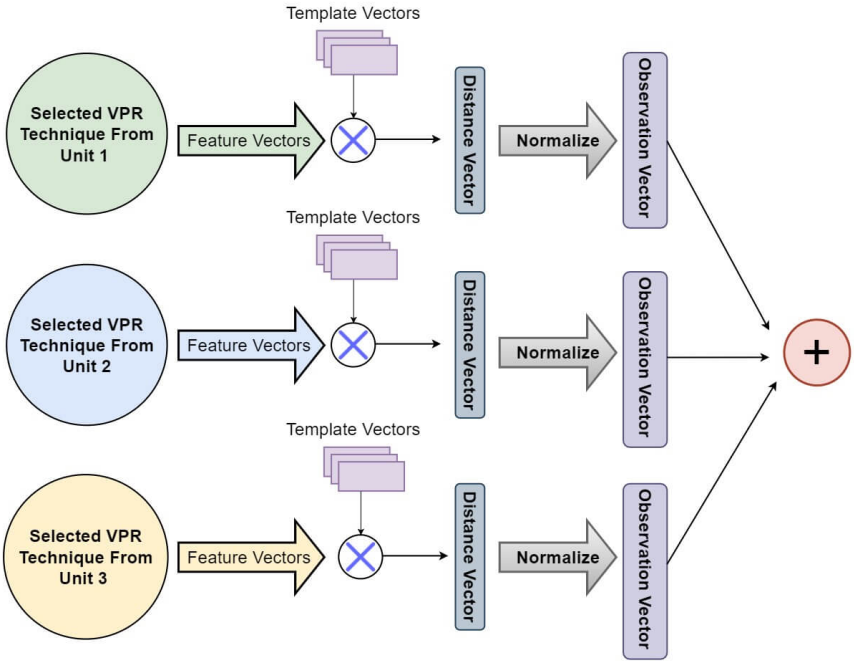

Fig. 5.3. The fusion step normalizes and sums distance vectors from selected VPR techniques to enhance matching accuracy.

5.2.2 Fusion

As illustrated in Figure 5.3 the fusion step incorporates how each selected VPR technique produces a distance vector with the distance scores between the query image and reference images in the database. After normalizing these distance vectors before fusion, which is performed by taking the summation of these normalized/observation vectors, the selected VPR techniques are combined to provide the results. As discussed above fusion is the last and final step of the system. Generally, to perform VPR each query image is compared against a database of prior images mainly by using different feature extractions and image matching techniques. This process results in a D dimensional similarity vector, which is a list of similarity scores between the query and reference images in the database. The similarity vector can be interpreted as the bigger the similarity score the

stronger the chances of a correct match.

In the fusion step, the selected VPR techniques with the highest probability of correctly matching the image are used simultaneously to fuse their similarity scores. In the currently tested system the number of selected techniques is limited to three but depending on computational resources and a different pool of techniques the number can be increased or decreased. Each *n* produces a similarity vector with the similarity scores between the query image and reference images in the database. Furthermore, as the set of techniques is arbitrary hence the distribution of these similarity scores within each technique may not be consistent with the distribution of other techniques. So it is important to normalize prior to fusing the set of *N* to maximize the likelihood that each similarity vector has a minimum and maximum value of -0.001 and 0.999 respectively where $\epsilon = 0.001$. (In case of normalized values falling under threshold a value of 0.001 is forcefully assigned.) This equation is inspired by the methodology presented in [34].

$$\hat{D}_n(i) = \frac{\hat{D}_n(i) - min(D_n)}{max(D_n) - min(D_n)} - \epsilon, \forall_i, n \in N \tag{5.6}$$

The final step is to produce the combined similarity vector for fusion that is the D_F. The matched image is the one with the maximum D_F score.

$$D_F = \sum_{n=1}^{N} \hat{D}_n \tag{5.7}$$

5.3 SwitchFuse Experimental Setup

The SwitchFuse system allows for the selection of the most suitable VPR techniques for fusion given any query image, over different VPR data sets. TABLE 5.1 lists all the

TABLE 5.1: VPR-BENCH DATASETS TESTED FOR SWITCHFUSE

Dataset	Conditional-Variation
GardensPoint	Day-Night
ESSEX3IN1	Illumination
Cross-Seasons	Dawn-Dusk
Nordland	Seasonal
Corridor	None
Living-room	Day-Night

data sets along with their variations types including Corridor [11], Living room [12], ESSEX3IN1 [5], GardensPoint [3], Cross-Seasons [4], and Nordland [2]. A wide variety of datasets are selected to maximize the likelihood that each type of conditional variation is tested for the system.

TABLE 5.2: VPR TECHNIQUES EMPLOYED IN EACH CONDITIONAL VARIATION UNIT OF THE SWITCHFUSE SYSTEM

Conditional Variations		
Seasonal	**Illumination**	**Day-Night**
AlexNet	HybridNet	NetVLAD
AMOSNet	CoHOG	RegionVLAD
HOG	CALC	HybridNet

Table 5.2 presents the structure in which VPR techniques have been employed in the proposed SwitchFuse system for the purpose of this thesis. Each unit of different conditional variation is provided with three techniques that are theoretically known to be complementary pairs for the respective variation type in Chapter 3. The first unit consists of AMOSNet [127], HOG [58],

and AlexNet [74]. The second unit employs CoHOG [130], HybridNet [127] and CALC [74]. Finally, the last unit consists of HybridNet [128], NetVLAD [129] and RegionVLAD [63]. This combination ensures the inclusion of complementary pairs and all techniques widely used for experiments. The implementation and experimental details

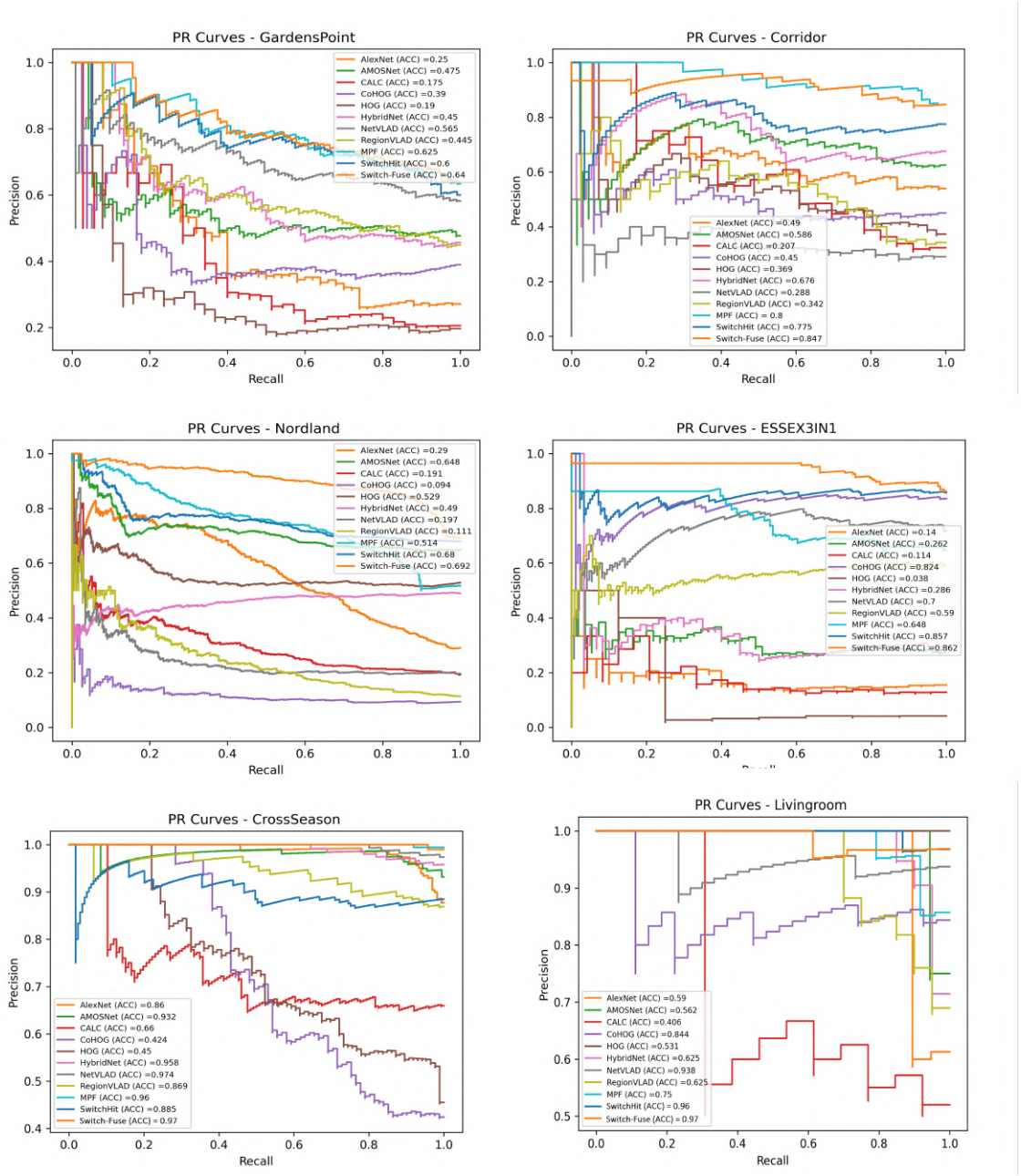

Fig. 5.4. Precision-Recall curves showcasing performance of SwitchFuse in comparison to SwitchHit, MPF and other VPR methods on GardensPoint, Corridor, Nordland, Cross-Season, ESSEX3IN1 and Livingroom.

of all VPR techniques, including the dataset splitting strategy, follow the same approach as outlined in Chapter 4, ensuring comprehensive evaluation across the entire dataset.

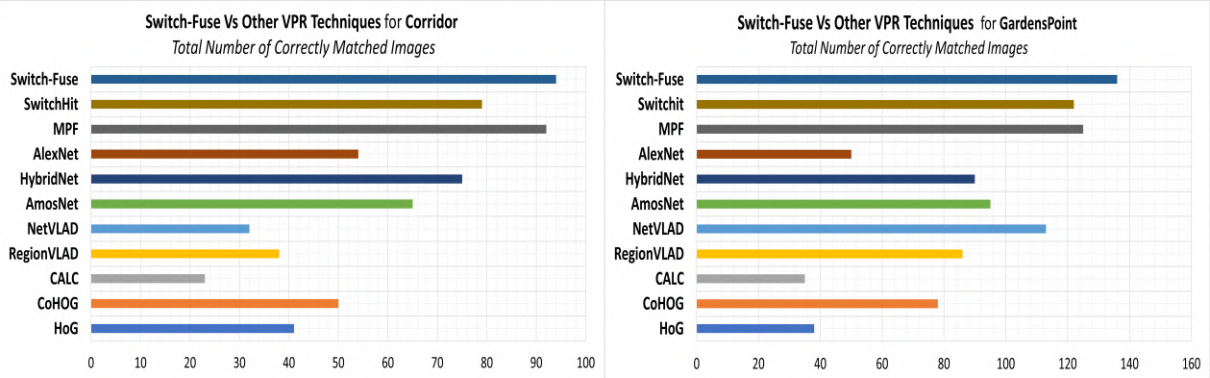

Fig. 5.5. Performance improvement in terms of correctly matched images by SwitchFuse in comparison to SwitchHit, MPF and other VPR methods on GardensPoint and Corridor dataset.

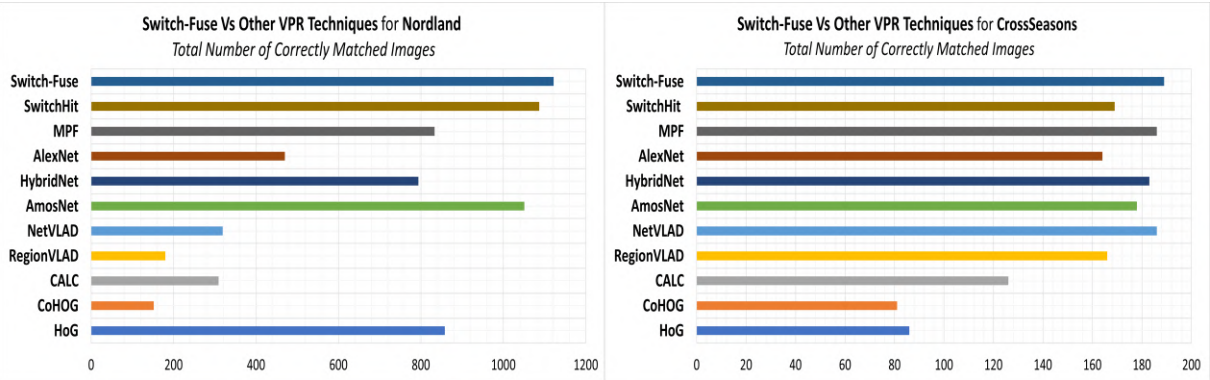

Fig. 5.6. Performance improvement in terms of correctly matched images by SwitchFuse in comparison to SwitchHit, MPF and other VPR methods on Nordland and CrossSeasons dataset.

5.4 SwitchFuse Results and Analysis

The SwitchFuse System like the testing of other VPR techniques considered in this thesis, multi-fusion and the SwitchHit system has been tested using the same widely employed datasets to ensure standardization with level and type of variations encountered. Further the results are assessed using the same evaluation metrics are employed to draw accurate comparison among all these systems and observe the level of difference in performance. The commonly employed evaluation metrics, firstly on an image-by-image basis to show the increase in total correct matches and then the accuracy and PR-curves as described

in Chapter 3. The improvement produced in the results via SwitchFuse is due to the accurate prediction and selection of the best suited VPR techniques and then their fusion, as further explained via different examples.

Figure 5.4 compares the SwitchFuse system to the SwitchHit system [131], Multi-Process Fusion (fusing three best performing VPR techniques overall) [31] and as well as other VPR techniques to showcase the difference in performance using PR curves. The results and improvements vary over different data sets due to their extremely varying environments and sizes. For example testing out the GardensPoint an accuracy of 0.64 is observed over the data set and this is a significant improvement over SwitchHit, MPF or any other VPR techniques in comparison. Similarly, for the Corridor data set SwitchFuse produces an overall accuracy of 0.84 which again is not only higher in comparison to all single VPR techniques but both SwitchHit and MPF as well. SwitchFuse is able to achieve an accuracy of almost 0.7 for Nordland data set which again is significantly higher than both SwitchHit and MPF. It is also a good example to observe how in some cases SwitchHit outperforms MPF and vice versa. Empirical data for different techniques that should be suitable together for fusion is not always true for all data sets or query images and the SwitchFuse system, as evident by the results, helps predict VPR techniques which are actually useful to be fused. Similar results for other data sets including CrossSeasons, ESSEX3IN1 and Livingroom are presented depicting the higher accuracy SwitchFuse was able to achieve in comparison to SwitchHit or MPF on these data sets. After testing SwitchFuse for a series of varied data sets it can be concluded that the system is in fact able to boost accuracy performance over different environmental variations by performing informed switching, and then fusing these selected techniques only.

Figure 5.5 to 5.7 depicts a comparative analysis between overall performance on each

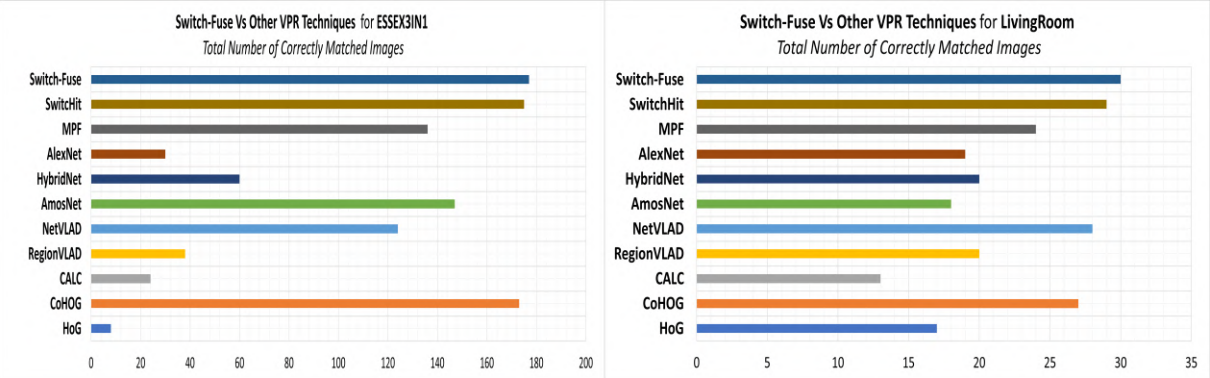

Fig. 5.7. Performance improvement in terms of correctly matched images by SwitchFuse in comparison to SwitchHit, MPF and other VPR methods on ESSEX3IN1 and Livingroom dataset.

data set tested on the basis of total increase in number of correctly matched images. It is a simple way to observe over an image-by-image basis how the overall accuracy of the system is better than other systems in comparison. Figure 5.5 presents the GardensPoint and Corridor data set where an improvement of 15 or more images than SwitchHit [131], MPF [31] and even significantly more images than other individual VPR techniques can be observed. Figure 5.6, the SwitchFuse system has around 60 more correctly matched images for the Nordland data set than any other option available throughout experimentation. While for the CrossSeason in Figure 5.6 and ESSEX3IN1 data set in Figure 5.7 an improvement of 3 to 4 images can be observed which is still higher than even the best performing VPR technique, MPF or SwitchHit. ESSEX3IN1 is one of the examples to show the capability of SwitchFuse where MPF has lower performance than a single technique (CoHOG) which is trained specifically for the data set, while SwitchHit has very minor improvement but together they outperform any options available including CoHOG. A pattern of significant improvement in performance by the SwitchFuse system helps conclude that the hybrid model of SwitchFuse allows for an intelligent switching to the most suitable VPR techniques to be fused and the fusion methodology further boosts

performance.

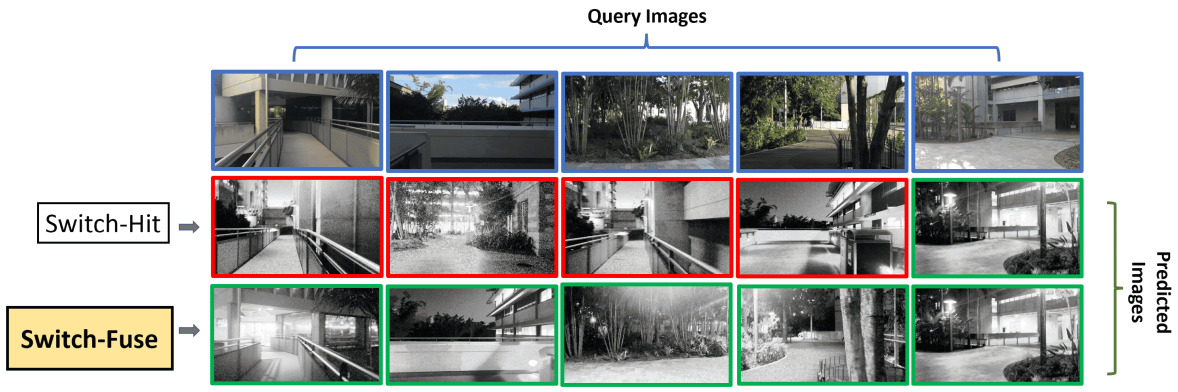

Fig. 5.8. Example of the SwitchFuse System's performance in various scenarios. These examples were specifically chosen to illustrate the system's strengths in correctly matching query images under different environmental conditions.

Figure 5.8 and Figure 5.9 are actual representations of examples taken from the experiment to showcase how the SwitchFuse system performs. Figure 5.7 explains this over the GardensPoint data set to show the different VPR techniques selected, on each query image, to be fused. It is important to mention that although mostly a combination of the same techniques can be observed over a data set this does change, more in some cases than the others. Although it is possible for each query image to have its own combination of VPR techniques for fusion, a certain level of uniformity can be observed over the data set with mostly the same combination being selected. Furthermore, Figure 5.9 gives an example illustrating a case where none of the three individual VPR techniques are able to correctly match the query but the selection of the three specific techniques and their fusion results in a successful match. Many similar cases can be observed overall on multiple queries that result in the performance improvement seen using SwitchFuse.

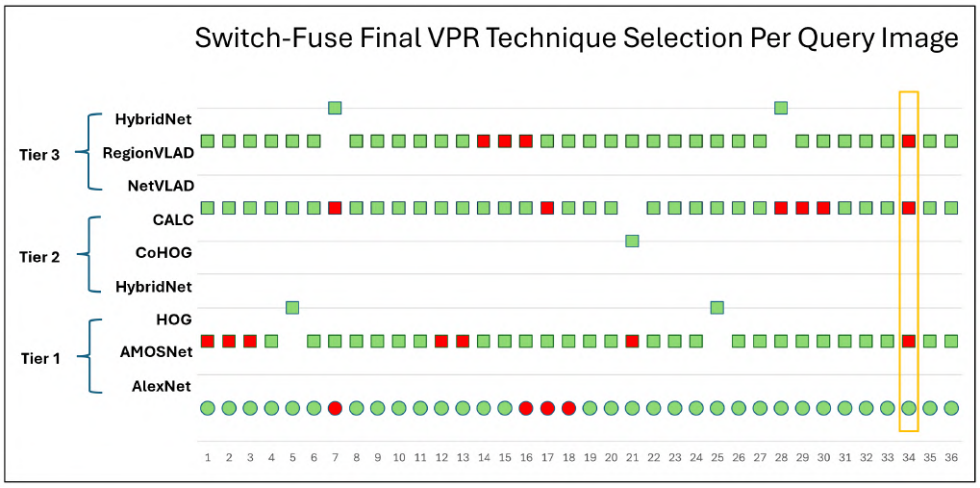

Fig. 5.9. Displays SwitchFuse's final selections from an example of GardensPoint dataset: green and red blocks for individual matches or mismatches, circles for SwitchFuse outcomes, and a yellow window showing a successful match where individually all techniques failed yet SwitchFuse is successful.

5.5 SwitchFuse Summary

This chapter presents SwitchFuse, a hybrid system designed to incorporate the properties of both dynamic switching and fusion systems. The proposed method not only attempts to overcome the shortcomings that both systems possess individually but forms an amalgamation of the strengths of these two otherwise discrete approaches. SwitchFuse combines the adaptability of switching with the robustness of fusion, efficiently selecting and combining VPR techniques based on their complementarity and environmental variations. The results demonstrate a significant improvement as evident by the increase in overall performance accuracy, outperforming several other methods like SwitchHit and multi-process fusion. SwitchFuse is an example showcasing the wide possibilities of attempting innovative methods to solve the VPR problem and has been tested on a certain number of datasets. In the larger scope, SwitchFuse holds potential for further

deeper study into its computational and storage efficiency. While this research primarily focused on performance improvements in terms of accuracy through efficient switching and fusion, the avoidance of brute force methods suggests potential advantages that merit further exploration in future work exploring improvement in terms of computational improvement. An important observation to be made is that the basis of switching and selection of different VPR techniques in this work is complementarity, bringing us full circle back to Chapter 3 and the utility of the knowledge on VPR complementarity. However, this also points in other directions such as whether a different metric other than complementarity could help aid in the selection process for developing ensemble methods; or if other than switching, fusing, or SwitchFuse, there are other innovative approaches yet to be explored in terms of ensemble VPR setups.

Chapter 6

Universal Voting Schemes for Improved VPR Performance[1]

The previous three chapters predominantly focus on gathering and utilizing complementarity knowledge and exploring different ideas for ensemble VPR setups. However, in concluding Chapter 5, it is important to discuss how complementarity or specific methods towards building ensemble setups, such as switching or fusion, are not the only components of ensemble VPR setups that remain unexplored. Other factors also influence the performance of any VPR setup consisting of multiple VPR techniques. One such factor that frequently appears in research involving multiple VPR methods is voting. Voting is a major aspect common to many strategies involving multiple techniques, as highlighted in recent studies [96, 132]. This makes it an extremely relevant topic to explore in terms of its application and significance for any ensemble VPR setup.

Drawing inspiration from various voting schemes widely employed in other fields

[1]The work presented in this chapter has been accepted and presented at 2023 International Conference on Robotics and Automation Workshop (ICRA 2023), London, UK, 2023.

such as politics and sociology, this chapter explores the impact of different voting methods on VPR performance in terms of accuracy. The idea for this chapter stems from the observation that different voting methods can lead to varying outcomes for the same data. This phenomenon has been extensively researched in other fields, and each voting scheme is utilised for specific cases in different academic contexts. The following sections will delve into various voting schemes and their potential applications in enhancing the robustness and accuracy of ensemble VPR systems.

This chapter hence analyses several universal voting schemes to test if this observation stands true for VPR tasks involving voting as well. And if so, it would be worthwhile to maximise the place detection accuracy of a VPR ensemble set up and determine the optimal voting schemes for selection. In this chapter, a wide variety of voting schemes are tested to demonstrate the improvement in VPR results for several datasets. Furthermore, it is established whether a single optimal voting scheme exists or, as observed in other fields of research, the selection of a voting technique is relative to its application and environment. The chapter additionally presents these different voting methods to determine the best or worst or satisfactory cases in order to make informed decisions while selecting a voting mechanism. The results collected are presented in this chapters with the help of several illustrations, such as depicting the performance bounds of a voting mechanism in terms of radar charts, PR curves to showcase the difference in performance and finally a comparison methodology using a McNemar-like test variant to determine the statistical significance of the differences. This test is performed to further confirm the reliability of outcomes and reiterate the significance of using a certain voting method over another. Finally comparisons are drawn for better and informed selection of a voting technique.

The reminder this chapter follows this organizational set up, beginning from Section 6.1 providing an introduction to the current voting practises and knowledge. Section 6.2 presents the several universal voting schemes selected for testing and their methodologies explaining their employment in terms of VPR. Section 6.3 describes the experimental setup designed to test each of the voting mechanism. The results based on testing these different voting schemes are presented in Section 6.4 and the final conclusions and insights entering voting world for VPR are given in section 6.5.

6.1 Introduction to Universal Voting Schemes

With many excellent VPR techniques available to the robotics community to tackle the VPR task, a problem that remains is lack of a universal VPR technique that performs equally well in all types of variations encountered. Chapter 1 to 5 discuss this problem in detail stating the several methods existing to work around this problem and tackle this issue using innovative approaches introducing different ensemble setups. From taking inspiration from Multi-sensor use in robotics to Multi-fusion of VPR techniques and discovering complementarity in VPR methods to switching among different methods based on their level of complementarity, overall several endeavours have been made in the right direction. However, the journey towards any one ensemble VPR set up is only beginning as several other factors are yet to studied and explored. One such factor commonly recurring among many of the so far designed ensemble set ups is some sort of voting that occurs at different stages in different systems proposed. [96]

Innovative solutions such as the concept of multi-process fusion between different VPR techniques [31], [34], and SwitchHit [131] highlight the capabilities of various

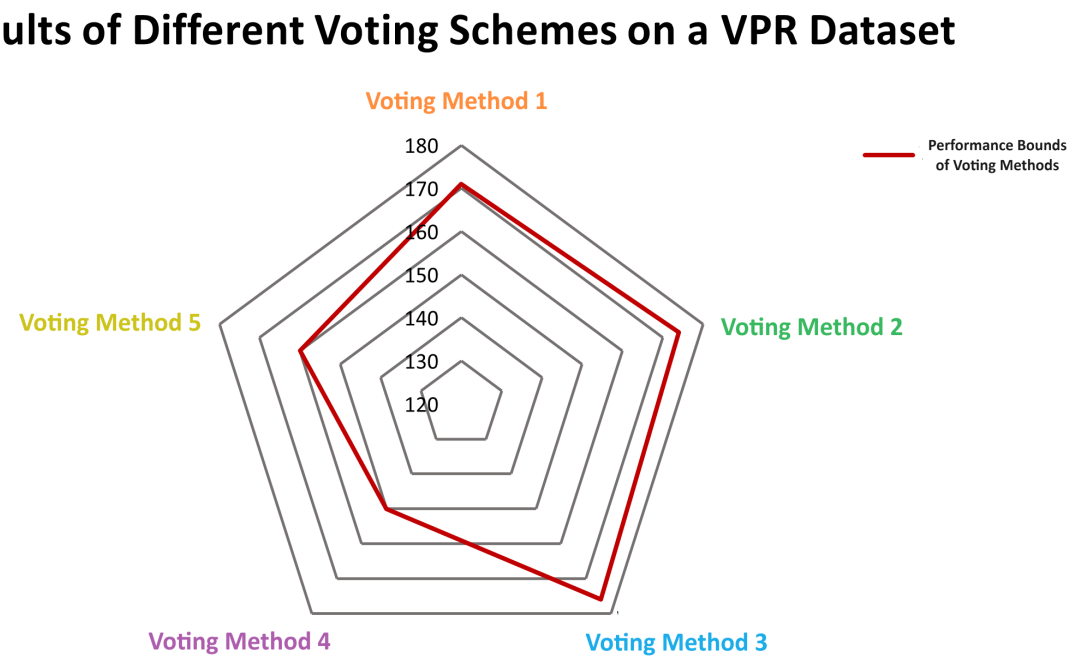

Fig. 6.1. Sample radar chart from the experimental setup shows performance bounds of voting methods; a red line closer to the boundary indicates better performance.

VPR techniques by switching to a more suitable option. Further interesting work presented by [96] involves gathering a collection of very small CNN voting units to enhance VPR performance in terms of accuracy. Additionally, [132] discusses the idea of probabilistic voting for VPR using the nearest neighbor descriptors. These approaches, whether conventional or unconventional, incorporate voting mechanisms at different stages to improve the overall performance of VPR systems.

Exploring work conducted for voting in other fields such as politics and sociology [133, 134], it is evident many researchers have attempted to create a uniform and standardized voting system that can be considered fair or optimal in any scenario. For instance, research in political science has examined various voting methods like the Borda count and their impact on electoral outcomes [134]. Moreover, even within the VPR field, there are examples of comparing and contrasting voting schemes such as in the work presented

in [135]. This has elevated the choice of voting selection method in other fields and VPR, from being a random or trivial task to a well-thought and curated decision. However, the exploration of different voting schemes to determine which is the optimal approach to use in an ensemble VPR setup is an area that has not yet been thoroughly explored and can help produce some interesting insights.While voting is an essential component of ensemble setups, other fusion approaches such as feature mixing, random Gaussian projections, and hyperdimensional computing are also relevant. Feature mixing integrates features at various levels, enhancing overall performance [136]. Random Gaussian projections reduce the dimensionality of feature vectors, making fusion more efficient [137]. Hyperdimensional computing uses high-dimensional binary vectors, offering robustness to noise and efficient handling of large-scale data [120]. These methods provide alternative ways to enhance VPR accuracy. However, the exploration of different voting schemes to determine the optimal approach for an ensemble VPR setup remains an area that has not been thoroughly explored and can yield interesting insights. This chapter presents and evaluates the applications of universally employed voting schemes on a standard VPR ensemble setup with multiple VPR techniques to observe the difference in performance and results across a varied set of VPR datasets. This setup involves running several VPR methods in parallel to generate the top matches for a given query image, which are then combined using various voting schemes to determine the final match. As illustrated in Figure 6.2, the standard VPR ensemble setup includes multiple VPR techniques (VPR Technique A, B, C, and D) that each produce top-matched reference images. These matches are subjected to different voting schemes to select the final reference image, combining the strengths of individual techniques to enhance overall performance.

In addition it could show how the common practice of selecting the basic and com-

monly used type of voting might not always be the wise decision to take in terms of the VPR task. Figure 6.1 shows an example of a sample output from the proposed experimental setup to evaluate the difference in performance bounds between each voting methodology. The red line in the radar chart is representative of the performance bounds (images correctly matched) and the axis represent the total number of query images in the data set. The various dimensions are useful to interpret the difference in the performance of a voting scheme, in comparison to the other voting schemes, such that the closer the red line is to the boundary the better the performance,

Voting as a concept when employed universally offers a pool of options each with its own set of individual characteristics. Furthermore, each voting scheme due to its unique methodology produces different results hence it is correct to assume that selection of a voting scheme in any field is not a trivial task. Researchers in different fields have made the effort to tackle the lack of standardization among voting schemes and determine the optimal voting schemes for different types of tasks. However, an attempt to employ universal voting methods or an exploration to determine which of the voting schemes is optimal to use in an ensemble VPR set up is an area that has not yet been attempted.

This chapter presents and evaluates the applications of universally employed voting schemes on a standard VPR ensemble setup with multiple VPR techniques to observe the difference in performance and results across a varied set of VPR datasets. This setup involves running several VPR methods in parallel to generate the top matches for a given query image, which are then combined using various voting schemes to determine the final match. As illustrated in Figure 6.2, the standard VPR ensemble setup includes multiple VPR techniques (VPR Technique A, B, C, and D) that each produce top-matched reference images. These matches are subjected to different voting schemes to select

the final reference image, combining the strengths of individual techniques to enhance overall performance.

In previous chapters, it has been argued that a fusion of all available VPR techniques is undesirable due to inefficiencies and the potential for diminished performance. Instead, more efficient methods have been showcased, such as using complementarity for switching or the Switch-Fuse system. However, for the sake of this experiment, it was necessary to maintain a standard ensemble setup to fairly compare the performance differences caused by different voting methods. This approach controls for variables other than the voting method, ensuring that the observed differences in performance can be attributed solely to the voting methodologies employed.

A variety of voting systems with unique characteristics, widely employed in other fields such as politics and sociology, have been selected for testing. The goal is to determine whether a single best voting method exists among these schemes or if, much like in other fields, the optimal voting method is case-specific. The results are presented in terms of performance bounds of voting schemes on different datasets, precision-recall (PR) curves for comparing accuracy, and statistical significance of performance differences via McNemar-like test's test and Z-scores values.

Figure 6.2 illustrates this ensemble setup in detail. The ensemble consists of multiple VPR techniques (denoted as VPR Technique A, B, C, and D), each producing a set of top matched reference images. These images are then subjected to different voting schemes to observe the differences in results and determine the final selected reference image, which is the most likely correct match for the query image.

6.2 Universal Voting Schemes Methodology

This section presents the different universally employed voting schemes and describes their methodologies in detail to understand their employment for VPR setup. These voting methods have been employed in a series of different VPR data sets to observe how the use of different voting mechanism effects overall results in a basic ensemble VPR system. These voting schemes are tested in a VPR set up that is simultaneously employing all state-of-the-art VPR techniques available and the final step involves selecting the correct reference image by using a voting methodology. The total number of VPR techniques for this experiment is the *voters* while all reference images are the *candidates* and the selected top reference image/images by each VPR technique represents their *votes*.

The structure of this methodology section is based on analyzing different voting schemes that have been carefully selected to include unique voting systems to be tested. This work will help determine whether there is a clear winner when selecting the type of voting technique to employ or if its a relative choice dependent on other factors for example dataset type.

6.2.1 Voting Scheme I: Plurality Voting

Plurality voting mechanism belongs to the family of positional voting, which involves different ranks for different candidates, with each rank holding a different priority. For plurality voting, the *candidate* with the most first-place *votes* is selected as the final match. This is the most common and basic type of voting, similar to hard voting used in classification problems. When dealing with an ensemble of VPR techniques, each of which results in different reference images selected to match with the query image, it is not al-

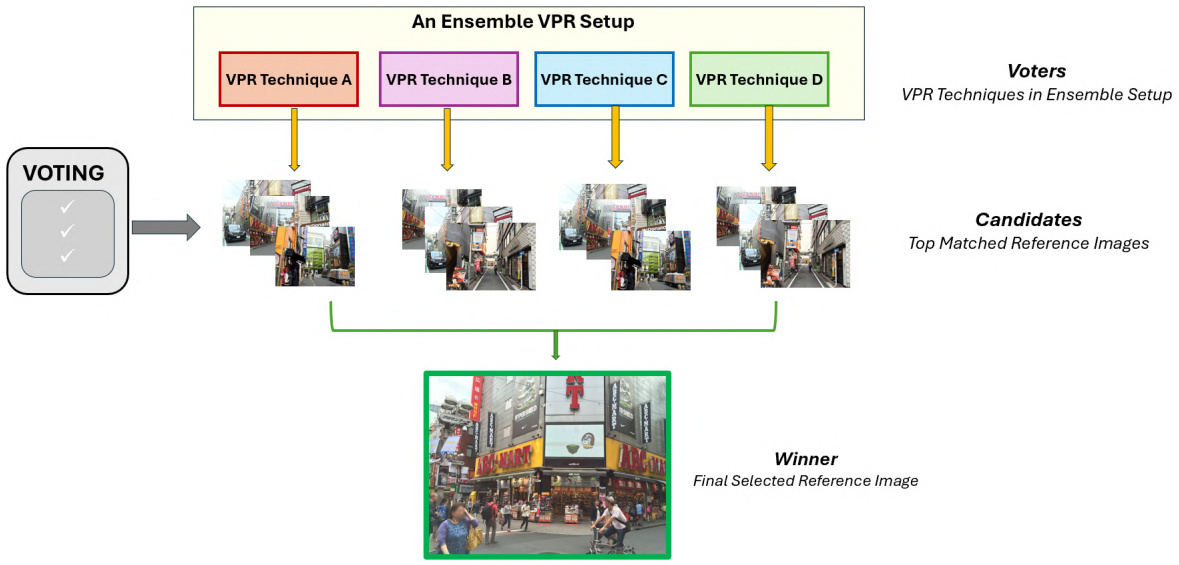

Fig. 6.2. A standard VPR ensemble setup employing several VPR methods simultaneously, which produces the top best matches by each method. These matches are subjected to various voting schemes to observe differences in results. The image shows the VPR techniques as voters, the top matched reference images as candidates, and the final selected reference image as the winner.

ways simple or obvious which image is the correct match to the query. Here, the plurality voting mechanism is employed to test the results it produces over different VPR datasets.

Let c be the retrieved image and v_c be the number of votes each retrieved image acquires. $Argmax$ returns the reference image with the maximum votes, which is the final selected matched image to the query.

$$\text{argmax}; c \in 1, 2, ..., n; v_c \tag{6.1}$$

In some cases, multiple v_c may evaluate to the same value, leading to a tie. To address this, the tie can be resolved by selecting the image with the highest combined confidence score from the VPR techniques. This approach leverages additional information to make a more informed decision. Alternatively, other methods such as randomly selecting one

of the tied images or conducting a secondary round of voting using a different voting scheme can be considered.

6.2.2 Voting Scheme II: Condorcet Voting

Condorcet is a ranked type of voting method that attempts to determine the overall selection of the *candidate*, reference image, by comparing the results of each *voter*, VPR technique, from the ensemble techniques in one-on-one match-ups. Each VPR technique produces a sequence of potential correct images that are ranked based on matching scores. A pairwise comparison matrix M is created where each element M_{ij} represents the number of times reference image c_i is ranked higher than reference image c_j by the ensemble of VPR techniques.g

Let C be the set of reference images where n is the total number of reference images.

$$C = \{c_1, c_2, ..., c_n\} \tag{6.2}$$

Let's denote the ranked positions as B where v is the total number of ranks.

$$B = \{b_1, b_2, ..., b_v\} \tag{6.3}$$

Each b_v in B contains a ranked preference order for the reference images in C. For example, if we have 3 reference images c_1, c_2, c_3, a ranked position could be $b_v(c_1, c_2, c_3)$, indicating that c_1 is the technique's first choice, c_2 is the second choice, and c_3 is the third choice. To calculate the pairwise victories for each pair of candidates c_i and c_j in C, we sum the number of times c_i is ranked higher than c_j across all VPR techniques.

The Condorcet winner is the reference image that would win in a head-to-head matchup against any other candidate. The reference image that is ranked higher than all other images is the final selected reference image. However, in the case of a tie, where two or more candidates have equal pairwise victories, the candidate with the highest overall rank is selected as a tiebreaker. This approach was chosen because it leverages the existing ranking information provided by the VPR techniques, ensuring consistency in the decision-making process and reducing the need for additional steps or methods.

The matrix M is formally defined as follows:

$$M_{ij} = \sum_{k=1}^{v} \delta(b_k(c_i) < b_k(c_j)) \tag{6.4}$$

where δ is the function which equals 1 when the condition is true and 0 otherwise. The Condorcet winner is the image c_i such that:

$$\sum_{j \neq i} M_{ij} > \sum_{j \neq i} M_{ji} \tag{6.5}$$

Equation 6.4 thus sums over the comparisons $c_i > c_j$ to find the image c_i that is preferred over all others.

6.2.3 Voting Scheme III : Broda Count Voting

Broda Count is another positional voting system utilised for various types of electoral tasks. Similar to the plurality voting, it also belongs to the family of positional voting where the *candidates*, potential matches to the query image, are ranked in descending order based on their matching scores. The position or rank of the reference image is

important as a higher rank (i.e., higher points/score) suggests a higher chance or preference for the particular image being selected as the final match in the ensemble VPR setup. However, in the event of a tie, where two or more reference images have the same total score, the tie is resolved by selecting the image with the highest individual rank from one of the VPR techniques.

Let c be the reference image selected while i represents the total number of techniques being considered. j represents the rank of each reference image and n is the value of points or score for a given rank. Finally, S_c is the summation of all the points for a single reference image. Here, $x_{i,j}$ represents the score assigned by the i-th technique to the reference image at rank j.

$$S_c = \sum_{i=1}^{i=n} x_{i,j} \tag{6.6}$$

The image with the highest sum is selected as the final match to the query.

$$\text{argmax } i \in \{1, 2, ..., n\}\ S_c \tag{6.7}$$

6.2.4 Voting Scheme IV: Contingent Voting

Contingent Voting is a form of rank-choice voting in which the *candidates*, reference images, are ranked while each rank represents a different priority. This type of voting on an ensemble VPR setup allows ensuring that the final selected image has the broadest possible support/votes from among all the VPR techniques being employed.

In a contingent voting system, the first-choice votes, meaning the top-ranked candidates or images, are tallied. The images with the highest first-place votes are considered for further stages of the voting process. If a reference image has an absolute major-

ity (more than 50%) of the first-choice votes, it is simply selected as the final image. If, however, no reference image has an absolute majority, the images with the lowest votes are eliminated, and their ranks and scores are transferred to the image with the highest votes. Lastly, the steps for recount and redistribution are performed by taking into account the recounted votes, and the process is repeated until a single best winner for a reference image is found. Although there are several other ways for counting and redistributing the votes for a contingent system, the core principle is the same one as employed in this setup, where the VPR techniques select the possible reference image matches and rank their preferences, and the voting aims to select a reference image that has the broader votes/support of selection among all the employed VPR methods.

Contingent voting does not lend itself to a single mathematical formula like some other voting systems, but it involves a series of steps that can be explained mathematically. Each VPR method employed in an ensemble VPR setup puts forth a ranked ballot of reference images to match, and these are assigned preferences such as "1" for the first choice, "2" for the second choice, and so on. Additionally, in the event of a tie, where two or more images have the same number of votes after redistribution, the tie is resolved by eliminating the image with the lowest combined score across all ranking positions. This method ensures that the reference image with the strongest overall support remains, which maximizes consensus while maintaining fairness.

For representation, let C be the set of reference images where n is the total number of reference images.

$$C = \{c_1, c_2, ..., c_n\} \tag{6.8}$$

Let's denote the ranked positions as B where v is the total number of ranks.

$$B = \{b_1, b_2, ..., b_v\} \tag{6.9}$$

For counting first-choice votes, let N be the total number of VPR techniques and let V_i be the number of first-choice votes recorded by the candidate i. For checking for an absolute majority, if any reference image receives more than 50% of $N/2$ of the first-choice votes, that reference is the final selected image to match. To explain, if a reference image d exists such that V_d is greater than $N/2$, then the reference image is selected.

For elimination of the lowest ranked image, the image with the least first-choice votes is eliminated. The votes from the eliminated image are then redistributed to the remaining images. This redistribution is weighted by the rank of the images; higher-ranked images receive a greater proportion of the redistributed votes. To recount and recalculate, the votes are tallied again with the redistributed votes, and the process is repeated to check for an absolute majority winner. This cycle continues until an image receives more than 50% of the votes or until only one image remains to be selected.

6.2.5 Voting Scheme V: Instant RunOff Voting

Instant Runoff Voting (IRV) is another type of ranked voting used to ensures the selection of a candidate with the broadest support/votes is selected. Images are selected in order of preference, in this case in terms of highest to lowest similarity scores. Then for tabulation of the votes, if any reference image receives an absolute majority it is simply selected as the final reference image, similar to Contingent voting. If, however, no such reference image exists, the votes are recounted and redistributed to their second-choice candidates.

After redistribution, the votes are recounted, and the process repeated until a candidate with a majority is selected.

Let N be the number of VPR techniques being employed and let C be the set of reference images where n is the total of images. Each N ranks all reference images from 1 to n.

$$C = \{c_1, c_2, ..., c_n\} \tag{6.10}$$

The next step is to find the first-choice votes for each image and determine whether an absolute winner exists. If it does, the process is stopped and a final image is selected. If no such winner is found, the candidate with the fewest first-choice votes, let's call this reference image D, is identified. Reference image D is then eliminated and all votes for D are redistributed to the next-highest-ranked reference image still in the running. This redistribution is achieved by transferring each vote to the highest-ranked image on each ballot that has not yet eliminated. This ensures that each vote is transferred to the next available preference.

The votes after this redistribution are recounted without the eliminated reference images to determine if an image with an absolute majority now exists. This process continues until a reference image achieves a majority. However, in the case of a tie where two or more images have the same number of votes after redistribution, the tie is resolved by selecting the image with the highest cumulative rank.

6.3 Universal Voting Schemes Experimental Setup

This briefly describes the two components involved in the testing of these voting schemes. As described in Figure 6.2, each voting scheme is tested on the same basic ensemble

TABLE 6.1: VPR-BENCH DATASETS TESTED FOR DIFFERENT VOTING SCHEMES

Dataset	Environment	Query	Ref. Images
GardensPoint	University	200	200
ESSEX3IN1	University	210	210
CrossSeasons	City-Like	191	191
Corridor	Indoor	111	111
17Places	Indoor	406	406
Livingroom	Indoor	32	32

VPR setup consisting of all the same techniques and run on each data set individually. This ensures the performance difference observed is solely based on the difference in the voting methodology and nothing else such as datasets or individual performance of a VPR technique. The VPR set up consists of eight state-of-the-art VPR techniques including AMOSNet [127], HOG [58], and AlexNet [128], HybridNet [127], NetVLAD [129], RegionVLAD [63], CoHOG [130], and CALC [74]. These have been selected not only for their wide use overall but also taking into consideration that these VPR techniques have been the focus for experimental designs for Chapter 3 to 5 as well. This gives us some predictability for the type of results expected from each technique individually and roughly provides more standardization for the testing for these voting techniques. The data sets chosen for testing the performance differences observed are also commonly used VPR datasets and also the ones utilised so far in Chapter 3 to 5. Table 6.1 shares some details about the datasets utilised such as the type of their environment and number of query and reference images. Table 6.2 shares some details for knowledge about the different characteristics of the universal voting schemes, these try to provide a clear and intuitive comparison of the voting schemes.

Voting Scheme	Methodology	Advantages	Drawbacks
Plurality	Most first-place votes	Simple implementation	Potential ties
Condorcet	Pairwise comparisons	Comprehensive preference consideration	May not yield a definitive winner
Borda Count	Rank-based scoring	Incorporates all ranked preferences	Moderately complex
Contingent	Rank-choice with elimination	Broad acceptance	Multi-step process
Instant Runoff	Rank-choice with redistribution	Ensures majority support	Multi-round counting

TABLE 6.2: COMPARISON OF VOTING SCHEMES

6.4 Universal Voting Schemes Results and Analysis

This section discusses and presents the results produced by testing each of the selected voting methodologies over various VPR datasets, under the same ensemble VPR set up. The results are presented in three different categories to efficiently evaluate the utility of each voting method. Firstly, the results are presented in terms of the performance bounds of each voting method and PR curves for accuracy as explained in Chapter 3, and finally Z-score for testing the significance in the performance difference that is observed. Z-score and the use of the McNemar-like test's test are discussed in detail below to highlight its importance and contribution to the results collected.

6.4.1 PR curves and Radar Charts

Figure 6.3 presents the results that were produced starting from the 17Places dataset, for which two of the voting methods, Instant Runoff and Condorcet Voting, have the highest performance bounds, followed by Borda Count and then Plurality and Contingent Voting methods. For the Livingroom dataset, an overall uniform performance bound is observed for most methods, but with Plurality, Instant Runoff, and Contingent Voting slightly out-

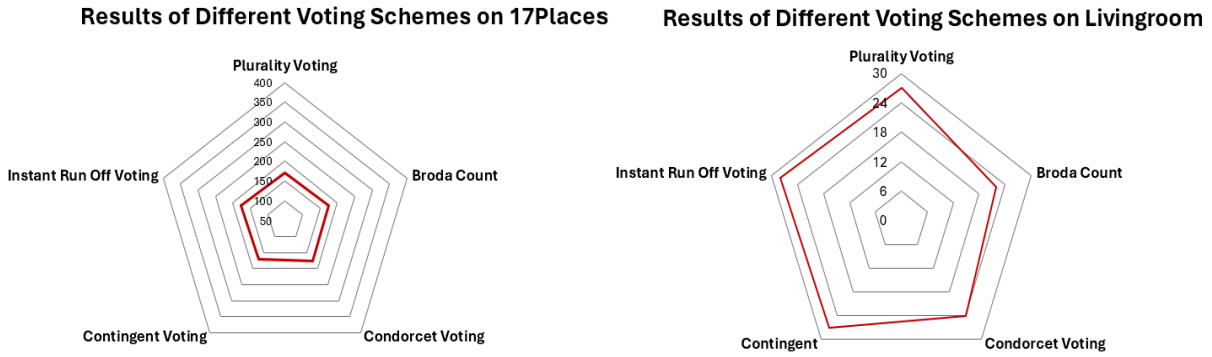

Fig. 6.3. Difference in performance bounds of each voting methodology in terms of query images correctly matched for 17Places and Livingroom Dataset

ranking the others. The next two datasets presented in Figure 6.4 are Corridor and CrossSeasons, for both of which Borda Count ranks the highest among all methods, followed by Contingent Voting and then the others. The last two datasets in Figure 6.5 tested for their performance bounds are ESSEX3IN1 and GardensPointWalking. ESSEX3IN1 has similar performance bounds for most methods, although Contingent Voting outranks the others slightly. GardensPointWalking has the least favorable performance bounds for Contingent and Plurality Voting, while Condorcet, Borda, and Instant Runoff have better performance bounds overall. The x-axis ranges in these figures differ due to the large variations in dataset sizes, allowing for clear and accurate representation of each dataset's performance metrics.

Next, Figure 6.6 presents results in terms of PR curves to showcase the performance difference observed for the different voting methods tested for multiple VPR data sets. Varied results in performance are observed beginning from the Corridor Data set where Instant RunOff and Condorcet voting outperform the remaining three substantially. The livingroom dataset has a more uniform performance in terms of the precision-recall observed over this dataset with most voting methods performing similarly. The

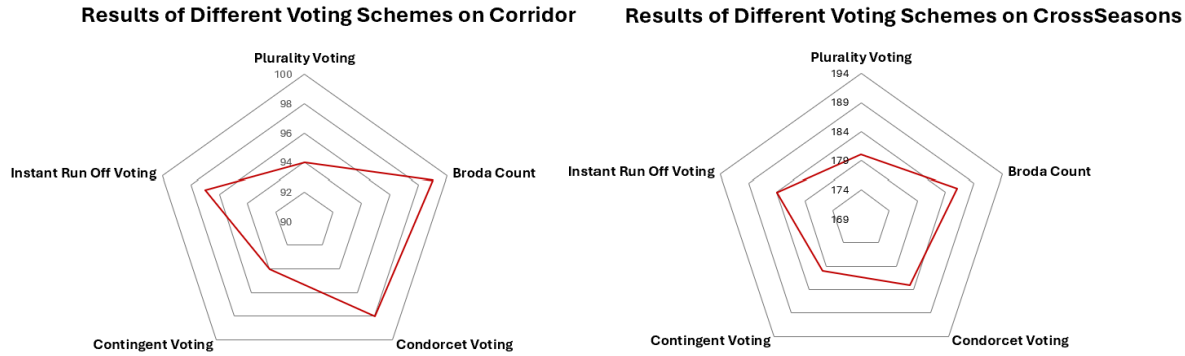

Fig. 6.4. Difference in performance bounds of each voting methodology in terms of query images correctly matched for CrossSeasons and Corridor Dataset

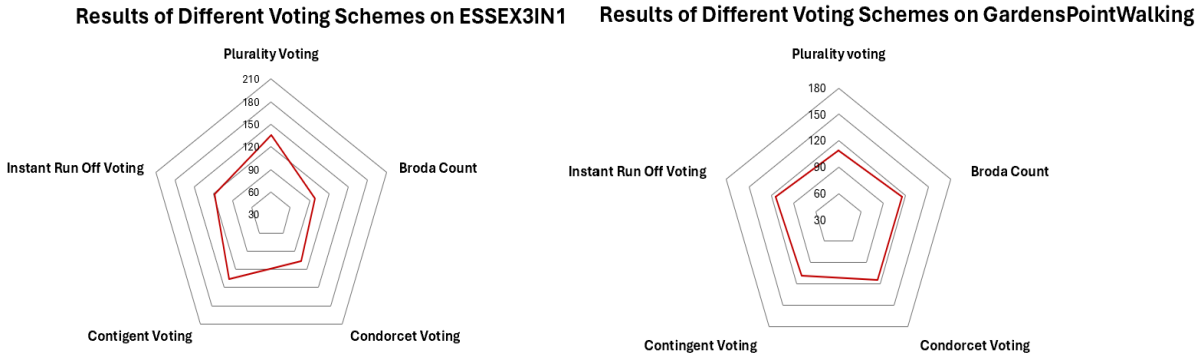

Fig. 6.5. Difference in performance bounds of each voting methodology in terms of query images correctly matched for ESSEX3IN1 and GardensPointWalking Dataset

Corridor and CrossSeasons datasets both have the highest performance when utilizing Broda Count as the voting method followed by Condorcet Voting. For the ESSEX3IN1 dataset plurality and Contingent voting appear to produce better accuracy. While GardensPointWalking data has the opposite results as it has the least favourable results when employing plurality and Contingent Voting. However, Condorcet Voting appears to produce better results in comparison to other voting methods tested.

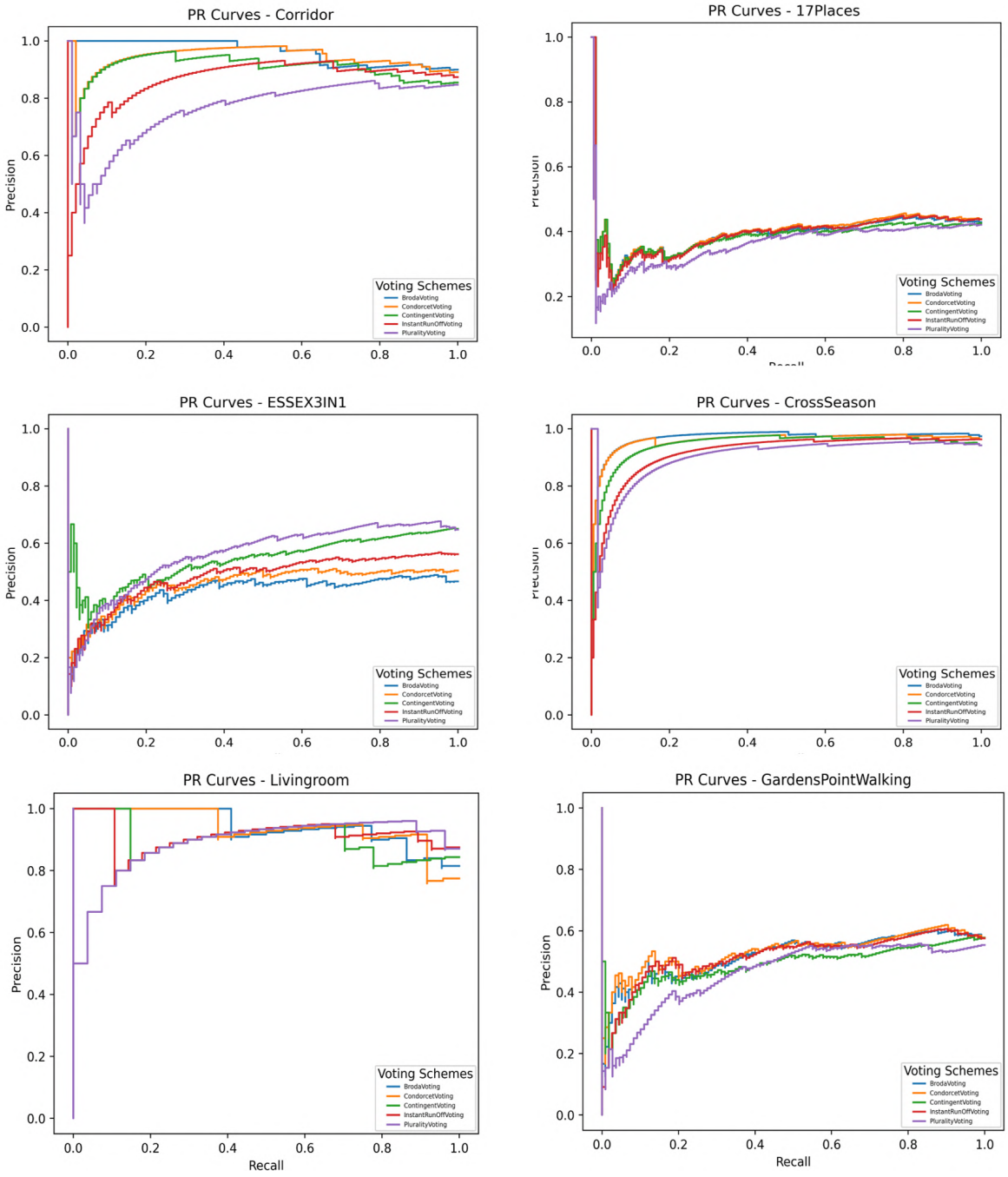

Fig. 6.6. PR curves for voting methods i.e Plurality, Condorcet, Contingent Voting, Broda Count and Instant Run Off voting for datasets 17Places Livingroom, Corridor, CrossSeasons, ESSEX3IN1 and GardensPointWalking).

6.4.2 McNemar-like Test's

To further confirm that the results presented in Figures 6.3 to 6.5 are not a chance occurrence but rather conclusive evidence of the significant difference between different voting schemes, utilising a variant of the McNemar test [138], inspired by the original statistical test references [122, 123]. This variant helps identify the statistically relevant performance differences, with a confidence interval, between the various voting schemes, distinguishing the best cases from the worst cases. By adapting the approach described in [138] to perform a pairwise evaluation of voting methods using a series of frame-by-frame matches/mismatches on the same dataset. Since this approach cannot compare more than two VPR methods simultaneously, a series of independent pairwise tests are necessary to compare multiple voting methods. These results are presented in the form of a Z-Score table that states the different Z-scores corresponding to their confidence intervals to showcase the statistical significance. In particular, a 95% significance level corresponds to Z = 1.96 and presents a highly significant performance difference between the two compared voting schemes.

$$X^2 = \frac{|N_{sf} - N_{fs}|}{\sqrt{N_{sf} + N_{fs}}} \tag{6.11}$$

The value to generate the Z-score utilizing the two-tails table is generated as N_{sf} denotes the number of trials where the voting 1 succeeded, and 2 failed; N_{fs} denotes the number of trials where 1 failed and 2 succeeded. X_2 is distributed, to a good approximation, as chi-squared with one degree of freedom. Where the confidence interval associated with Z can be determined using tables [190]. The results are presented in Figure 6.7 in form of a heat map to efficiently locate pairs and voting methods that

are significantly better than their compared counterparts. The table can be interpreted using the Z scores mentioned and the corresponding confidence interval to understand whether the difference is significant or not.

Z-Scores

Figure 6.7 presents results in terms of different data sets tested such as ESSEX3IN1 for which overall Contingent voting scheme is significantly better in performance than the others. The second best option for selection can be concluded as IRV over Plurality or Broda and lastly Broda is a better selection over Plurality voting. Again, the results demonstrate how substantial the difference can be in selection if voting schemes are con-

	ESSEX3IN1	Livingroom	GardensPoint	CrossSeasons	Corridor	17Places
Contingent vs Condorcet	5	1.341640786	-1.341640786	-1.732050808	-1.632993162	-1.88982237
Contingent vs IRV	3.53009043	-0.447213595	-1.386750491	-1	-1.732050808	-1.88982237
Contingent vs Plurality	0	0	0	0	0	0
Contingent vs Broda	5.72871555	1.889822365	-1.341640786	-2.449489743	-1.666666667	0
Condorcet vs IRV	0	-2	0.25819889	-1	0.577350269	0
Condorcet vs Pluarlity	-5	-1.341640786	1.341640786	1.732050808	1.632993162	1.889822365
Condorcet vs Broda	1.78885438	1	0	-1.732050808	-0.377964473	0.816496581
IRV vs Pluarlity	-3.53009043	-1	1.386750491	2	1.732050808	2.449489743
IRV vs Broda	3.9223227	2.121320344	-0.25819889	-1.414213562	-0.816496581	0.816496581
Broda vs Plurality	5.72871555	1.889822365	-1.341640786	-2.449489743	-1.666666667	-0.90453403

Fig. 6.7. Pairwise voting method comparisons use a sign convention: positive Z indicates the first method outperforms the second, and negative Z the opposite. The legend's color ranges from green (highest confidence intervals) to red (lowest).

sidered based on performance rather than simply selecting the conventional approach. Considering the results for Livingroom Dataset IRV is a far better selection over Broda with a confidence interval of over 95% and Contingent voting is again better than Broda with over 90% confidence interval. The differences in outcomes among various voting schemes can sometimes be significant, but the value lies not just in identifying signi-

ficant differences but in providing a guideline for better selection. By considering all possible voting options, a more informed decision can be made. As shown in Figures 6.3 and 6.4, the research demonstrates that conventional methods are not always the best choice. For instance, in the GardensPointWalking dataset, the results suggest that while the differences are substantial, identifying better options such as IRV over Contingent and Plurality provides further valuable insights.

CrossSeasons dataset is another example of why random selection of a voting method is not the best idea given Contingent voting, which is the ideal scenario for a case like ESSEX3IN1, consistently underperforms in comparison to other available methods such as Condorcet, IRV, Broda with a confidence interval ranging from even 90% to 95%. Furthermore, in comparison to IRV and Broda, Plurality voting significantly performs better for this case with a z score of over 1.96. Similarly, Corridor dataset again does not stick to the pattern observed for ESSEX3IN1 and Contingent voting is not the better option to be selected over any of the other schemes. In fact Condorcet, IRV and Broda all are the better choice while IRV is better than Plurality but Plurality is better than Broda. Lastly for 17Places the results vary where Condorcet is better than Contingent, and so is IRV. Some cases the difference is not significant to consider while Plurality is better to consider than some of the other choices. Overall, the heat-map is a useful reference point in navigating through the selection of a voting scheme in different ensemble cases to improve performance.

6.5 Universal Voting Schemes Summary

This chapter explores and implements the different universally used voting schemes that can be extended to an ensemble VPR setup to discover whether they provide any performance improvement over the conventional practices of voting. Furthermore, evaluating the performance difference via McNemar's test-like approach using a pairwise analysis to determine whether the difference in performance is statistically significant. With a selection of five employed and diverse voting methods which are very commonly used among other fields of research, the experimental setup tests these schemes on multiple datasets to provide the results presented. The selected methods tested include Plurality voting, Condorcet voting, Contingent Voting, Broda count, and IRV voting. Several insights were collected, the first and foremost being that the employment of different voting schemes, much like in other fields, produces varied results when only the voting method is different. This finding confirms that the selection of a voting method for a VPR ensemble setup is not a trivial task but rather a process of careful selection, with different voting schemes standing out in performance for different types of datasets or variations in surroundings. Furthermore, among the tested voting methods, the closest to the conventionally employed method is the Plurality method. Interestingly, the results showcase that the assumption that this common practice method might be a good choice for all cases is not true. For example, in most experiments, the Condorcet, IRV, and Broda methods were almost always the better option compared to Plurality. Within these methods, IRV outperformed Plurality in some cases, and Plurality outperformed Broda in specific scenarios. These conclusions indicate that selecting the appropriate voting method depends on specific requirements and context. Overall, when a fundamental ranking

system without prerequisites related to dataset type, variation, or size is required, Condorcet Voting demonstrates overall consistency and produces significant results, followed by IRV and Contingent Voting, with Broda Count showing relatively lower performance. To further confirm the findings, the results are supported with a statistical analysis confirming the statistical significance of the results. These can be seen in the Z-score table in Figure 6.5, presenting Z-score values that correspond to the confidence intervals to showcase how our results presented in Figure 6.3 are significant in difference. It is important to note that the primary goal of this experiment was to highlight the impact of selecting different voting methods on VPR performance. Hence, computational efficiency was not the focus of this study. The experiment aimed to determine how varying voting schemes influenced VPR accuracy, and future work could explore the computational efficiency of these voting schemes once accuracy differences have been established. Moreover, the experiment utilized a standard ensemble VPR setup, as discussed earlier, which relies on a brute-force approach. While this method is not computationally efficient, it was deliberately chosen to ensure that any observed differences in VPR accuracy are solely driven by the selection of different voting methods, rather than by computational optimizations. Future research could explore how to improve computational efficiency while maintaining the accuracy gains observed through careful voting scheme selection.

In conclusion, the investigation of various voting systems within an ensemble VPR setup highlights their substantial potential to enhance VPR accuracy and robustness. The findings demonstrate that the choice of voting scheme significantly influences the overall performance of VPR systems, with Condorcet, IRV, and Broda methods consistently outperforming the conventional Plurality method. This insight emphasizes that the selection of an appropriate voting method is a critical decision, directly impacting

the efficiency and effectiveness of VPR ensembles. By employing adaptive and scalable voting algorithms, future research can further refine the integration of multiple VPR techniques, leading to more resilient and efficient VPR systems. The potential for innovative approaches to VPR through the thoughtful selection and application of voting methods offers promising avenues for future exploration, aiming to maximize performance.

Chapter 7

Conclusions and Future Directions

7.1 Overview

Visual Place Recognition (VPR), a critical component of autonomous navigation systems, faces significant challenges due to environmental variabilities, such as changes in lighting and weather conditions; viewpoint alterations from different navigation angles; dynamic obstacles like moving vehicles and pedestrians; and perceptual aliasing, where different scenes appear confusingly similar. These issues complicate the task of VPR, severely limiting the efficacy and robustness of traditional methodologies. While these traditional methods are potent, their adaptability and efficiency diminish when confronted with these diverse and unpredictable variations. Responding to these challenges, this thesis explores the concept of complementarity among various VPR techniques. Traditional VPR systems, relying on single-method approaches, often struggle to address all operational environments effectively. This research posits that a synergistic approach, which leverages the distinct advantages of multiple VPR methods, can significantly enhance

system performance.

The exploration starts by introducing and systematically evaluating the complementary nature of various Visual Place Recognition (VPR) methods. This initial phase is crucial as it assesses how different techniques can enhance each other's strengths and mitigate weaknesses when combined. This understanding of complementarity is pivotal, as it lays the foundational groundwork for developing more sophisticated ensemble systems like SwitchHit and SwitchFuse. SwitchHit introduces a dynamic switching mechanism that selects the most suitable VPR technique based on environmental inputs, thus optimizing performance adaptively. Building on this, SwitchFuse combines the strengths of selected VPR techniques through a sophisticated fusion process, further enhancing the recognition accuracy. Moreover, the thesis examines the role of universal voting schemes in ensemble VPR setups. This exploration highlights how different voting mechanisms can influence the decision-making process of VPR systems, potentially leading to improvements in accuracy and reliability. This novel exploration highlights how different voting mechanisms can influence the decision-making process of VPR systems, potentially leading to improvements in accuracy and reliability. By systematically analyzing these various approaches, the research presented in this thesis offers a different perspective on solving VPR tasks, paving the way for the development of autonomous systems that can navigate more effectively in complex and changing environments.

While this thesis does not claim to completely solve the longstanding challenges in VPR, it provides significant insights and methodologies that improve upon existing techniques. The proposed systems and frameworks are designed to address specific issues related to environmental variability, adaptability, and decision-making processes, thereby contributing to the advancement of VPR technology.

7.2 Summary of Contributions and Significance

The contributions highlighted in this thesis mark significant steps forward in the field of Visual Place Recognition (VPR). The research introduces innovative frameworks and systems in each chapter, providing new approaches to address longstanding challenges. These contributions present not only valuable theoretical knowledge but also demonstrate practical implications that could enhance how autonomous systems navigate. The following is a glimpse into the work each chapter contributes, pushing the boundaries of what's possible in VPR.

- The thesis presents a novel framework for assessing the complementarity among different Visual Place Recognition (VPR) techniques, offering a systematic approach to enhance VPR systems' performance by leveraging the unique strengths of various algorithms. This framework represents an attempt to quantify and utilise the concept of complementarity in VPR, addressing the challenge of environmental and viewpoint variability that has affected the reliability of existing VPR methodologies. The introduction of complementarity provides valuable insights for building efficient and robust ensemble VPR methods. The framework, tested on a wide array of VPR datasets and state-of-the-art methods, offers detailed insights into the complementarity behaviour among these widely employed methods. It is adaptable and can be tested to study the complementarity of any other VPR methods.

- The thesis introduces SwitchHit, a novel ensemble setup designed to dynamically select the most suitable Visual Place Recognition (VPR) technique based on the concept of complementarity among different VPR methods. This chapter advances the field of VPR by addressing the limitations of existing systems when confronted

with dynamic and varied real-world environments. SwitchHit, through its probabilistic, complementarity-based switching system, enables a VPR system to intelligently adapt its technique selection in response to specific environmental cues and variations in the query images, optimizing recognition performance.

- Building upon the success of SwitchHit, the SwitchFuse system is introduced, incorporating both switching and fusion strategies to further optimize VPR accuracy. SwitchFuse represents a hybrid model that utilises the complementarity framework to select the best techniques for fusion and incorporates a tripartite model to adapt dynamically to various environmental variations. This approach to VPR system design sets new benchmarks for accuracy and reliability, outperforming existing multi-fusion and standalone VPR systems, as well as SwitchHit itself.

- The introduction and exploration into universal voting schemes for VPR opened another promising and previously under-researched avenue. By evaluating the impact of different voting methodologies on ensemble VPR setups, it was revealed that the choice of voting scheme could significantly influence an ensemble VPR system's performance. This research underscores the potential for refining ensemble VPR methods further, highlighting the need for careful selection and application of voting schemes based on the specific requirements of the VPR task. This was demonstrated by applying several universal voting methods on the same experimental setup to show the difference in performance by merely changing the voting method and then analyzing the significance of this difference.

While this thesis does not claim to completely solve all longstanding challenges in VPR, it provides significant insights and methodologies that improve upon existing tech-

niques. The proposed systems and frameworks are designed to address specific issues related to environmental variability, adaptability, and decision-making processes, thereby contributing to the advancement of VPR technology.

7.3 Impact on Computer Vision and Robotics

Reviewing the significant contributions made to Visual Place Recognition (VPR) through this thesis unveils several new ideas and their widespread implications. These advancements in VPR set new benchmarks in computer vision and robotics, transforming theoretical concepts into practical, real-world applications.

The evaluation of complementarity among VPR techniques addresses a critical gap in existing methodologies. By systematically quantifying and leveraging the strengths of various VPR algorithms, this framework enhances the adaptability and robustness of autonomous systems against environmental and viewpoint variabilities. The significance of this framework extends beyond VPR, offering profound implications for computer vision by suggesting a method for combining diverse algorithms to optimize system performance. For instance, complementarity-based approaches can significantly enhance the robustness and reliability of autonomous vehicles navigating through complex, dynamic environments by leveraging the strengths of multiple recognition techniques.

In the realm of robotics, this framework provides a robust strategy for enhancing autonomous navigation, particularly in environments where conditions can change unpredictably. It encourages the development of intelligent systems that can dynamically adjust their processing based on real-time inputs, which is crucial for deploying robots in complex scenarios such as rescue missions or unpredictable urban landscapes. For

example, in rescue and disaster response scenarios, complementarity-based approaches enable rescue robots to dynamically adapt to changing conditions, improving their effectiveness in search and rescue missions.

The development of systems like SwitchHit, which dynamically selects the most appropriate VPR technique based on complementarity, represents a significant advancement in VPR. SwitchHit enables real-time adaptability to the dynamic and varied nature of real-world environments. This system showcases the potential of probabilistic models to facilitate intelligent decision-making in real-time, thereby optimizing the recognition performance of autonomous systems. In robotics, this translates to a framework for autonomous vehicles and mobile robots to adapt their navigational strategies based on environmental cues, ensuring effective operation in diverse conditions such as urban navigation, disaster recovery, or varied industrial environments.

Exploring hybrid systems like SwitchFuse, which integrate switching and fusion strategies to enhance VPR performance in terms of accuracy, sets new benchmarks for accuracy and reliability. These systems utilise a strategic fusion of technologies to optimize the strengths of individual VPR techniques. The broader impact on computer vision involves demonstrating how hybrid systems can be designed to leverage the complementary strengths of different algorithms, offering a new model for systems that require robust decision-making capabilities under varied conditions. In robotics, SwitchFuse provides a blueprint for the development of advanced autonomous systems that can utilise a combination of sensory inputs and analytical methods to achieve superior operational effectiveness. This is particularly relevant for applications requiring high levels of precision and reliability, such as autonomous drones in surveillance tasks or robotic systems in manufacturing settings, where environmental conditions and operational demands can

vary extensively.

The exploration of universal voting schemes unveils their significant impact on the performance of ensemble VPR systems. Strategic voting scheme selection can greatly enhance the accuracy and reliability of ensemble methods. The implications for computer vision are broad, as nuanced algorithmic adjustments can refine the decision-making processes of computer vision systems, enhancing their application in real-world scenarios. For robotics, these insights contribute to the development of more sophisticated collaborative systems, where multiple autonomous agents must synchronize and make decisions in real-time. This could revolutionize applications such as swarm robotics, where effective consensus mechanisms are crucial for coordinated action in complex tasks like exploration, mapping, or coordinated transport and assembly operations.

By targeting these specific subfields, the research underscores the practical implications and potential impact of complementarity-based approaches in advancing the state-of-the-art in computer vision and robotics.

7.4 Future Directions

Building on the research previously discussed, the following section explores the potential future directions for this work in Visual Place Recognition (VPR). Each chapter proposes transformative changes and outlines a roadmap for their implementation, aiming to make significant progress in practical applications in both computer vision and robotics.

- The innovative framework for assessing complementarity among Visual Place Recognition (VPR) techniques, introduced in Chapter 3, sets the stage for transformat-

ive advancements in the VPR field. A promising future direction entails leveraging machine learning and artificial intelligence to automate the identification of complementarity, enabling VPR systems to dynamically adapt to environmental changes without manual tuning. Additionally, there's significant potential in developing adaptive systems that not only optimize VPR performance in terms of accuracy based on contextual cues but also conserve computational resources by intelligently applying the most effective techniques for the task at hand. Further exploration into the environmental and contextual factors affecting complementarity could lead to VPR systems capable of operating in extreme conditions, enhancing their resilience and versatility.

- The development and introduction of SwitchHit in Chapter 4, as a dynamic, complementarity-based selection system for Visual Place Recognition (VPR) techniques, suggests a promising advancement in intelligent, adaptable VPR systems. Future research could focus on enhancing SwitchHit's predictive capabilities through the integration of more sophisticated machine learning models that consider a wider range of environmental variables and data inputs. This enhancement could further refine the system's ability to make accurate, context-aware decisions on technique selection, thereby optimizing performance across an even broader spectrum of scenarios. Additionally, investigating the potential for SwitchHit to operate in conjunction with sensor fusion technologies could open up avenues for creating VPR systems that are not only more accurate but also more resilient to extreme environmental conditions or sensor degradation. Unlike traditional methods such as the Extended Kalman Filter (EKF), which rely on predefined models, or recent transformer-based solutions, which are computationally intensive, SwitchHit aims

to provide a more flexible and contextually adaptive approach by leveraging real-time data and learning from the environment dynamically.

- The SwitchFuse system, in Chapter 5, which intelligently combines switching and fusion strategies for Visual Place Recognition (VPR), lays a robust groundwork for advancing VPR technology. Future explorations could delve into harnessing deep learning algorithms to automate the fusion process, enabling SwitchFuse to dynamically learn the most effective fusion strategies tailored to specific environmental contexts and query images. This approach could significantly enhance the system's adaptability and performance in diverse conditions. Additionally, extending the SwitchFuse concept to integrate with multi-sensory data, beyond visual inputs, presents a thrilling prospect. By incorporating data from lidar, radar, or auditory sensors, SwitchFuse could achieve a new level of environmental understanding and recognition accuracy. This differs from standard fusion approaches like EKF or transformer-based solutions by focusing on the dynamic and adaptive integration of complementary sensory inputs, thus providing enhanced robustness and flexibility. There is also potential in exploring the scalability of SwitchFuse in large-scale applications, such as city-wide navigation systems or search-and-rescue operations, where its advanced fusion capabilities could provide critical improvements in operational efficiency and reliability.

- Lastly, Chapter 6 and its exploration of universal voting schemes in ensemble Visual Place Recognition (VPR) systems opens new dimensions for refining the accuracy and reliability of VPR ensemble methods. Future directions could concentrate on a deeper investigation into adaptive voting mechanisms, where the selection of

voting schemes is dynamically adjusted based on the context or specific characteristics of the environment. This adaptive approach could leverage machine learning models to analyse the performance impact of different voting strategies over time, leading to a more intelligent system that optimizes its decision-making process. Unlike traditional sensory fusion methods, this approach focuses on leveraging the complementarity and dynamic adaptability of the ensemble, which can significantly enhance performance in real-time applications. There's also significant potential in applying the insights gained from universal voting schemes to collaborative robotics systems, where multiple autonomous agents must make collective decisions based on shared visual information. This application could vastly improve the coordination and efficiency of robot swarms in complex tasks, from exploration to disaster response, by ensuring optimal consensus-building mechanisms are employed.

By targeting these specific subfields, the research underscores the practical implications and potential impact of complementarity-based approaches in advancing the state-of-the-art in computer vision and robotics. Future work will continue to build upon these foundational advancements, pushing the boundaries of VPR systems and their applications in increasingly complex and dynamic environments.

7.5 Closing Remarks

This thesis marks a significant milestone in the evolution of Visual Place Recognition (VPR), laying down the foundational work that paves the way for future advancements. By introducing the concept of complementarity, and the innovative development of systems like SwitchHit and SwitchFuse, along with the examination of universal voting

schemes, this research has made considerable progress in addressing some of the long-standing challenges in VPR. This research stands as a testament to the potential of combining various VPR techniques to enhance system performance, highlighting its importance in the broader context of technological advancement.

However, the journey does not end here; there is much more to explore and refine. The field of VPR is ripe with opportunities for further innovation. By delving deeper into machine learning algorithms, expanding our understanding of environmental dynamics, and exploring the integration of these systems into real-world applications, we can continue to build on this foundation. The path forward involves collaborative efforts across disciplines, harnessing the power of technology to develop smarter, more adaptive systems. This thesis not only contributes to the academic discourse but also lights the way for practical applications that could transform our interaction with technology, making the dream of fully autonomous navigation a closer reality.

Appendix A

Detailed Complementarity Analysis

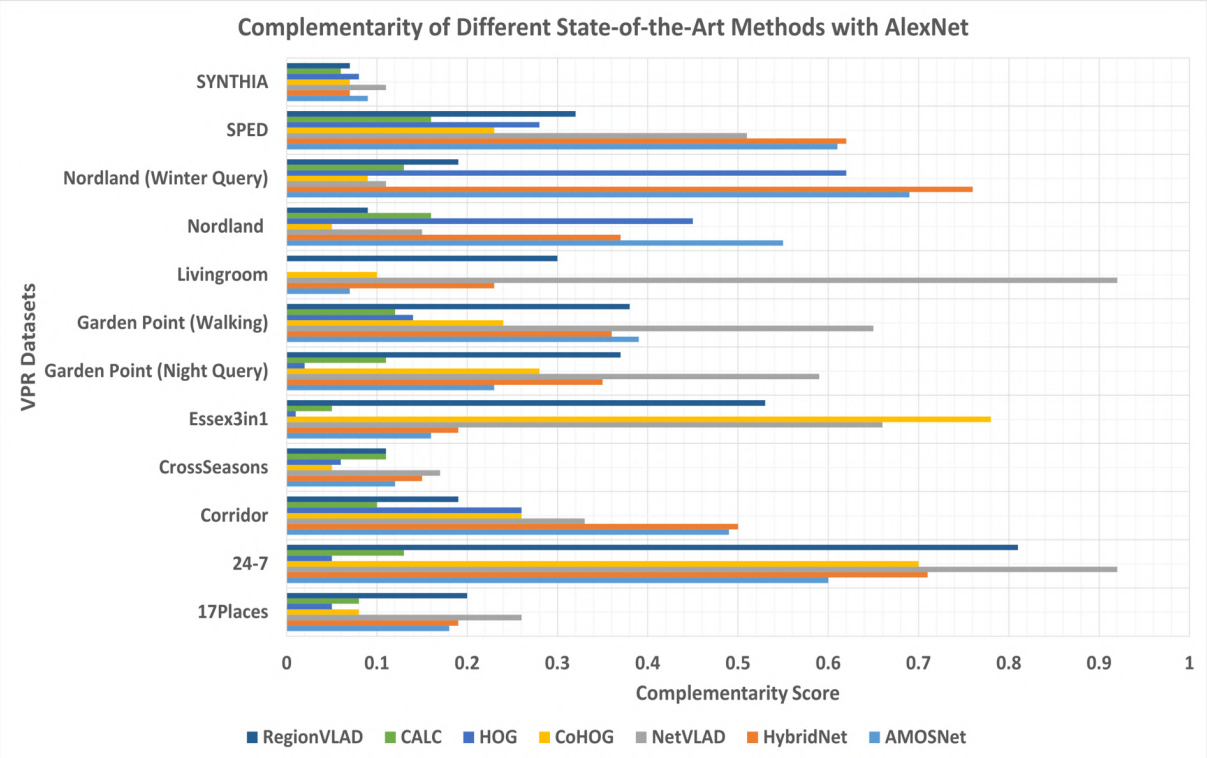

Fig. A.1. Complementarity of VPR methods with AlexNet on Multiple VPR datasets.

This section presents the results generated by utilizing the proposed framework over a set of eight VPR techniques on various standard VPR datasets. Figures 3.3 to 3.10

depict the complementarity scores of different VPR methods with each other across these datasets, allowing for visual analysis of how different pairs of VPR methods exhibit varied complementarity levels.

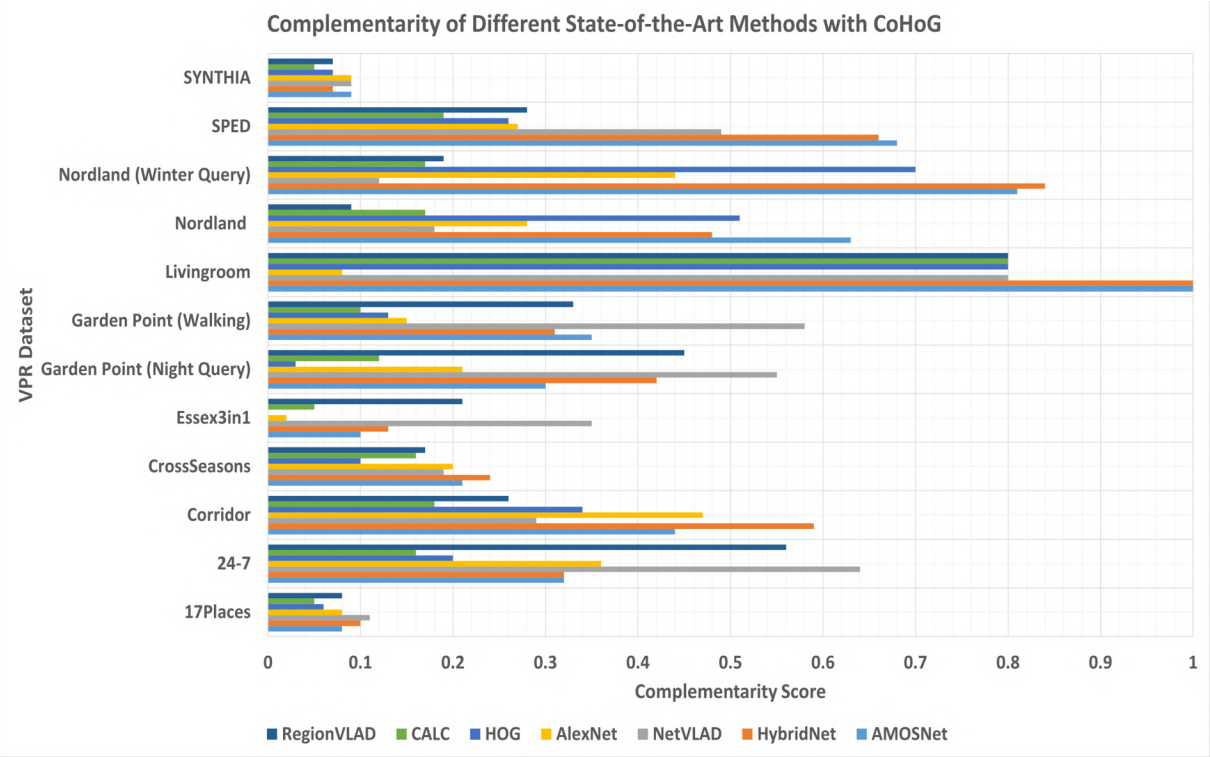

Fig. A.2. Complementarity of VPR methods with CoHOG on Multiple VPR datasets.

One of the most intriguing findings is how certain VPR techniques, which perform poorly on their own, show remarkably high complementarity scores when paired with others. This phenomenon occurs when the errors made by one technique are systematically different from the errors made by another, allowing them to effectively cover each other's weaknesses. For instance, CALC, despite its relatively lower individual performance, demonstrates high complementarity with NetVLAD across several datasets. As shown in Table 3.2, their combination achieves a complementarity score of 0.9 on the Essex3in1 and Livingroom datasets. This high score indicates that CALC and NetVLAD

make different errors on these datasets, and their combination significantly boosts overall performance.

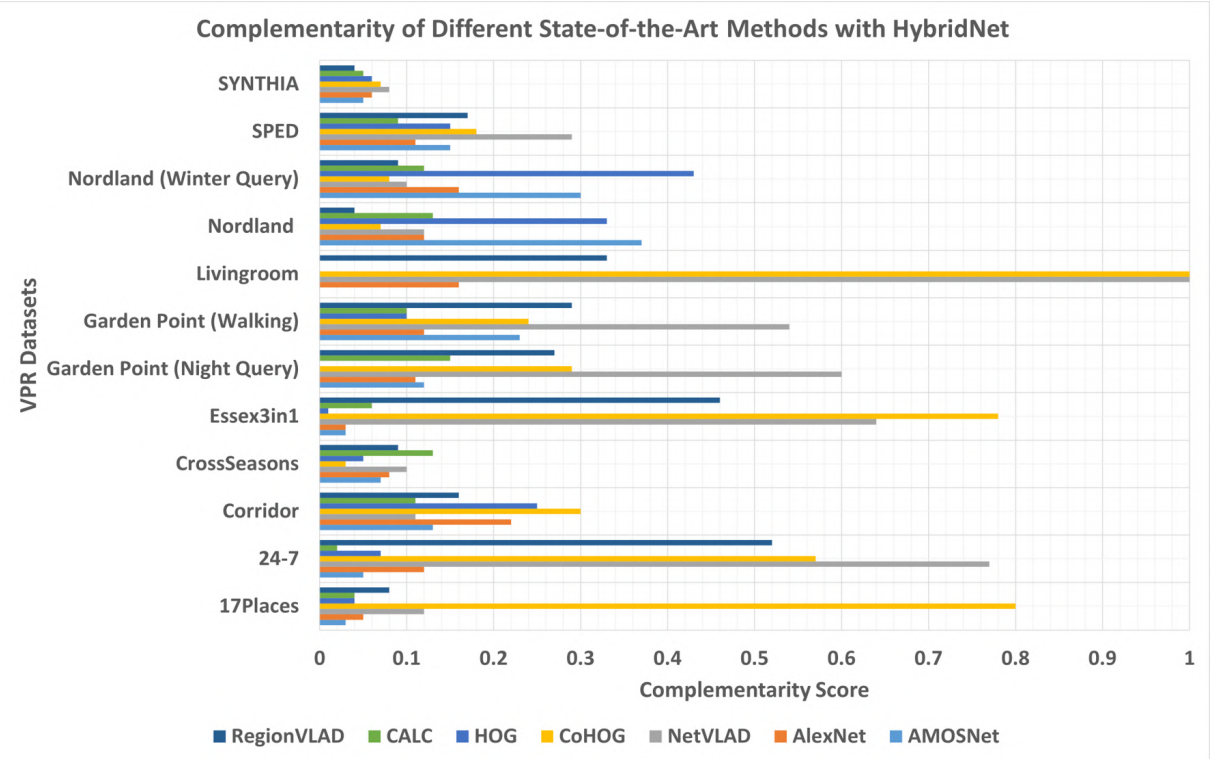

Fig. A.3. Complementarity of VPR methods with HybridNet on Multiple VPR datasets.

Taking a closer look, Figure 3.3 illustrates the levels of complementarity AlexNet has with other methods. Notably, AlexNet demonstrates high complementarity with NetVLAD, HybridNet, and RegionVLAD on several datasets. For example, NetVLAD achieves complementarity scores of 0.9, 0.65, 0.65, and 0.9 on the 24-7, Essex3in1, GardenPoint, and Livingroom datasets, respectively. Similarly, HybridNet shows strong performance on the 24-7, Corridor, Nordland, and SPED datasets, while RegionVLAD achieves the highest scores on the 24-7 and Essex3in1 datasets.

In another interesting example, Figure 3.4 reveals that CoHoG and NetVLAD are the only methods that complement AMOSNet well. For instance, AMOSNet and CoHoG

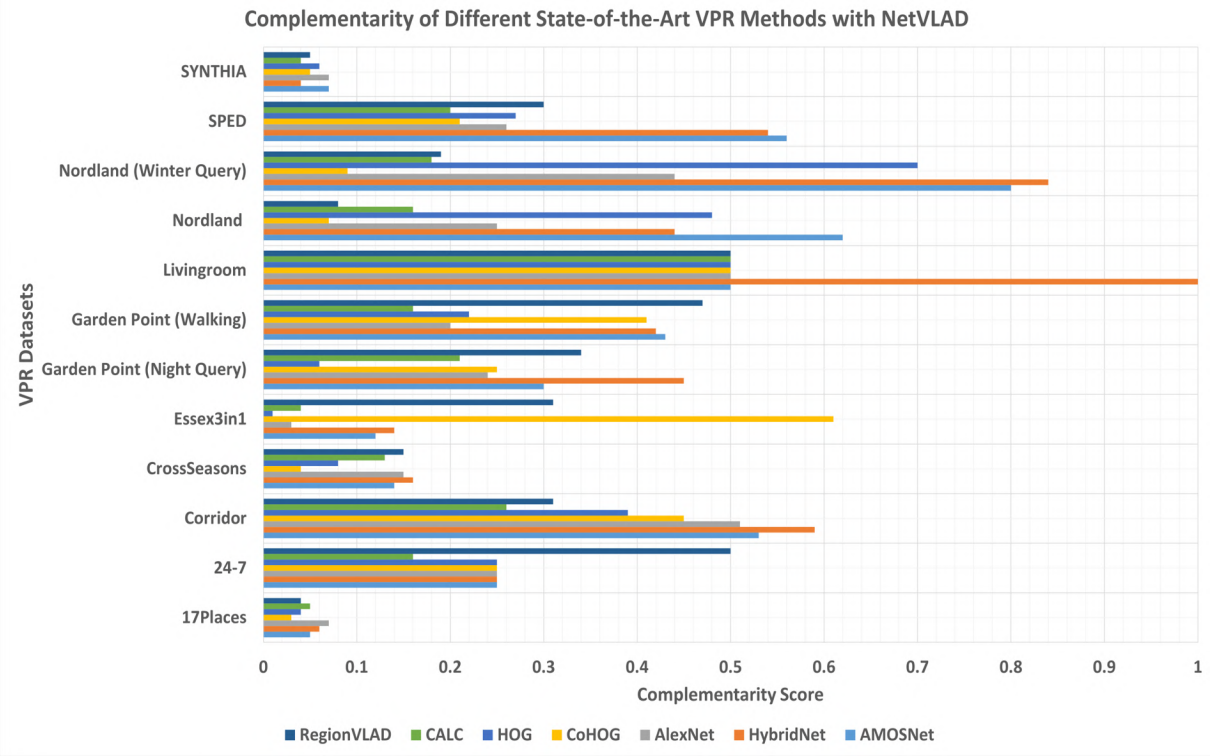

Fig. A.4. Complementarity of VPR methods with NetVLAD on Multiple VPR datasets.

achieve high complementarity scores of 0.7, 0.8, and 1 on the 24-7, Essex3in1, and Livingroom datasets, respectively. Similarly, the combination of AMOSNet and NetVLAD reaches scores of 0.85, 0.65, and 0.9 on the same datasets. This suggests that combining AMOSNet with CoHoG or NetVLAD can lead to a robust VPR system. On the contrary, CALC consistently scores low, indicating poor complementarity with AMOSNet.

When considering CALC as the primary VPR technique, Figure 3.5 shows that CoHoG and NetVLAD emerge as suitable partners. CoHoG exhibits high complementarity scores on the 24-7, Essex3in1, Livingroom, and SPED datasets, while NetVLAD matches CALC well on the 24-7, Essex3in1, GardenPoint, and Livingroom datasets. Interestingly, despite CALC's generally lower standalone performance, its combination with these techniques significantly boosts overall performance. This highlights the importance of evaluating

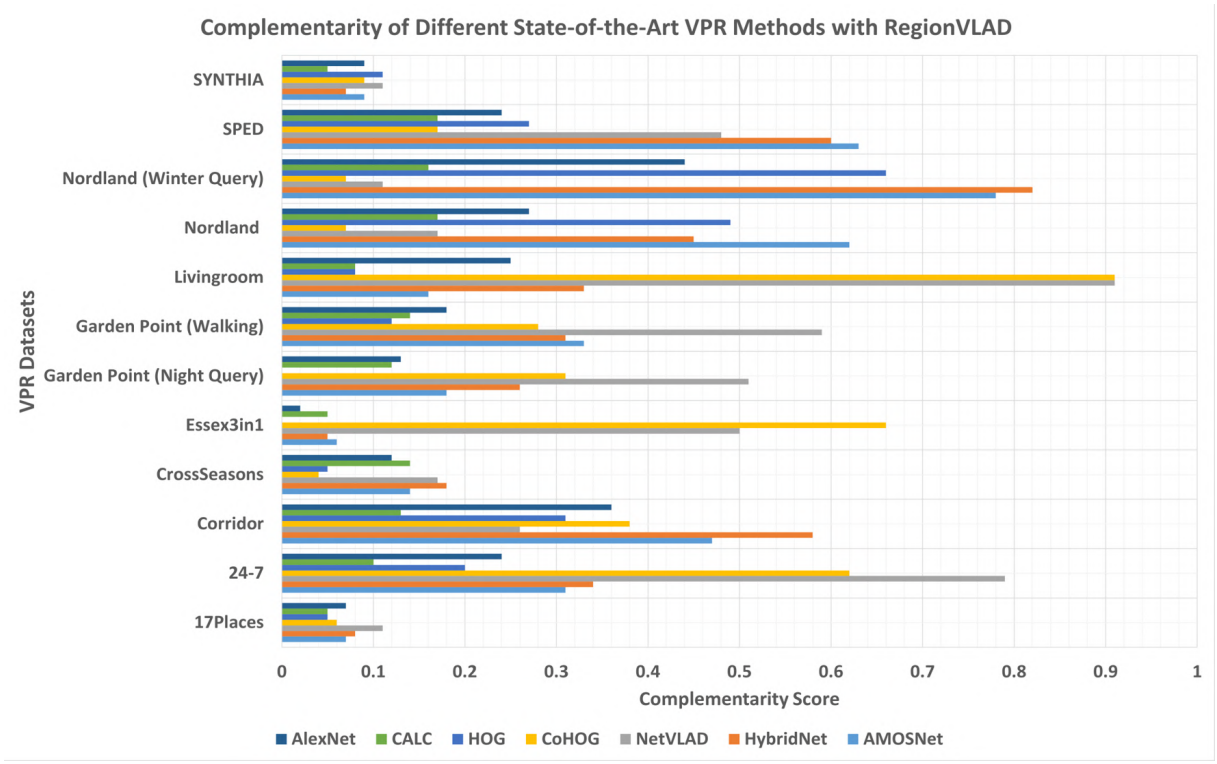

Fig. A.5. Complementarity of VPR methods with RegionVLAD on Multiple VPR datasets.

complementarity independently of individual performance metrics.

As we continue our analysis, Figure 3.6 highlights CoHoG's complementarity with HybridNet, AMOSNet, and NetVLAD, while showing that CALC is the least favourable option. This pattern is further reinforced in Figure 3.7, where HoG demonstrates strong complementarity with CoHoG, NetVLAD, HybridNet, and AMOSNet.

HybridNet, as shown in Figure 3.8, achieves high complementarity with CoHoG and NetVLAD, particularly on the 24-7, Essex3in1, and Livingroom datasets. However, combinations with AMOSNet and CALC generally have lower scores. Figure 3.9 illustrates that NetVLAD pairs well with HybridNet and AMOSNet, especially on the Livingroom and Nordland datasets, while RegionVLAD and CALC are less suitable partners.

Lastly, Figure 3.10 indicates that RegionVLAD complements well with CoHoG, NetVLAD,

and HybridNet, but not with CALC and AlexNet. Carefully considering these results, it can be concluded that NetVLAD is the most suitable VPR technique for forming viable combinations with other methods, while CALC is the least favourable due to its consistently low complementarity scores.

The complementarity levels are also presented in the form of radar charts (Figure 3.12), representing the lower and upper bounds of complementarity of each VPR technique with all other methods. These charts provide a holistic view of how much the complementarity levels vary among different techniques. For example, combinations with AlexNet show the largest upper bounds with NetVLAD, RegionVLAD, and HybridNet, while CALC has the smallest bounds. AMOSNet combinations exhibit the highest upper bounds with NetVLAD and CoHoG, whereas HybridNet and CALC have the smallest bounds.

Lastly, the research provides numerical data for an accurate estimation of the maximum achievable VPR performance using different complementarity pairs (Table 3.2). These results highlight the significant improvements that can be achieved by maximizing complementarity, offering insights into the improved selection of VPR techniques for ensemble setups. By analyzing these insights, it is evident that certain combinations, such as CALC with NetVLAD, can yield high complementarity and thus noticeably improve VPR performance, even if one of the techniques performs poorly on its own.

Appendix B

Detailed Performance Analysis of SwitchHit

This section presents the results generated by utilizing the SwitchHit framework over various standard VPR datasets. Figures 4.4 to 4.11 depict the switching patterns and performance improvements of different VPR method combinations across these datasets, allowing for a visual analysis of how SwitchHit enhances the performance of VPR techniques. The detailed numerical results and switching patterns, including the exact number of correctly matched images for each combination of VPR techniques across all datasets, are provided below. These results emphasize the substantial performance improvements achieved by SwitchHit through intelligent switching, demonstrating its effectiveness in enhancing VPR system performance beyond the capabilities of individual techniques.

Corridor Figure 4.4 presents the results for the Corridor dataset where the three combinations tested all present a varied switching pattern for the dataset. The three

combinations used are CALC, HoG, and NetVLAD; CoHoG, HybridNet, and CALC; and NetVLAD, AMOSNet, and CoHoG. All three combinations have varying switching patterns, correctly matching an average of three to four more images than any individual VPR technique.

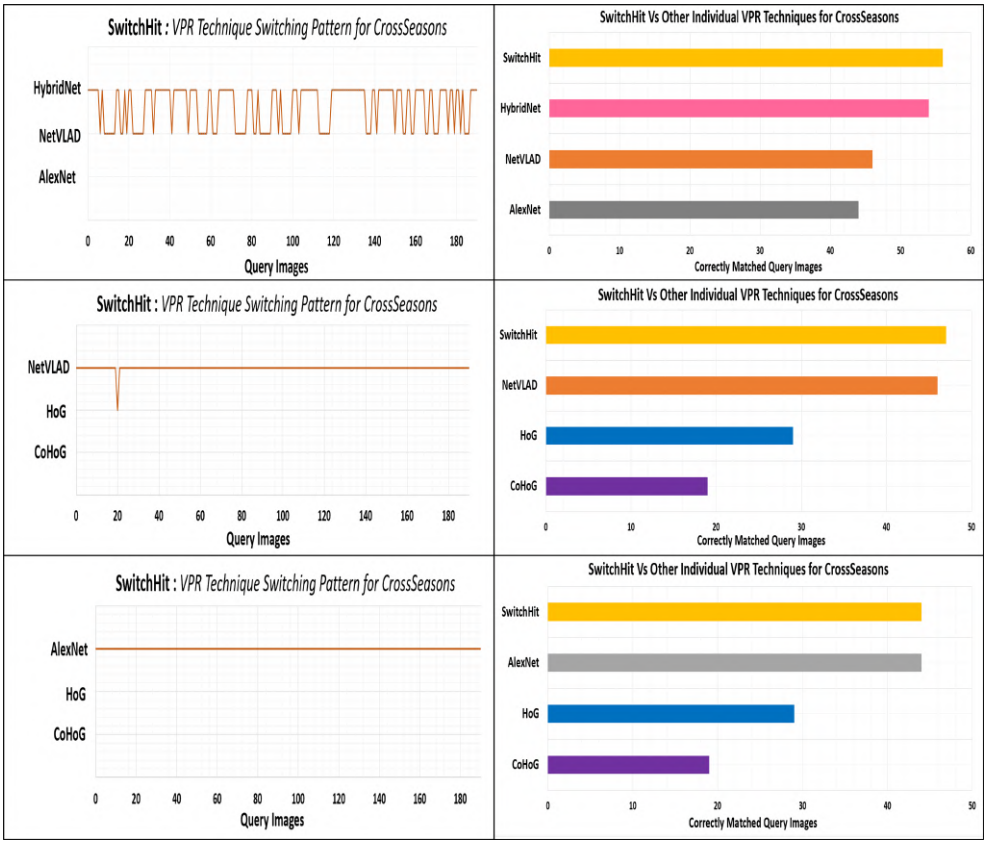

Fig. B.1. Switching patterns and total Number of correct matches for CrossSeasons dataset.

ESSEX3IN1 Figure 4.5 shows the results for the ESSEX3IN1 dataset where the first combination given to SwitchHit contains CALC, CoHoG, and HybridNet. CALC delivers the worst performance individually while CoHoG has the highest performance as a standalone technique. SwitchHit correctly matches four to five more images than CoHoG, thus outranking the best state-of-the-art option available otherwise.

Livingroom In Figure 4.6, the results for the Livingroom dataset reveal that SwitchHit

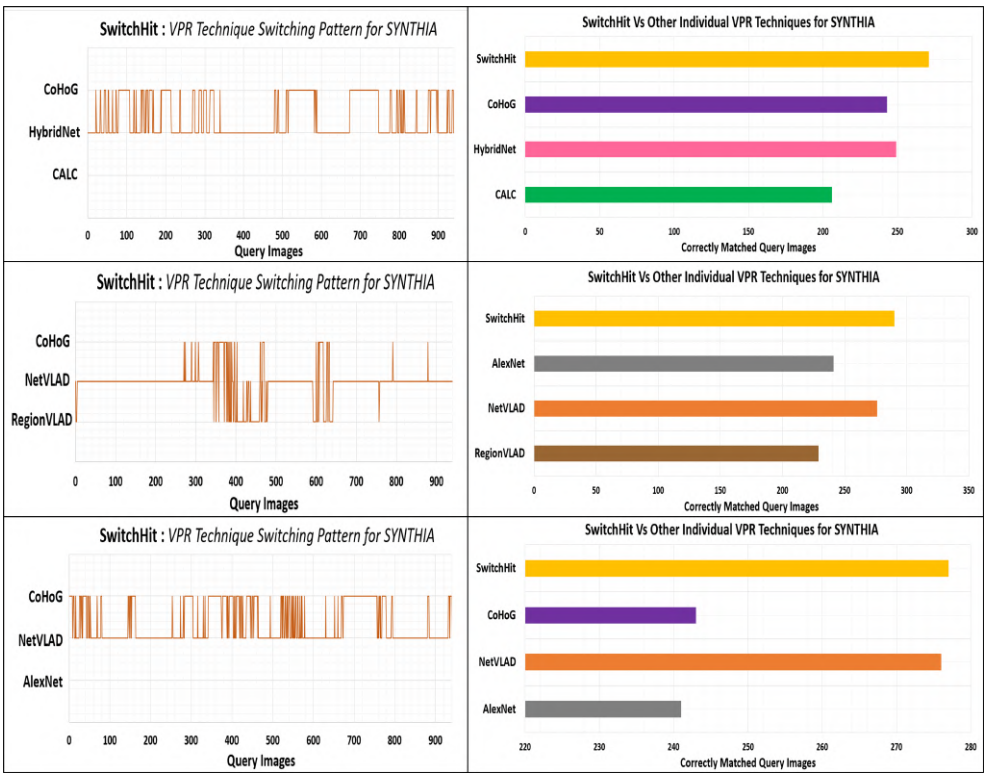

Fig. B.2. Switching patterns and total Number of correct matches for SYNTHIA dataset.

improves performance by two images while switching between AMOSNet and NetVLAD. The second combination contains AlexNet, NetVLAD, and RegionVLAD. This combination, which contains NetVLAD, known for its high performance on the Livingroom dataset, is outperformed by SwitchHit by matching three more images correctly. The last combination tested on the Livingroom dataset is CALC, CoHoG, and AlexNet, which are generally not the best VPR techniques for this dataset. Yet, SwitchHit improves performance by four images and matches NetVLAD's performance. **CrossSeasons** Figure 4.7 illustrates the results for the CrossSeasons dataset. In the first experiment, SwitchHit shifts between NetVLAD and HybridNet constantly and matches two more images correctly than HybridNet. The next combination switches only once from NetVLAD to HoG, leading to one more image correctly matched. The last combination for CrossSeasons

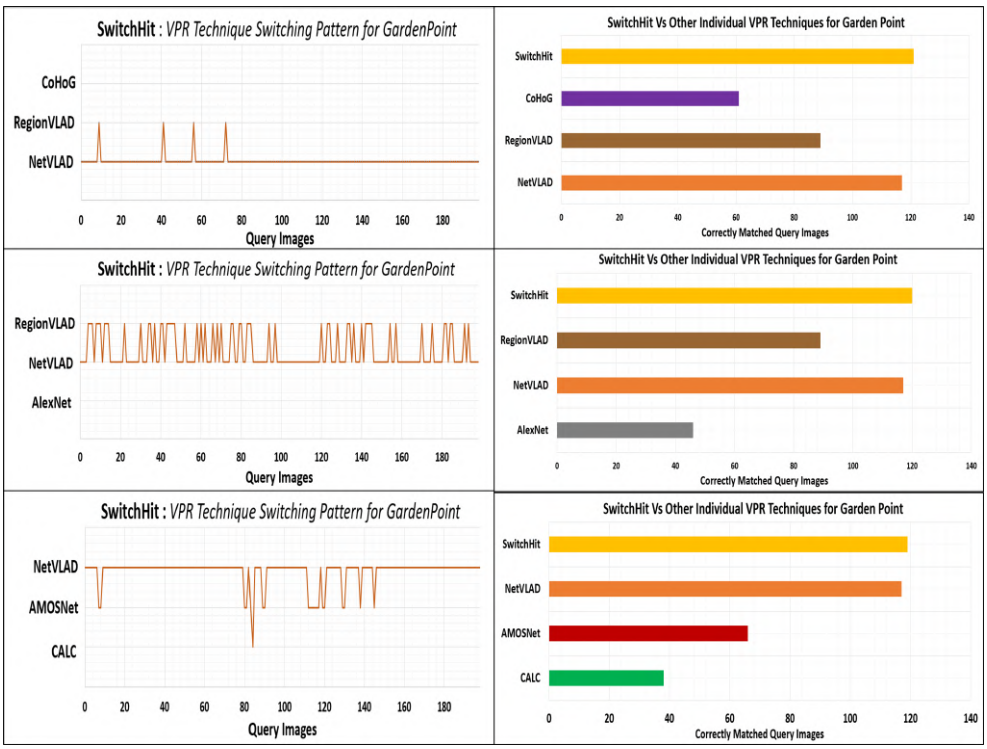

Fig. B.3. Switching patterns and total Number of correct matches for GardensPoint dataset.

presents a unique result as SwitchHit makes no switches at all and remains constantly on AlexNet, producing the same result as AlexNet, the best VPR technique available.

SYNTHIA The SYNTHIA dataset results in Figure 4.8 show that the first combination of CALC, HybridNet, and CoHoG results in 12 more correctly matched images than HybridNet. The next combination of RegionVLAD, NetVLAD, and AlexNet leads to SwitchHit switching between all three techniques, resulting in ten more correctly matched images than NetVLAD. The last combination tested for SYNTHIA correctly matches two more images than the highest performing individual VPR technique.

GardensPoint Figure 4.9 shows the results for the GardensPoint dataset. The first combination consists of NetVLAD, RegionVLAD, and CoHoG, making four successful switches from NetVLAD to RegionVLAD. The next combination of AlexNet, NetVLAD, and

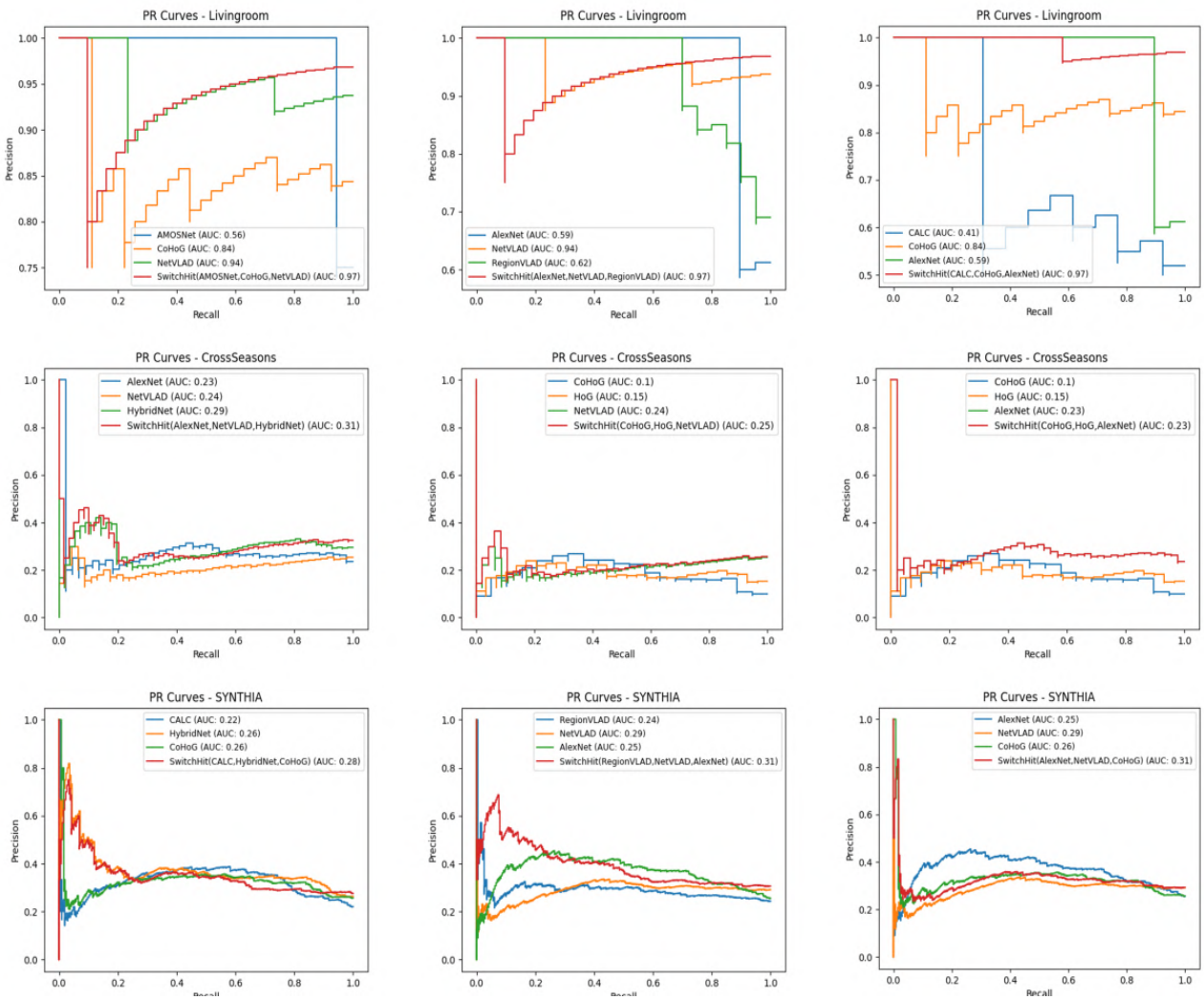

Fig. B.4. PR curves showcasing SwitchHit's performance on Livingroom, CrossSeasons and SYNTHIA datasets versus other VPR techniques.

RegionVLAD results in SwitchHit mostly shifting between RegionVLAD and NetVLAD, matching three more images correctly than the highest performing VPR technique present. The last combination between CALC, AMOSNet, and NetVLAD switches between all three options and matches two more images correctly.

PR Curves Analysis Figure 4.10 presents the PR-curves for the datasets SwitchHit is tested on, starting from Corridor, ESSEX3IN1, and GardensPoint datasets tested for

three different SwitchHit scenarios. For Corridor, SwitchHit manages to outperform all individual VPR techniques available. The results for ESSEX3IN1 and GardensPoint datasets show that SwitchHit performs better than any individual VPR technique, including CoHoG and NetVLAD, the highest performing VPR techniques for the datasets.

Figure 4.11 presents the PR curves for the Livingroom dataset, showing that SwitchHit outperforms NetVLAD with an AUC of 0.97 in both cases tested. For Cross-Seasons, SwitchHit outperforms other individual VPR techniques in the first two cases. For SYNTHIA, SwitchHit performs better than any individual VPR technique in all three cases tested, demonstrating its ability to improve overall performance by utilizing complementary VPR techniques.

Bibliography

[1] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "Netvlad: Cnn architecture for weakly supervised place recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 5297–5307.

[2] S. Skrede, "Nordlandsbanen: minute by minute, season by season," 2013.

[3] A. Glover, "Gardens point walking dataset," *wiki. qut. edu. au/display/cyphy/Open+ datasets+ and+ software*, 2014.

[4] M. Larsson, E. Stenborg, L. Hammarstrand, M. Pollefeys, T. Sattler, and F. Kahl, "A cross-season correspondence dataset for robust semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9532–9542.

[5] M. Zaffar, S. Ehsan, M. Milford, and K. D. McDonald-Maier, "Memorable maps: A framework for re-defining places in visual place recognition," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 12, pp. 7355–7369, 2020.

[6] 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.

[7] 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.

[8] N. Sünderhauf, S. Shirazi, F. Dayoub, B. Upcroft, and M. Milford, "On the performance of convnet features for place recognition," in *2015 IEEE/RSJ international conference on intelligent robots and systems (IROS)*. IEEE, 2015, pp. 4297–4304.

[9] 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.

[10] M. Zaffar, "Visual place recognition for autonomous robots," Ph.D. dissertation, University of Essex, 2020.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] S. Schubert, P. Neubert, S. Garg, M. Milford, and T. Fischer, "Visual place recognition: A tutorial," 2023.

[16] L. G. C. P. U. J. N. Tiago Barros, Ricardo Pereira, "Place recognition survey: An update on deep learning approaches," *ArXiv*, 2022.

[17] P. Yin, S. Zhao, I. Cisneros, A. Abuduweili, G. Huang, M. J. Milford, C. Liu, H. Choset, and S. A. Scherer, "General place recognition survey: Towards the real-world autonomy age." *CoRR*, vol. abs/2209.04497, 2022. [Online]. Available: http://dblp.uni-trier.de/db/journals/corr/corr2209.html#abs-2209-04497

[18] C. Masone and B. Caputo, "A survey on deep visual place recognition," *IEEE Access*, 2021.

[19] X. Zhang, L. Wang, and Y. Su, "Visual place recognition: A survey from deep learning perspective," *Pattern Recognition*, 2021.

[20] S. Garg, T. Fischer, and M. Milford, "Where is your place, visual place recognition?" *ArXiv*, 2021.

[21] X. Zhang, L. Wang, and Y. Su, "Visual place recognition: A survey from deep learning perspective," *Pattern Recognition*, vol. 113, p. 107760, 2021.

[22] S. Garg, T. Fischer, and M. Milford, "Where is your place, visual place recognition?" in *IJCAI*, vol. 8, 2021, pp. 4416–4425.

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

[24] P. Yin, S. Zhao, I. Cisneros, A. Abuduweili, G. Huang, M. Milford, C. Liu, H. Choset, and S. Scherer, "General place recognition survey: Towards the real-world autonomy age," *arXiv preprint arXiv:2209.04497*, 2022.

[25] 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.

[26] M. J. Milford and G. F. Wyeth, "Seqslam: Visual route-based navigation for sunny summer days and stormy winter nights," in *2012 IEEE international conference on robotics and automation*. IEEE, 2012, pp. 1643–1649.

[27] 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.

[28] P. Neubert, N. Sünderhauf, and P. Protzel, "Superpixel-based appearance change prediction for long-term navigation across seasons," *Robotics and Autonomous Systems*, vol. 69, pp. 15–27, 2015.

[29] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, pp. 91–110, 2004.

[30] A. Torralba and A. A. Efros, "Unbiased look at dataset bias," in *CVPR 2011*. IEEE, 2011, pp. 1521–1528.

[31] S. Hausler, A. Jacobson, and M. Milford, "Multi-process fusion: Visual place recognition using multiple image processing methods," *IEEE Robotics and Automation Letters*, vol. 4, no. 2, pp. 1924–1931, 2019.

[32] S. Lynen, M. W. Achtelik, S. Weiss, M. Chli, and R. Siegwart, "A robust and modular multi-sensor fusion approach applied to mav navigation," in *2013 IEEE/RSJ international conference on intelligent robots and systems*. IEEE, 2013, pp. 3923–3929.

[33] Q. Li, J. P. Queralta, T. N. Gia, Z. Zou, and T. Westerlund, "Multi-sensor fusion for navigation and mapping in autonomous vehicles: Accurate localization in urban environments," *Unmanned Systems*, vol. 8, no. 03, pp. 229–237, 2020.

[34] S. Hausler and M. Milford, "Hierarchical multi-process fusion for visual place recognition," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 3327–3333.

[35] M. Brown, D. G. Lowe *et al.*, "Recognising panoramas." in *ICCV*, vol. 3, 2003, p. 1218.

[36] M. Brown and D. G. Lowe, "Automatic panoramic image stitching using invariant features," *International journal of computer vision*, vol. 74, pp. 59–73, 2007.

[37] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proceedings of the seventh IEEE international conference on computer vision*, vol. 2. Ieee, 1999, pp. 1150–1157.

[38] T. Tuytelaars and L. Van Gool, "Content-based image retrieval based on local affinely invariant regions," in *Visual Information and Information Systems: Third*

International Conference, VISUAL'99 Amsterdam, The Netherlands, June 2–4, 1999 Proceedings 3. Springer, 1999, pp. 493–500.

[39] Sivic and Zisserman, "Video google: A text retrieval approach to object matching in videos," in *Proceedings ninth IEEE international conference on computer vision*. IEEE, 2003, pp. 1470–1477.

[40] J. Knopp, J. Sivic, and T. Pajdla, "Avoiding confusing features in place recognition," in *Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part I 11*. Springer, 2010, pp. 748–761.

[41] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," in *Workshop on statistical learning in computer vision, ECCV*, vol. 1, no. 1-22. Prague, 2004, pp. 1–2.

[42] Dorko and Schmid, "Selection of scale-invariant parts for object class recognition," in *Proceedings Ninth IEEE International Conference on Computer Vision*. IEEE, 2003, pp. 634–639.

[43] R. Fergus, P. Perona, and A. Zisserman, "Object class recognition by unsupervised scale-invariant learning," in *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings*, vol. 2. IEEE, June 2003, pp. II–II.

[44] A. Opelt, M. Fussenegger, A. Pinz, and P. Auer, "Weak hypotheses and boosting for generic object detection and recognition," in *Computer Vision-ECCV 2004:*

8th European Conference on Computer Vision, vol. 8. Prague, Czech Republic: Springer Berlin Heidelberg, May 2004, pp. 71–84, proceedings, Part II.

[45] S. Se, D. Lowe, and J. Little, "Mobile robot localization and mapping with uncertainty using scale-invariant visual landmarks," *The International Journal of Robotics Research*, vol. 21, no. 8, pp. 735–758, 2002.

[46] F. Schaffalitzky and A. Zisserman, "Automated location matching in movies," *Computer Vision and Image Understanding*, vol. 92, no. 2, pp. 236–264, 2003.

[47] S. Lazebnik, C. Schmid, and J. Ponce, "A sparse texture representation using affine-invariant regions," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Madison, Wisconsin, USA, 2003, pp. 319–324.

[48] ——, "Affine-invariant local descriptors and neighborhood statistics for texture recognition," in *Proc. The 9th IEEE International Conference on Computer Vision*, Nice, France, 2003, pp. 649–655.

[49] C. C. Wang and K. C. Wang, "Hand posture recognition using adaboost with sift for human robot interaction," in *Recent Progress in Robotics: Viable Robotic Service to Human*, 2008, pp. 317–329.

[50] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," in *Computer Vision–ECCV 2006: 9th European Conference on Computer Vision, Graz, Austria, May 7-13, 2006. Proceedings, Part I 9*. Springer, 2006, pp. 404–417.

[51] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "Orb: An efficient alternative to sift or surf," in *2011 International conference on computer vision*. Ieee, 2011, pp. 2564–2571.

[52] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *International journal of computer vision*, vol. 42, pp. 145–175, 2001.

[53] A. C. Murillo and J. Kosecka, "Experiments in place recognition using gist panoramas," in *2009 IEEE 12Th international conference on computer vision workshops, ICCV workshops*. IEEE, 2009, pp. 2196–2203.

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

[55] G. Singh and J. Kosecka, "Visual loop closing using gist descriptors in manhattan world," in *ICRA omnidirectional vision workshop*, 2010, pp. 4042–4047.

[56] Y. Liu and H. Zhang, "Visual loop closure detection with a compact image descriptor," in *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2012, pp. 1051–1056.

[57] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, vol. 1. Ieee, 2005, pp. 886–893.

[58] C. McManus, B. Upcroft, and P. Newman, "Scene signatures: Localised and pointless features for localisation," *Robotics: Science and Systems X*, pp. 1–9, 2014.

[59] H. Jégou, M. Douze, C. Schmid, and P. Pérez, "Aggregating local descriptors into a compact image representation," in *2010 IEEE computer society conference on computer vision and pattern recognition*. IEEE, 2010, pp. 3304–3311.

[60] J. Yue-Hei Ng, F. Yang, and L. S. Davis, "Exploiting local features from deep networks for image retrieval," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2015, pp. 53–61.

[61] N. Sünderhauf, S. Shirazi, A. Jacobson, F. Dayoub, E. Pepperell, B. Upcroft, and M. Milford, "Place recognition with convnet landmarks: Viewpoint-robust, condition-robust, training-free," *Robotics: Science and Systems XI*, pp. 1–10, 2015.

[62] Y. Kong, W. Liu, and Z. Chen, "Robust convnet landmark-based visual place recognition by optimizing landmark matching," *IEEE Access*, vol. 7, pp. 30 754–30 767, 2019.

[63] 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*, vol. 36, no. 2, pp. 561–569, 2019.

[64] S. Senthamizhselvi and A. Saravanan, "Intelligent visual place recognition using sparrow search algorithm with deep transfer learning model," *International Journal of Engineering Trends and Technology*, vol. 71, no. 4, pp. 109–118, 2023.

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

[66] M. A. Uy and G. H. Lee, "Pointnetvlad: Deep point cloud based retrieval for large-scale place recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4470–4479.

[67] S. S. Kannan and B.-C. Min, "Placeformer: Transformer-based visual place recognition using multi-scale patch selection and fusion," *arXiv preprint arXiv:2401.13082,* 2024.

[68] S. Zhu, L. Yang, C. Chen, M. Shah, X. Shen, and H. Wang, "R2former: Unified retrieval and reranking transformer for place recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 19 370–19 380.

[69] R. Wang, Y. Shen, W. Zuo, S. Zhou, and N. Zheng, "Transvpr: Transformer-based place recognition with multi-level attention aggregation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 13 648–13 657.

[70] A. Ali-Bey, B. Chaib-Draa, and P. Giguere, "Mixvpr: Feature mixing for visual place recognition," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 2998–3007.

[71] G. Berton, C. Masone, and B. Caputo, "Rethinking visual geo-localization for large-scale applications," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4878–4888.

[72] Y. Latif, R. Garg, M. Milford, and I. Reid, "Addressing challenging place recognition tasks using generative adversarial networks," in *2018 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2018, pp. 2349–2355.

[73] P. Yin, L. Xu, Z. Liu, L. Li, H. Salman, Y. He, W. Xu, H. Wang, and H. Choset, "Stabilize an unsupervised feature learning for lidar-based place recognition," in

2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2018, pp. 1162–1167.

[74] N. Merrill and G. Huang, "Lightweight unsupervised deep loop closure," *arXiv preprint arXiv:1805.07703*, 2018.

[75] Z. Wang, J. Li, S. Khademi, and J. van Gemert, "Attention-aware age-agnostic visual place recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019, pp. 0–0.

[76] R. Dube, A. Cramariuc, D. Dugas, H. Sommer, M. Dymczyk, J. Nieto, R. Siegwart, and C. Cadena, "Segmap: Segment-based mapping and localization using data-driven descriptors," *The International Journal of Robotics Research*, vol. 39, no. 2-3, pp. 339–355, 2020.

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

[78] C. Masone and B. Caputo, "A survey on deep visual place recognition," *IEEE Access*, vol. 9, pp. 19 516–19 547, 2021.

[79] R. Mereu, G. Trivigno, G. Berton, C. Masone, and B. Caputo, "Learning sequential descriptors for sequence-based visual place recognition," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 10 383–10 390, 2022.

[80] S. Garg and M. Milford, "Seqnet: Learning descriptors for sequence-based hierarchical place recognition," *IEEE Robotics and Automation Letters*, vol. 6, no. 3, pp. 4305–4312, 2021.

[81] F. Fu, J. Yang, J. Zhang, and J. Ma, "Matc-net: Learning compact sequence representation for hierarchical loop closure detection," *Engineering Applications of Artificial Intelligence*, vol. 125, p. 106734, 2023.

[82] J. Zhao, F. Zhang, Y. Cai, G. Tian, W. Mu, C. Ye, and T. Feng, "Learning sequence descriptor based on spatio-temporal attention for visual place recognition," *IEEE Robotics and Automation Letters*, 2024.

[83] S. Schubert, P. Neubert, and P. Protzel, "Fast and memory efficient graph optimization via icm for visual place recognition," in *Robotics: Science and Systems (RSS)*, 07 2021.

[84] F. Zhang, J. Zhao, Y. Cai, G. Tian, W. Mu, and C. Ye, "Learning sequence descriptor based on spatio-temporal attention for visual place recognition," *ArXiv*, 2023.

[85] R. Mereu, G. Trivigno, G. Berton, C. Masone, and B. Caputo, "Learning sequential descriptors for sequence-based visual place recognition," *IEEE Robotics and Automation Letters*, 2022.

[86] M.-A. Tomiță, M. Zaffar, B. Ferrarini, M. J. Milford, K. D. McDonald-Maier, and S. Ehsan, "Sequence-based filtering for visual route-based navigation: Analyzing the benefits, trade-offs and design choices," *IEEE Access*, vol. 10, pp. 81 974–81 987, 2022.

[87] S. Thrun, "Probabilistic robotics," *Communications of the ACM*, vol. 45, no. 3, pp. 52–57, 2002.

[88] 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 Transactions on robotics*, vol. 32, no. 6, pp. 1309–1332, 2016.

[89] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[90] S. Ehsan, N. Kanwal, A. F. Clark, and K. D. McDonald-Maier, "Measuring the coverage of interest point detectors," in *Image Analysis and Recognition: 8th International Conference, ICIAR 2011, Burnaby, BC, Canada, June 22-24, 2011. Proceedings, Part I 8*. Springer, 2011, pp. 253–261.

[91] B. Ferrarini, S. Ehsan, A. Leonardis, N. U. Rehman, and K. D. McDonald-Maier, "Performance characterization of image feature detectors in relation to the scene content utilizing a large image database," *IEEE Access*, vol. 6, pp. 8564–8573, 2018.

[92] S. Ehsan, A. F. Clark, B. Ferrarini, N. U. Rehman, and K. D. McDonald-Maier, "Assessing the performance bounds of local feature detectors: Taking inspiration from electronics design practices," in *2015 International Conference on Systems, Signals and Image Processing (IWSSIP)*. IEEE, 2015, pp. 166–169.

[93] S. Ehsan, A. F. Clark, and K. D. McDonald-Maier, "Rapid online analysis of local feature detectors and their complementarity," *Sensors*, vol. 13, no. 8, pp. 10 876–10 907, 2013.

[94] C. Malone, S. Hausler, T. Fischer, and M. Milford, "Boosting performance of a baseline visual place recognition technique by predicting the maximally comple-

mentary technique," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 1919–1925.

[95] J. Davis and M. Goadrich, "The relationship between precision-recall and roc curves," in *Proceedings of the 23rd international conference on Machine learning*. ACM, 2006, pp. 233–240.

[96] B. Arcanjo, B. Ferrarini, M. Milford, K. D. McDonald-Maier, and S. Ehsan, "An efficient and scalable collection of fly-inspired voting units for visual place recognition in changing environments," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 2527–2534, 2022.

[97] 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.

[98] B. Ferrarini, M. Waheed, S. Waheed, S. Ehsan, M. Milford, and K. D. McDonald-Maier, "Visual place recognition for aerial robotics: Exploring accuracy-computation trade-off for local image descriptors," in *2019 NASA/ESA Conference on Adaptive Hardware and Systems (AHS)*. IEEE, 2019, pp. 103–108.

[99] B. Ferrarini, S. Ehsan, A. Bartoli, A. Leonardis, and K. D. McDonald-Maier, "Assessing capsule networks with biased data," in *Scandinavian Conference on Image Analysis*. Springer, 2019, pp. 90–100.

[100] B. Ferrarini, M. J. Milford, K. D. McDonald-Maier, and S. Ehsan, "Binary neural networks for memory-efficient and effective visual place recognition in changing

environments," *IEEE Transactions on Robotics*, vol. 38, no. 4, pp. 2617–2631, 2022.

[101] A. Torii, R. Arandjelovic, J. Sivic, M. Okutomi, and T. Pajdla, "24/7 place recognition by view synthesis," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1808–1817.

[102] A. Torii, R. Arandjelović, J. Sivic, M. Okutomi, and y. T. Pajdla, journal=IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), "24/7 place recognition by view synthesis."

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

[104] G. Berton, C. Masone, and B. Caputo, "Rethinking visual geo-localization for large-scale applications," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[105] A. Ali-bey, B. Chaib-draa, and P. Giguere, "Gsv-cities: Toward appropriate supervised visual place recognition," *Neurocomputing*, 2022.

[106] B. Yildiz, S. Khademi, R. Siebes, and J. V. Gemert, "Amstertime: A visual place recognition benchmark dataset for severe domain shift," in *International Conference on Pattern Recognition (ICPR)*, 2022.

[107] G. Berton, V. Paolicelli, C. Masone, and B. Caputo, "Adaptive-attentive geolocalization from few queries: A hybrid approach," in *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2021.

[108] T. Weyand, A. Ara'ujo, B. Cao, and J. Sim, "Google landmarks dataset v2 – a large-scale benchmark for instance-level recognition and retrieval," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[109] F. Warburg, S. Hauberg, M. Lopez-Antequera, P. Gargallo, Y. Kuang, and J. Civera, "Mapillary street-level sequences: A dataset for lifelong place recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[110] T. Sattler, W. Maddern, C. Toft, A. Torii, L. Hammarstrand, E. Stenborg, D. Safari, M. Okutomi, M. Pollefeys, J. Sivic, F. Kahl, and T. Pajdla, "Benchmarking 6dof outdoor visual localization in changing conditions," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[111] W. Maddern, G. Pascoe, C. Linegar, and P. Newman, "1 year, 1000km: The oxford robotcar dataset," *The International Journal of Robotics Research*, 2017.

[112] N. Carlevaris-Bianco, A. K. Ushani, and R. M. Eustice, "University of michigan north campus long-term vision and lidar dataset," *The International Journal of Robotics Research*, 2016.

[113] A. Zamir and M. Shah, "Image geo-localization based on multiple nearest neighbor feature matching using generalized graphs," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 2014.

[114] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *The International Journal of Robotics Research*, 2013.

[115] T. Sattler, T. Weyand, B. Leibe, and L. Kobbelt, "Image retrieval for image-based localization revisited," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[116] D. M. Chen, G. Baatz, K. Köser, S. S. Tsai, R. Vedantham, T. Pylvänäinen, K. Roimela, X. Chen, J. Bach, M. Pollefeys, B. Girod, and y. R. Grzeszczuk, booktitle=IEEE Conference on Computer Vision and Pattern Recognition (CVPR), "City-scale landmark identification on mobile devices."

[117] J. Knopp, J. Sivic, and T. Pajdla, "Avoiding confusing features in place recognition," in *European Conference on Computer Vision (ECCV)*, 2010.

[118] M. Cummins and P. Newman, "Highly scalable appearance-only slam - fab-map 2.0," in *Robotics: Science and Systems (RSS)*, 2009.

[119] M. Milford and G. Wyeth, "Mapping a suburb with a single camera using a biologically inspired slam system," *IEEE Transactions on Robotics*, 2008.

[120] P. Neubert and S. Schubert, "Hyperdimensional computing as a framework for systematic aggregation of image descriptors," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 16 938–16 947.

[121] N. Sünderhauf, P. Neubert, and P. Protzel, "Are we there yet? challenging seqslam on a 3000 km journey across all four seasons," in *Proc. of workshop on long-term autonomy, IEEE international conference on robotics and automation (ICRA)*, 2013, p. 2013.

[122] Q. McNemar, "Note on the sampling error of the difference between correlated proportions or percentages," *Psychometrika*, vol. 12, no. 2, pp. 153–157, 1947.

[123] J. L. Fleiss, B. Levin, and M. C. Paik, *Statistical methods for rates and proportions*. john wiley & sons, 2013.

[124] H. Zhang, F. Han, and H. Wang, "Robust multimodal sequence-based loop closure detection via structured sparsity." in *Robotics: Science and systems*, 2016.

[125] M. Zaffar, S. Garg, M. Milford, J. Kooij, D. Flynn, K. McDonald-Maier, and S. Ehsan, "Vpr-bench: An open-source visual place recognition evaluation framework with quantifiable viewpoint and appearance change," *International Journal of Computer Vision (IJCV)*, 2021.

[126] 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.

[127] Z. Chen, A. Jacobson, N. Sünderhauf, B. Upcroft, L. Liu, C. Shen, I. Reid, and M. Milford, "Deep learning features at scale for visual place recognition," in *2017 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2017, pp. 3223–3230.

[128] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.

[129] Y. Zhou, W. Ping, S. Arik, K. Peng, and G. Diamos, "Hybridnet: A hybrid neural architecture to speed-up autoregressive models," 2018.

[130] M. Zaffar, S. Ehsan, M. Milford, and K. McDonald-Maier, "Cohog: A light-weight, compute-efficient, and training-free visual place recognition technique for chan-

ging environments," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 1835–1842, 2020.

[131] M. Waheed, M. Milford, K. McDonald-Maier, and S. Ehsan, "Switchhit: A probabilistic, complementarity-based switching system for improved visual place recognition in changing environments," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 7833–7840.

[132] M. Gehrig, E. Stumm, T. Hinzmann, and R. Siegwart, "Visual place recognition with probabilistic voting," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2017, pp. 3192–3199.

[133] W. Kulachai, U. Lerdtomornsakul, and P. Homyamyen, "Factors influencing voting decision: a comprehensive literature review," *Social Sciences*, vol. 12, no. 9, p. 469, 2023.

[134] D. G. Saari, "Selecting a voting method: the case for the borda count," *Constitutional Political Economy*, vol. 34, no. 3, pp. 357–366, 2023.

[135] T. Fischer and M. Milford, "Event-based visual place recognition with ensembles of temporal windows," *IEEE Robotics and Automation Letters*, vol. 5, no. 4, pp. 6924–6931, 2020.

[136] L. Huang, C. Chen, J. Yun, Y. Sun, J. Tian, Z. Hao, H. Yu, and H. Ma, "Multi-scale feature fusion convolutional neural network for indoor small target detection," *Frontiers in Neurorobotics*, vol. 16, p. 881021, 2022.

[137] E. Bingham and H. Mannila, "Random projection in dimensionality reduction: applications to image and text data," in *Proceedings of the seventh ACM SIGKDD*

international conference on Knowledge discovery and data mining, 2001, pp. 245–250.

[138] S. Ehsan, A. F. Clark, A. Leonardis, N. Ur Rehman, A. Khaliq, M. Fasli, and K. D. McDonald-Maier, "A generic framework for assessing the performance bounds of image feature detectors," *Remote Sensing*, vol. 8, no. 11, p. 928, 2016.