AN EFFECTIVE SCENE RECOGNITION STRATEGY FOR BIOMIMETIC ROBOTIC NAVIGATION

TEO CHING LIK

(B. ENG(Hons.), National University of Singapore)

A THESIS SUBMITTED

FOR THE DEGREE OF MASTER OF ENGINEERING

DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING

NATIONAL UNIVERSITY OF SINGAPORE

2007

Acknowledgements

This thesis would not have been possible without the guidance of my supervisor, Dr. Cheong Loong Fah. His foresight, enthusiasm and constructive criticisms created the environment that motivated me to produce the results presented here. I would like to say a big THANK YOU to you Sir, for your help and encouragement when the going got tough and I look forward to working with you soon.

I would like to take this opportunity to thank all my friends and colleagues who offered their help in whatever ways that will make me look back at my postgraduate days fondly. Thank you Daniel for the illuminating discussions we always have; lots of thanks to Chern-Horng for his advice and help with the cameras and finally to Hsiao Piau for his jokes and concern that made the lab a better place to work in. I thank Francis, the lab officer for his technical help and patience when I am

late with returning the hardware as I needed more time for experiments. Finally I thank my brother in Christ, Zachary, as well as Shimiao, Wen Cong and Daniel, for all their help in proof reading an initial draft of this thesis, God bless you all in your research too.

I thank my family for their love and support and I reserve my final words to Shujing: sorry for ignoring you at times when I am so busy with this work, and yet you are always there for me with your kind words and patience. May this thesis bear testimony to the sacrifice you have made for me.

TEO Ching Lik

29/12/2006

Contents

Summary

This thesis presents a novel Scene Recognition Strategy (SRS) suitable for biomimetic navigation. The proposed SRS decomposes the scene recognition problem into two phases. In the first phase, the scene in question is encoded into memory by an automatic selection of salient landmarks. The choice of these landmarks follows a modified computational model of human visual saliency to obtain initial salient regions of interest (ROIs) in the scene. These regions are then encoded using SURF (Speeded-Up Robust Features) keypoint descriptors over three colour spaces - grayscale, saturation and hue to enhance the robustness of the SRS against illumination changes. The SURF descriptors are then augmented with ordinal depth information obtained from optic flow arising from a specialised form of motion known as the Turn-Back-and-Look (TBL) flight, performed by certain species of

bees and wasps. The use of ordinal depth together with the spatial configuration information of these salient-SURF keypoints improves the robustness of the SRS against viewpoint changes. A set of salient-SURF descriptors in one colour space constitutes the Scene matrix. Combining the three Scene matrices together, one for each colour space, form the Scene matrix cell that completely represents the scene. The second phase is the scene decision phase. Given an input query or test scene, represented by its Scene matrix cell, an effective scene decision module is proposed to rapidly decide if the test scene matches one of the memorised scenes in the reference database using a novel measure of scene similarity known as the Global Configuration Coefficient. The final decision to accept or reject a candidate match is obtained by estimating an adaptive decision threshold from the statistics of the matches. Extensive tests and experimental results show that the proposed SRS is accurate even for challenging scenes in both indoor and outdoor environments.

List of Tables

List of Figures

List of Symbols

Ω	Set of finite ordinals	56
S_ρ	Spearman's ρ rank correlation	57
K_τ	Kendall's τ rank correlation	57
$\mathbf{C}_f^j$	Composite features	83
$\mathbf{P}_k^j$	Pyramid image	90
$\mathbf{P}_{diff}^j$	Difference image	91
$\mathbf{C}^j$	Conspicuity map	92
$\mathbf{S}_{dm}$	Depth weighted saliency map	94
$\hat{\mathbf{D}}_{prox}$	Dense ordinal proximity map	94
$\mathbf{L}_m$	Labelled map for salient ROIs	97

d_{ratio}	Distance ratio	104
$\mathbf{m}_{prox}$	Proximity matrix	108
Z_{ord}	Ordinal depth	116
f	Camera focal length	116
ω_y	Rotation in the y axis	116
d_{prox}	Ordinal proximity	119
$\mathbf{D}_{prox}$	Sparse ordinal proximity map	119
$\mathbf{m}_s$	The Scene matrix	121
$\mathbf{M}_s$	Scene cell matrix	123
$\dot{\mathbf{m}}_{kp}^{j}$	Matched salient-SURF keypoints in the j^{th} colour space	130
$\dot{\mathbf{M}}_{kp}$	Matching matrix	131
G_c	Global Configuration Coefficient	132
$\overline{S_\rho}$	Mean of Spearman's ρ in the three spatial directions	132
$\overline{K_\tau}$	Mean of Kendall's τ in the three spatial directions	132
N_{ref}	Number of reference scenes in image database	134
$\mathbb{D}_{ref}$	Reference database	134
$\mathbf{M}_s^{test}$	Test scene matrix cell	134
$\mathbf{\Pi}_s$	Match statistic matrix	135
$\mathbb{G}_{cand}$	Candidate match	136
G_{cand}	Best match score	136

D_t	Scene decision threshold	136
D_f	Final scene decision	136
$\mathbf{\Delta}_s$	Decision matrix	138
$\mathbf{\Xi}_s$	Threshold vector	139
$\mathbf{\Sigma}_{\rho\Delta}$	Composite Spearman's rank correlation matrix	140
$\mathbf{\Lambda}_{\tau\Delta}$	Composite Kendall's rank correlation matrix	140
t_{rank}	Threshold for significant rank correlations	140
D_{min}	Absolute minimum threshold for scene decision	144
D^*_{min}	Modified absolute minimum for decision threshold	146
N_{pos}	Number of positive test scenes	149
N_{neg}	Number of negative test scenes	149
P_{acc}	Positive acceptance rate	155
P_{rej}	Positive rejection rate	155
N_{test}	Number of test scenes	155
N_{iter}	Number of iterations for computing recognition accuracy	156
$P_{overall}$	Overall recognition accuracy	157
G_{cw}	Weighted Global Configuration Coefficient	199
$\overline{S_{\rho w}}$	Weighted mean of Spearman's ρ	199
$\overline{K_{\tau w}}$	Weighted mean of Kendall's τ	199

Chapter 1

General Introduction

This introductory chapter presents the problem of *scene recognition* - its definition; what properties are desired of a scene recognition algorithm; and the main challenges in designing a reliable algorithm to perform scene recognition (sections 1.2–1.4). As scene recognition has important applications in *biomimetics* - an emerging field that uses results from biology to construct working computational models - the implications of this work are highlighted in the context of biomimetic navigation (section 1.1). The scope of this thesis is then defined in section 1.5 together with a brief presentation of its main contributions (section 1.6).

1.1 Biomimetic navigation

Navigation is one of the most fundamental behaviours of animals. Animals have evolved various strategies for effective navigation and this involves the development of abilities such as to recognise a previously visited place. The latter forms an integral component of what is known as the place (or scene) recognition-triggered response [39, 113] in the domain of biomimetic navigation - the animat or biological agent has a set of places in memory that is linked with a learnt set of actions that it must take once it recognises that it has returned to the same place again. By following a sequence of these actions that leads on from one learnt place to the next, the agent successfully navigates from one point to another. This offers a simple, yet elegant solution to the successful navigation of certain insects such as bees [19]. An overview of insect navigation strategies can be found in [23] and more recently in [24].

A reliable scene recognition system is crucial as the place recognition-triggered response strategy, described above is classified as a low-level local navigation strategy in [113] (Fig. 1.1). Each level, starting from *homing* to *metric navigation*, increases in complexity and is built upon the successful implementation of the strategy at the lower levels (*i.e.* before one can implement metric navigation, topological navigation must have been implemented). From Fig. 1.1, the complex navigation strategies such as topological and metric navigation depend on the

successful implementation of the place recognition-triggered response. An effective solution to solving the scene recognition problem will thus pave the way for more high-level strategies to be implemented. Furthermore, low-level navigation is interesting as it is a *common* strategy employed by diverse groups of animals, from humble bees that navigate between their nests and foraging sites to migratory birds that fly across vast continents. Animal behavioural studies and human psychophysical studies of navigation provide a wealth of information in designing a successful biomimetic navigation strategy; and in this thesis, a few of these ideas are used to achieve this goal.

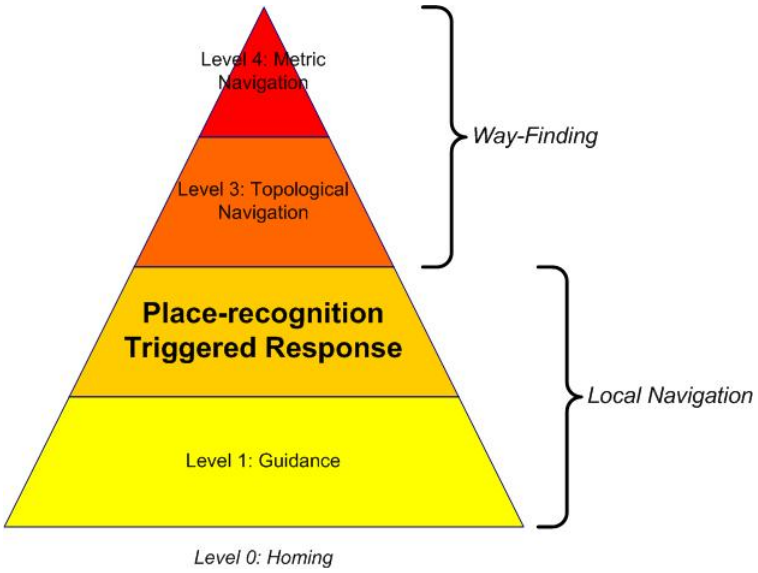

Figure 1.1: 4 level hierarchical organisation of biomimetic navigation.

1.2 Scene recognition

Scene or place recognition is defined as the ability, given an input query (test) image and an image database containing several reference images, to recognise if a match can be found between the test and one (or several) reference image(s). Although this task may seem simple to humans, scene recognition remains in fact one of the most *difficult* problems in computer vision due to the inherent complexity and large variety of scenes that need to be taken into account. One may be able to recognise a previously visited place with ease in the afternoon, even though the place was first visited in the late evening many weeks before under different lighting and weather conditions. How humans (or animals) are able to reliably recognise a scene viewed under very different conditions remains one of the most challenging problems in psychophysics. Modelling this behaviour to achieve a robust and general *scene recognition strategy* (SRS) remains an open question in computer vision. This thesis attempts to use several ideas from computer vision and biomimetics to propose a novel and reliable SRS suitable for robotic navigation.

1.3 Characteristics of a good SRS

A successful SRS on a practical mobile system must possess two important characteristics. Firstly, the strategy must be able to tolerate various types of image distortions for the given test scene and find the correct match in its memory in

spite of the distortions. Common image distortions considered in this thesis are viewpoint and illumination changes (Fig. 1.2(left)) as well as changes in the scene content (Fig. 1.2(right)) itself. This requirement is fundamental as practical systems suffer from wheel slippages and accumulative drift errors such that more often than not, the agent upon returning to a previously visited place is presented with a slightly distorted view of the same scene. In an outdoor environment, the change in the position of the sun, the effect of clouds and the resultant movement of shadows cast in the scene produces dramatic changes. Revisiting the same scene several days or weeks later presents further challenges due to the dynamic nature of the scenes. For example, natural erosion and human intervention can cause significant differences in the scene content. An effective SRS that tolerates such changes is said to be *robust*.

Figure 1.2: Various common image distortions.

Secondly, the same SRS must be able to *discriminate* dissimilar scenes from

those found in the memory. This is an important aspect which many other authors have ignored. The discriminatory power of the SRS is particularly important for outdoor natural scenes where common features appear over several different scenes (for instance, the same type of trees and bushes for a particular environment). A naive method of matching only these features will certainly fail. The ambiguity problem occurs in indoor scenes as well - man-made structures are often repeated in the same environment such that different locations may possess a large number of similar looking features that will easily confuse an algorithm based on simple matching (Fig. 1.3). The ability to discriminate dissimilar scenes is also important during the learning phase of the agent - any scenes that are rejected are 'new' and should be added to the memory.

1.4 Challenges of scene recognition

The challenge of scene recognition is that the two desirable characteristics - robustness and discriminatory power - are unfortunately mutually *antagonistic*. A SRS that is too discriminatory is often not robust enough to tolerate even slight changes in viewpoint and illumination. On the other hand, a SRS that is too robust will not be discriminatory enough, leading to numerous false positive matches. A compromise between these two characteristics is often needed for most practical SRSs and this is often set by the user or determined by a separate learning algorithm.

Figure 1.3: Ambiguous scenes with similar features: outdoor natural (top) and indoor (bottom).

This need to balance between robustness and discriminatory power is analogous to the *overfitting problem* that is well known in machine learning [77] defined as: the preference of a hypothesis that does not have the true lowest error of the considered hypothesis, but that by chance has the lowest error on the training data. The performance of the scene classifier depends on how it is trained. If the training set of scenes have only very small differences, the classifier will be too sensitive to such small changes, and is too discriminatory. If instead the training

set are too varied, the sensitivity drops significantly and the classifier will be be too robust for large changes which is also undesirable. The crucial problem is the selection of the training set such that it captures just the right amount of variability and consistency to train a balanced classifier. Nonetheless, the selection of an optimal training set remains an open problem.

Designing a SRS that is *general* enough for a variety of environments (*e.g.* indoors and outdoors) is especially difficult. Different environments have different requirements such as the choice of a good landmark - an indoor scene can use strong corners while corners in a natural scene may be unreliable due to the foliage and vegetation. Another factor that needs to be considered is the effect of natural erosion that is more pronounced in a natural setting than in an indoor laboratory. For example, trees may fall or tides may change over time and weather conditions can dramatically change the scene content compared to the relative stability of the scenes in an indoor environment. Changes in illumination which are less pronounced indoors than outdoors provide another set of varying requirements that needs to be taken into account (see Fig. 1.2 for good examples).

The simplifying assumptions in an indoor scene are the main reasons why research in the past two decades had been focused on indoor robotic navigation. 'Outdoor navigation' have been limited to structured environments such as road following [31]. The same authors in [31] concluded that for a robot to

> *...stop at a stop sign under various illumination and background conditions, we are still eons away.*

This is a clear indication of the challenges that outdoor scene recognition pose.

1.5 Scope of the thesis

This thesis is concerned with the design of an effective SRS used in other applications such as biomimetic navigation. This thesis is inspired from various biological models but does not propose a plausible model that describes how biological agents perform scene recognition. The main idea is to use the clues available in nature to design an effective solution to scene recognition, *not* to propose a radically new model of animal navigation, which would be beyond the scope of this thesis. A single calibrated camera with a limited field of view is used to capture the images. The only input used in the work are the RGB images obtained from the camera. No other imaging devices or sensors are used. The solution proposed here is thus entirely limited to vision in the visible spectrum, perceivable by humans. The learning phase of the algorithm, where the SRS constructs the reference image database is not considered here and is assumed to be available. Finally, it is assumed that the image databases are of reasonably small sizes, so that a simple database query system can be used without affecting the efficiency of the algorithm.

1.6 Contribution of the thesis

This thesis addresses the problem of scene recognition from an entirely new perspective. Inspired from the domain of human psychophysics and animal behavioural studies, a novel SRS that is robust to common image distortions and is general enough for both indoor and outdoor environments is proposed. Fig. 1.4 illustrates the various components of the proposed SRS, which are briefly presented in the next paragraph.

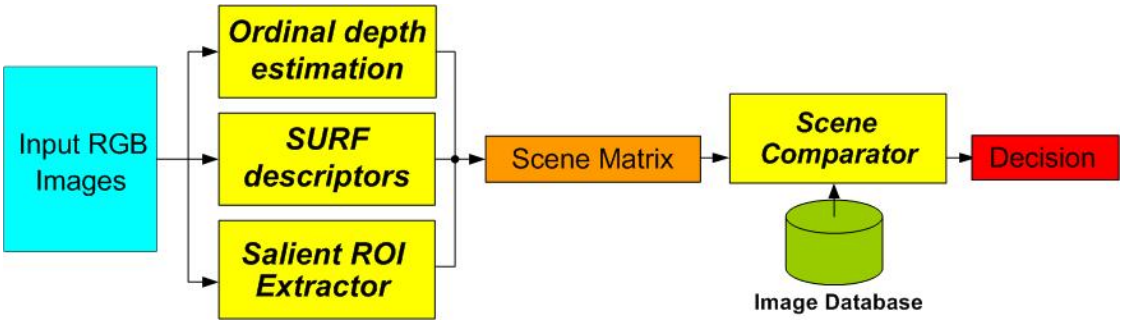

Figure 1.4: The various components of the proposed SRS.

In this work, a modified computational model of visual saliency inspired from [51] that includes several new composite feature cues is implemented to provide an initial 'mask' to efficiently reduce the number of salient ROIs (regions of interests) extracted from the scene. These ROIs are further encoded using SURF (Speeded-Up Robust Features) [10] to obtain 'salient-SURF' keypoints/descriptors for reliable matching. Motivated from special TBL (Turn-Back-and-Look) flights observed in certain species of flying hymenopterans [61, 116], the descriptors are

augmented with ordinal depth information computed from optical flow. In this work, optical flow is induced by a camera that simulates the TBL. Other authors [14, 63] have only used TBL to extract reliable landmarks for navigation and have completely ignored the robustly obtainable ordinal depth. Combining the spatial position (x, y, z) of the landmarks encodes the global spatial configuration of a scene into a *Scene matrix*. By extracting these keypoints from the HSV colour space and comparing their rank correlations, a simple measure of scene similarity that is invariant to illumination [35, 92] and viewpoint changes is proposed. Finally, a novel *scene decision* module compares an input query test scene with a database of reference scenes to arrive at a final decision to accept or reject the test scene.

The work focuses particularly on outdoor *natural* environments that do not contain man-made structures. Man-made objects often simplify the problem of scene recognition because certain obvious and unique features exist in these objects making the discriminating component of a SRS inconsequential. Instead, this thesis applies ideas taken from animal and insect navigation strategies and formulates a SRS that achieves a recognition performance far exceeding what current state of the art systems achieve in both accuracy and generality.

The ultimate aim of this work is to model how these animals and insects achieve robust and reliable scene recognition in natural outdoor environments. This is a problem that is largely untouched by robotics and vision researchers due to its

apparent complexity that often overwhelms many traditional algorithms.

Finally, this thesis highlights to the research community the importance of testing the effectiveness of their SRS or navigation systems with challenging outdoor scenes so that further progress in practical outdoor navigation can be made. The availability of several large image databases online[1] of predominantly outdoor scenery taken under various weather and lighting conditions serve this purpose.

1.7 Mathematical notation

Throughout the thesis, a set of standard mathematical notations is used. *Scalar* values are denoted by italicised non-bold letters such as G_c or d_{thresh}. *Matrices* are denoted by bold non-italicised upper case letters such as $\mathbf{S_m}$. Symbols that are used to represent *semantic objects* are denoted by blackboard bold uppercase letters. For example, $\mathbb{G}_{cand}$ refers to the candidate match in a typical scene decision situation. Other notations will be specified when required throughout the thesis. A list of mathematical symbols can be found in page xvii.

1.8 Outline of the thesis

The rest of this thesis is organised as follows. Several recent works related to scene recognition, focusing on applications related to navigation are reviewed in chapter

[1] `http://www.ece.nus.edu.sg/stfpage/eleclf/robust_SRS.htm`

2. In chapter 3, important concepts related to the design of the proposed SRS are explained. The next few chapters introduces the various subcomponents of the proposed SRS. The use of visual saliency to extract useful landmarks in the scene is described in chapter 4. The extracted landmarks or salient ROIs are then encoded with SURF descriptors augmented with ordinal depth to form a *Scene matrix*, described in chapter 5. Next, a simple scene decision module, where an input test scene is compared with a database of reference scenes, is described in chapter 6. The performance of the proposed SRS is then evaluated and analysed using several image databases in chapter 7. Finally, chapter 8 concludes the thesis and suggests future research directions, based on this work.

Chapter 2

Literature Review

The problem of scene recognition has been explored by authors in diverse fields such as visual SLAM (Simultaneous Localisation and Mapping), CBIR (Content-Based Image Retrieval) and biomimetic navigation. In this chapter, recent works from these fields are reviewed (sections 2.1–2.3) respectively, with a focus on biomimetic navigation techniques that addresses the scene recognition problem. Since the problem of determining scene equivalence is common in these three domains, it is not surprising to see many works in the literature with solutions that are suitable for multiple applications. The main aim of this chapter is to present what is the current state of the art in scene recognition algorithms. At the same time, certain shortcomings in these works are also discussed (section 2.4) that this thesis attempts to address with the proposed SRS.

2.1 Related work from visual SLAM

In the domain of visual SLAM, the problem of determining scene equivalence is posed in the current robotics literature as the 'loop closing problem' or 'robust data-association problem'. Knowing that the mobile agent has returned to the same location is crucial as SLAM requires that the *uncertainty* associated with a current position is small in order to create a stable closed loop system. If scene recognition fails, the robot is essentially *lost*, since the uncertainty of the robot's location grows out of bounds.

In the work of Newman and Ho [87], the loop closing problem is specifically addressed in an indoor setting using a mobile robot that performs visual SLAM along a corridor. The visual front end consists of the detection and extraction of salient features using the Kadir-Brady scale saliency algorithm [54] that is combined using MSER (Maximally Stable Extremal Regions) [71] to detect regions that display both saliency and wide baseline stability. These regions are then encoded using Lowe's SIFT (Scale Invariant Feature Transform) descriptors [68] for reliable matching. The decision to determine if loop closing has occurred is based entirely on the number of SIFT matches between the input query scene and the reference scenes in the database (created after one loop). A fixed threshold is used to either accept or reject the best matches. This threshold completely arbitrary and can result in false positives given the large number of ambiguous features in

an indoor environment.

The major problem with their approach is the use of two very different region detection algorithms to extract stable 'salient MSE' regions that do not have much overlap (Fig. 2.1). The results in Fig. 2.1 show that the number of SIFT descriptors

Figure 2.1: Solution from [87]. Kadir-Brady salient regions (left), MSER (middle) and SIFT descriptors (right).

extracted from the full sized image (640x480) is very small, and only four are matched in the example shown with another frame taken two seconds apart. The authors do not explicitly explore (or show) the possibility of incorrect SIFT matches that would have made the scene recognition difficult. As the authors have admitted in their conclusions, the use of a fixed threshold to reject bad matches is not satisfactory in practical applications and they propose to use supervised learning techniques to determine the value of this important parameter.

As an extension to [87], laser scanners are employed in [86] to detect loop closing in outdoor urban environments. A method to detect loop closing is proposed that uses a *similarity matrix* that summarises the L_2 distances of Harris-Affine

Detectors [73] described by SIFT between any two image pairs taken in sequence as the robot navigates. The authors suggest a method using *rank reduction* to remove ambiguous and repetitive scenes in the similarity matrix while attempting to fit a probabilistic model of scene similarity so as to detect a reliable loop closure. The use of the 3D laser information is limited to recovering the current pose of the robot, and it does not serve any purpose in determining loop closure. The possibility using the valuable depth information obtained from the laser scanner is completely ignored. Furthermore, the authors do not provide details on the success of the loop closure detection in various situations and environments, and the only example shown is a completely built-up scene with no natural vegetation (Fig. 2.2). Furthermore, the authors do not discuss or present any results under weather and illumination changes, which are the main challenges to outdoor visual SLAM [31].

Figure 2.2: Two image sequences from [86]. Loop closure is detected for the corresponding scene pairs between the top and bottom rows.

A large number of other works in the visual SLAM literature follows a similar

framework described in [86, 87] to detect loop closure. Most of them ([3, 9, 76, 100]) use a combination of various SLAM algorithms and SIFT descriptors. For an overview of SLAM and robotic navigation, refer to [34, 72]. A recent paper [65] surveyed the current state of the art in visual SLAM and presented various solutions using monocular and stereo camera systems.

2.2 Related work from CBIR

Image retrieval has grown in importance over the past two decades due partly to the tremendous increase in information size and availability. This increase is the result of the growth in information storage capacity (*e.g.* hard disks, DVD optical drives) and the growth of the World Wide Web. The need to organise the increasing amount of information and to retrieve them in the shortest time possible is a topic of intense research. Database searching techniques, including CBIR, are thus developed to address these issues.

A comprehensive review of CBIR techniques in [105] describes the general framework of how an effective CBIR can be implemented by separating the description of the image content into two phases. Firstly an image processing step is used to effectively choose regions of interests in the image to reduce the amount of data to be manipulated. The second step provides unique descriptions of these extracted regions. A decision is made from the amount of similarity between a pair

of images using their descriptors. It is not surprising that certain authors have used ideas in CBIR to solve the loop closing problem in visual SLAM (section 2.1). For an application to be useful in CBIR, the major consideration is the *efficiency* in database search techniques, which is equally important for real-time visual SLAM.

The work presented in [60] proposes a reduced SIFT feature descriptor by assuming that the robot navigates in an indoor office/lab environment and the camera is orthogonal to the walls. The authors also claim that the majority of SIFT features are extracted from the textured walls and not from the floors or ceilings that are usually textureless. Reduction of the complexity of the SIFT descriptors is based on removing the rotational components of the algorithm which becomes redundant under these assumptions. However, the assumptions are based on simplistic observations from two locations described in their paper (Fig. 2.3) and may not be applicable even in general indoor scenes where the walls may be devoid of texture. As the authors have admitted, although slight bumps may not affect the effectiveness of their descriptors, a slope greater than twenty degrees will reduce the performance of the algorithm. This algorithm is only effective in a very restricted set of environments, and cannot be used in general environments.

Other well known solutions to reduce the complexity of the SIFT descriptors exist and they had been explored and compared with other competing descriptors in a comprehensive review in [75]. One of them is PCA-SIFT proposed in [55]. PCA-SIFT attempts to reduce the computational complexity of SIFT by applying

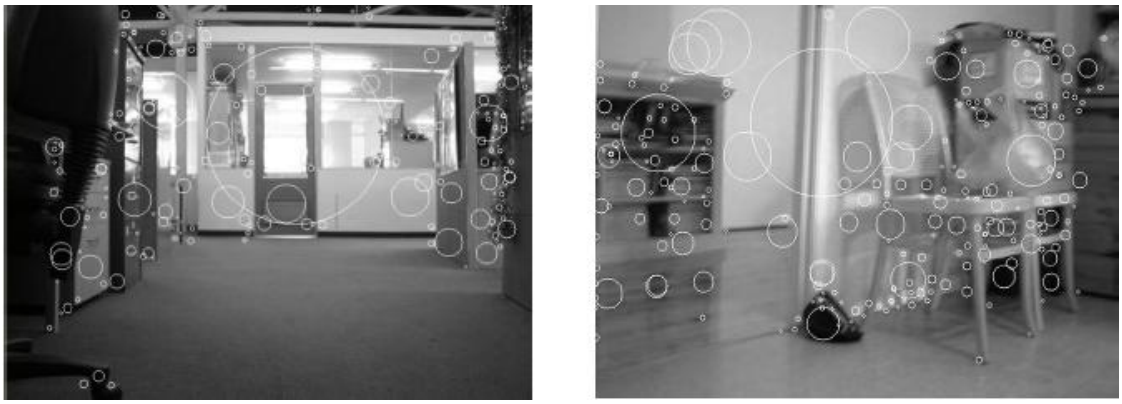

Figure 2.3: Two example scenes with reduced SIFT features from [60].

Principal Components Analysis (PCA) on the eigenspace produced just before the final descriptor assignment step of SIFT. The PCA reduced eigenspace is computed from a diverse image database of 21000 image patches which are not used in any of their matching experiments. In the evaluation framework of [75], PCA-SIFT only displayed an average performance and do not perform as well as SIFT in terms of *recall* and *precision* [29], which are common evaluation metrics in machine learning. The reduction in computational complexity using PCA-SIFT is however significant.

Another work in [118] uses a localised colour histogram technique adapted from [104] to group the detected features together to represent a scene. Monte-Carlo localisation techniques are then applied in the context of visual SLAM. The detected features are integrated with non-linear functions over a range of Euclidean motions that are shown in [104] to be invariant to rotation and translation. A similarity score between the query image and reference images is computed from

the intersections of the histograms normalised by the number of bins in the query image. As is shown in [118], the system returns a number of resulting images ranked by their level of similarity for a given query image (Fig. 2.4). This similarity score is then integrated into a Monte-Carlo localisation algorithm to determine the weights of the different returned image samples.

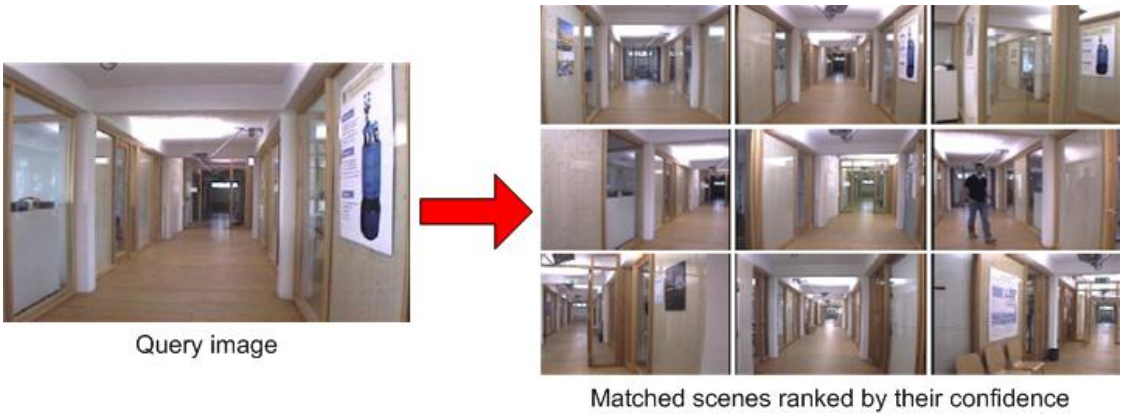

Figure 2.4: An input query image returns several closest matches. Data from [118]

Another SRS (or visual SLAM) application motivated from CBIR techniques used panoramic images [66]. A modified Harris detector is used to detect interest points, which are then encoded by a scale invariant descriptor similar to SIFT. The authors propose a novel technique of managing the growing image database so that a reasonable database size is always assured for efficient recognition. This is done by indexing the reference images with a set of image statistic data, derived from the first and second derivatives of the Gaussian which are stored as separate histograms. A similarity score based on the χ^2 distance of the histograms is used

to detect loop closure. The authors reported very good localisation results with reliable loop closing using this method in an indoor environment.

2.3 Related work from biomimetics

In this section, the focus is on solutions in the literature inspired from biology which provide clues in designing an efficient SRS. Such solutions are interesting as animals possess remarkable scene recognition abilities that perform better than many artificial solutions. Since the thesis is concerned about developing a SRS suitable for biomimetic navigation, related works in the literature with close links to the work in this thesis are presented.

The recent work of [42, 43] selects salient ROIs by constructing a general saliency map that combines the computational model of *bottom-up* visual saliency of Itti *et al.* [51] with a novel *top-down* saliency map constructed from prior knowledge of preselected target locations known as the *VOCUS* system [41] (Fig. 2.5). Note that the top-down saliency map is only used in an active search task for selected targets used for loop closure detection. The salient ROIs are extracted from the saliency map by choosing a rectangle of the same height and width of the *most salient regions* (MSaRs) which are shown as crosses in the bottom right image of Fig. 2.5. The authors encode the ROIs by detecting stable Harris-Laplace features

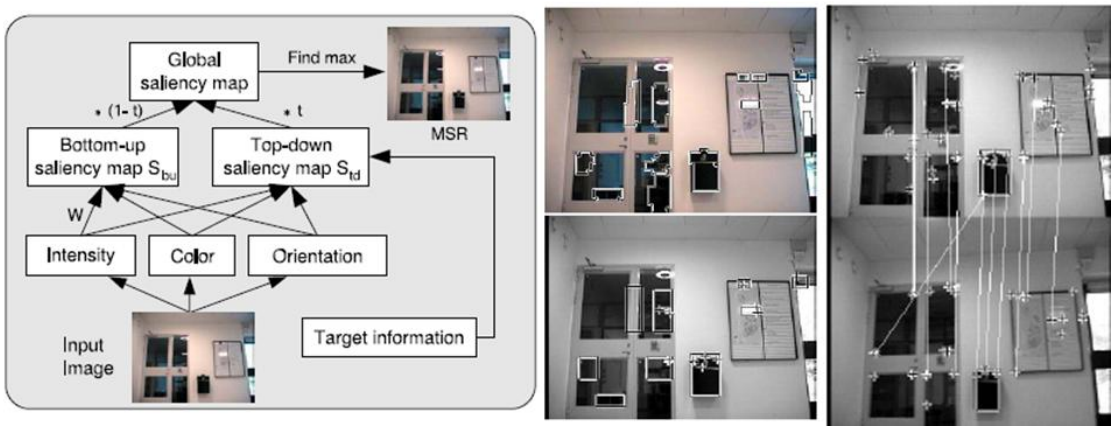

Figure 2.5: Saliency map creation using VOCUS (left), extraction of salient ROIs from MSaRs (middle) and loop detection by active search for a target (right), from [43].

[73] and SIFT descriptors for reliable matching. Experiments in detecting loop closure are conducted in a small indoor hallway with constant ambient lighting that does not pose much of a problem for the SIFT descriptors. The main objective of the paper is to show that the reduction of features detected using salient ROIs maintains a high detection rate for loop closure but there are no false positives in the image database. From the experimental data shown (Fig. 2.5 (right)), the number of Harris-Laplace features detected in the ROIs is certainly small and the authors use only one single region, the dustbin in this case, containing only three matched features for localisation. This is because the authors use regions that are predicted to be at that position using odometry to perform an active search for possible targets (*e.g.* the dustbin) to determine that the place was previously visited [42]. The target regions are selected based on the criterion of their ability

Figure 2.6: Examples of loop closure detection using a tracked target, from [42].

to be tracked over a large number of frames. The experimental results specific to loop closure detection show that very small (and possibly unreliable) landmarks may be used (Fig. 2.6). The loop closure experiments are conducted by simply driving the robot around in circles. They do not consider what happens when the environment contains many similar features which will have certainly made loop closure detection more difficult. This is especially true when the robot takes a different path and returns to the same place, rendering the prediction using odometry unreliable. Furthermore, the simplistic assumption of finding a unique set of targets for loop closure will fail in a dynamic outdoor environment that contains numerous repeated features with significant changes in scene content.

The series of work by Bianco *et al.* [13–15, 63] that exploits the *Turn-Back-and-Look* (TBL) behaviour observed in certain species of flying hymenopterans (bees and wasps) [61] are closely related to the work presented in this thesis. The

TBL motion is also known as *zig-zag flights* [116] or *learning flights* [121]. TBL has been observed when young honeybees leave their nests for the first few times and when they discover important feeding sites (*e.g.* flowers with abundant honey) [61, 116, 121] and is believed to be important for the bees to recognise these scenes on their return trip. The details of TBL and how it is exploited in this work is discussed in section 3.5.

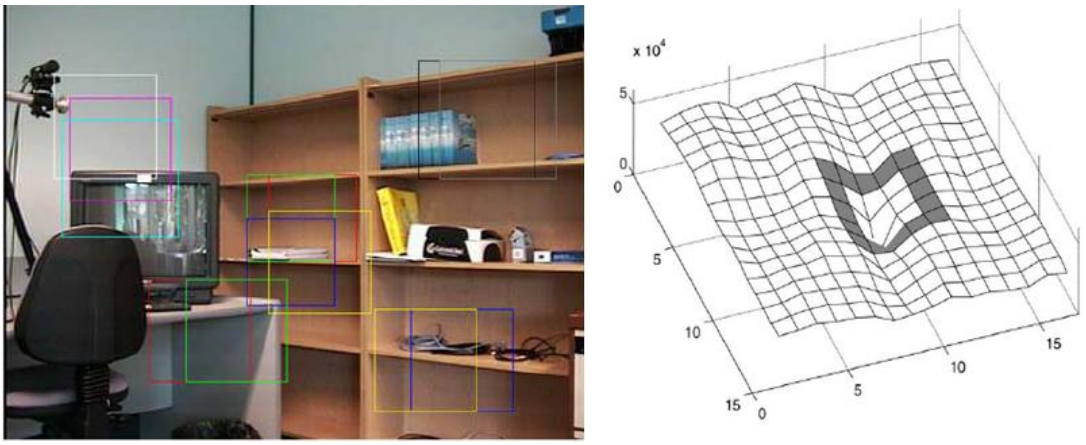

Figure 2.7: Preselected targets (boxed) from a static scene using the valley method of [83], data from [15].

Motivated by the importance of TBL for scene recognition in insects, the work of Bianco *et al.* hypothesised that this special motion allows the insect to perform a sort of *testing* procedure on the pre-selected landmarks by perturbing the possible return paths with several arcs that resembles a typical TBL (Fig. 2.8). Only the landmarks that are stable throughout the whole TBL phase executed by the robot will be retained for navigation use. The landmarks are pre-selected

from static images of the camera using a modified *valley method* for computing a correlation metric of reliability for landmark selection [83] (Fig. 2.7). However,

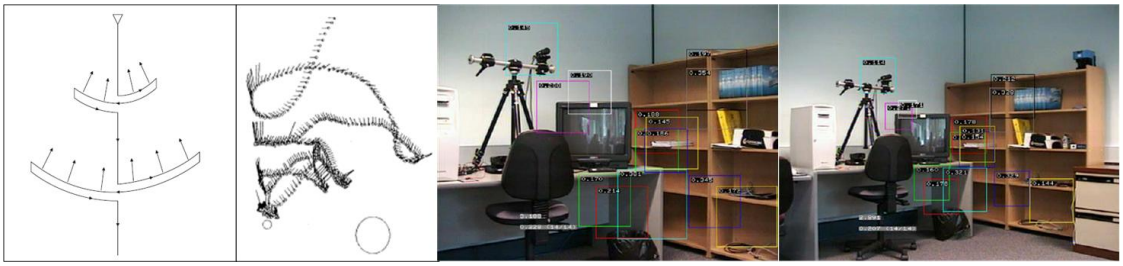

Figure 2.8: Comparing the TBL motion of the artificial robot *vs.*a real TBL motion of a wasp (left). Image frames captured with selected landmarks during the TBL phase (right), data from [63].

the use of TBL as a testing framework is unlikely to be its main use. The insect does not *need* to perform TBL in order to memorise how a target scene appears from various approach angles. One can easily envisage other forms of flight (moving backwards and forwards for example) or even randomly stopping at various positions to memorise the scene to check if the original landmarks are reliable or not. The important question to ask is *"Why are TBL flights designed in such a fashion?"*. The discussion of this question is deferred to section 3.5 where the importance of this motion in extracting depth information from the scene structure is highlighted.

Although insects and animals navigate in natural outdoor environments, there are only a few related works in the biomimetic literature that attempt to propose models that function in outdoor conditions. The work of Lambrinos *et al.* [58]

employ navigational strategies for scene recognition inspired from the desert ant *Cataglyphis*. They propose a new landmark navigation model known as the *Average Landmark Vector* (ALV) model that is modified from the original *snapshot model* proposed by Cartwright and Collett [19]. This model is also explored in [78] in a robot that uses analog electronic components. Basically, the snapshot model as it name implies, captures a 1D snapshot of the scene memorised by the animat. The detected landmarks are represented as dark patches in a circular ring that indicate the spatial position of the selected landmarks in terms of their angular separation. Scene recognition using the snapshot model is very simple. Landmarks of the same scene viewed with a slight distortion have slightly different angular separation. Comparing the current scene and the stored snapshot is achieved by simple vector additions/subtractions using radial and tangential vector components that represent the position of the landmarks. The final resultant vector *guides* the animat in a direction that reduces the difference between the currently viewed scene and the stored snapshot (Fig. 2.9 (left)). The ALV model simplifies the computations by storing an averaged vector of all the detected landmarks instead of storing the individual radial and tangential components in the original snapshot model (Fig. 2.9 (right)).

This model has been successfully tested on an outdoor mobile robot, *Sahabot2*, that is equipped with several sensors that mimics that of the Saharan desert ant. In order to validate the ALV model in a desert environment with virtually no

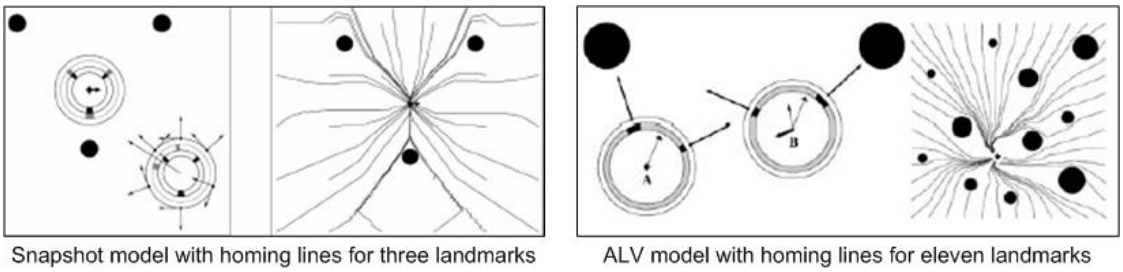

Figure 2.9: Illustration of the snapshot model (left) and ALV model (right) shown with the homing paths from different starting positions, data taken from [58, 78].

obvious visual cues, the experiments used artificial blocks to serve as landmarks to test the homing strategies of this model (Fig. 2.10). The main problem with the

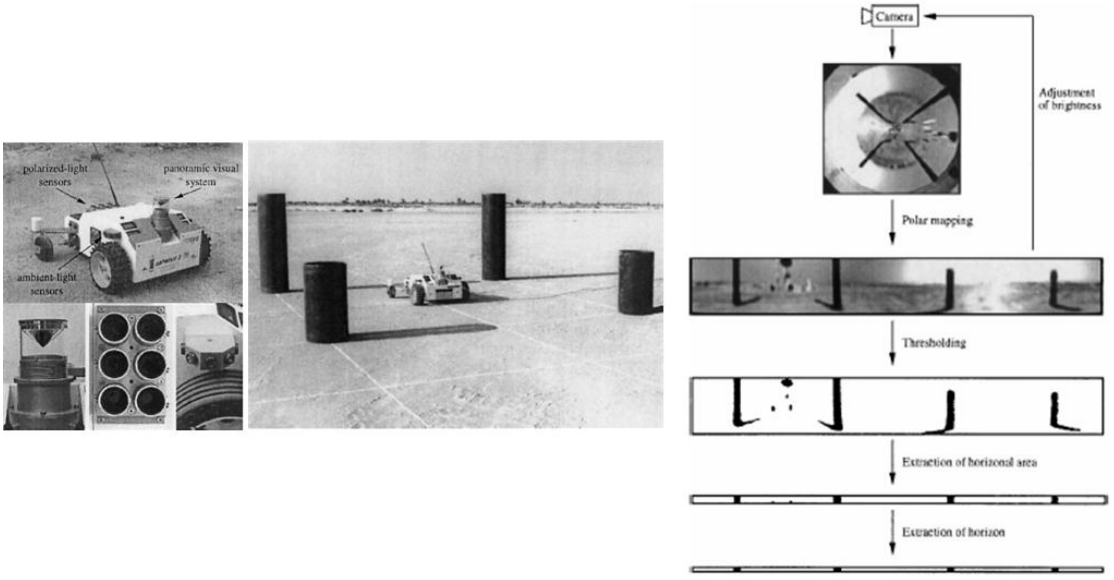

Figure 2.10: The *Sahabot2* with various sensors (left). The testing arena with artificial landmarks (middle). The simple visual processing used to extract the landmarks (right), data from [58]

proposed ALV model is that it has only been validated in artificially manipulated outdoor environments with obvious landmarks. The same holds for the snapshot

model where in [19], only computer simulations are used to validate this method in theory. The main problem is that real insects navigate in complex outdoor environments and the selection of landmarks from these scenes is much harder. An attempt is made to validate the ALV model in an indoor environment [80] where the same *Sahabot2* is placed in a hallway with numerous features. The authors use a similar method to extract landmark features, this time using an adaptive threshold from a low-pass filtered strip of the unwrapped panoramic image obtained from the environment (Fig. 2.11). The homing result of using the ALV in this relatively complex environment is quite satisfactory and is surprising given its extreme simplicity.

The ALV model in this indoor setup has a major unresolved problem, highlighted by the authors in [80]. The simple visual processing method is unable to address the effects of lighting changes that make landmark detection unstable. The main reason is due to the use of the adaptive threshold that extracts landmarks from the filtered image strip. Landmarks are selected at locations that display long segments of consistent intensities. A change in lighting conditions, however, violates the consistency assumption resulting in unreliable landmark detection. This leads to wrong homing decisions by the robot. To the best knowledge of this author, there has been no subsequent work done on the ALV model in other real environments.

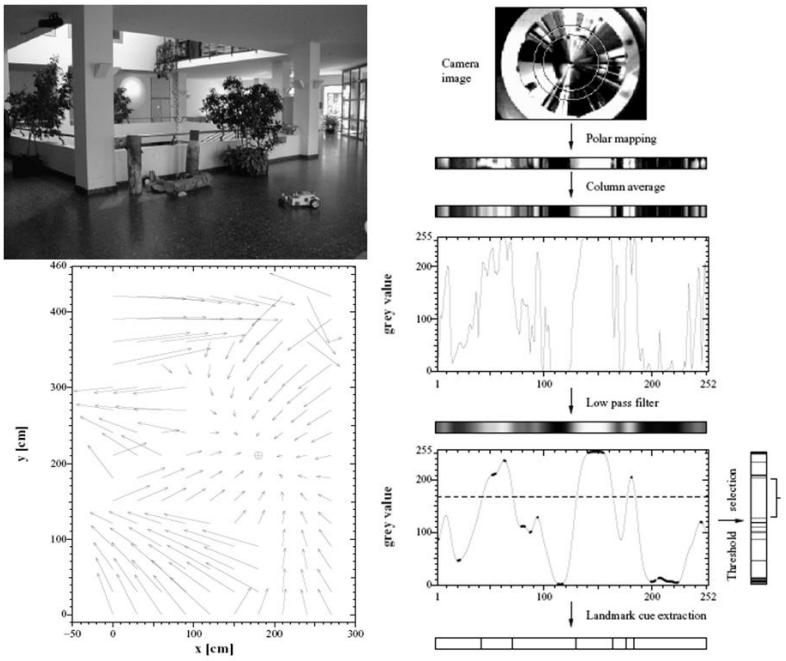

Figure 2.11: The *Sahabot2* tested in a hallway (top left). The modified visual processing (right). The resulting homing vectors generated using ALV (bottom left). All figures from [80].

2.4 Conclusion

This chapter has reviewed several related works in scene recognition from different domains in the literature - visual SLAM, CBIR and biomimetics. This review is not comprehensive but is meant to showcase a representative subset of related work so that an idea of the current state of the art in scene recognition is established. Several shortcomings of these works are evident from this review, and will be discussed here.

Almost all of the works reviewed focused on indoor environments where the complexity in terms of image distortions is greatly reduced. Even the few works [58,

86] that explored scene recognition in an outdoor environment limited themselves to built-up urban areas [86] or have manipulated the environment with artificial landmarks [58]. In [86], although the robot did traverse a short distance on one side of a park, the authors did not perform loop closure detection at that place and even added that the natural scenery, considered as a "homogeneous foliage", causes problems in loop closure detection and are removed by rank reduction techniques on the similarity matrix shown in Fig. 2.12 (left). The loop closure detection is in fact tested in a purely built-up environment, where the smaller yellow loop overlaps the larger white loop shown in Fig. 2.12 (right).

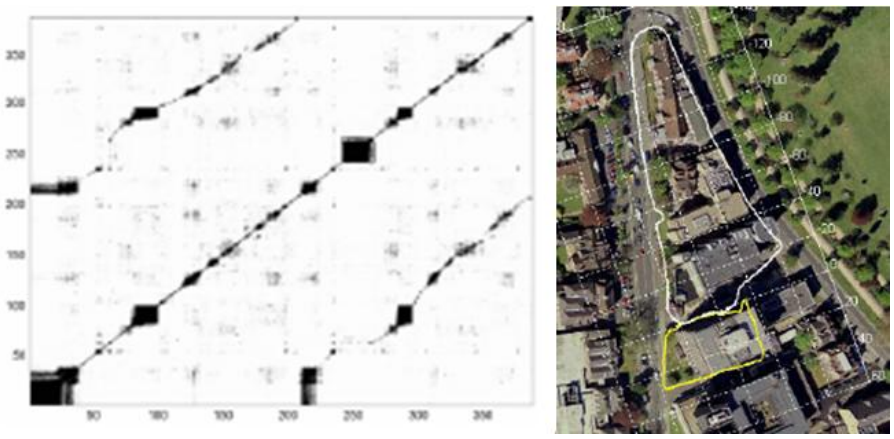

Figure 2.12: A *similarity matrix* with many off diagonal streaks that represents false loop closure (left). The two loops performed by the robot (right). Figures from [86].

The major problem of performing scene recognition in an outdoor environment is the change in illumination that reduces ability of these algorithms to determine reliable correspondences. This is because scene equivalence is determined solely

by the *number* of correspondences which will fail in an outdoor natural environment with numerous similar features. Most of the works simply *assumed* that correspondences will be reliable which is only true in an indoor environment with consistent lighting and static man-made structures. This assumption is obviously too simplistic in an outdoor environment. Changes in illumination, non-uniform lighting as well as shadows from foliage cause significant distortion not addressed by these works. Furthermore, almost all of the papers reviewed are tested in ideal conditions where the scenes have zero dynamic content, for example, an empty corridor or hallway. Natural environments are extremely dynamic with changes caused by different weather conditions, natural erosion and human intervention. Using the same algorithms in these challenging scenarios will fail as the original assumptions in the design of the algorithms are violated.

It is interesting to see that almost all of the related work in the literature concerning scene recognition is limited to the 2D pixel data of the image. There is virtually no work done in extending scene recognition to include the z or depth dimension. Although [86] tried to use depth information for '3D SLAM', it is entirely laser based which is extremely slow. A major contribution of this thesis is to highlight the importance of depth information for scene recognition which is discussed in section 3.5.1 and will be incorporated in the proposed SRS, described in section 5.3.

Another interesting observation from this review is that most researchers have

ignored the importance of the second criterion of a good SRS described in section 1.3, that is, the *discriminatory* power of the SRS against difficult and ambiguous scenes. Works from the visual SLAM domain (*e.g.* [86, 87]) regarded this as a highly improbable event as the robot odometry is able to give a prediction of the expected location of the robot so that the search space for detecting loop closure is greatly reduced. This is however not the case in the aptly named "kidnapped robot" problem where the robot visits a new location for the first time. In this case, the robot must realise that it has moved to a new environment and this involves discriminating the new scenes from the old ones in memory. Other works such as [42] only drove the robot in circles to simulate the kidnapped robot problem yielding unconvincing results (see section 2.3). The rest of the other papers reviewed simply ignored the possibility that false positives can occur.

The fact that scene recognition (and its variants) are explored by many different researchers from various domains highlight the inter-disciplinary nature of scene recognition. The proposed SRS presented in this thesis is thus not only limited to navigation but has potentially many other uses in different domains. The contribution of this thesis is thus very general and extensions of the proposed SRS to other domains such as visual SLAM or CBIR can be achieved with minor modifications. The next chapter describes the important concepts introduced by the proposed SRS that address the above shortcomings of these as well as many other state of the art works in the literature.

Chapter 3

Important Concepts

This chapter provides preliminary information required in understanding the proposed SRS detailed in chapters 4–6. This is done so by introducing the core concepts used in the SRS, presented in separate sections. The concept of *visual saliency* and its use as a generalised landmark selector is first described in section 3.1. The concept of *image descriptors* is then introduced in 3.2. This is followed in section 3.3 by a discussion of how *ordinal measures* improve robustness against viewpoint changes. Since the proposed SRS is designed to be used in an outdoor natural environment susceptible to illumination changes, the input RGB images are converted to the HSV (hue, saturation, value) colour space which displays a degree of *illumination invariance* in section 3.4. Finally the addition of *depth information* so as to improve the performance of scene recognition is explained in section 3.5, motivated from the TBL motion introduced in section 2.3. As the aim

of this chapter is to make clear the *concepts*, the details of how they are actually implemented in the proposed SRS are reserved in the later chapters which will reference the appropriate sections here as necessary.

3.1 Selecting good landmarks using visual saliency

This section introduces the notion of a *landmark* as well as the desirable characteristics of a "good" landmark in section 3.1.1. The concept of *visual saliency* from human psychophysics is then introduced as a tool to determine initial landmarks or salient ROIs in section 3.1.2. Finally, several computational models of visual saliency are introduced in section 3.1.3 where a *saliency map* is produced. This map highlights the most salient regions in the scene as potential landmarks.

3.1.1 What makes a good landmark?

Landmarks are important in scene recognition as they serve as the basic components that optimally represent the scene. This representation tries to optimise certain important characteristics that define a *good* landmark. Using *all* the pixel information to represent a given scene is not a practical solution as it is too memory intensive when many scenes need to be stored. Such a representation is also not robust. A change in a few pixel intensities will make any comparison (usually by correlation of the two images) yield unreliable results. This is likely to occur

under the various image distortions this thesis considers - viewpoint changes, illumination changes and changes in the scene content. Hence choosing intelligently the important parts of the image as landmarks reduce significantly the memory needed for scene representation and allow for faster recognition.

Certain regions of the image may be more tolerant to various forms of distortions and are therefore useful for scene recognition. Such regions serve as good potential landmarks. For example, a region of open space or a clearing in the distance that reveals the distant *skyline* (section 4.3) is robust to viewpoint changes if there are no significant occluding objects in the foreground. Good landmarks are thus robust to these distortions and can be detected with high *repeatability*.

Another important characteristic of a good landmark is its *uniqueness*. A landmark that significantly stands out from the rest of the scene can be used simply on its own to link this landmark with the scene in question. This is analogous of landmarks that represent a particular city in the World - The Eiffel tower is linked to Paris and the Big Ben is linked to London for example. Humans do this association naturally as these landmarks uniquely identify the particular city. The same reasoning goes to an unique and special landmark that identifies the scene. However, the use of a single landmark to reliably identify a scene is very rare in both indoor and outdoor environments, and even more so in an outdoor natural environment where there are really no obvious or unique features to use (Fig. 3.1). The lack of uniqueness in such scenes can be overcome by considering

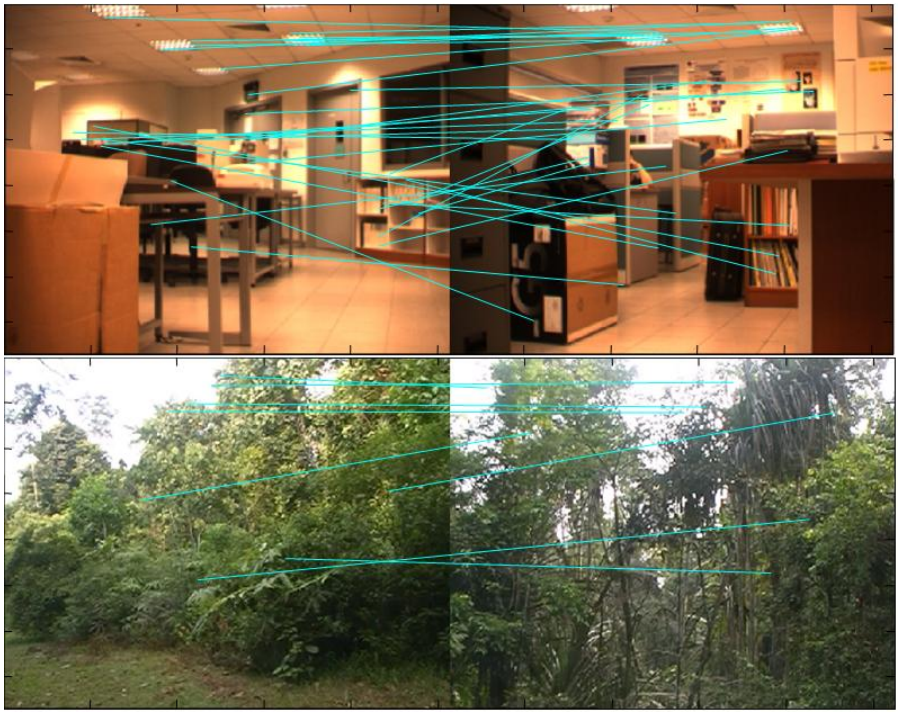

Figure 3.1: Dissimilar ambiguous indoor (top) and natural (bottom) scenes with mismatched (cyan lines) but similar features.

several landmarks at once and how the landmarks are related to one another in terms of their spatial arrangement. This is explored in section 3.3.

In conclusion, good landmarks are regions in a given scene that allow the scene to be robustly identified under various image distortions. To this end, the landmarks must be themselves robust to such distortions so that they display high repeatability. Furthermore, the landmark or groups of landmarks must represent uniquely the scene for reliable recognition.

3.1.2 Visual saliency as tool for landmark selection

Knowing the characteristics of good landmarks, this section introduces the concept of visual saliency that provides a tool for selecting potential landmarks in any given scene. These initial landmarks are known as *salient Regions of Interest* (ROIs) for the rest of the thesis.

Visual saliency is a concept from the domain of psychophysics, which is the scientific study of psychology. Psychophysical studies of human perception attempt to explain in a quantitative manner how humans perceive the environment. In the case of visual saliency, psychophysicists explain how *attention* can be modelled in perception so that the most visually *important* or salient regions are efficiently detected. In the human visual system (HVS), visual perception begins in the retina and follows on to two other separate pathways (Fig. 3.2) [45, 101]:

- *What* or *Retino-geniculate* pathway is where the majority (ninety percent) of the visual signals go. These signals go to the Lateral Geniculate Nucleus (LGN) that performs low level visual processing of the data and acts as a relay station of these processed signals before they continue on to the primary visual cortex (V1). V1 is located at the back of the head, near the occipital lobe where further processing of the visual signals (edge detections, orientation assignments) is done.

- *Where* or *Collicular* pathway is where the remaining signals go to. It involves

the Superior Colliculus and is responsible for controlling eye movements and visual attention in humans.

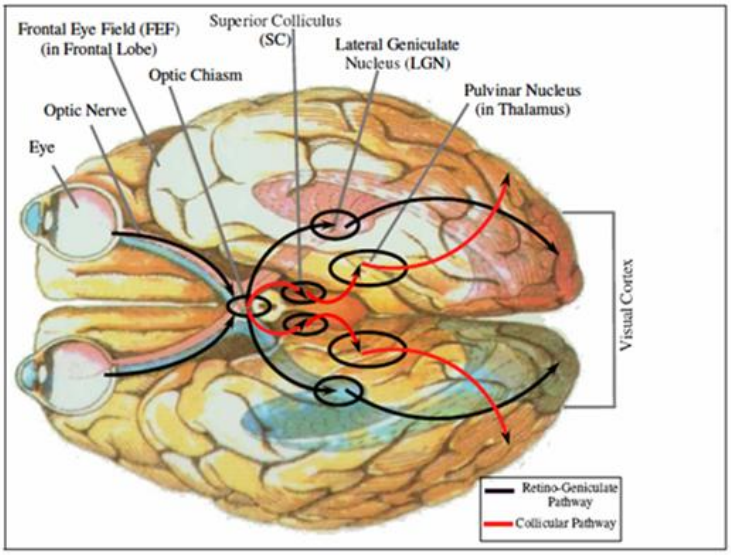

Figure 3.2: The two different visual pathways in the HVS: Retino-geniculate (black) and Collicular (red) pathways. Adapted from [101].

When presented with a scene, human subjects perform rapid eye movements known as *saccades* to move the focus of attention from one part of the scene to the next. Between these saccades are moments of *fixation* where the eyes stop moving to analyse the region where the attention is directed to. These saccadic movements are necessary as the retina has a foveated, multi-resolution structure (Fig. 3.3). The fovea contains only cone receptors needed for detailed colour vision and occupies only 0.02% of the retinal surface area. This fovea is however responsible for 30% of the signals that go to V1. Since the fovea is located at the end of the visual

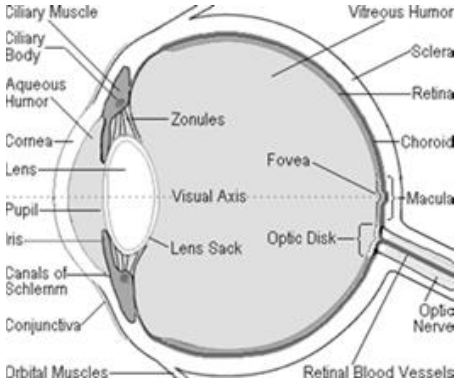

Figure 3.3: Structure of a human eye showing the position of the fovea (macular) and the visual axis.

axis of the eye, objects that are in the centre of focus are projected directly on the fovea for detailed processing to be done. These saccades thus serve the purpose of bringing the region of focus to the fovea so that detailed visual information can be percieved. Such movements have been recorded in pyschophysical experiments using a headmounted eye tracking device (Fig. 3.4(left)). The resulting scanpath made by a human subject viewing a scene is recorded and superimposed over the original image (Fig. 3.4(right)). From the scanpaths, one can distinguish between the fixations and the saccades that bring the eye from one fixation point to the next.

Having two separate visual pathways, one to decide where the eye should move to focus attention on a particular region in a scene (*Where* pathway) and another one to do further processing on that focused region (*What* pathway), together with a foveated retina are part of an elegant solution to effectively reduce the

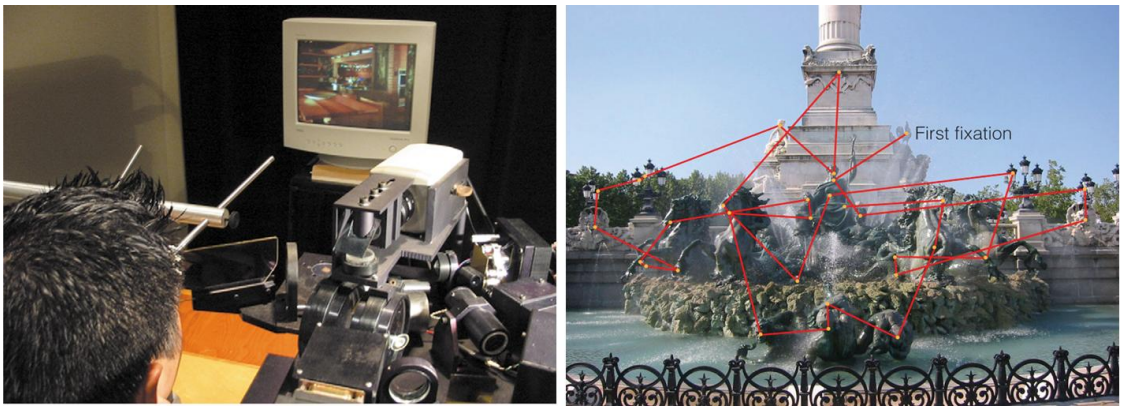

Figure 3.4: A camera based eye tracker (left)[48]. A typical scanpath showing saccades (red lines) between fixations (yellow points) (right)[88].

incoming visual information to only the most important regions needed for scene understanding. These critical regions are thus the salient ROIs that contain the most informative parts of the scene. Such regions can be exploited by biomimetic visual systems to optimally reduce the incoming visual information by dedicating resources to analyse these regions only. These regions should ideally possess similar characteristics that define a good landmark: they are unique, robust and catch our attention. The next crucial step is to determine a computational model of visual saliency so that these salient ROIs can be extracted.

3.1.3 Computational model of visual saliency: Saliency Map

Given a scene, can an algorithm predict the fixation points when a human observer scans the scene? Furthermore, can the algorithm predict the order of the saccades that the human observer will makes? In this thesis, the first question is of concern

since the link between the fixation points and salient ROIs was made clear in the previous section while the order at which the fixation points (or a subset of them) are selected are of lesser concern.

Significant research efforts from cognitive psychologists have proposed several computational models to explain how human observers select these salient regions. For an overview, please see [57, 88, 91, 112, 119]. Saliency is often defined in terms of *bottom-up* or *top-down*. Bottom-up saliency, which is the focus of this thesis, usually occurs at the onset of the presented scene to the human observer ([88, 89]). The observer makes eye saccades and fixates at certain regions of the image due to certain characteristics that are linked to how the fixation regions are chosen. Top-down saliency usually follows thereafter and is affected by the *mental state* of the observer: what he/she 'likes' to see or is instructed to see. Top-down saliency is thus linked to the *order* at which the fixation points are viewed and is harder to model than bottom-up saliency. It is therefore ignored in this work.

Central to the idea of a computational model of visual saliency is the formation of a *saliency map* ([51, 88, 119]). A saliency map encodes the 2D spatial position of the most conspicuous regions in an image. The higher the conspicuity of that location, the brighter it will appear on the map. These regions are the salient ROIs that serve as the initial landmarks for scene recognition.(see Fig. 3.5). The bright regions correspond to locations on the image that have a high salience. These regions are further processed by various image processing techniques to extract the

salient ROIs.

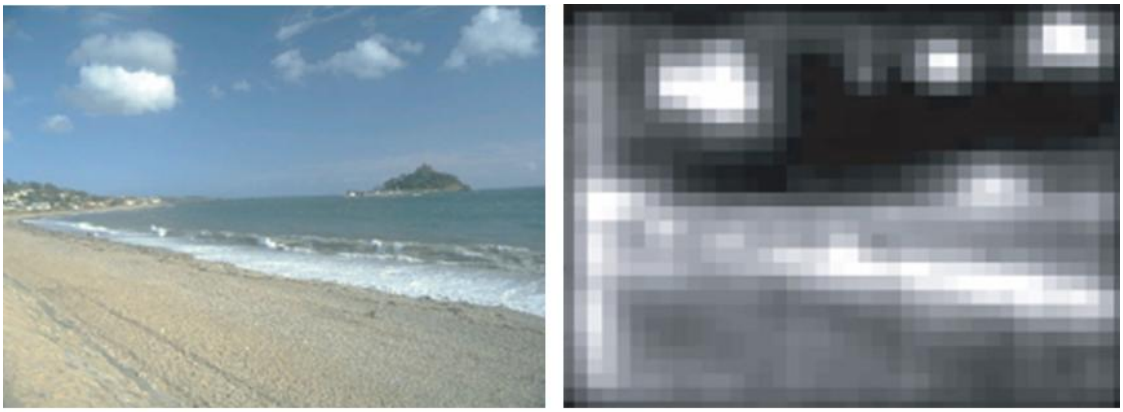

Figure 3.5: A computed saliency map using the computational model of [88].

Several computational models of visual saliency have been proposed by different authors. The work in this thesis is based on a modified computational model of Itti *et al.* [51]. This model includes several new *composite features* suitable for robust scene recognition in an outdoor environment. The details of this model, as well as the details of how the salient ROIs are extracted from the resulting saliency map are found in section 4.1. Note that the extracted salient ROIs are simple regions with no additional information that facilitates identification of that region in a matching procedure. This task of identifying and encoding a useful and robust representation of the scene is undertaken by *keypoints*, described in the next section.

3.2 Image keypoint descriptors

In this section, the concept of *keypoints* and their corresponding *descriptors* is introduced. An overview of keypoints detectors and descriptors is detailed in section 3.2.1. The need to use *both* salient ROIs extracted from the saliency map (section 3.1.3) and keypoints for reliable scene recognition is highlighted in section 3.2.2. A brief review of the state of the art on keypoint detection and extraction in the literature is found in section 3.2.3. The SURF keypoint descriptor used in this thesis, introduced recently by Bay *et al.* [10], is also described.

3.2.1 Keypoints detectors and descriptors

Keypoints can be seen as the pixel (and sometimes even sub-pixel) equivalent of the salient ROIs described earlier. They thus encode saliency *locally*, restricted to a few pixels in general. Keypoints are locations in an image that possess desirable characteristics similar to that of a good landmark. They are robust and invariant to viewpoint and scale changes; they should be unique for correct identification and they should also be highly repeatable and detectable under various forms of image distortions. In [73], a set of these keypoints produces a *covariant* region. These regions transform and change their shape covariantly with the camera movement under various viewpoint distortions as shown in Fig. 3.6.

Keypoints are extracted from the image by *keypoint detectors* and the output

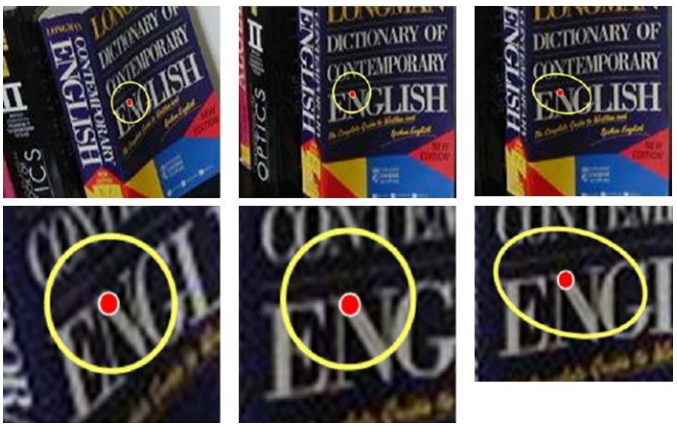

Figure 3.6: A simple fixed circle (left and middle) is not sufficient to ensure that the same region continues to be encoded under a typical viewpoint (affine) transformation. Using an oval is necessary (right). From [73].

of these detectors is a set of *keypoint descriptors*. A typical descriptor usually encodes its position and a vector that summarises and describes the keypoint based on certain computations done on the image and/or a part of the image near the keypoint (see Fig. 3.7 and Fig. 3.8 for illustrations of this).

The descriptors are designed so as to balance between the conflicting goals of efficiency and uniqueness. A high-dimensional descriptor makes it unlikely that another keypoint will possess the *same* descriptor, and this may cause erroneous matching. However, such a descriptor may take an excessively long time to compute and match with another descriptor. Descriptors thus provide a robust and unique encoding for finding correspondences of the same scene under various image distortions. This encoded information is shown in the next section to supplement the initial salient ROIs detected.

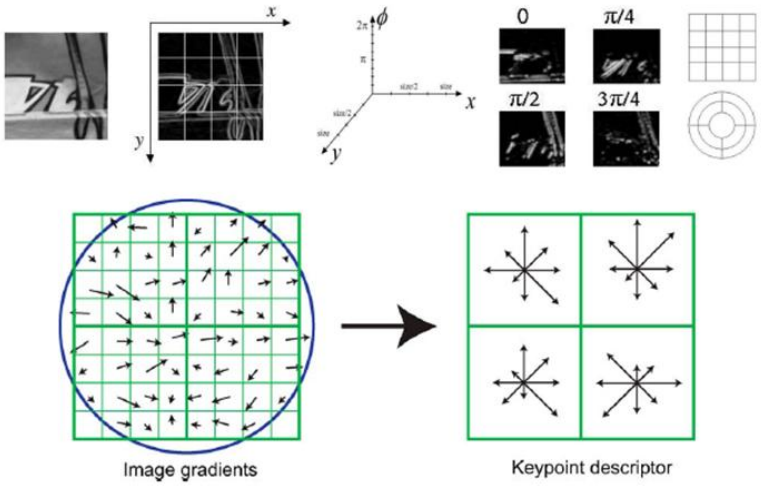

Figure 3.7: (Top row) The various stages in detecting SIFT descriptors: The detected region. Gradiant image with a 4x4 location grid superimposed. Dimensions of the 3D histogram containing the location and orientation of image sample points around the keypoint. Two possible location grids are possible, Cartesian or log-polar (from [75]). (Bottom row) Details of the computation of the SIFT descriptor. The orientation and gradient magnitudes of image samples near the keypoint regions are denoted as arrows and weighted by a Gaussian filter (left). Summing up the orientation and gradient magnitudes of the image samples produces a set of descriptors (right) (from [68]).

3.2.2 Salient ROIs versus covariant keypoints

Since salient ROIs and keypoints are two different methods of representing a scene, this section considers an important question: why is there a need for two apparently *similar* representations of the same scene? At first glance, either one of the representations should be sufficient to encode the scene properly for accurate recognition. The problem is illustrated using salient ROIs and SURF keypoints of the same scene in Fig. 3.9.

Using only keypoints at a fixed threshold of sensitivity for the detector on the

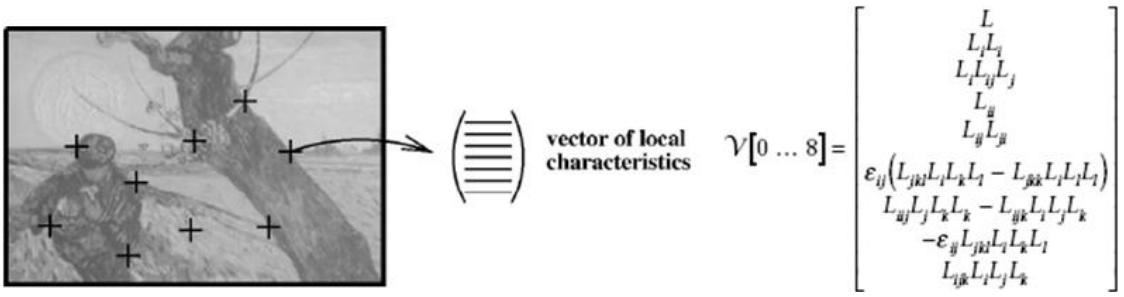

Figure 3.8: A vector of local characteristics is used to represent the detected interest points (keypoints) on an image (denoted by '+') (left). The authors used a set of local differential invariants to compute the descriptors, summarised as a vector of 'local jets' (right). Adapted from [98].

complete image produces a large number of keypoints, many of which may not even be useful for scene recognition. The reason is due to the fact that the keypoint detectors work on certain simplifying assumptions on the type of distortions that it is designed to be robust against. The most common form of distortion can be modelled by an *affine* transform as it can be easily evaluated using homographies. However, this assumption is only true if the scene is largely planar or the scene content is far away for the affine model to be valid [40, 81]. A large number of redundant keypoints makes recognition less efficient as the computational effort to compute correspondences over a pair of images is increased significantly. With more keypoints, mismatches are also more likely to occur as the limited dimensionality of the descriptors means that there is a higher chance of wrong correspondences for very similar features. This reduces the reliability of the scene recognition. Instead, by focusing the keypoint detection at the initial salient ROIs, only the most salient

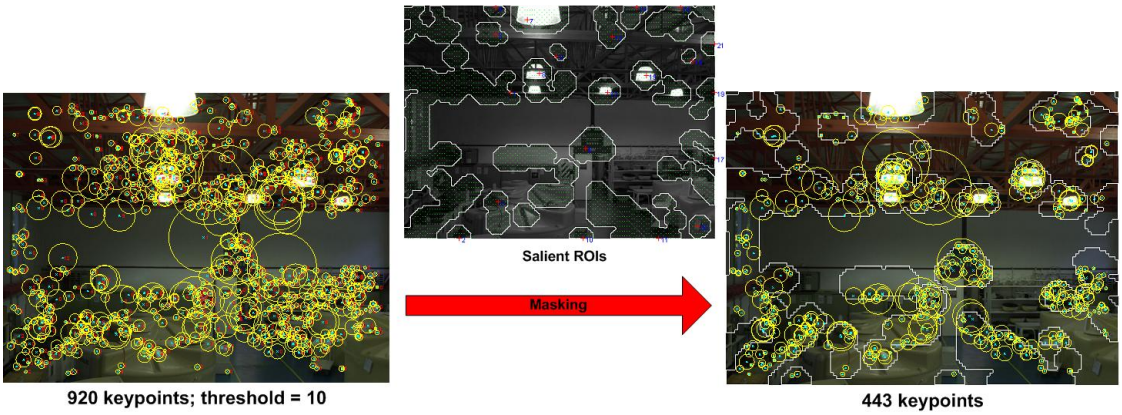

Figure 3.9: Comparison of keypoints and salient ROIs. Initially detected 920 SURF keypoints (left). Detecting SURF keypoints at the salient ROIs only produces 443 keypoints (right). Using less keypoints located at the most salient regions reduces computational time with insignificant loss in accuracy.

and stable keypoints are detected and used for scene recognition. Furthermore, the reduction in the total number of keypoints used improves the efficiency and reduces the probability of mismatches. There is thus an overall improvement in scene recognition performance.

It is possible to reduce to the number of detected keypoints by modifying the detection threshold of sensitivity. This in turn reduces the computational effort of the recognition algorithm. The main problem is that the *density* of the keypoints is greatly diminished as can be seen in Fig. 3.10. Having fewer and sparser keypoints means that the importance associated with each keypoint is increased since an important landmark may be encoded by just a few keypoints. Any mismatches, occlusions or deformations at that keypoint will have an increased detrimental effect on the performance of the scene recognition algorithm. Extracting keypoints

from salient ROIs allows keypoints of high density to be detected without much overhead in computational effort. This can be seen in Fig. 3.9 where 443 dense keypoints are detected with a threshold of 10. A comparable but sparsely distributed number using SURF keypoints alone can only be found by increasing the threshold to 50 (Fig. 3.10 (middle)).

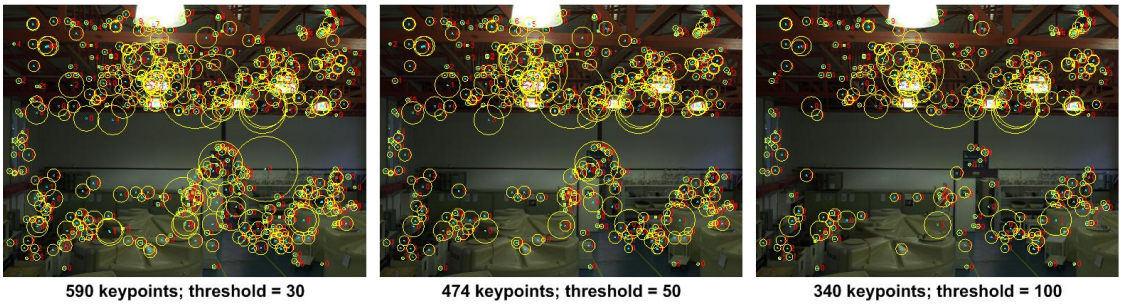

Figure 3.10: Increasing threshold values reduces the sensitivity of the SURF keypoint detector, so that there are less initial keypoints detected.

This section has shown that instead of being two redundant representation of the same scene, salient ROIs and keypoints are in fact *complementary* to each other. This strategy of combining both salient ROIs and keypoints is detailed in section 5.1.4.

3.2.3 State of the art on keypoint detectors and descriptors

Having introduced the concept of image keypoints, this section reviews some current works in this domain.

The amount of work in the literature dedicated to finding stable and robust

regions that are *repeatable* and *unique* over a large viewpoint change (or a change modelled by an affine transformation) is immense. The highly cited work of Mikolalczyk *et al.*[73] surveyed and compared several *affine-invariant* region detectors; the follow-up survey by Mikolalczyk and Schmid [75] on *local descriptors* compared several algorithms proposed by different authors. A new descriptor termed *GLOH* (Gradient Location and Orientation Histogram) is also introduced. GLOH is shown to outperform SIFT which was the best performing descriptor in an earlier survey by Mikolalczyk and Schmid [74]. Both GLOH and SIFT are 128D vectors but GLOH is more computationally expensive than SIFT as it applies an additional PCA (Principal Components Analysis) step to arrive at the descriptor vector. The work in this thesis uses the recently introduced SURF descriptors that further improves on these works.

An important extension of the evaluation framework described above to 3D keypoints, proposed by Moreels and Perona [81, 82], addresses the fact that affine-invariant keypoints may not perform as well as what the authors claimed in [73–75]. The main criticism is that the evaluation data, available online at `http://www.robots.ox.ac.uk/~vgg/research/affine/index.html`, consists of planar images that allows the affine model of deformation to be valid. For 3D objects, however, this assumption no longer holds and there exists no simple mathematical model to predict these deformations (Fig. 3.11). The evaluation in [81, 82] concludes that a combination of the *Hessian-affine detector* (a variant of a corner detector) and

SIFT descriptor give the best results. The work of Fraundorfer and Bischof [40] have also explored and evaluated the usefulness of such descriptors in 3D objects and confirms that the conclusion in [73] is only valid for planar scenes.

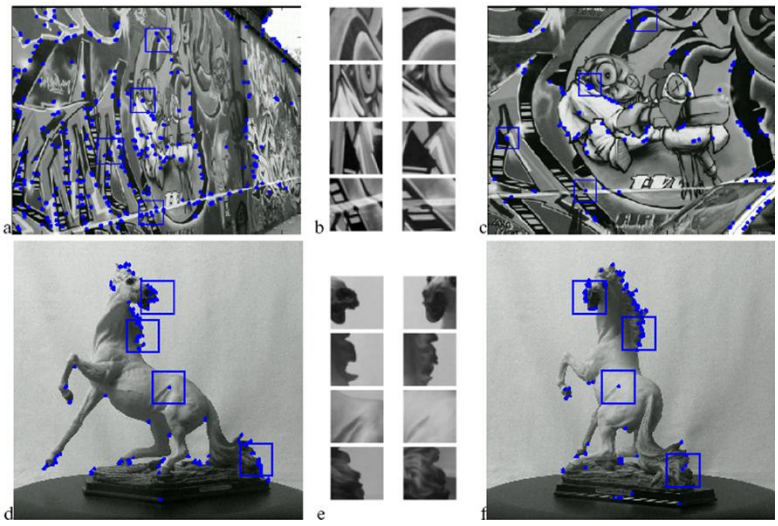

Figure 3.11: (Top row) Two views of the "grafitti" scene from [73] under significant viewpoint change (left and right). Matched points are shown in the middle and can be modelled by an affine transform. (Bottom row) Two views of a complex 3D object (left and right). Matched points from the two views are significantly different and is hard to model (middle). Data from [82].

SURF, introduced by Bay *et al.* [10], combines the results of these previous works and attempts to improve the efficiency of SIFT by combining a *Fast-Hessian* detector together with a descriptor based on the distribution of *Haar-wavelet* responses limited to 64 dimensions. The speed of the SURF algorithm draws mainly from the concept of *integral images* introduced in [115] where the time needed to compute the SURF keypoints by convolving the image with large box filters are reduced significantly. Experimental results in [10] showed that

SURF outperformed the current state of the art (SIFT and GLOH as well as many others reviewed in [75]) in terms of recognition accuracy and speed for CBIR applications. This makes the SURF algorithm the descriptor of choice in the proposed SRS. For this work, the latest version of SURF available online (`http://www.vision.ee.ethz.ch/~surf/`) is used. The structure of SURF descriptors and how they are exploited in the proposed SRS are detailed in section 5.1.

Terminology

Since section 3.1, the terms *features*, *keypoints*, *salient ROIs* and *landmarks* have been used, and it is important to differentiate between them. For clarity, their definitions are given below.

Definition 3.1. *Feature* A feature is any pixel location on the image, occupying one or many pixels. It is the most general term that does not convey any specific meaning and it is used when no such meaning is needed.

Definition 3.2. *Keypoint* A keypoint is a special feature in an image that has been chosen by a keypoint detector, and encoded by a corresponding descriptor. Keypoints possess certain desirable characteristics such as affine-invariance and good localisation.

Definition 3.3. *Salient ROI* A salient ROI is a particular region in an image

detected from a saliency map. It is thus formed from a group/ensemble of features. In the proposed SRS, these salient ROIs are the initial 'landmarks' that are further validated by other intermediate steps.

Definition 3.4. *Landmark* A landmark is a particular region in an image that possesses special characteristics and meaning. A landmark or group of landmarks are used to identify a scene uniquely. A landmark is composed of a group/ensemble of features. In the proposed SRS, landmarks are validated salient ROIs initially extracted from a saliency map, encoded by keypoints descriptors.

Concluding this section, the important concept of image keypoints as the pixel equivalent of salient ROIs is introduced. The importance of using keypoints and salient ROIs together is also highlighted by the improvement in the reliability and efficiency of the proposed SRS. Finally, a short review on the current keypoint detectors/descriptors in the literature and an evaluation of the different algorithms available led to the choice of using SURF descriptors in this work.

3.3 Ordinal measures of spatial configuration

This section highlights one of the most crucial concepts that this work exploits so as to improve the robustness of the proposed SRS against viewpoint distortions. As was pointed out in section 3.1, real scenes (indoors and outdoors) often possess a lot of similar features such that matching a landmark or even several landmarks

alone is insufficient and unreliable to determine the equivalence between two scenes. Instead, the *spatial configuration* of the landmarks provides better discriminatory information for scene recognition to be successful (section 3.3.1). In order to improve the reliability of the proposed SRS, this thesis uses the concept of *ordinal measures* that detects the spatial configuration of the landmarks and at the same time improves the robustness of the proposed SRS to viewpoint changes. A mathematical treatment of ordinal numbers and measures is found in section 3.3.2. The robustness of ordinal measures is reviewed in past works in section 3.3.3. This thesis proposes a novel idea of ensuring viewpoint invariance by extending the use of ordinal measures to the spatial configuration of the landmarks. This concept is illustrated in section 3.3.4.

3.3.1 Spatial configuration of landmarks

Using only the landmarks alone in determining scene equivalence lacks the *global* understanding of the complete scene structure when the spatial configuration of the landmarks with respect to one another is considered instead. Although a certain feature (a particular leaf from a particular tree, or a particular man-made object) may be found in many scenes of the same environment, it is highly unlikely that this feature from two different scenes will be found with the *same* spatial relationship with other neighbouring features (Fig. 3.12).

Figure 3.12: Two indoor scenes with the same features with very different spatial configuration.

In other words, although the uniqueness of a single feature may not be reliable, by invoking the configuration of the feature together with its neighbours, their uniqueness is greatly enhanced. This idea of using a local configuration of landmark features has been explored in [98] which termed it as a *semi-local* constrained matching that gave good image retrieval results. In the proposed SRS, scene equivalence is determined by the preservation of the global landmark configuration between matching scene features. Note that the landmark configuration extends into the z (depth) direction. The inclusion of depth information, explained later in section 3.5, plays an equally important role in improving the discriminatory power of the proposed SRS.

3.3.2 Ordinal numbers and rank correlation metrics

This section presents a brief mathematical treatment of ordinal numbers and how they can be compared using rank correlation measures. These concepts are exploited in the design of the proposed SRS so as to provide robustness against

viewpoint changes (section 3.3.3).

A *cardinal* number is a number that indicates quantity or size but not order except by comparison with another cardinal number. For example, the cardinal number three represents a specific quantity, but it is only by comparison with another cardinal number say five that one can conclude that three is smaller than five. For cardinal numbers to exist, there must be a measurable scale for quantification.

An *ordinal* number on the other hand indicates order or position in a series, such as first, second etc. Formally ordinal numbers in set theory are defined as the order type of a *well-ordered set* and is an extension of whole numbers proposed by Cantor [28]. In this thesis, all ordinal numbers belong to the set of finite ordinals, Ω defined by Rubin [94]:

Definition 3.5. Let Ω denote the set of finite ordinals or ordinal numbers where $\Omega = \{1, 2, 3, \cdots\}$. □

This thesis will thus denote ordinal numbers using positive integers as defined in 3.5. Although this notation does not distinguish between ordinal and cardinal numbers, the context of the problem should make clear which number space the measurements are placed. When objects are arranged in an order according to some continuous and measurable quality, they are said to be *ranked* with respect to that quality. The whole arrangement of these objects is called a *ranking*. Hence ordinal numbers indicate the respective rank or position of an object in the ranking.

Ordinal numbers are thus used when the precise quality cannot be measured or cannot be measured with reliability for practical or theoretical reasons.

Ordinal measures are defined as mathematical computations using ordinal numbers. In particular, this thesis is interested in the *similarity* of the two rankings. *Rank correlations* of ordinal numbers are used for this purpose. In this thesis, two common rank correlations are used: Spearman's ρ, S_ρ and Kendall's τ, K_τ [56, 70] defined respectively in (3.1) and (3.2).

$$S_\rho = 1 - 6\frac{\sum d^2}{n^3 - n} \tag{3.1}$$

$$\begin{aligned} K_\tau &= \frac{2S}{n(n-1)} = \frac{2(P-Q)}{n(n-1)} \\ &= 1 - \frac{4Q}{n(n-1)} = \frac{4P}{n(n-1)} - 1 \end{aligned} \tag{3.2}$$

In (3.1), d is denoted as the *difference* between two rankings. For example, if the ranking of a certain quality X is 3 while that of another quality Y is 8, the difference in ranks is simply $8 - 3 = 5$. This may seem at first illogical as how can one subtract "third" from "eighth"? The implication of this difference can be inferred as the difference in *preference* of one quality over another one. A rank of 3 for X means that there are two other members in priority over it in that particular ranking. Similarly for Y, there are seven members preferred over it in

that ranking. Hence the difference of 5 between the two rankings shows by how much the number of preferences in Y exceeds that of X. This number is thus a cardinal number as it has a specific value that arises from counting [56]. n is the *maximum* number of ranks in the data considered. For example, given a class of twenty students ranked by their height, $n = 20$ in this case. S_ρ is often considered in the literature as a *coefficient of weighted inversion* between the ranked data sets [56]. See Example A.2 for a demonstration of the computation of S_ρ.

In (3.2), as in (3.1) n is the maximum number of ranks in the data considered. S is the *score* obtained by comparing the rank of each element in the two rankings. This score is obtained by considering the rank of each element in both sets with respect to the other elements in the 2 sets. Comparison is done pair-wise and if the comparisons are concordant with respect to the *natural order* which is defined as increasing ranks, a score of $P = +1$ is given and a score of $Q = -1$ is given if the comparisons do not respect the natural ordering. Summing up the individual scores yields S. Hence $S = P - Q$. The number of possible pairwise combinations possible is given by $P + Q = \binom{n}{2} = \frac{1}{2}n(n-1)$. The different variations of K_τ are thus derived as shown in (3.2). K_τ is often seen as a *coefficient of disarray* between the data sets considered [56]. Refer to Example A.3 for a demonstration of how K_τ can be computed from sample ranked data.

As with all standard correlation metrics, both S_ρ and K_τ range continuously from $\{-1 \cdots 1\}$. It is expected that both rank correlations will *not* give the same

results except for the cases of perfect *positive* correlation (with a score of 1) and perfect *negative* correlation (a score of -1). From [56], it was shown that the differences between the two rank correlations lie in their scale of representation - S_ρ gives greater weight than K_τ to inversions of rank which are further apart. In practice, it is found that for values that are not too close to unity, S_ρ is 50% larger in magnitude than K_τ but this is not a fixed rule.

Other rank correlation measures exist in the statistical literature, such as Ulam's distance, τ_B, that was used in [11] as well as Kemeny and Snell's distance, d_{ks}, in which a normalised version was proposed in [69]. These measures present possible future extensions of the proposed SRS.

3.3.3 Robustness from ordinal measures

In order to further improve the robustness of the SRS against viewpoint distortions, a certain amount of tolerance to changes in the spatial relationship of the matched features must be allowed. In [8], templates of features are stored in a local database and are robustly matched to a potential scene using the idea of *elastic template matching*. The idea is to allow the templates (or features) a restricted amount of freedom to deform their spatial configuration resulting from viewpoint changes. Similarly in the proposed SRS, this 'freedom' for the features to deform their spatial configuration is incorporated by measuring a similarity of the spatial relationship

in an *ordinal* scale.

The use of ordinal scales of measurement had been adopted to improve the robustness of 2D pixel correlations, most of them in stereo matching problems (for a good overview see [16, 52]. An extensive evaluation can be found in [96]). The conversion to ranked pixels values is demonstrated in example 3.1 below:

Example 3.1. Suppose an image patch, $\mathbf{W}$ of size 3×3 containing raw pixel intensity values is converted to a ranked pixel image patch, $\mathbf{R}$. Since there are 9 elements in the patch, the ranked pixel values range from $\{1, 2, 3, \cdots, 9\}$:

$$\mathbf{W} = \begin{bmatrix} 10 & 30 & 70 \\ 20 & 50 & 80 \\ 40 & 60 & 100 \end{bmatrix} \Rightarrow \mathbf{R} = \begin{bmatrix} 1 & 3 & 7 \\ 2 & 5 & 8 \\ 4 & 6 & 9 \end{bmatrix}$$

Now lets suppose $\mathbf{W}$ is subjected to certain perturbations that distorts one of the raw pixel intensity value such that $\mathbf{W} \rightarrow \mathbf{W}'$. Converting $\mathbf{W}'$ to ranked pixels values, $\mathbf{R}'$ reveals the robustness of ordinal measures:

$$\mathbf{W} \rightarrow \mathbf{W}' = \begin{bmatrix} 10 & 30 & 70 \\ 20 & 50 & 80 \\ 40 & 60 & \mathbf{255} \end{bmatrix} \Rightarrow \mathbf{R}' = \begin{bmatrix} 1 & 3 & 7 \\ 2 & 5 & 8 \\ 4 & 6 & 9 \end{bmatrix}$$

Since $\mathbf{R}' = \mathbf{R}$, this shows that the conversion to an ordinal scale improves the robustness of the pixel representation against random perturbations compared to using raw pixel intensity values. □

Zabih and Woodfill [120] introduced the *rank* and *census transforms* to increase the robustness of standard image correlation algorithms against occlusions and depth disparities while Bhat and Nayer[12] introduced κ as a robust measure of correlation between two image patches by converting to an ordinal scale. The κ in [12] is improved further by making it more discriminatory in [97]. Other authors

(*e.g.* [69, 103]) have also similarly introduced other measures for computing image similarity in an ordinal scale.

In the next section, an example scene containing very simple landmarks is used to show the robustness of ordinal measures against viewpoint changes.

3.3.4 Viewpoint invariance from ordinal measures

The robustness of ordinal measures in providing a viewpoint invariant representation of the scene is illustrated here. For the sake of clarity, a pair of simple images that contain only four distinct landmarks set against a white background are used. The only difference between the two images is a slight viewpoint change to the right (Fig. 3.13).

For the moment, one can assume that the SRS is able to reliably detect the four landmarks and that reliable correspondences can be found between the two images. The arrows in Fig. 3.13 show how the positions of the four landmarks have changed as the camera shifts to the right. Although the landmarks have changed in their *absolute* positions in the (x, z) directions, their ordinal positions remain *invariant* - the landmarks maintain the same order with respect to one another. The preservation of this order can be detected using ordinal measures of the detected landmarks. In other words, all the metric information of the landmark's position is converted to an ordinal scale. Hence only the ranks of the relative positions of the

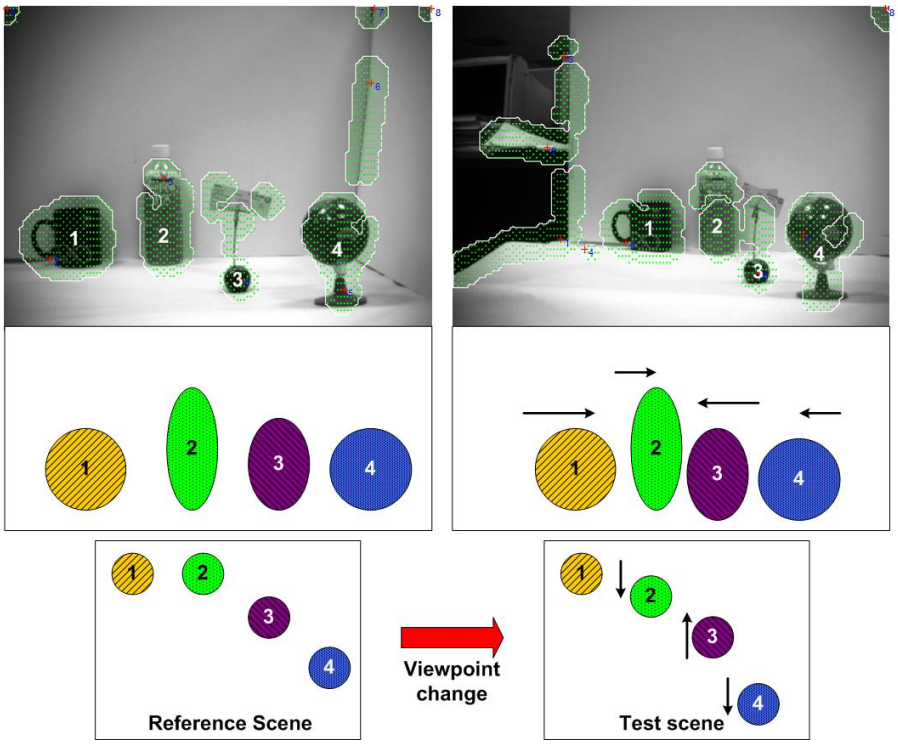

Figure 3.13: Slight viewpoint change between two images. The numbers indicate the detected features that are highlighted.

landmarks are of concern. *Rank correlations* of the matched landmarks, introduced in section 3.3.2 as a measure of similarity, can then be used. The proposed SRS computes S_ρ and K_τ (3.1,3.2) rank correlations over the three spatial directions (x, y, z) (section 6.1.2 and (6.5)). A significant degradation in correlations scores indicate that the ordinal positions of the matched landmarks are not preserved and the scenes could be different. The case of a positive match is illustrated in Fig. 3.14. As some mismatches are likely to occur, it is expected that the value of $S\rho$ and K_τ will not be exactly 1 (perfect correlation), even for similar scenes.

In conclusion, this section has highlighted the importance of using the spatial

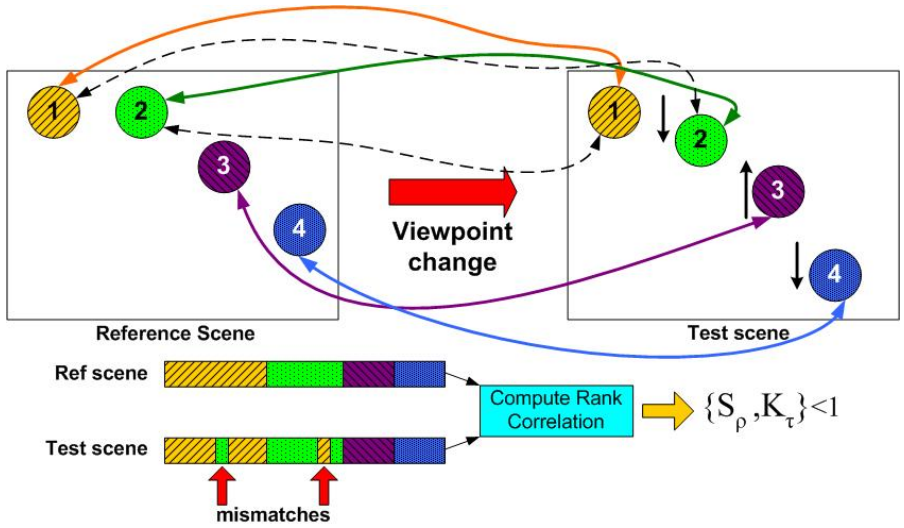

Figure 3.14: Computing the rank correlations of a positive test scene. The solid arrows represents correct matches while the dashed arrows represents wrong matches. The colour blocks shows the matched features of the two scenes. The rank correlations are computed over the three spatial directions (x, y, z).

configuration of landmarks that provides a complete and global view of the scene structure for effective recognition. The concept of ordinal measures and its usefulness in providing a robust measure in image processing applications are also briefly presented. Finally, an illustration of how rank correlations provide a robust measure to detect changes in the spatial configuration of scenes under various viewpoint changes is shown. The details of how ordinal measures of spatial configuration are incorporated in the proposed SRS is found in section 5.3 as part of the *Scene matrix*, $\mathbf{m}_s$. The use of rank correlations as a measure of similarity between two scenes is detailed in section 6.1.2 using the *Global Configuration Coefficient*, G_c.

3.4 Illumination invariance using HSV colour space

This section describes the concept of an *illumination-invariant* representation of a scene as well as its importance to outdoor scene recognition. The main challenges of illumination changes are described in section 3.4.1 and a short review of past works that motivated the use of the HSV colour space in this thesis is presented in section 3.4.2.

3.4.1 Challenges of illumination changes in outdoor scenes

For the proposed SRS to be effective in natural outdoor environments, it must be robust to a variety of changes in illumination caused by the change in the position of the sun as well as changes in weather conditions. The effects of illumination changes on two similar scenes taken at different times of the day are shown in Fig. 3.15. The top scene was taken under bright sunlight while the lower scene was taken under diffused lighting on two different days. Furthermore, the two scenes were taken under different weather conditions - a clear sunny day (top) *vs.*a hazy overcast sky (bottom).

Analysing the RGB images of the two scenes reveal several challenges caused by illumination changes that make outdoor scene recognition particularly difficult. Firstly, there is a *global* change in illumination levels due to different weather conditions. The top scene has pixel intensity values that are on the average higher

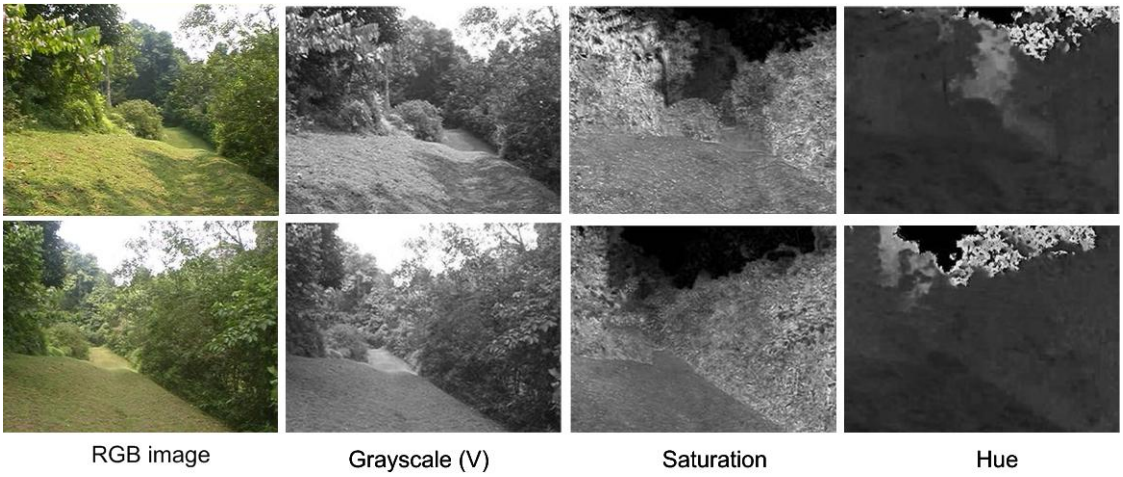

Figure 3.15: Two RGB images and their HSV components under different illumination conditions. Slight viewpoint distortions occur between the two scenes as they were taken on different days.

than those in the bottom scene. This is due to the fact that the overcast sky significantly reduces the amount of sunlight in the bottom scene. The haze particles in the air further diffuse the sunlight, resulting in an overall reduction in intensity values in the bottom scene. Secondly, the effects of shadows in the top scene caused by the dense natural foliage and strong sunlight further compounds the problem if one attempts to match it with the bottom scene which has virtually no shadows. Shadows in natural environments produce patches of *non-uniform* or *local* intensity changes that are unstable due to the movement of the sun. Such periodic and dynamic changes make it necessary to incorporate an *illumination-invariant* representation of the scene, described in the next section.

3.4.2 Illumination-invariant representations

Because of the problems posed by global and local illumination changes in natural outdoor environments, several methods of removing shadows from a scene are proposed so that an *illumination-invariant* representation can be achieved. The highly cited works of Finlayson and other researchers have explored several techniques to perform shadow removal. One of their work [36] showed how shadows can be removed from a RGB image so that a shadowless grayscale image is produced. This is achieved by determining a single scalar function derived from the RGB image that is invariant to changes in colour and intensity. Applying this function results in a 1D invariant image that depends only on its reflectance. By detecting the shadow edges where the only changes are in colour and intensity, shadows are easily removed from the invariant image. An extension of this work is shown in [37] where the final result is a shadowless RGB image. This is done by extracting edges from the original RGB image such that material edges are preserved while shadow edges are removed, using the shadowless grayscale image as a mask. The results are shown in Fig. 3.16.

Other works extended [36] by extracting the shadowless chromaticity image [32] while another work [38] removed shadows by combining Land's Retinex algorithm [59] together with information on the shadow's edge. Other methods transform the original RGB image to other colour spaces such as hue [35] or saturation [92]

Figure 3.16: The shadow removal algorithm of [36, 37]. Top row: Original RGB image (left). Grayscale illumination invariant (middle). Grayscale normal (right). Bottom row: Edge map of illumination-invariant image (left). Edge map of normal grayscale (middle). Final shadowless RGB image (right).

which are shown to be robust against illumination changes. This is the method that this thesis will adopt.

It is clear from Fig. 3.15 that merely using grayscale images is *not* robust against the effects of shadows caused by the foliage and changes in overall illumination levels. Any information concerning the landmarks, such as the SURF descriptors (section 3.2), derived from the grayscale image are affected by shadows. As a result, there will be few and incorrect correspondences if two scenes are matched. By comparing the saturation and hue images of the two scenes, one can see that the effect of illumination changes is almost completely removed in these two colour spaces. That is, the SURF descriptors in these colour spaces are almost illumination-invariant. However, the monotonic nature of the saturation

and hue images may reduce the uniqueness of the descriptors and this may cause mismatches to occur. In the proposed SRS, *all* the three colour spaces in HSV are used to achieve a reasonable compromise.

This section has presented the concept of illumination invariance and its importance for a robust scene recognition algorithm to function effectively in a dynamic outdoor environment. A short review of recent works in producing an illumination invariant representation, mainly through the removal of shadows, highlighted the usefulness of working in a different colour space other than RGB. This thesis uses the HSV colour space to improve the performance of the proposed SRS (see section 5.1.1) and the improvement in recognition accuracy compared to using grayscale alone is discussed in section 7.5.10.

3.5 Importance of depth information obtained from TBL motion

This section continues from section 2.3 where the question *"Why are TBL flights designed in such a fashion?"* is posed to address the purpose of TBL flights in bees and wasps. Unlike Bianco's hypothesis [13–15, 63] that TBL flights serve as a testing framework to determine stable landmarks across a TBL arc, this thesis argues that the special structure of TBL flights actually aids in depth recovery.

More precisely, such flights aid in *ordinal depth* recovery of the scene structure that is shown to be crucial for a viewpoint invariant representation of the scene (section 3.3.1). In section 3.5.1, the importance of depth information for scene recognition motivates the need to find a robust solution for its recovery. Section 3.5.2 gives an overview on TBL flights and shows how the special motion of this flight can be used by bees and wasps to recover ordinal depth information.

3.5.1 Importance of depth information

The importance of depth information for outdoor scene recognition was briefly mentioned in section 3.3.1 where it is required to describe the complete spatial configuration of an ensemble of feature points (or landmarks). This is especially true if the agent performing scene recognition is navigating in a confusing environment where only a few plant species dominate as shown in Fig. 3.17, resulting in a distinct lack of highly unique features. In this environment, the discriminating component for these ambiguous scenes are not the features themselves but how the features are spatially related to one another.

Therefore, depth information plays an important role in enhancing the discriminatory power of the proposed SRS for ambiguous scenes. Natural scenes that the SRS encounters are often devoid of highly unique features on its own (Fig. 3.17) such that 2D features alone, may not provide sufficient information to tell

Figure 3.17: Various scenes obtained at *different* positions along a footpath in an enclosed mangrove forest. Notice the strong similarity in these scenes due to the fact that only a few dominant plant species thrive in the harsh mangrove environment. The ambiguity of these scenes makes recognition very difficult.

the scenes apart. Rather it is often the spatial arrangement or configuration of an ensemble of features that defines a scene uniquely. Depth, being an integral component that defines this spatial configuration should thus be included as it enhances the information of each scene by a third dimension, the z dimension.

Besides reducing the ambiguity of difficult scenes, depth or more precisely ordinal depth was shown in section 3.3.4 to be a crucial component in providing a viewpoint invariant representation of a scene. From Fig. 3.13, one can see that changes in viewpoint for similar scenes preserve the relative orders of the spatial positions of the landmarks, including ordinal depth. Using rank correlations in the three spatial directions (x, y, z) enable the proposed SRS to effectively detect any

degradation in the spatial ordering of the landmarks, which could indicate that the scenes may be dissimilar.

Depth is also an important component in *prioritising* the importance of various landmarks in the scene. As was mentioned in section 3.1.1, certain regions in the scene provide important information that serves as good landmarks for scene recognition. Landmark selection at different depths from the agent is an important factor that should be taken into consideration. Landmarks that are *far* from the agent in the *background* such as the skyline provide stable and robust features that are relatively invariant to viewpoint changes and changes in scale (by moving forward and backward) (Fig. 3.18).

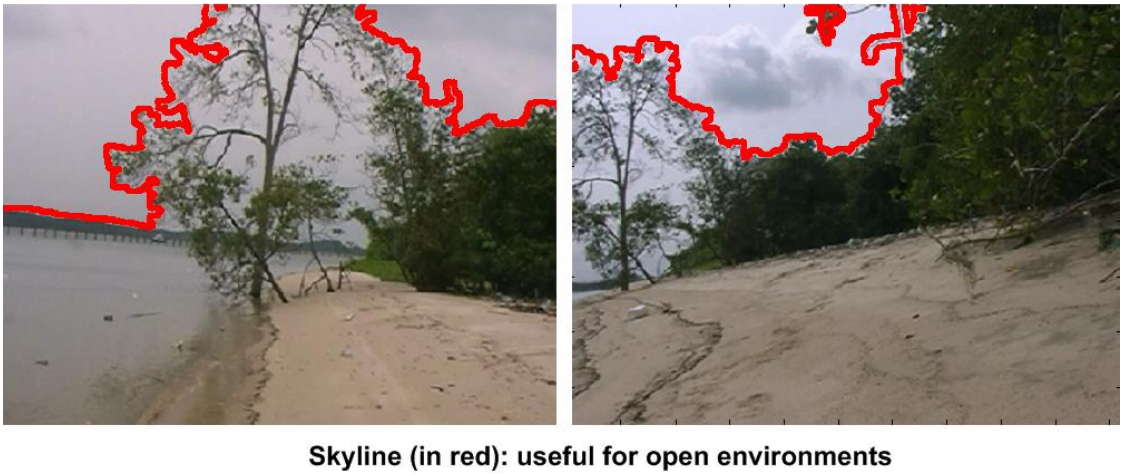

Figure 3.18: The skyline (indicated in red) serves as a very useful and important region for scene recognition in a relatively open beach environment.

However, unless the skyline is extremely obvious, the presence of foreground

foliage *near* the agent may easily obstruct the skyline. The skyline is thus vulnerable to occlusions. Furthermore, features that are far away may be so stable that they may appear in multiple instances of *different* scenes taken in the same environment. This may confuse a SRS that matches different scenes based on the far features alone (Fig. 3.19).

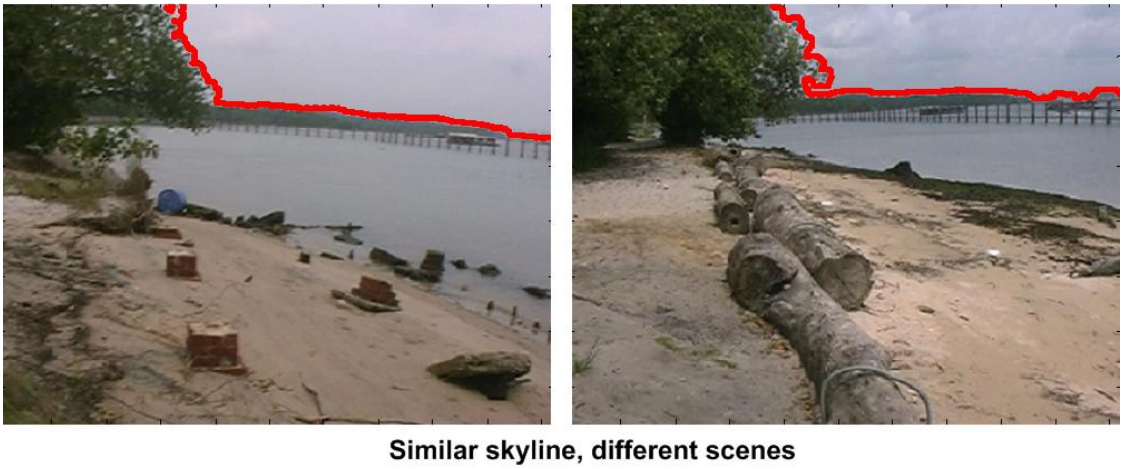

Figure 3.19: The far features in the skyline may not be useful in this case where two dissimilar scenes possess remarkably similar skylines. Notice that the foreground, however, is significantly different.

In order to discriminate between the two scenes shown in Fig. 3.19, it is necessary to detect landmarks in the *foreground*, near the agent. Near features close to the agent often identify the particular scene uniquely, and are important for the discrimination of scenes that share common background features. Furthermore, foreground features are also less likely to be occluded than background features since they are nearer to the agent. However, because of their proximity to the agent, even slight changes in the agent's pose will induce significant distortions to

these features. Near features are thus more vulnerable to viewpoint changes and changes in scale.

The inclusion of depth information allows the landmark extraction stage of the SRS to select and possibly *weigh* the importance of the landmarks differently based on their spatial positions in depth. For scene recognition that requires good resolution, the SRS must be highly discriminatory and sensitive to small changes so that closely separated scenes can be distinguished. This can be done by giving a higher weight or more importance to landmarks that are near. If the resolution required for the application is not high, then far landmarks can be given a higher weight so that a more global idea of the agent's location can be used. The work in providing a differential weighting in this thesis is described in section 4.4 in the design of a *depth-weighted* saliency map for landmark extraction.

Depth also provides supplementary information important for successful navigation. *Obstacle avoidance* is an important ability that requires the agent to determine if a nearby object is on a collision course. By computing the time to collision, the agent can change its original path to avoid possible dangers. Detecting an open area in the immediate surroundings helps in path-planning and this requires the agent to obtain reliable depth information. Before the agent plans the route to use, it must also ensure that the current path is clear and free from obstacles. If that is not the case, an alternate path that is safe and accessible must found. Without depth information, the agent will not be able to perform these

tasks effectively.

The next question is *how* can an artificial agent robustly recover this important depth information? The TBL motion, discussed in the next section, provides a possible solution.

3.5.2 Ordinal depth from TBL flight

Having motivated the importance of depth information for scene recognition, this section explains how the *motion* of TBL flights, introduced in section 2.3 where related works from biomimetics were reviewed, can be used as a robust and efficient mechanism for ordinal depth extraction.

TBL flights

TBL flights, also known as *zig-zag flights* [116] or *learning flights* [121], are observed in certain species of flying hymenopterans; (bees and wasps) (Fig. 3.20) which they perform the first few times they approach or leave important sites, such as their nests or new feeding grounds. The first two wasps (*Bembix* sp. and *Bembecinus* sp.) shown in Fig. 3.20 practice what is known as *progressive provisioning*. That is, the parent wasps have to provide the larvae with several helpings of the insect prey to complete their life cycle. In order to do so, the parent wasps relocate and open their nest several times to deliver the food. The wasps are remarkable in

Figure 3.20: Three examples of common species of wasps from Singapore, all taken along the sandy shores of Pulau Ubin. All the wasps are from the Digger Wasp Family (Family *Sphecidae*). Left to right: *Bembix* sp., *Bembecinus* sp., *Tachysphex* sp. Thanks to Dr. Cheong L.F. for the beautiful pictures.

locating their well-hidden nests. After the nests are dug in the sand, the parent wasp will carefully fill the entrance with sand, levelling the sand so that there is no obvious marking. *Bembix* sp. hunts fierce insects such as Robber Flies and other wasps. *Bembicinus* sp. hunts Homopteran insects. In Fig. 3.20 (middle), the wasp is opening up its nest with the paralysed prey (a leafhopper) firmly clasped under its abdomen. *Tachysphex* sp. usually hunts for Orthopteran insects such as crickets. In Fig. 3.20 (right), the wasp has paralysed a cockroach which is partially occluded by the body of the wasp.

The motion of several TBL flight paths are shown in Fig. 3.21. In a detailed and illuminating study of TBL motion, Lehrer [61] conducted extensive studies on the physiology and characteristics of TBL flights in both controlled indoor and outdoor natural feeding dishes. Her observations on this particular form of flight performed by the honeybees inspired her to coin the term "Turn-Back-and-Look",

because that exactly describes what the bee's movement. From Fig. 3.21, it is clear that not only does the bee turn back and look at the goal site, it also performs a beautiful motion that consists of a *series of increasing arcs* near the vicinity of the goal. For indoor targets where there is only a single entrance to the feeding dish housed in a dark container, all the TBLs are performed in front of the entrance. For open feeding dishes situated outside in a garden, the bees are observed to perform TBL at various directions with respect to the sun's position in the sky - TBL is always performed with the sun behind the bees. The reason for this behaviour is not exactly know, but a possible explanation could be that flowers are often phototropic, that is, they follow the sun's movement over the course of the day. Bees could be using this strategy so that the flowers can be viewed at the best position and illumination for recognition later.

Another interesting observation is that TBL is performed not only once but is *repeated* several times. From Fig. 3.21(left), the first few TBLs were very long and consist of very elaborate arcs around the target site lasting often as long as 8s. The last few TBLs were very short (2s) with few arcs. The conclusion from this observation is that TBL serves functions other than to learn the position of the goal target, since the bee had to return successfully in order to perform subsequent TBLs. It is likely that the bees were learning something about the target that may not have been immediately important for knowing "where" the target is but is crucial for future return trips. This is believed to be linked closely to the arc-like

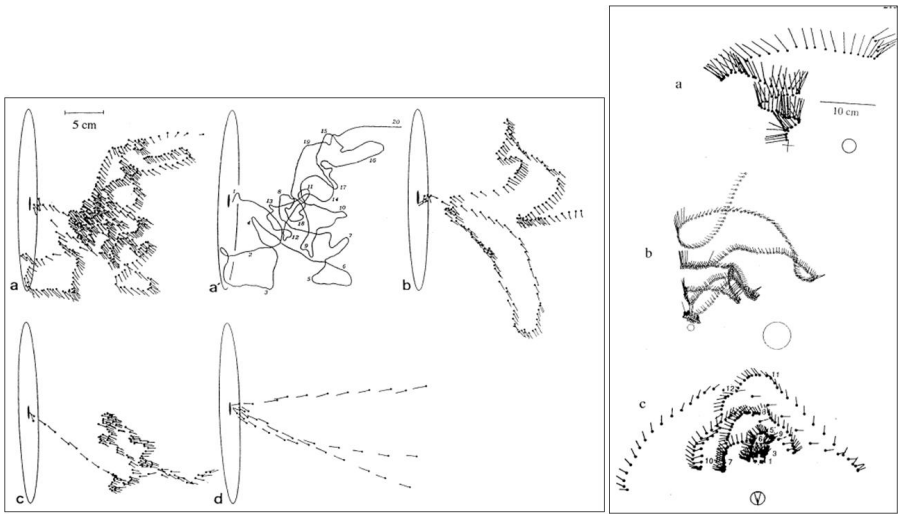

Figure 3.21: Several TBL flight paths from bees (left) and wasps (right). The left panel shows the different stages (a) to (d) of a honeybee performing TBL repeatedly in front of an indoor feeding dish, shown as a black dot surrounded by an oval. The first few TBLs (a) and (b) are longer and contains more arcs and repetitions while the final TBL (c) is much shorter. Once the TBL phase is over, the bees fly directly away from the feeding dish (d). The right panel shows the similarity of TBL flights with landmarks denoted as circles of several species of wasps. Notice the significant arc-like motion made by the wasps. All data from [63].

motion of the TBL flight itself shown in Fig. 3.21(right).

Recovery of ordinal depth from TBL motion

The TBL motion consists of a series of arcs centred about an object of interest or target, with the direction of translation almost *perpendicular* to the line of sight of the insect. The characteristics of TBL motion was measured in precise studies reported in [116] (Fig. 3.22). It is thus highly likely that the significant translational component in the TBL flight enables the insects to recover depth

information of the target scene, shown to be important for scene recognition in section 3.5.1.

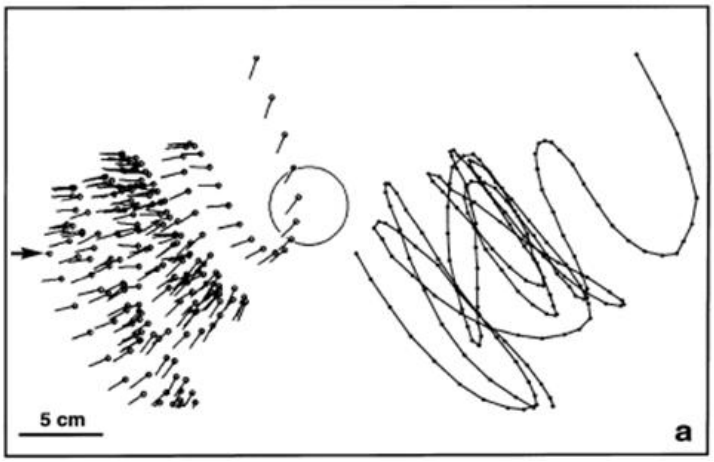

Figure 3.22: The TBL of a wasp recorded in [116]. The circle is the target. Notice the significant translational motion that is almost perpendicular to the target at each arc formed. The complete path is shown on the right.

Note that there exists another kind of flight known as *orientation* or circular flight [61, 121] that has important technical differences to TBL flights. Such flights are longer and more obvious as the insects fly around in large increasing circles above the target goal site. From [61], such orientation flights are performed as soon as the bees fly out of the laboratory after collecting the food reward from the indoor feeding dish. Lehrer hypothesised that the orientation flights serve to help the bees in locating landmarks that are *far* from the goal site. Although both TBL and orientation flights are apparently dissimilar, this thesis argues that the essence of these motions and thus their main purpose seems to be the same. In both cases, there is significant *translational motion*, with a greater rotation about the y axis

in an orientation (circular) flight than a TBL arc. Once again, this translational motion aids in depth recovery as there is no forward motion at all, which would have otherwise made the recovery very difficult [22].

The *real* purpose of TBL is not known, but numerous observations from the literature [62, 64] have suggested the usefulness of TBL in choosing suitable landmarks for scene recognition. Besides using *apparent size*, it was concluded in [106] that bees do in fact obtain *distance* information from selected features so as to perform precise landing at feeding sites and at their nests [107]. Since TBL is also performed at these locations, it is highly likely that TBL is used to obtain distance information related to the selected features [25, 61, 121]. Since insects such as bees do not have stereo vision, the most likely way to compute an estimate of this depth is through optic flow arising from TBL flights [61, 121]. Based on computer simulations of TBLs performed by wasps, Voss and Zeil [116] provide a qualitative analysis of the optic flow vectors induced by such a motion and suggest how 3D information of the scene might be obtained (Fig. 3.23).

As the TBL arc contains both translational and rotational components, any depth information computed from optical flow measurements will be imprecise unless the rotational components can be accurately estimated. From [22], it was shown that although *metric* depth information may be unreliable under such conditions (lateral translation mixed with rotation), the *depth orders* can be obtained with great robustness even when there are significant errors in the motion estimates

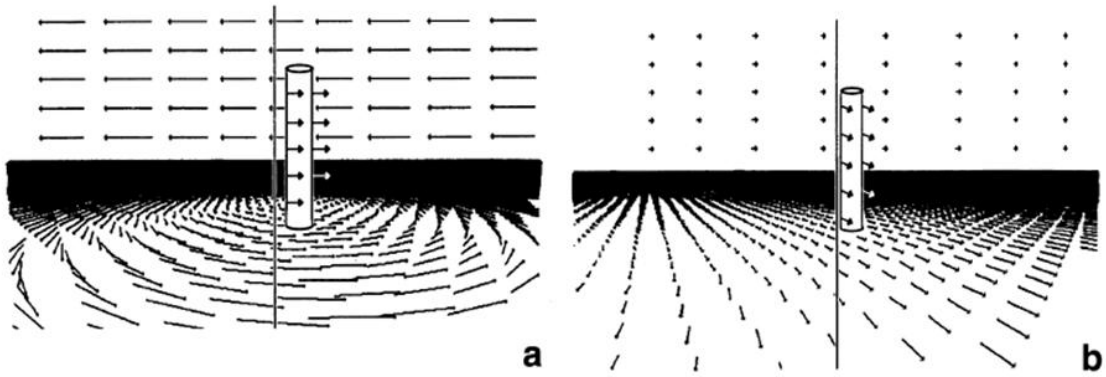

Figure 3.23: Simulated optical flow as viewed from the head of a wasp undergoing TBL motion. The target is a simple cylinder in the middle of the scene. (a)Optical flow induced as rotational and translational vectors are adjusted to place the object in the centre. (b)Optical flow induced at the start and end of a TBL arc where there is nearly pure translation, as the wasp is about to change direction. Data from [116].

or camera focal length. This motivates the use of ordinal depth in the proposed SRS.

For the artificial agent, TBL thus provides a possible method to robustly recover the depth information so important for effective scene recognition (section 3.5.1). Although many depth cues are available in an image, in the context of navigation, *motion* information provides the most natural cue for depth recovery. Unfortunately the structure from motion problem is a notoriously ill-posed problem, from which it is very difficult to recover accurate *metric* depth information. This problem can be circumvented by mimicking the TBL motion of flying hymenopterans discussed in this section, as well as using only qualitative *ordinal* depth information, thereby eschewing the need for accurate metric depth recovery.

This section has provided an overview on the importance of depth information for effective scene recognition. A robust and efficient method of ordinal depth recovery, inspired from the TBL motion of bees and wasps is also presented. How the proposed SRS exploits TBL motion to enhance the salient ROIs with ordinal depth information is detailed in section 5.2.2. The contribution of ordinal depth to the recognition accuracy of the SRS is discussed in section 7.5.5.

3.6 Final remarks

This chapter has introduced several important concepts related to the proposed SRS. It is helpful to link these core concepts, summarised below, to the parts of the thesis for easy reference. All the references link to the specific sections where the concepts are first introduced.

- *Visual saliency* for reliable landmark detection (section 4.1)
- Encoding of the salient ROIs using *SURF descriptors* (section 5.1.4)
- *Ordinal spatial configuration* in the Scene matrix (section 5.3)
- Determining scene similarity using *rank correlations* (section 6.1)
- *Illumination invariance* using HSV colour space (section 5.1.1)
- Recovering ordinal depth from *TBL motion* (section 5.2.2)

The rest of the thesis will focus on presenting the design of the proposed SRS using the concepts introduced. The next chapter begins this description by detailing how the proposed SRS selects the initial salient ROIs using visual saliency.

Chapter 4

Visual saliency for landmark extraction

This chapter describes the first stage of the proposed SRS with the selection of visually salient initial landmarks, or *salient ROIs* from a computed saliency map. The computational model of visual saliency adopted in this work is described in section 4.1. This computational model includes several enhancements such as the inclusion of new *composite feature* maps described in sections 4.2 and 4.3. The final output is a *depth-weighted* saliency map (section 4.4) where the regions in the foreground are made more salient than regions in the background using an estimated *dense ordinal proximity map*. The depth-weighted saliency map indicates regions of high salience as potential salient ROIs which are then extracted using simple image morphological operations (section 4.5).

4.1 Modified Itti's computational model of visual saliency

The usefulness of visual saliency in selecting initial salient ROIs for scene recognition had been explained in sections 3.1.2 and 3.1.3. In this section, an overview of the modified computational model of visual saliency by Itti is detailed.

The computational model of visual saliency proposed by Itti *et al.* [51] is one the most popular models of the human attention system. The original work made predictions of how human attention shifts from one salient region to another, by applying a Winner-Takes-All (WTA) algorithm on the resulting saliency map that is produced. The original model is shown in Fig. 4.1 (left). In this work, a modified version of this model that uses several new *composite features*, $\mathbf{C}_f$, suitable for scene recognition in outdoor environments is proposed (Fig. 4.1 (right)).

Definition 4.6. *Composite features* Composite features, denoted as $\mathbf{C}_f^j$ for the j^{th} composite feature, are image regions that are deemed to be potentially salient. They are used by the saliency algorithm to determine if salient ROIs exist in these regions. □

Apart from the original intensity and orientation composite features, a modified *opponent-colour* composite feature using the concept of *colour constancy* for an illumination invariant representation (section 3.4.2) proposed in [111] is used. This

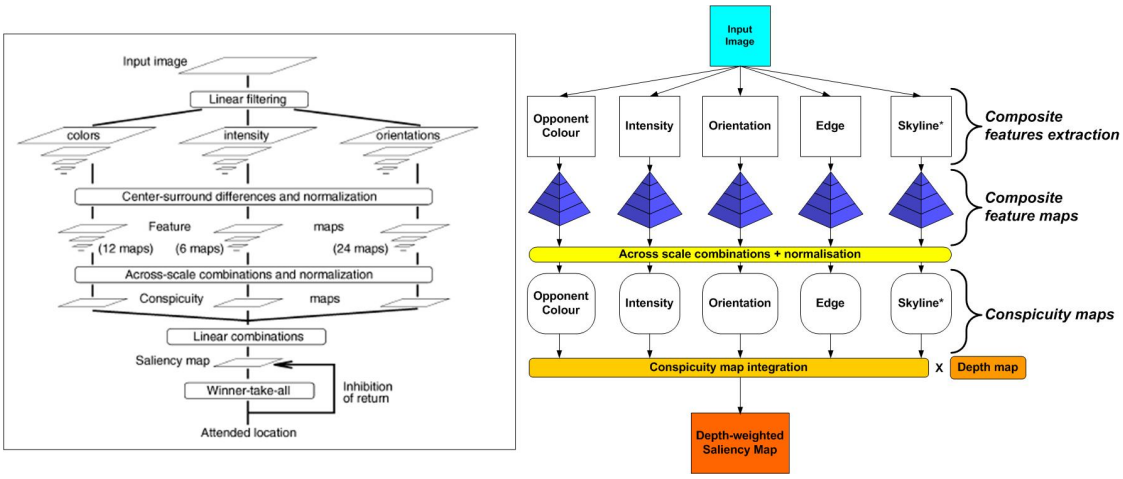

Figure 4.1: Left: Itti's original computational model of visual saliency [51]. Right: The modified computational model of visual saliency with several new composite features.

representation was shown in [111] to be able to robustly extract salient regions under varying illumination. Furthermore, two new composite features, namely *long edges* and *skyline*, are extracted from the image if they exist. The salience of each composite feature is computed and represented as a *conspicuity map*. Unlike the equal weights that are used in [51] to combine the conspicuity maps together, the *entropy* of the individual conspicuity map is used instead to weigh each map before combination to form the saliency map. A final additional step involves modulating the saliency map with an estimated *dense ordinal proximity map* so as to finally obtain a *depth-weighted* saliency map.

The three composite features - intensity, orientation and opponent-colour, are extracted by various algorithms detailed in [51, 111]. These features are well-known in the psychophysics literature as potentially salient regions that model well the

human attention system. The intensity features are extracted by simply converting the original RGB image to grayscale while the orientation features are obtained by convolving the intensity (grayscale) image with a set of Gabor filter banks [85]. The four opponent-colour channels R', G', B', Y' are computed directly from the *normalised* RGB channels, r, g, b. See [51] for details of the computations. The various composite features extracted are shown in Fig. 4.2. The next two sections focus on the remaining two novel composite features introduced in this thesis.

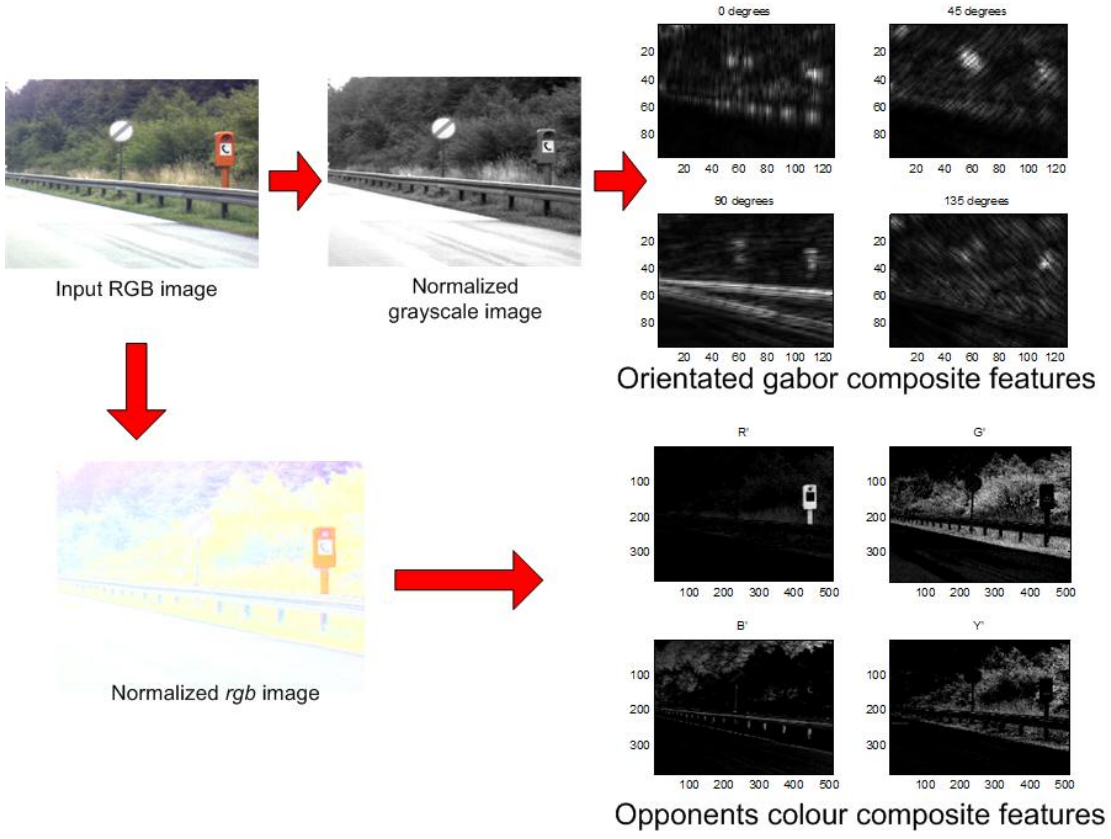

Figure 4.2: Various composite features extracted from the input RGB image: intensity (grayscale), orientation (gabor) and opponent colours.

4.2 Detecting long edges as composite features

The contribution of this thesis is the addition of two new composite features, *long edges* and *skyline*, so as to enhance the robustness and accuracy of the final detected salient ROIs in natural outdoor environments. Long edges are known in the literature as an extremely useful and viewpoint invariant salient feature [47, 105] that are robust against illumination changes and occlusions. Such long edges were exploited in natural sceneries to reliably detect tree trunks for outdoor visual SLAM [5, 6] and to detect certain species of trees for navigation [17]. In this work, the edge composite map is extracted by applying Canny's method [18] on the intensity image (Fig. 4.3).

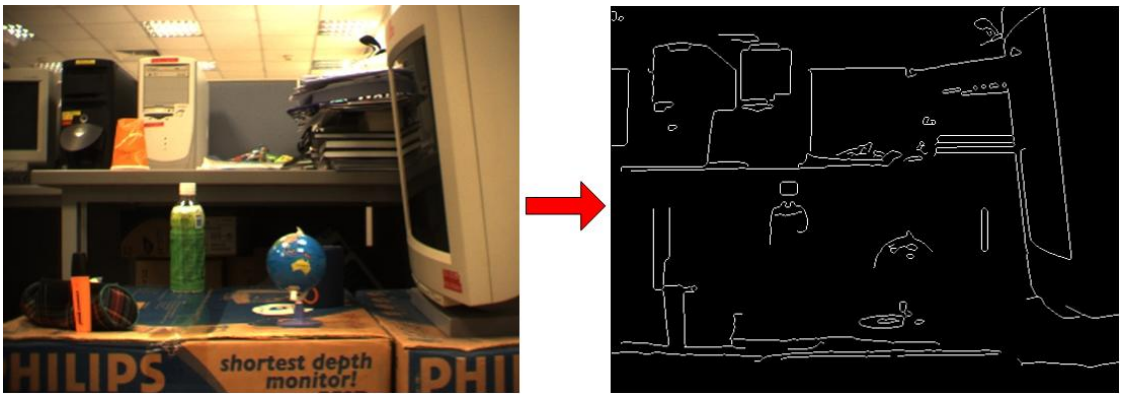

Figure 4.3: An edge map (right) detected for the saliency algorithm using Canny's method.

4.3 Skyline as useful composite features

The skyline is used by several authors in past works for scene recognition in navigation. This is especially true for flying vehicles where the segmentation of the sky from the ground is important in determining the bank and yaw angles, which are needed to ensure the stability of the vehicle [109, 110]. For land-based navigation, faraway landmarks such as mountain peaks are detected and used for localisation ([7, 26, 27, 108]). A recent and original approach in [53] showed that the skyline of buildings in an urban environment offers a simple but effective representation for scene recognition. The use of the skyline for scene recognition has also been hypothesised by behavioural scientists for certain species of bees and wasps [79]. It is known that these flying hymenopterans have a visual system that is sensitive to UV, green and blue colours. A colour-contrast mechanism is proposed that enables such insects to extract the skyline reliably under a variety of illumination conditions. In particular, the sky contains a larger proportion of UV than the (typically) green vegetation which absorbs the UV from the sky. The UV-green contrast mechanism was shown in experiments by [79] to be reliable in segmenting the sky (which looks brighter) from the vegetation (which looks darker) from a variety of lighting conditions at different times of the day (from dawn till dusk).

Motivated from these results, the utility of the skyline in the context of our proposed SRS is very clear. The skyline is in fact one of the most *distinctive*

and *robust* features that conveys a coarse idea of the agent's position. As was briefly mentioned in section 3.1.1, the skyline remains unchanged for significant distortions in the agent's pose. Lateral translations do not affect it much and it is almost invariant to changes in depth since the skyline is composed of objects in the far distant background. The drawback of using the skyline lies in the fact that it compromises the discriminatory power of the SRS (see section 3.5.1) - it is possible that two very *different* locations possess similar skylines. Nevertheless, it is shown in the experimental results using real outdoor scenes (section 7.4) that the complexity of the various features (trees, bushes and other objects) makes it *highly unlikely* that such an ambiguous scenario will occur.

The skyline is detected from an image by assuming that the sky has the following properties: 1) It is in the top half of the image; 2) It is more luminous (brighter) than the ground; 3) It contains a higher percentage of the blue colour component when compared to the ground. Furthermore, as the sky contains relatively few objects, it is relatively textureless compared to the ground that contains abundant vegetation. The skyline is thus defined as the intersection between the sky and ground. Fig. 4.4 summarises the various stages in the skyline detection algorithm for a natural outdoor scene.

The first step of the algorithm detects edges using Canny's method, similar to section 4.2. Since the sky has less textures, there should be less edges extracted from the potential sky region. Assuming furthermore that the sky is at the top

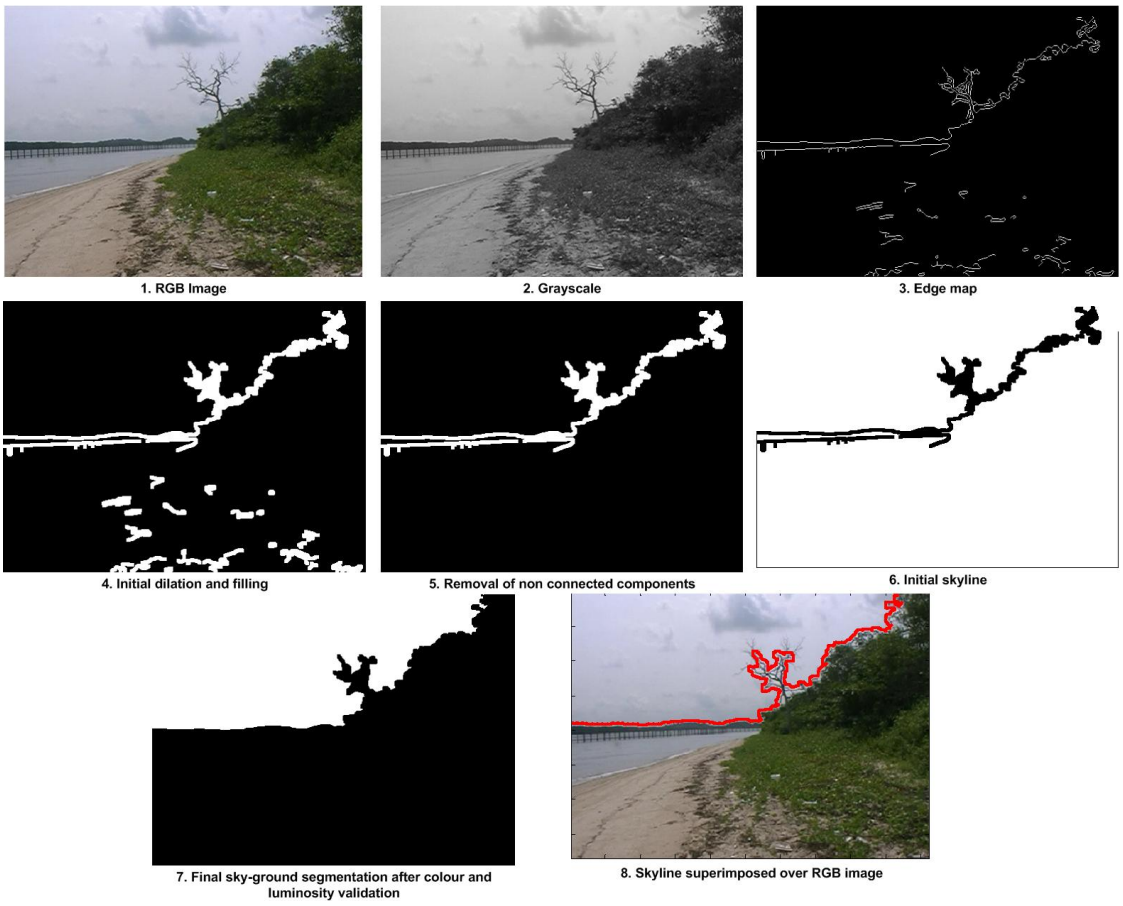

Figure 4.4: Steps (1–8) of the skyline detection algorithm. See text for details. The final skyline is shown in red superimposed over the original RGB image.

half of the image, the algorithm performs several image morphological operations of dilation and filling to create a *labelled image* that should represent the segmented sky-ground regions. In order to obtain the skyline, pixel columns are extracted and the first pixel counting from the top that shows a significant change in luminosity and blueness is classified as the skyline. The process is repeated until the full width of the image is processed. Note that the accuracy of this algorithm is based entirely on the assumptions stated above. There are some cases where these assumptions

fail, such that the skyline is detected with errors (Fig. 4.5). Nonetheless, such errors are rare and they do not degrade the performance of the proposed SRS significantly.

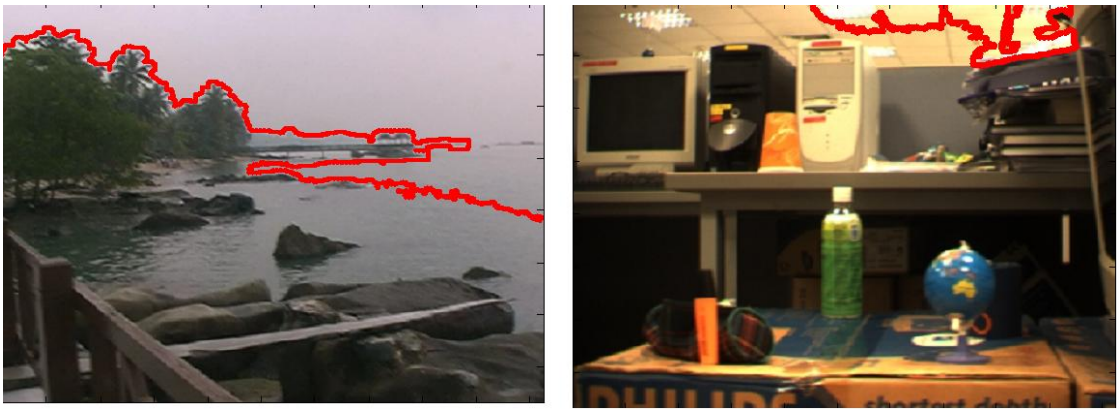

Figure 4.5: Two examples of erroneous skylines detected. Left: The presence of haze and the sea makes the sky almost indistinguishable. Right: Artificial lighting in the ceiling mimics the sky that confuses the algorithm.

4.4 From image pyramids to saliency maps

With the five composite features extracted, the saliency algorithm creates a set of *Gaussian image pyramids* [93] by filtering the composite features repeatedly with a low pass Gaussian filter, **G** over several spatial scales (Fig. 4.1). All the initial image sizes are resampled to 512×384 to give the largest spatial scale 1 together with seven other smaller scales. Each scale produces a subsampled image that contains half the resolution of the image at the preceding scale. This procedure

can be summarised as

$$\begin{cases} \mathbf{P}_1^j & = \mathbf{C}_f^j, j \in \{col, int, ort, edge, sky\} \\ \mathbf{P}_{k+1}^j & = \downarrow 2(\mathbf{P}_k^j * \mathbf{G}), k \in \{1 \cdots 7\} \end{cases} \tag{4.1}$$

where $\{col, int, ort, edge, sky\}$ and k are the five composite features and the spatial scales respectively. $\mathbf{P}_k^j$ is the *pyramid image* of the j^{th} composite feature at the k^{th} spatial scale. $*$ and $\downarrow 2$ are the convolution and down-sampling operator respectively. A set of $\mathbf{P}_k^j$ is created for each composite feature, forming a set *composite feature maps*. An example of the intensity composite feature maps at eight different scales is shown in Fig. 4.6.

Figure 4.6: Image pyramids of the intensity composite feature maps formed by the repeated application of the $\downarrow 2$ operator and convolution with a Gaussian.

For each set of composite feature maps, a set of six normalised *difference maps*, $\mathbf{P}_{diff}^j(i)$, for the i^{th} difference, is obtained by computing the difference between

two pyramid images at different scales (c, s). This procedure models the *centre-surround difference* mechanism found in the *receptive fields* of the retina and LGN [45, 51, 101]:

$$\begin{aligned} &\mathbf{P}^j_{diff}(i) = \mathcal{N}_1(\|\mathbf{P}^j_c - \mathbf{P}^j_s\|) \\ &c \in \{2,3,4\}, s \in \{c+d\}, d \in \{3,4\}, i \in \{1 \cdots 6\} \end{aligned} \tag{4.2}$$

where j represents the composite features as before and $\mathcal{N}_1$, known as *content based global non-linear amplification* [50], is used as a normalisation procedure to promote salient features in each of the difference maps. Maps with isolated salient features which are more conspicuous are promoted while maps with numerous comparable peak responses are suppressed. $\mathcal{N}_1$ consists of two simple steps:

1. Each $\mathbf{P}^j_{diff}$ is normalised to a fixed range $0 \cdots M_g$ where M_g is the global maximum of the difference maps.

2. Multiply each $\mathbf{P}^j_{diff}$ by $(M_g - m_{av})^2$ where m_{av} is the mean of the local maxima of the difference maps.

Hence maps with a local maxima that are near to the value of the global maximum will yield a small $(M_g - m_{av})^2$ that effectively suppresses this map while maps with a few isolated salient features are globally promoted since $(M_g - m_{av})^2$ is large. The effect of $\mathcal{N}_1$ is shown in Fig. 4.7.

Summing up the difference maps for the j^{th} composite feature and applying $\mathcal{N}_1$

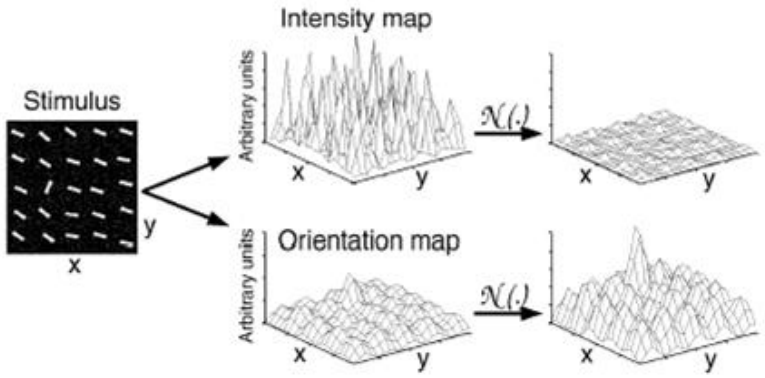

Figure 4.7: Normalisation using $\mathcal{N}_1$ to two $\mathbf{P}^j_{diff}$, intensity and orientation composite features. Data from [51].

again yields the various *conspicuity maps*, $\mathbf{C}^j$, that represents the saliency of each composite feature:

$$\mathbf{C}^j = \mathcal{N}_1(\sum_i \mathbf{P}^j_{diff}(i)) \tag{4.3}$$

These conspicuity maps, $\mathbf{C}^j$, each weighted by the *entropy* of the individual $\mathbf{C}^j$, are then combined and normalised by $\mathcal{N}_2$, by a procedure that recursively applies a large *Difference of Gaussian*(DoG) filter over the conspicuity maps. The main idea is to model the *centre-on* ganglion cells that approximate the feature combination strategy of the visual cortex [45, 50, 101]. The output is known as the *depth-free*

saliency map, $\mathbf{S}_m$:

$$\mathbf{S}_m = \mathcal{N}_2(\sum_j \mathcal{H}(\mathbf{C}^j) \times \mathbf{C}^j) \tag{4.4}$$

where $\mathcal{H}(\cdot)$ is the entropy operator defined by Shannon [117], with larger weights given to conspicuity maps that have a higher entropy. Following the work in [33], it was shown that weighing the conspicuity maps by their entropy gave a more distinctive and robust representation of the final saliency map, $\mathbf{S}_m$, that is similar to the saliency maps produced in [33, 43, 51]. From section 3.5.1, the importance of the depth information in *prioritising* which parts of the scene are more important was highlighted. In the proposed SRS, this depth information is integrated into $\mathbf{S}_m$ to form the *depth-weighted* saliency map, $\mathbf{S}_{dm}$ by

$$\mathbf{S}_{dm} = \hat{\mathbf{D}}_{prox} \times \mathbf{S}_m \tag{4.5}$$

where $\hat{\mathbf{D}}_{prox}$ is a *dense ordinal proximity map* that estimates the depths of the 3D points associated to the image features. In addition to the importance of depth in encoding the spatial configuration of the landmarks (section 3.5.1), needed for a viewpoint invariant representation of the scene (section 3.3.4), the motivation of adding a depth component to the SRS manifold comes from human psychological studies. These studies show that humans tend to focus attention at a particular

depth plane [2, 84]. By increasing the salience of features that are nearer, the SRS tends to pick out features that are of immediate relevance to various tasks such as recognising near obstacles and immediate dangers. In the case of insect navigation, experiments from [21] found that bees tend to select objects and features nearer to the target goal site as such features define the scene more accurately. This means that a mechanism to determine the *distance* of the selected feature is necessary. The estimation of $\hat{\mathbf{D}}_{prox}$ from optical flow is detailed in section 5.2.3 using simulated TBL motion. Examples of conspicuity maps and $\mathbf{S}_{dm}$ are shown in Fig. 4.8.

With the depth-weighted saliency map obtained, the next section shows how the salient ROIs are extracted from $\mathbf{S}_{dm}$.

4.5 Salient ROIs from the saliency map

The computed $\mathbf{S}_{dm}$ represents a *one-to-one* mapping of the salient regions associated with the input RGB image. This is clearly shown when the saliency map is compared side by side with the RGB image (Fig. 4.9).

Simple image morphological operations are then applied to extract the salient ROIs from $\mathbf{S}_{dm}$. The algorithm can be summarised in the following steps:

1. Edges are detected from the input $\mathbf{S}_{dm}$ using Canny's method [18] with an initial predefined threshold, t_{edge} to form an edge map.

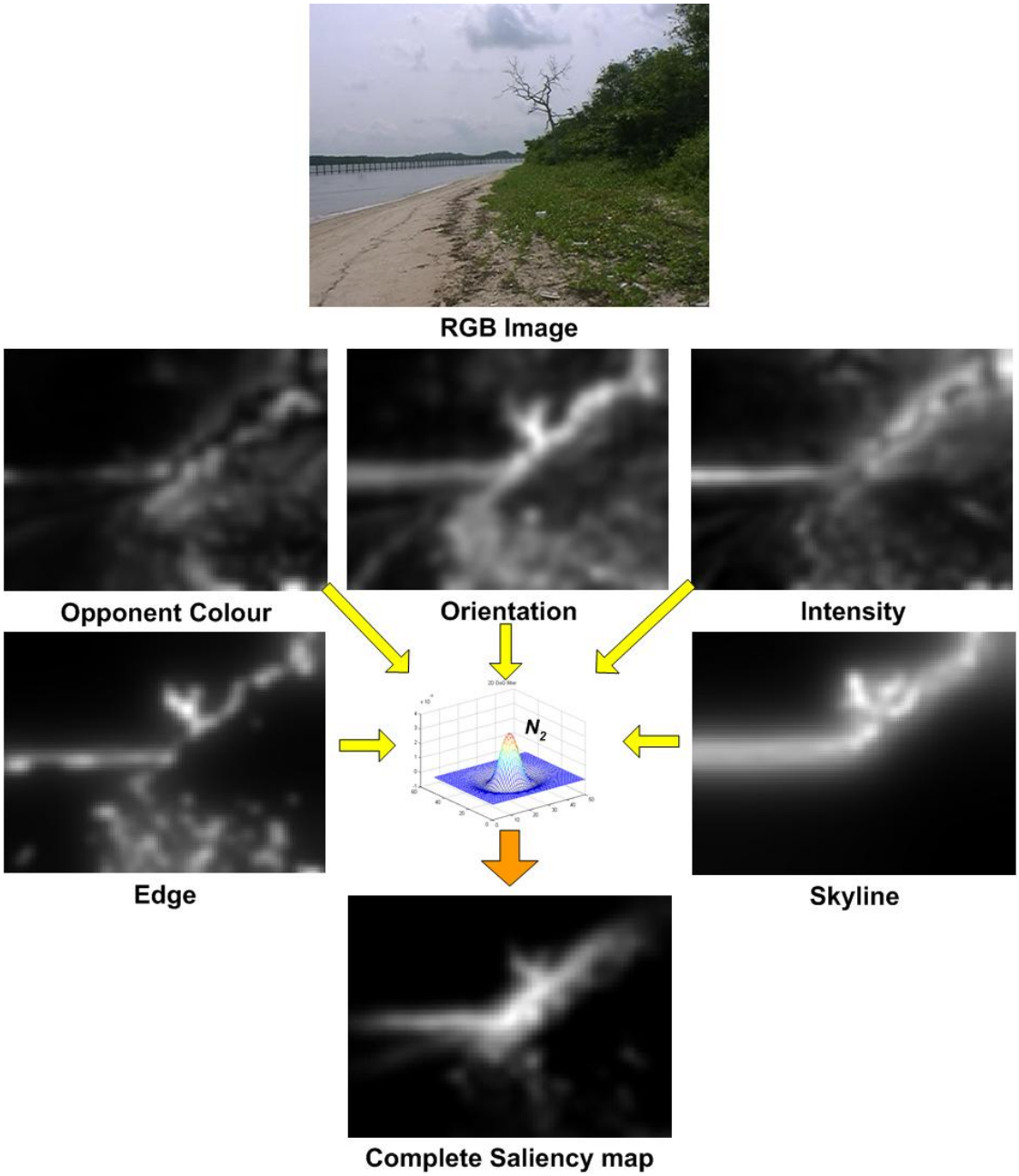

Figure 4.8: The five conspicuity maps derived from the RGB image are combined by $\mathcal{N}_2$ to form $\mathbf{S}_{dm}$. $\hat{\mathbf{D}}_{prox}$ is not shown.

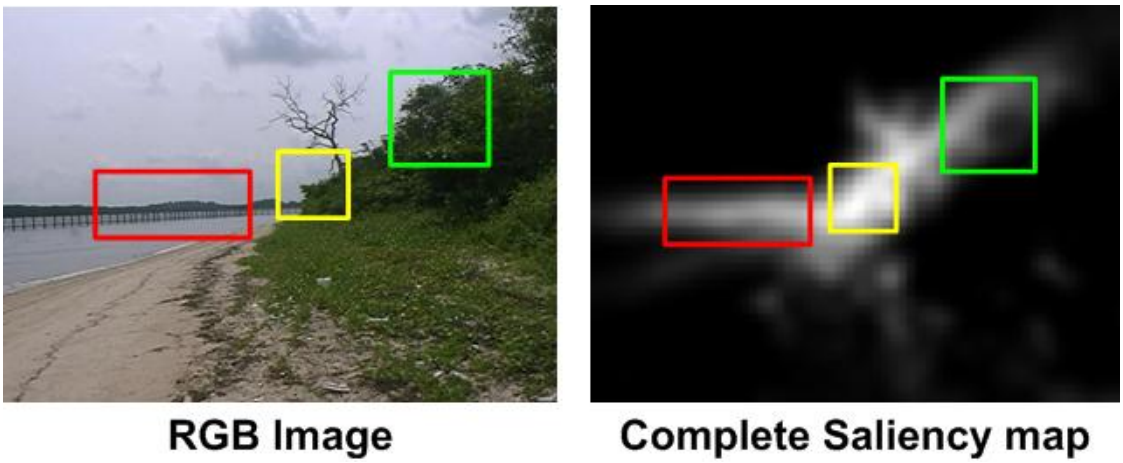

Figure 4.9: A one-to-one mapping of the most salient regions (boxed) can be found in the original RGB image (boxed in the same colour).

2. Dilate the edges with a suitably chosen structuring element such that broken edges are connected together. In this work, disk and cross shaped elements are used.

3. The connected edges delimit the edge map into the salient ROIs which are then filled and counted. The number of ROIs detected is returned and if this number is not between the minimum and maximum number of salient ROIs desired, the algorithm goes back to step 1 with a suitably adjusted t_{edge}.

4. The filled edge map with the salient ROIs is resized to the same scale as the original RGB image so that the coordinates of the ROIs are comparable with the original image. The perimeters of the resized edge map are then extracted which delimit the salient ROIs.

Fig. 4.10 illustrates the various intermediate steps in extracting the salient ROIs from the input $\mathbf{S}_{dm}$. The output is a labelled map, $\mathbf{L}_m$, that identifies each salient ROI with a particular number. This map is used as a mask to indicate which regions of the image are salient for further by SURF in the next chapter.

Notice that the salient ROIs occupy only specific regions on the image. From this step onwards, the proposed SRS will focus *all* processing at these regions only, improving the efficiency of the algorithm since less data is processed. This models well the attentional strategy employed by the HVS (section 3.1.2) that similarly focuses the processing of the incoming visual information only at the most salient locations.

4.6 Final remarks

This section has described the initial step in the proposed SRS that uses a modified computational model of visual saliency to produce a depth-weighted saliency map, $\mathbf{S}_{dm}$. Using this saliency map, salient ROIs are extracted which is represented as labelled regions in $\mathbf{L}_m$. The next chapter shows how $\mathbf{L}_m$ is used to indicate which regions of the image are encoded with SURF features for the final representation of the scene into the *Scene matrix*.

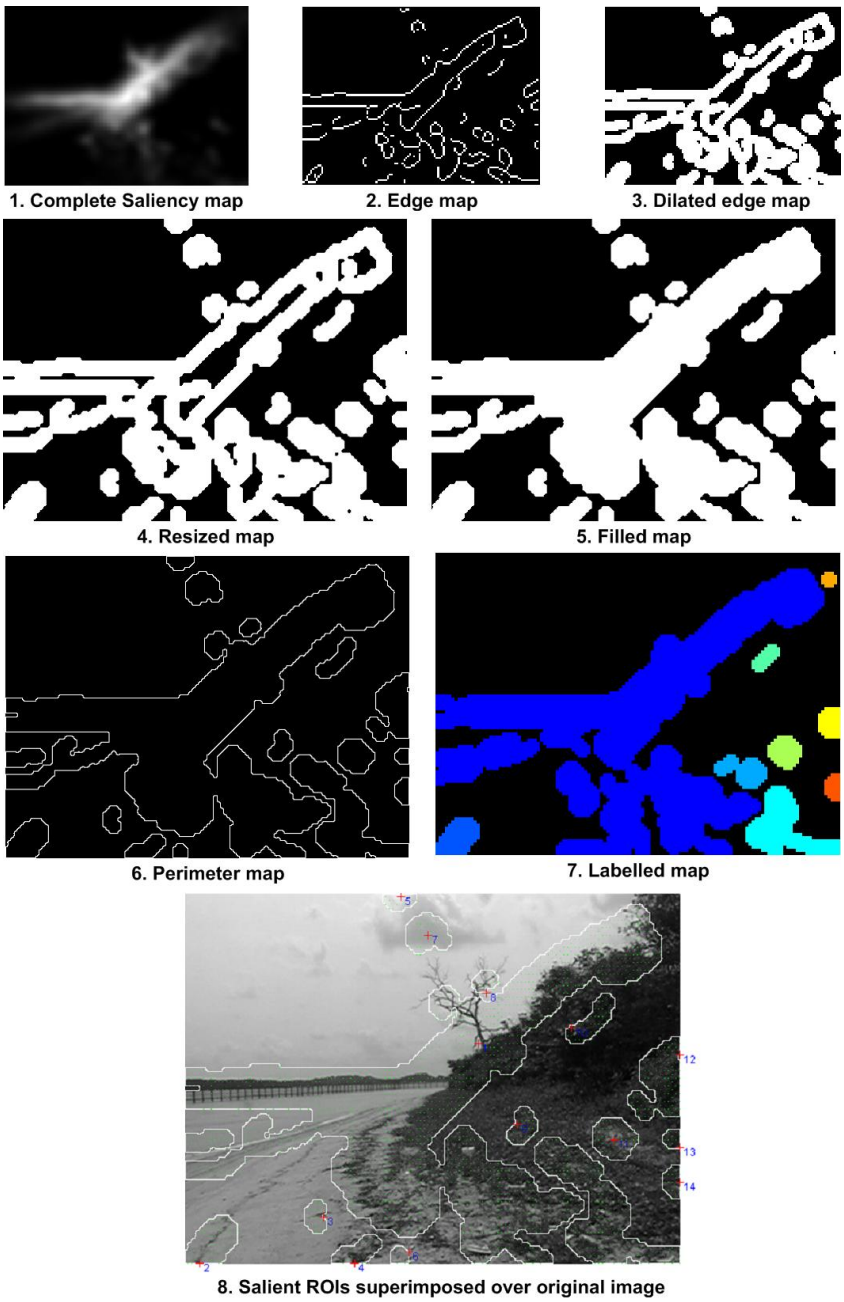

Figure 4.10: Steps 1–8 that describe the various stages of extracting the salient ROIs using various image morphological operations. The iterative method to determine t_{edge} is not shown here. $\mathbf{L}_m$ is shown in false colours and the extracted salient ROIs are boxed in white and highlighted in green (step 8).

Chapter 5

The Scene Matrix

This chapter continues from chapter 4 where the salient ROIs are extracted from the depth-weighted saliency map $\mathbf{S}_{dm}$. These ROIs are represented as labelled regions $\mathbf{L}_m$. How these regions are encoded by *SURF descriptors* is detailed in section 5.1. *Ordinal depth* obtained from TBL motion is then estimated so as to augment the descriptors with important depth information (section 5.2). The augmented SURF descriptors at the salient ROIs, termed as *salient-SURF* descriptors are then validated and combined into a compact *Scene matrix*, $\mathbf{m}_s$ that represents the scene completely (section 5.3).

It is important to note that at this stage of the algorithm, *two* $\mathbf{S}_{dm}$ are created together with two resulting $\mathbf{L}_m$ from two *closely* separated frames (with a small change in position) of the *same* scene. These two frames will be denoted as $\mathbf{S}_m^i, \mathbf{L}_m^i$ for $i \in \{1, 2\}$ for the first frame and second frame respectively. For simplicity, the

description will drop the superscript i when no distinction is needed between these two frames. Using two frames allows the algorithm to robustly recover ordinal depth (section 5.2) and to validate the resulting salient-SURF keypoints before incorporating them into the scene matrix, $\mathbf{m}_s$ (section 5.3).

5.1 Encoding the salient ROIs using SURF descriptors

The concept of image(keypoint) descriptors and their usefulness for scene recognition was discussed in section 3.2. Combining salient ROIs and such descriptors was shown in section 3.2.2 to be complementary in improving the general performance of the proposed SRS. From these concepts, this section details how SURF keypoints and its corresponding descriptors are used and combined with salient ROIs for a robust and reliable representation of the scene.

5.1.1 Illumination invariance in HSV colour space

The importance and robustness of using the HSV colour space was shown in section 3.4. In section 3.4.2, it was shown that illumination changes distort scenes significantly. Thus, it is crucial to use an illumination-invariant representation of

the scene as input to the proposed SRS. The current work achieves illumination invariance by encoding the salient ROIs with SURF descriptors over the HSV colour space. As was shown in section 3.4.1, using grayscale images alone for SURF descriptors is very sensitive to illumination changes and this causes the descriptors to be very *different* for the same scene under different illumination, leading to poor recognition (see Figs.3.15 and 5.1).

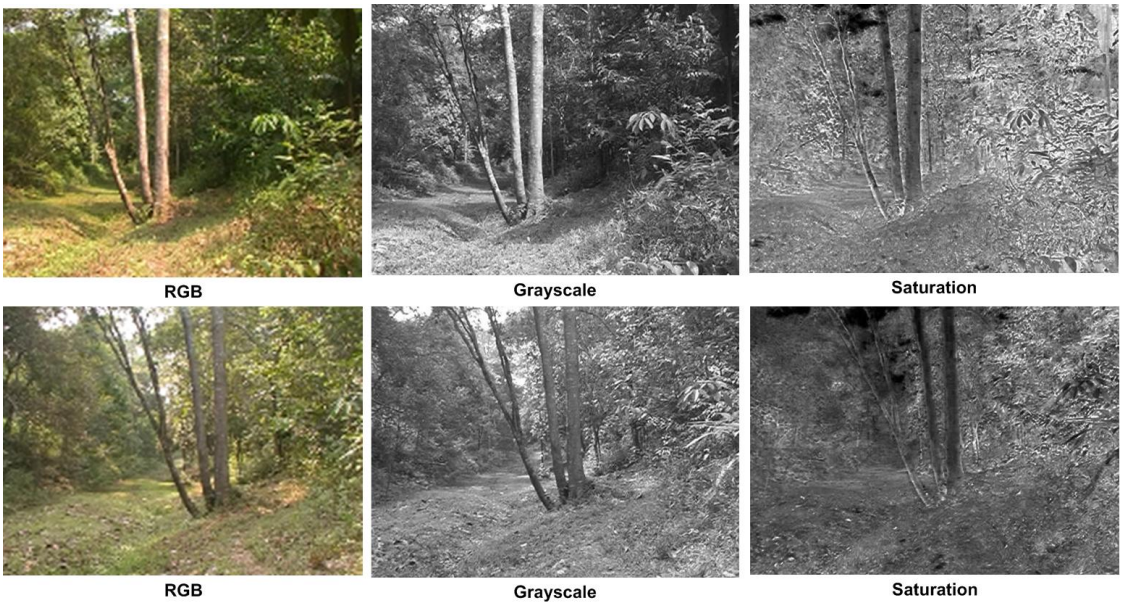

Figure 5.1: A scene with varying illumination and their grayscale and saturation components. The hue component is not shown. Top: Morning scene with the sun shining in the foreground. Bottom: Evening scene with the sun blocked by the foliage.

From Fig. 5.1, it is evident that the grayscale images of the two scenes differ significantly whereas the saturation image is stable over many features (*e.g.* the tree trunk). Of course, the saturation space is not *always* invariant to illumination

changes and in some cases is less stable than the grayscale or hue spaces. To achieve generality and accuracy against various forms of illumination distortions, *all three* colour spaces are used to compute the SURF keypoints (section 3.4.2). An important point to note is that the SURF keypoints detected for each colour space are kept *separate* during the SURF matching process described in section 5.1.3. A keypoint detected in say, the hue space should not be matched to a keypoint in the saturation space. The reason is that the SURF descriptors are unique to each colour space - mixing the SURF descriptors over the different colour spaces may lead to even more false matches!

5.1.2 Structure of the SURF descriptor

The SURF descriptor associated with a SURF keypoint is made up of a 6D *localisation* and a 64D *description* components [10]. The structure of this 70D descriptor is as follows:

Definition 5.7. *Structure of SURF descriptor* The *complete* SURF descriptor of a SURF keypoint is a 70 element vector: $[x\ y\ a\ b\ a\ l\ \mathbf{desc}]$ where (x, y) are the x and y coordinates(subpixel) of the position of the keypoint. a represents the *scale* at which the keypoint is detected. b represents the *corner strength* of the keypoint which is detected by a Hessian matrix (section 3.2.3). l is the sign of the Laplacian $[+1, -1]$ that allows for rapid matching. The first six elements form

the localisation component while the 64D **desc** vector forms the actual description component that is used for determining correspondences. □

Since a scene contains a number of keypoints, the set of *all* the descriptors are grouped into a *Descriptor matrix* defined as:

Definition 5.8. *Descriptor matrix* A scene containing N_d keypoints are grouped into a $N_d \times 70$ Descriptor matrix with the following structure: [**x y a b a l Desc**] where the elements are all matrices and **Desc** is the set of SURF description components from the N_d keypoints. □

The next section describes how this descriptor is used for determining correspondences between keypoints from two scenes.

5.1.3 Determining correspondences from descriptors

In order to match the SURF descriptors between two scenes, Lowe's *nearest neighbour ratio* method [68] for matching SIFT descriptors is used. The use of the nearest neighbour distance ratio threshold was found to improve the robustness of SIFT and to yield acceptable matching accuracy in the performance evaluation of [75]. As SURF is basically a SIFT-based descriptor, using this threshold also benefits SURF; in fact there are even more gains in performance due to the reduced vector size of SURF (64D versus 128D for SIFT). Note that in this work, only the 64D **desc** component of the complete SURF descriptor is used for matching.

The nearest neighbour ratio method is interesting as it does not use a simple distance measure to determine if two keypoints are matching using their descriptors. Instead, the *ratio* of the Euclidean distances of the closest match, L_{close}, and the second closest match, L_{2close}, known as the *distance ratio*, d_{ratio}, is used as the measure of similarity:

$$d_{ratio} = \frac{L_{close}}{L_{2close}}, d_{ratio} \in \{0 \cdots 1\} \tag{5.1}$$

As was explained in [68], using this measure allows the matching algorithm to discard keypoints that do not have good matches. The main idea is that for matches that are correct, the closest match is significantly closer than the closest incorrect match (second closest neighbour) for reliable matching. On the other hand, false matches are likely have their second closest neighbour nearer to the closest match, brining d_{ratio} closer to unity. Hence a d_{ratio} *threshold* that is near to 1 allows for more relaxed matching at the expense of a higher false detection rate, while a small threshold only allows for very constrained matching with very low false detection rates. In this thesis, the d_{ratio} threshold is fixed at 0.83. This value achieves good positive matches and rejects the majority of bad matches in practice.

Enforcing uniqueness constraint in matching

As Lowe's original MATLAB® code for SIFT descriptor matching (available online at `http://www.cs.ubc.ca/~lowe/keypoints/`) does *not* guarantee against repeated matches (a many-to-one mapping) of the SIFT descriptors, mismatches can occur between a SIFT feature with *many* SIFT features. This problem is even more likely to occur for SURF due to its reduced vector dimensions that results in a possible reduction in uniqueness of the descriptor (Fig. 5.2).

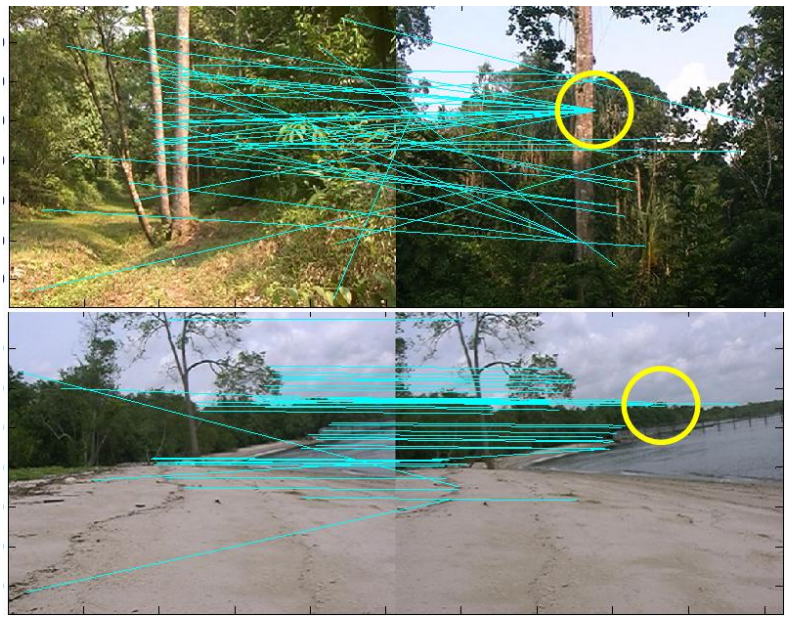

Figure 5.2: Numerous many-to-one SURF matches (cyan lines) using the original Lowe's matching algorithm (circled) result in unpredictable scene recognition: Wrong recognition (top) and correct recognition with some mismatches (bottom).

This *uniqueness* constraint had been posed by Ullman [114] in his *minimal*

mapping theory to effectively solve the correspondence problem using three intuitive local criteria to establish good global mapping between any two image frames. These three criteria are:

1. *Principle of Similarity*: similar features are matched;
2. *Principle of Proximity*: close features should be matched;
3. *Principle of Mutual Exclusion*: only one-to-one mappings are allowed, which is the uniqueness constraint.

In a ground breaking work, Scott and Longuet-Higgins [99] proposed an algorithm to match point features that encompasses Ullman's second and third principles (proximity and mutual exclusion) using *Singular Value Decomposition*(SVD). This work was extended by Sharpiro and Brady [102] who used an eigenvalue approach to further constrain the matchings. In [90], Pilu improved the original Scott-Longuet-Higgins (SLH) algorithm by the use of SVD over a *correlation-weighted proximity matrix* that contains the cross-correlation values of the image features that enforces Ullman's first principle (the feature similarity principle) that complements the SLH algorithm. In this thesis, Pilu's algorithm is adapted to SURF descriptors by constructing a similar correlation matrix *without* computing the SVD which is computationally expensive since there could be many SURF keypoints (> 1000). The justification for applying SVD is to impose the uniqueness constraint by making the proximity matrix *orthogonal* which is useful for perfect image registration.

Since this possibility is extremely unlikely in a practical SRS system on a mobile agent, the uniqueness constraint can be reliably approximated for most scenarios without performing SVD.

A brief description of the matching strategy follows. Note that the description below applies only to one particular colour space which can be extended to all the three colour spaces that the descriptors are encoded in (section 5.1.4). Denoting the two SURF descriptors as $(\mathbf{Desc}^1, \mathbf{Desc}^2)$ containing $(N_d(\mathbf{Desc}^1), N_d(\mathbf{Desc}^2))$ SURF keypoints respectively as inputs:

1. Construct the *dot-product matrix*, $\mathbf{m}_{dotp}$ of the SURF descriptors by

$$\mathbf{m}_{dotp} = \mathbf{Desc}^1 \bullet (\mathbf{Desc}^2)^T \tag{5.2}$$

   where $\bullet$ is the dot product operator and T represents the transpose of the matrix. $\mathbf{m}_{dotp}$ optimises the computations since MATLAB® is optimised for matrix operations.

2. Compute the arccos of $\mathbf{m}_{dotp}$ which is a close approximation of the ratio of Euclidean distances when the angles between the input vectors are small. This forms the *proximity matrix*, $\mathbf{m}_{prox}$:

$$\mathbf{m}_{prox} = -\arccos \mathbf{m}_{dotp} \tag{5.3}$$

where the negative sign is necessary for the algorithm to perform a maximum search for the closest match that has numerically the smallest distance in a non-negative $\mathbf{m}_{prox}$.

3. Using Pilu's algorithm [90], a search for the maximum of each row and column in $\mathbf{m}_{prox}$ is performed with the indices saved as separate variables. Comparing the indices of the maximal elements, only the indices that are maximum in *both* the rows and columns are accepted as potential correspondences. This step ensures a one-to-one matching as shown in Fig. 5.3.

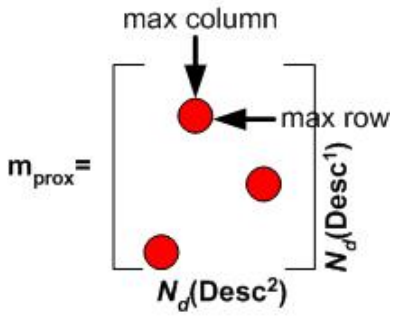

Figure 5.3: Ensuring one-to-one correspondences using $\mathbf{m}_{prox}$. The potential matches (red elements) are the maximum values in both the columns and rows of $\mathbf{m}_{prox}$.

4. Finally the potential correspondences are accepted if their values are *smaller* than the predefined threshold for d_{ratio}, with smaller values representing better matches.

Using this algorithm results in better SURF correspondences due to the uniqueness constraint as can be seen in Fig. 5.4

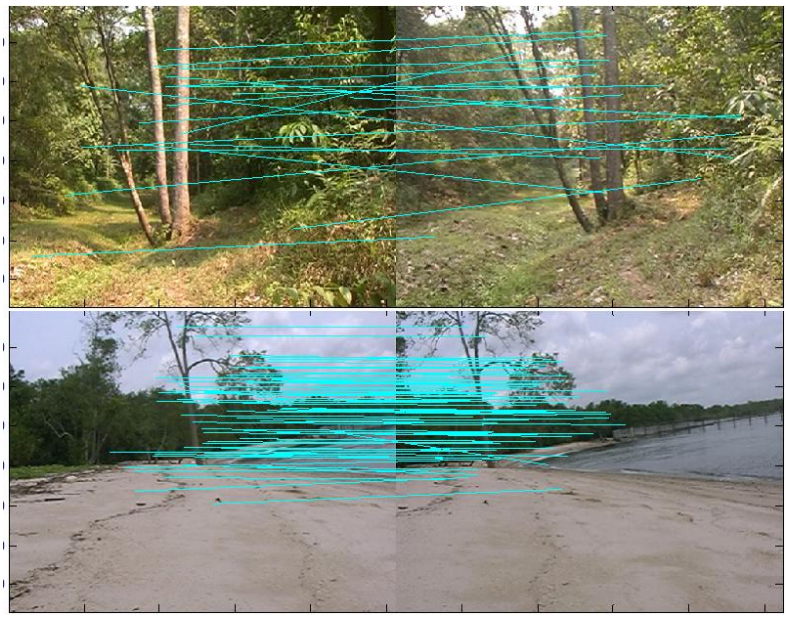

Figure 5.4: The same matching examples of Fig. 5.2 are shown here after invoking the uniqueness constraint. The top scene is correctly recognised and the bottom scene has fewer mismatches.

5.1.4 Combining SURF and salient ROIs

The process of combining the SURF descriptors and salient ROIs is very straightforward and can be summarised by the following steps, given the $\mathbf{L}_m$ of the salient ROIs and the original RGB image as inputs:

1. The input RGB image is converted to the HSV colour space.

2. For each colour space, SURF keypoint detection is applied *only* to the areas indicated by $\mathbf{L}_m$. Hence only the salient ROIs are processed in order to detect SURF keypoints.

3. Finally the associated SURF descriptors of the detected SURF keypoints at

the salient ROIs are extracted and saved into a *cell matrix*, with each cell occupied by the SURF descriptors in one of the three colour spaces.

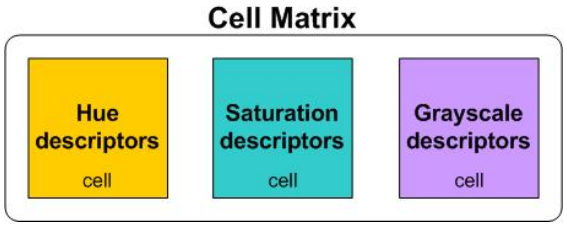

Figure 5.5: The cell matrix containing SURF descriptors from three colour spaces are kept separated in each cell.

A cell matrix is a special data structure in MATLAB® that combines different data structures into a single complex *superstructure*. This allows easy manipulation of complex data that often contains incompatible formats by storing the data into separate cells which can be of differing dimensions. In this work, since the SURF descriptors in the three colour spaces must be kept separate, the cell matrix contains three cells, one for the descriptors from each colour space (Fig. 5.5). The SURF keypoints detected over the three colour spaces are shown in Fig. 5.6.

With SURF keyoints extracted and encoded by their descriptors at the salient ROIs, only the *depth* information is lacking that describes the complete spatial configuration of the scene structure. This depth information is obtained from two closely separated frames of the same scene described in the next section.

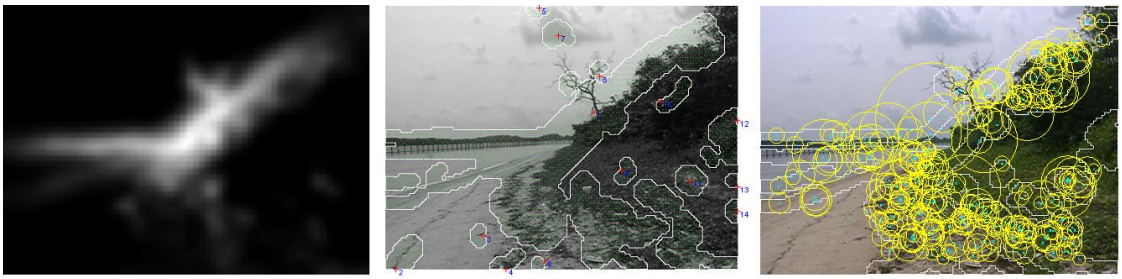

Figure 5.6: Right: SURF keypoints detected superimposed over the original RGB image. The keypoints are marked by the cyan 'x' while the yellow circles represent the scale at which the keypoints are detected. The saliency map (left) and salient ROIs (middle) are shown together for comparison.

5.2 Ordinal depth from simulated TBL motion

Depth has been shown to be an important component of the proposed SRS (section 3.5.1) that should be included for effective outdoor scene recognition of natural environments. Being an integral component that describes the scene structure, the inclusion of *ordinal depth* was shown in section 3.3.4 to be useful in providing a viewpoint invariant representation of the scene. The link between TBL motion and the possible recovery of ordinal depth information was discussed in section 3.5.2 due to the large translational components in this motion. In this section, the details of how the proposed SRS recovers this ordinal depth information from two closely separated frames of the same scene are presented.

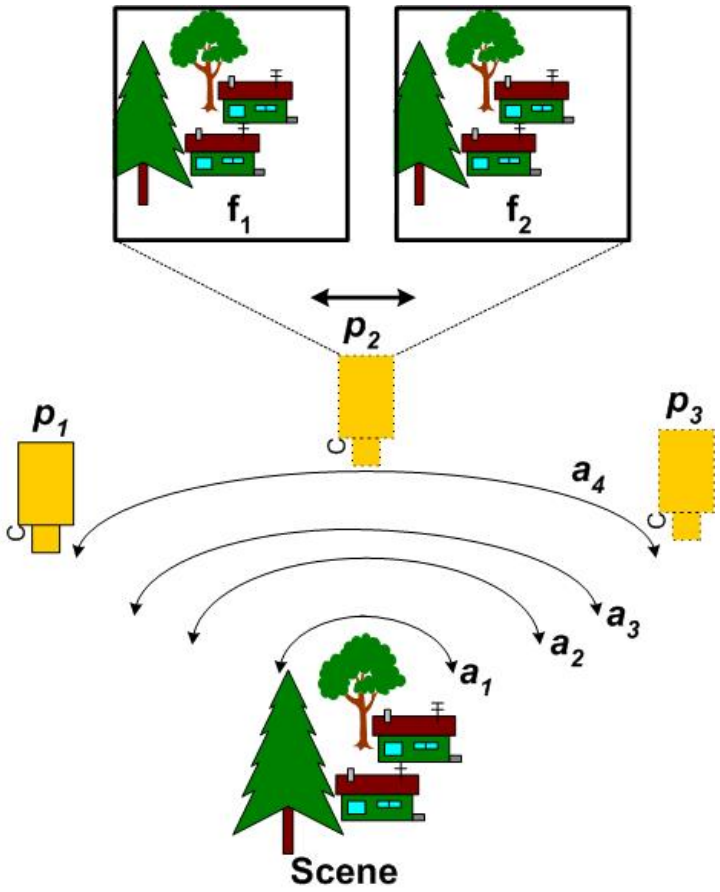

Figure 5.7: TBL arcs created by moving the video camera along increasingly bigger arcs (a_1 to a_4). At each arc, the camera is moved in a continuous fashion from position p_1 to p_3. At each position, two slightly displaced frames (f_1, f_2) are obtained for ordinal depth estimation.

5.2.1 Inducing optic flow from TBL

In order to recover the ordinal depth of a scene, artificial TBL motion is induced by moving a digital video camera in a series of arcs and taking a pair of image frames at specific locations along each arc (Fig. 5.7).

From the image frames, *optic flow* can be computed using several methods. In

this thesis, a *correlation-based* technique employing SURF keypoints extracted in the previous step (section 5.1.4) as features are used to determine the optic flow. As the descriptors are designed to be affine and illumination invariant (section 3.2.1), this makes the computation of the optic flow more robust against image distortions arising from viewpoint changes, which is a known problem in standard correlation techniques. Furthermore, optic flow derived from correlation-based techniques does not smooth over flow discontinuities unlike traditional differential methods [49].

From the conclusion in [25], it was suggested that the *end points* of the TBL arcs (positions p_1 and p_3) are more likely to be memorised by the insect during TBL due to its enhanced stability when the insect slows down and begins to change its direction in a new arc (see Fig. 3.23(b) for the optic flow vectors induced at that moment). Motivated from these observations, most of the image frames are taken at three locations corresponding to positions p_1 to p_3. Although it is not known whether position p_2 is really used by the insects to memorise scene information, bees and wasps are observed to approach the target directly [25, 61] once the TBL phase is over. The scene data captured at position p_2 is thus useful as an input query scene for the proposed SRS's scene decision module (section 6.2). Fig. 5.8 shows the same scene taken from the three different positions of the TBL arc that are used to construct an image database for testing the proposed SRS (section 7.1.4).

Figure 5.8: Three views of the same scene along a simulated TBL arc. Left to right: positions p_3 to p_1. Notice the significant occlusions for this enclosed mangrove environment as the viewpoint changes.

5.2.2 Estimating ordinal depth from optic flow

This section details how optic flow, obtained from simulated TBL motion, can be computed from SURF correspondences. Once the optic flow vectors are obtained, ordinal depth can easily estimated.

Obtaining optic flow information from two image frames is very simple and direct. The procedure is summarised in the followinf steps. Denoting the two sets of Descriptor matrices, $(\mathbf{Desc}^1, \mathbf{Desc}^2)$ (definition 5.8), with the superscripts representing the two image frames (f_1, f_2) respectively as inputs:

1. Using the procedure described in section 5.1.3, initial SURF correspondences are determined from $\mathbf{Desc}^1$ and $\mathbf{Desc}^2$ for each colour space.

2. Since there is always a small number of wrong matches, the proposed SRS attempts to remove these wrong matches by applying the well-known RANSAC

(RANdom SAmple Concensus) on the 8-point algorithm [46] so as to determine a *likely* geometric transformation between the keypoints in the two image frames. The geometry of this transformation is encoded as the *Fundamental Matrix*, $\mathbf{F}_{1\rightarrow 2}$. Keypoints that do not satisfy the epipolar geometry are subsequently removed by RANSAC. SURF matches that remain at the end of this procedure are said to be *epipolar-verified*.

3. Denoting (x_1, y_1) as the coordinates of a SURF keypoint in f_1 that is matched to a keypoint with coordinates (x_2, y_2) in f_2, the optic flow vector (u, v) in the (x, y) directions from f_1 to f_2 is then given as:

$$\begin{cases} u &= x_1 - x_2 \\ \\ v &= y_1 - y_2 \end{cases} \tag{5.4}$$

Repeating (5.4) for all the epipolar-verified SURF matches yields the set of optic flow vectors that describes the motion of the SURF keypoints from f_1 to f_2.

The above procedure is illustrated and summarised in Fig. 5.9.

From Appendix B, it can be shown that the depth recovered from optic flow under TBL motion possess ordinal invariance and hence this depth is termed the

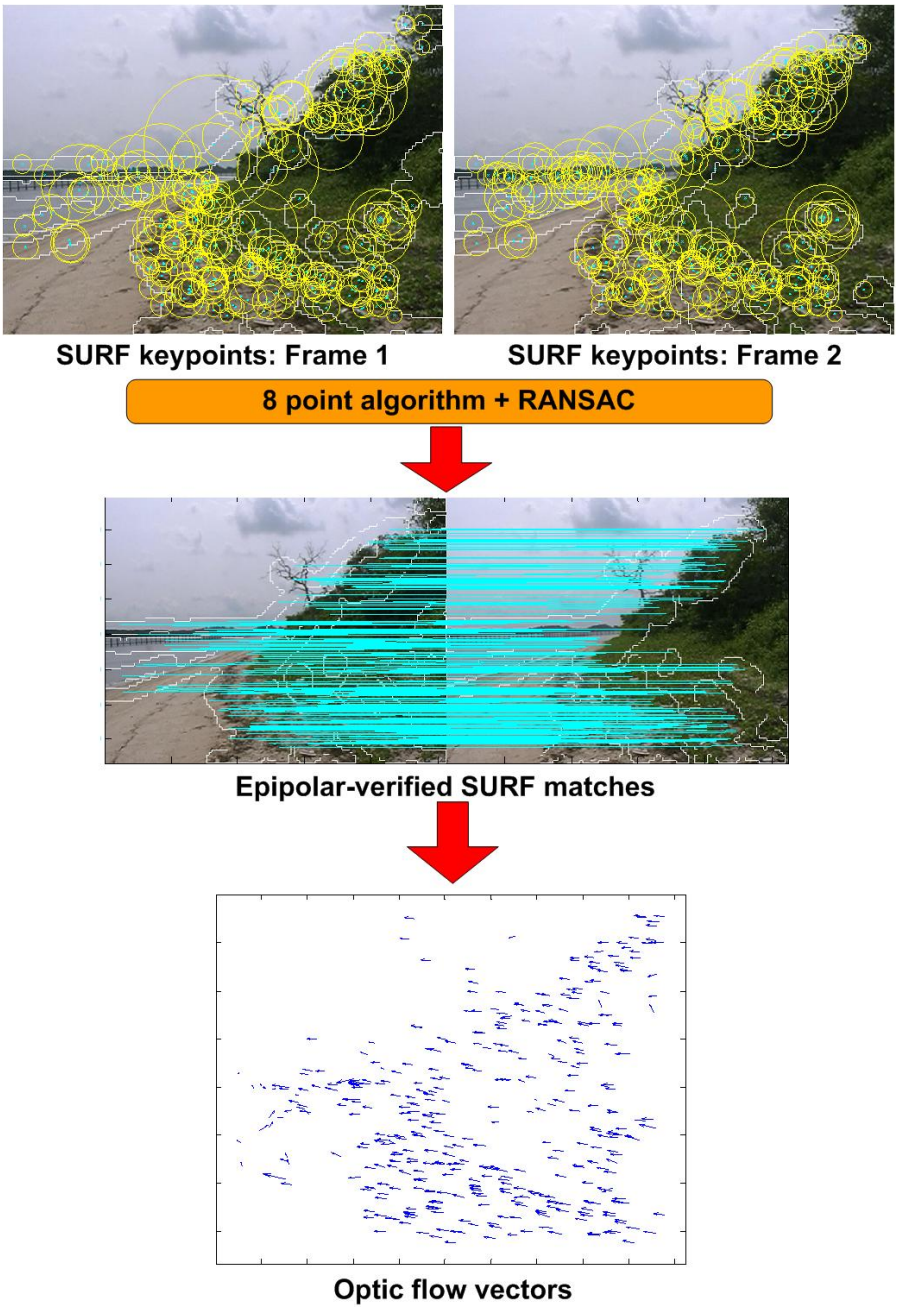

Figure 5.9: Steps summarising the computation of the optic flow between two image frames. Top: SURF keypoints from both frames, $(\mathbf{Desc}^1, \mathbf{Desc}^2)$. Middle: Epipolar-verified SURF matches after RANSAC, with the correspondences shown as cyan lines. Bottom: The optical flow vectors between the matched keypoints are illustrated as arrows.

ordinal depth, Z_{ord}:

$$Z_{ord} = \frac{-f}{u + \omega_y f} \tag{5.5}$$

where ω_y is the estimated rotation along the vertical y axis (usually small) and f is the estimated focal length of the camera in pixels. [22] proved that even if the estimates for ω_y and f are erroneous, the recovered scaled depth is related to the true scaled depth by a relief transformation that preserves the order of the depths. Since the exact value of ω_y is not crucial (see Appendix B for more details), it is acceptable to approximate $\omega_y = 0$. Thus Z_{ord} provides the ordinal depth estimate required by the propose SRS. In order to keep to a certain comparable scale, the computed Z_{ord} is normalised between $[1\ Z_{max}]$ where Z_{max} is the expected maximum optical flow for the camera motion. Any flow that is larger than Z_{max} is discarded before normalisation.

5.2.3 Ordinal depth adjustment using AHC

Although the depth orders recovered are invariant to errors in ω_y and f, *noise* and *inaccuracies* in determining the optic flow will affect the validity of the depth order recovered. This is especially true for points that have very close depth values and are susceptible to have their estimated depth orders reversed. To circumvent this problem, no attempt is made to resolve depth orders for depths that are very

close together. Instead, these depths are clustered into various depth layers by applying a distance based *agglomerative hierarchical clustering*(AHC) algorithm to the initial estimated ordinal depths.

The *number* of clusters, N_{clust} to be formed is estimated from the number of epipolar-verified SURF matches used to compute optical flow. In this work, N_{clust} is set between 20% to 25% of the number of SURF matches. The *adjusted* ordinal depth of a certain keypoint uses the *mode* value of the cluster that it belongs. This is illustrated clearly in Fig. 5.10. Notice that the depths of the various parts of the tree in Fig. 5.10 are more consistently ordered except at one keypoint. This reduces the errors incurred in the later stages when the ordinal depth is compared (section 6.1.2 and (6.5)). AHC is simple but depends a lot on the choice of the number of intial clusters, N_{clust}, to form and is not guaranteed to produce perfect depth layer segmentation. It is however a reasonable compromise when one needs to have a simple but efficient method to adjust the overly refined initial ordinal depths.

Note that the numbers in red associated with the SURF keypoints in Fig. 5.10 are not the ordinal depth, Z_{ord} introduced previously (5.5) but is the reciprocal of Z_{ord} known as *ordinal proximity*, d_{prox} where

$$d_{prox} = \frac{1}{Z_{ord}} \tag{5.6}$$

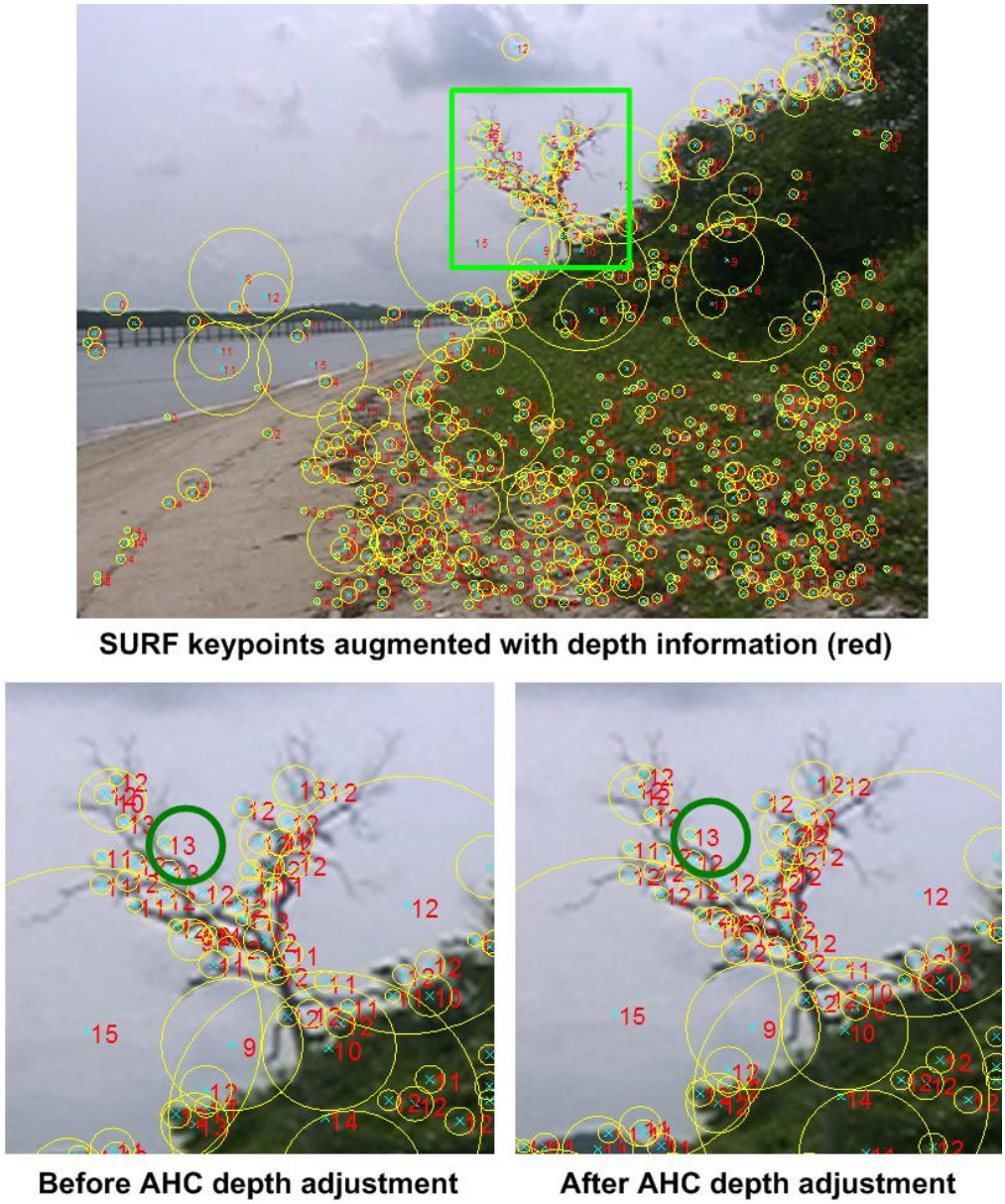

Figure 5.10: Using AHC to remove inconsistencies in the *proximity* (red numbers) associated with each SURF keypoint. The top area in the green box is expanded to highlight the effects of the depth orders before and after AHC (bottom). Except for one keypoint (circled), the other keypoints associated with the distant tree have consistent depths after applying AHC.

The reason for using d_{prox} is due to the need to *promote* regions that are nearer to the camera as they contain more unique information related to the scene (section 3.5.1). Hence the weights of keypoints near to the camera, with a smaller Z_{ord}, must be *larger*, and the weights of faraway keypoints are *smaller*. This is integrated into the proposed SRS by the formation of the *dense ordinal proximity map*, $\hat{\mathbf{D}}_{prox}$ from the optic flow computed to form the depth-weighted saliency map, $\mathbf{S}_{dm}$ (4.5). $\hat{\mathbf{D}}_{prox}$ is constructed by convolving a large Gaussian filter, $\mathbf{G}_{large}$, to spread the d_{prox} into regions that do not have any detected keypoints. This effectively converts the originally sparse proximity map into a pseudo-dense proximity map:

$$\hat{\mathbf{D}}_{prox} = \mathbf{D}_{prox} * \mathbf{G}_{large} \tag{5.7}$$

where $\mathbf{D}_{prox}$ is the set of all the d_{prox} computed from (5.6) that represents the *sparse ordinal proximity map* associated with the detected keypoints. The transformation from $\mathbf{D}_{prox}$ to $\hat{\mathbf{D}}_{prox}$ is illustrated in Fig. 5.11.

The SURF descriptors after AHC are then augmented with the adjusted d_{prox} to form the so called *salient-SURF* descriptors used in the formation of the Scene matrix, described in the next section.

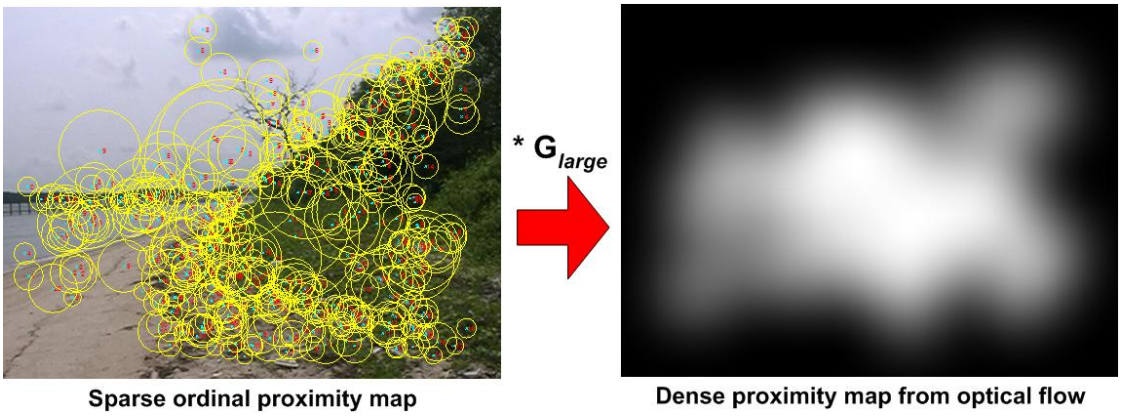

Figure 5.11: From the *sparse ordinal proximity map* $\mathbf{D}_{prox}$ estimated from the optic flow vectors (Fig. 5.9 (bottom)), the *dense ordinal proximity map* $\hat{\mathbf{D}}_{prox}$ is obtained by convolution with a large Gaussian filter, $\mathbf{G}_{large}$.

5.3 Constructing the Scene Matrix

The sparse ordinal proximity map, $\mathbf{D}_{prox}$, that contains the ordinal proximity, d_{prox}, associated with each SURF keypoint detected at the salient regions can now be combined with the Descriptor matrix (definition 5.8) to form the Scene matrix, $\mathbf{m}_s$, that contains a set of salient-SURF descriptors.

Definition 5.9. *The Scene matrix and salient-SURF keypoints/descriptors* The Scene Matrix extends the Descriptor matrix by incorporating the ordinal proximity, d_{prox} associated with each epipolar-verified SURF keypoint into a $N_d \times 71$ matrix with the following structure: $\mathbf{m}_s = [\mathbf{x}\ \mathbf{y}\ \mathbf{d}_{prox}\ \mathbf{a}\ \mathbf{b}\ \mathbf{a}\ \mathbf{l}\ \mathbf{Desc}]$ for a scene with N_d keypoints. The SURF keypoints are now localised in the three spatial directions (x, y, z) and are called *salient-SURF keypoints*. Each row in $\mathbf{m}_s$ forms a complete descriptor known as the *salient-SURF descriptor*. □

From the structure of $\mathbf{m}_s$, one can see that the inclusion of d_{prox} completely describes the position of the keypoints in an *ordinal scale*. This has been shown to be robust against viewpoint changes in section 3.3.4.

Since the proposed SRS encodes the salient-SURF keypoints separately for the three (HSV) colour spaces, three different scene matrices are equally created, denoted as $\mathbf{m}_s^j, j \in \{H, S, V\}$. A cell matrix (similar to section 5.1.4) is used to combine the Scene matrices together to form the *Scene cell matrix* denoted as $\mathbf{M}_s$:

Definition 5.10. *The Scene cell matrix* The Scene cell matrix, $\mathbf{M}_s$ combines the scene matrices from the three colour spaces together with the following structure: $\langle \mathbf{m}_s^H, \mathbf{m}_s^S, \mathbf{m}_s^V \rangle$ where (H, S, V) represent the hue, saturation and value (grayscale) colour spaces respectively. The notation $\langle \cdot \rangle$ represents a cell matrix. □

The structure of $\mathbf{M}_s$ is illustrated in Fig. 5.12.

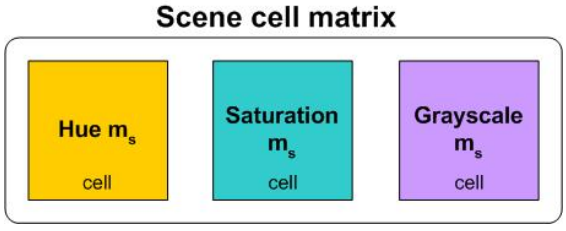

Figure 5.12: Combining three different $\mathbf{m}_s$ from different colour spaces into a Scene cell matrix.

The final set of salient-SURF keypoints (from all three colour spaces) is illustrated in Fig. 5.13 and the Scene cell matrix is stored into the local memory of

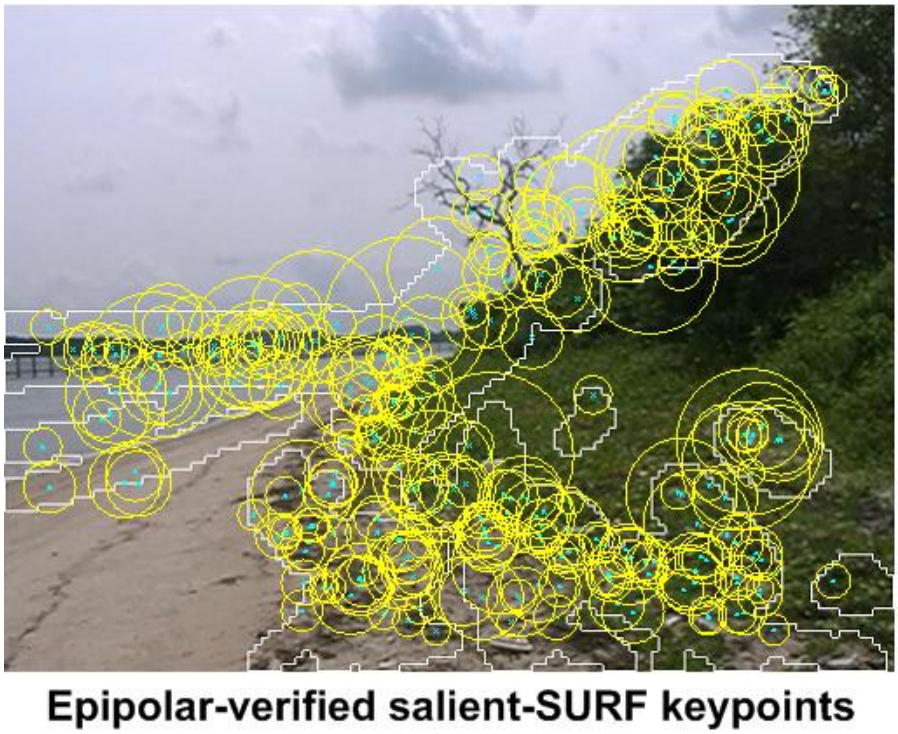

Figure 5.13: The final set of salient-SURF keypoints saved in the scene matrices. These keypoints are epipolar-verified and have their ordinal proximities adjusted by AHC.

the agent after this procedure. From this point onwards, the scene is completely represented by its three scene matrices $\mathbf{m}_s^j$, which will be used in determining the similarity score for scene recognition. This novel similarity metric, as well as the details of how an input query scene is processed by the proposed SRS is detailed in section 6.1.2.

5.4 Final remarks

This chapter has detailed how salient ROIs are combined with SURF keypoints to form salient-SURF keypoints/descriptors used to create the Scene matrix $\mathbf{m}_s$. The ordinal depth Z_{ord}, and ordinal proximity d_{prox}, are recovered from the optic flow of simulated TBL motion which are then integrated into the salient-SURF descriptors for a complete description of the scene structure in the three spatial directions (x, y, z). The three scene matrices, constructed from separate colour spaces, are combined into the Scene cell matrix that represents the scene. This Scene cell matrix is used in the *Scene decision module* of the proposed SRS, described in the next chapter.

Chapter 6

The Scene Decision module

This chapter describes the process of *scene decision*, that is, given an input query (test) scene and a database of reference scenes, can the proposed SRS decide if the test scene matches one (or some) of the reference scenes? In this work, the final decision of the proposed SRS is binary - *accept* the test scene as a reliable match or *reject* the test scene as unreliable.

The fundamental requirement in the design of a reliable scene decision module is a measure of scene similarity between two scenes. In this work, a novel scene similarity metric, known as the *Global Configuration Coefficient*, G_c, is formulated in section 6.1. This metric is computed from the Scene matrix cells, $\mathbf{M}_s$, of the two scenes that are being compared (definition 5.10). Using this measure, the scene decision procedure with a reference database is described in section 6.2 as a two step process. A *candidate match* is first determined after comparing the test scene

with all of the scenes in the reference database (section 6.2.1). This is followed by validating the candidate match using an *adaptive decision threshold* computed from the statistics of the matches (section 6.2.2). The final decision of accepting or rejecting the candidate match is based entirely on this estimated threshold. A discussion on how this threshold actually performs for true positives and negatives is presented in section 6.2.3. Finally, a modification to the decision threshold is proposed for the case of a difficult ambiguous scene in section 6.2.4.

6.1 A novel scene similarity metric

In this section, a novel measure of scene similarity that is fundamental to the scene decision module is formulated. The inherent shortcomings of using only feature matches alone as a similarity measure are first discussed in section 6.1.1. Using the concept of ordinal measures introduced in section 3.3, *rank correlations* of the spatial configuration of the salient-SURF keypoints encoded in the Scene matrix are used to improve the reliability of the proposed similarity measure against viewpoint changes, which is presented in section 6.1.2.

6.1.1 Using matches alone for similarity is unreliable

Given two scenes, a *reference* scene stored in the agent's memory and an input *test* scene, encoded by their respective Scene matrix cells, $\mathbf{M}_s^1$ and $\mathbf{M}_s^2$, is it possible to

compare and conclude that these two scenes are taken from the same location or not, even under significant viewpoint and illumination distortions (see Fig. 1.2)?

A simple solution would be to attempt to match the salient-SURF keypoints of the two scenes (section 5.1.3). The larger the number (or percentage) of matches, the more likely the scenes are from the same location. This simple method has several disadvantages. Firstly, the *correctness* and *uniqueness* of the matches are assumed. In cases where the image distortions are small, the salient-SURF matches are quite reliable and this assumption holds. This may be untrue, however, when difficult scenes are presented. Ambiguity occurs when the keypoints are considered independently from the entire scene context and mismatches of very similar looking features occur. For example, repeated structures such as shelves, posters (indoor), branches and leaves (outdoors) are often not unique and distinct enough for reliable recognition (Figs.1.3,3.1 and see section 3.1.1).

Secondly, although a descriptor may be salient (and possibly unique) within the same image, its uniqueness is not guaranteed over the full range of images that the agent may encounter in its environment. This is not a fault but a *limitation* of the descriptor that only has *local* knowledge of the scene it encodes but lacks the *global* knowledge of the scenes in the whole database. This means that relying on the number of matches alone is not going to work since the matches may be *one-to-many*. That is, a keypoint may be matched to a number of similar looking keypoints from different locations.

Thirdly, another problem with using this simple method is the *threshold* needed to reliably reject bad matches. A fixed threshold is obviously not going to work for the large variety of scenes that the SRS encounters (indoors and outdoors) with various degrees of distortions. The computation of an *adaptive* threshold is also not obvious using just the number (or percentage) of matches alone. A positive match, for example, could occur with only a small number of matches for scenes that possess significant image distortions (Fig. 6.1).

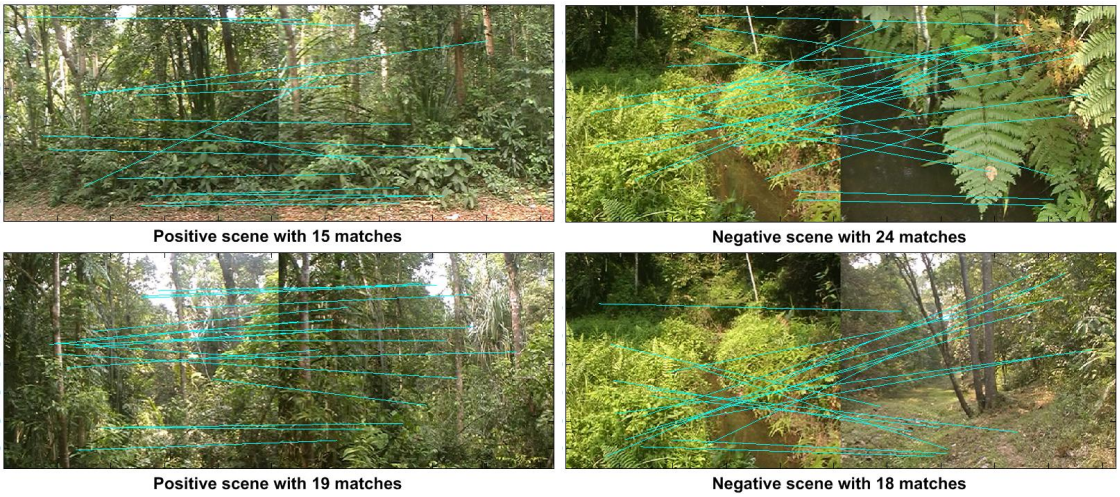

Figure 6.1: Using the number of matches alone as a measure of scene similarity is highly unreliable. The correspondences are shown as lines connecting matched keypoints across the images. Left: True positives with 15 and 19 matches, Right: Mismatched false positives (negatives) with 24 and 18 matches.

From Fig. 6.1, one can see that dissimilar scenes may have more matches (which are of course wrong) than true positives. Using the number of matches alone as a similarity metric is thus highly *unreliable*. Instead, the *discriminating* information is in the spatial configurations of the keypoints which are preserved for

true positives and non-existent for dissimilar cases (see section 3.3.1). Graphically, the lines connecting true positives are more *ordered* than negative scenes since the spatial configurations of the correctly matched keypoints are preserved. Such orderliness is not present at all for dissimilar (negative) scenes.

6.1.2 The Global Configuration Coefficient, G_c

This section continues from the discussion of the previous section by formulating a novel measure of similarity known as the *Global Configuration Coefficient*, G_c. This similarity metric addresses the shortcomings of using the number of matches alone by including the global landmark/keypoint configuration information into its design. This configuration information exploits the previously introduced concepts of rank correlation measures. From section 3.3.4, the usefulness of an ordinal scale for viewpoint invariant scene recognition is highlighted - as long as the viewpoint change is not too extreme, the spatial configuration of the matched keypoints is preserved. Using rank correlations of the ordinal positions of the matched keypoints (section 3.3.2), a measure of similarity sensitive to the rank orders of the spatial configuration is introduced.

Suppose two scenes are presented, represented by their individual Scene matrix cells ($\mathbf{M}_s^1$, $\mathbf{M}_s^2$). The first step is to match the salient-SURF keypoints separately over the three colour spaces using the procedure described in section 5.1.3. These

initial matches are then validated by RANSAC so as to remove any erroneous matches that do not respect the epipolar constraint (section 5.2.2):

$$\dot{\mathbf{m}}_{kp}^{j} = \mathbf{m}_{1s}^{j} \leftrightarrow \mathbf{m}_{2s}^{j}, j \in \{H, S, V\} \tag{6.1}$$

where the scene matrices of the first and second scenes from the j^{th} colour space are denoted by $(\mathbf{m}_{1s}^{j}, \mathbf{m}_{2s}^{j})$ respectively. The symbol $\leftrightarrow$ denotes the salient-SURF matching procedure together with verification by RANSAC. Next, the matched keypoints in the three colour spaces, denoted as $\dot{\mathbf{m}}_{kp}^{j}$ for the j^{th} colour space, are grouped together to form the *Matching matrix*, $\dot{\mathbf{M}}_{kp}$:

$$\dot{\mathbf{M}}_{kp} = \left[\dot{\mathbf{m}}_{kp}^{H} \dot{\mathbf{m}}_{kp}^{S} \dot{\mathbf{m}}_{kp}^{V}\right] \tag{6.2}$$

Grouping the matches together into one matrix loses all information concerning the colour space from which the keypoints originate. This is justifiable as the main reason for separating the keypoints and scene matrices into separate cells is to prevent mismatches of the keypoints across incompatible colour spaces (section 5.1.1). Since no more matching is required after this step, combining the matches together into $\dot{\mathbf{M}}_{kp}$ simplifies the implementation of the proposed SRS significantly. The structure of $\dot{\mathbf{M}}_{kp}$ is defined as:

Definition 6.11. *Matching matrix* The Matching matrix $\dot{\mathbf{M}}_{kp}$, is a $N_{match} \times 6$

matrix where N_{match} denotes the number of matches between two scenes with the following structure: $\dot{\mathbf{M}}_{kp} = [\mathbf{x}_1\ \mathbf{y}_1\ \mathbf{d}_{1prox}\ \mathbf{x}_2\ \mathbf{y}_2\ \mathbf{d}_{2prox}]$ The matrix retains only the *localisation* information of the matched keypoints of the two scenes, denoted by the subscripts $(1, 2)$ respectively. □

Next, a novel similarity metric known as the *Global Configuration Coefficient*, G_c, is defined with $\dot{\mathbf{M}}_{kp}$ as the input:

$$G_c(\dot{\mathbf{M}}_{kp}) = \frac{N_{\%test}}{200} \times (\overline{S_\rho} + \overline{K_\tau}) \tag{6.3}$$

where $N_{\%test}$ is the *percentage* matches with respect to the test (first) scene given by:

$$N_{\%test} = \frac{N_{match}}{N_{1d}} \times 100 \tag{6.4}$$

and N_{1d} denotes the original number of salient-SURF keypoints in the test scene. $(\overline{S_\rho}, \overline{K_\tau})$ are the *means* of the positive Spearman's ρ (3.1) and Kendall's τ (3.2) rank correlations in the three spatial (x, y, z) directions (see Fig. 3.14 for an illustration) given as:

$$\begin{aligned} \overline{S_\rho} &= \frac{1}{3}\sum_i S_\rho^i, i \in \{x, y, z\} \\ \overline{K_\tau} &= \frac{1}{3}\sum_i K_\tau^i, i \in \{x, y, z\} \end{aligned} \tag{6.5}$$

where Spearman's ρ and Kendall's τ of a particular direction are denoted as $(S^i_\rho, K^i_\tau), i \in \{x, y, z\}$. The rank correlations are computed from the elements of the Matching matrix, $\dot{\mathbf{M}}_{kp}$ (definition 6.11), given as:

$$\begin{cases} S^x_\rho &= S_\rho(\dot{\mathbf{M}}_{kp}(\mathbf{x}_1), \dot{\mathbf{M}}_{kp}(\mathbf{x}_2)) \\ S^y_\rho &= S_\rho(\dot{\mathbf{M}}_{kp}(\mathbf{y}_1), \dot{\mathbf{M}}_{kp}(\mathbf{y}_2)) \\ S^z_\rho &= S_\rho(\dot{\mathbf{M}}_{kp}(\mathbf{d}_{1prox}), \dot{\mathbf{M}}_{kp}(\mathbf{d}_{2prox})) \end{cases} \tag{6.6}$$

and

$$\begin{cases} K^x_\tau &= K_\tau(\dot{\mathbf{M}}_{kp}(\mathbf{x}_1), \dot{\mathbf{M}}_{kp}(\mathbf{x}_2)) \\ K^y_\tau &= K_\tau(\dot{\mathbf{M}}_{kp}(\mathbf{y}_1), \dot{\mathbf{M}}_{kp}(\mathbf{y}_2)) \\ K^z_\tau &= K_\tau(\dot{\mathbf{M}}_{kp}(\mathbf{d}_{1prox}), \dot{\mathbf{M}}_{kp}(\mathbf{d}_{2prox})) \end{cases} \tag{6.7}$$

Using the mean values penalise the rank correlation when one (or more) of its spatial configuration does not preserve the ordering constraint of the matched keypoints. This is the usually the case when the scenes are dissimilar. Although mismatches occur in all cases, the degradation in the rank correlations is expected to be less pronounced when two scenes are similar.

The formulation of G_c (6.3) combines both the *local* keypoint similarity and the *global* configuration of the matched keypoints together into a simple measure

of similarity. The local similarity is indirectly captured by $N_{\%test}$ that measures the percentage of the keypoint matches stored in $\dot{\mathbf{M}}_{kp}$. The global configuration of the matched keypoints is captured in the mean rank correlations, $\overline{S_\rho}$ and $\overline{K_\tau}$. G_c is thus close to 1 for a perfect match with a high $N_{\%test}$ that preserves the overall spatial configuration. For dissimilar scenes, G_c is near to zero with very few matches (small $N_{\%test}$) and the rank correlations of the mismatched keypoints are *likely* to be small too as the spatial configuration is not preserved.

The incorporation of $N_{\%test}$ is important as the use of rank correlations is highly *dependent* on the number of matched keypoints. This is because a small number of matches is often not statistically significant for the computed rank correlations to be useful. For example, if only three keypoints are matched in the test scene, it is very likely that the three keypoints will have a similar configuration with many scenes in the reference database. One can thus view the formulation of (6.3) as weighing the *confidence* of the rank correlations by $N_{\%test}$.

In practice, the effects of wrong correspondences and occlusions due to image distortions often degrade G_c significantly (between 0.3 to 0.4), even for positive scenes. This degradation is, nonetheless, usually more pronounced in negative scenes. As the amount of image distortions increases, this degradation will get even worse. This means that using a *fixed* threshold for scene decision is not feasible in practice. An *adaptive* threshold, estimated from (6.3) and (6.5), is presented in section 6.2.2 as a reliable alternative.

6.2 Determining scene equivalence from a database

This section extends the use of G_c with a *reference database*, $\mathbb{D}_{ref}$, of N_{ref} scenes each defined by their individual Scene matrix cell, denoted as $\mathbf{M}_s^i, i \in \{1, 2, \cdots N_{ref}\}$ for the i^{th} reference scene. Given an input query scene, represented by $\mathbf{M}_s^{test}$, the objective of this section is to show how a final decision that either accepts or rejects $\mathbf{M}_s^{test}$ is made. This is done in two phases detailed in the following two sections.

6.2.1 Determining the candidate match

The first phase proceeds by making N_{ref} pairwise comparisons between the input test scene, $\mathbf{M}_s^{test}$, and the N_{ref} reference scenes in the database, $\mathbb{D}_{ref}$. Each comparison computes G_c using (6.3) which is stored in a $N_{ref} \times 8$ *Match statistic matrix*, $\mathbf{\Pi}_s$. Besides G_c, each row of $\mathbf{\Pi}_s$ contains $N_{\%test}$ and the rank correlations in the three spatial directions denoted as $(S_\rho^i, K_\tau^i), i \in \{x, y, z\}$. $\mathbf{\Pi}_s$ has the following structure:

$$\mathbf{\Pi}_s = \left[\mathbf{N}_{\%test}\ \mathbf{S}_\rho^x\ \mathbf{S}_\rho^y\ \mathbf{S}_\rho^z\ \mathbf{K}_\tau^x\ \mathbf{K}_\tau^y\ \mathbf{K}_\tau^z\ \mathbf{G}_c\right] \tag{6.8}$$

where each term on the right hand side is a column vector of size N_{ref} and corresponds to the statistic collected over all pairwise comparisons.

The *candidate match*, $\mathbb{G}_{cand}$, is the reference scene that yields the largest G_c in $\mathbf{\Pi}_s$:

$$G_{cand} = \max(\mathbf{G}_c) \tag{6.9}$$

G_{cand} thus represents the *best* match score that is produced by the pairwise comparisons. It is the reference scene that resulted in the most matches with $\mathbf{M}_s^{test}$ with the least distortion in the global configuration of the matched keypoints. The whole process of extracting $\mathbb{G}_{cand}$ is illustrated in Fig. 6.2.

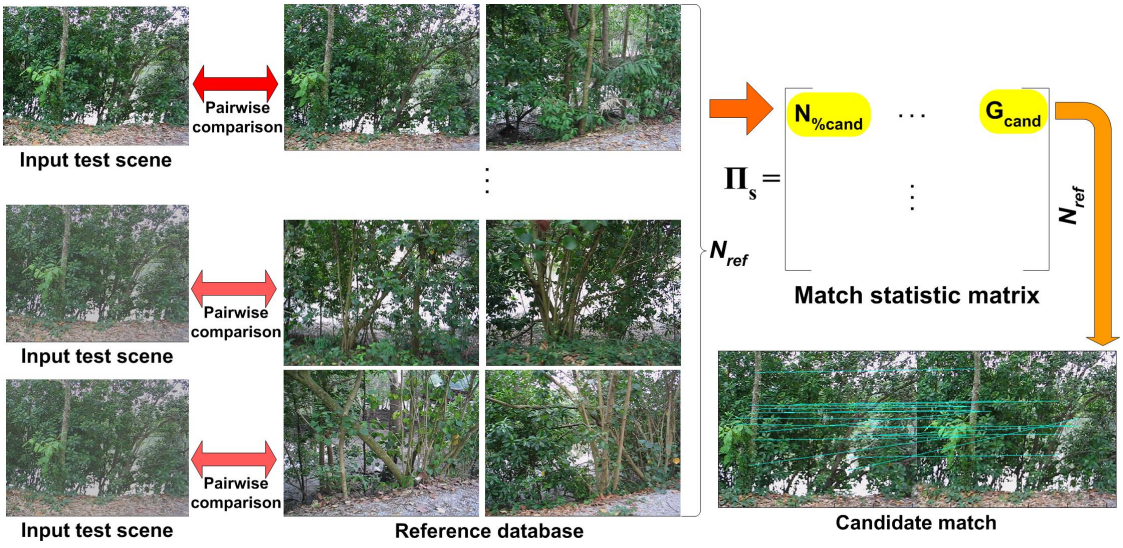

Figure 6.2: Extracting $\mathbb{G}_{cand}$ from a reference database. Multiple pairwise comparisons are made with each reference scene in the database to form a Match statistic matrix, $\mathbf{\Pi}_s$. The best score in $\mathbf{G}_s$ represents the candidate match score, G_{cand}, which is then selected (highlighted in yellow).

Since the test scene can be a *positive* with a matching scene in the database, or a *negative* with no matching scenes in the database, $\mathbb{G}_{cand}$ must be validated by a decision module. Hence, the remaining crucial problem is to propose a *decision threshold*, D_t, such that a decision, D_f on G_{cand}

$$D_f = \begin{cases} \text{ACCEPT} & \text{if } G_{cand} \geq D_t \\ \text{REJECT} & \text{if } G_{cand} < D_t \end{cases} \tag{6.10}$$

can be made. This is shown in the second phase of the decision module, where an *adaptive* decision threshold is proposed.

6.2.2 Adaptive decision threshold

In this section, a decision threshold D_t, is estimated so that (6.10) can be used to accept or reject the candidate match, $\mathbb{G}_{cand}$. D_t can be a *fixed* threshold learnt by presenting the SRS with a series of training scenes of the environment before actual scene recognition or it can be *estimated* from $\mathbf{\Pi}_s$. As this work is concerned with natural outdoor scenes with a large dynamic range, the latter method is chosen as it is *adaptive* to various environments and does not require any training images from the environment which may be totally unknown or outdated. Furthermore, some positive scenes may have only a few matches due to large distortions or dim illumination that degrade G_c so much so that using a fixed D_t becomes impractical,

as is shown in Fig. 6.1.

The first step in estimating D_t is to construct the *Decision matrix*, $\mathbf{\Delta}_s$ from the *best few* matches in $\mathbf{\Pi}_s$. These matches are determined based on two criteria, in terms of their G_c and also in terms of $N_{\%test}$:

1. From the first column vector of $\mathbf{\Pi}_s$ that contains $\mathbf{N}_{\%test}$, the N_{ref} matches are sorted based on $N_{\%test}$. The matches with $N_{\%test} > t_{\%}$, where $t_{\%}$ is a fixed percentage threshold, are retained.

2. Next, using the last column vector of $\mathbf{\Pi}_s$, the elements in $\mathbf{G}_c$ are ranked so that only the N_{top} G_cs are retained. N_{top} is a fixed number that determines how many best few matches are retained.

3. Finally, $\mathbf{\Delta}_s$ is obtained by combining the results in the first two steps so that only matches that are significant (the intersections of the first two steps) in *both* $N_{\%test}$ and G_c are retained for the estimation of D_t.

In this work, the values for the fixed parameters $(t_{\%}, N_{top})$ are set at $(10\%, 5)$ respectively. The number of rows that remain is denoted as N_{best} and this forms the $N_{best} \times 8$ $\mathbf{\Delta}_s$ where the structure is detailed below:

Definition 6.12. *Structure of* $\mathbf{\Delta}_s$ The Decision matrix, $\mathbf{\Delta}_s$ is a $N_{best} \times 8$ matrix with the following structure: $\mathbf{\Delta}_s = \left[\mathbf{N}_{\%\Delta}\ \mathbf{S}^x_{\rho\Delta}\ \mathbf{S}^y_{\rho\Delta}\ \mathbf{S}^z_{\rho\Delta}\ \mathbf{K}^x_{\tau\Delta}\ \mathbf{K}^y_{\tau\Delta}\ \mathbf{K}^z_{\tau\Delta}\ \mathbf{G}_{c\Delta}\right]$ where the subscript Δ is added to emphasise the difference in the column vectors

between $\mathbf{\Delta}_s$ and $\mathbf{\Pi}_s$. □

A *Threshold vector*, $\mathbf{\Xi}_s$, containing 7 elements is constructed from $\mathbf{\Delta}_s$ that has the same structure as a row in $\mathbf{\Pi}_s$ (without the G_c):

$$\mathbf{\Xi}_s = \left[N_{\%\Xi}\ S^x_{\rho\Xi}\ S^y_{\rho\Xi}\ S^z_{\rho\Xi}\ K^x_{\tau\Xi}\ K^y_{\tau\Xi}\ K^z_{\tau\Xi}\right] \tag{6.11}$$

The elements in $\mathbf{\Xi}_s$, once determined, are used to compute directly the estimate of the D_t defined as:

$$D_t = \frac{N_{\%\Xi}}{200} \times (\tilde{S_{\rho\Xi}} + \tilde{K_{\tau\Xi}}) \tag{6.12}$$

where $(\tilde{S_{\rho\Xi}}, \tilde{K_{\tau\Xi}})$ are derived from the means of the rank correlations in $\mathbf{\Xi}_s$:

$$\begin{cases} \tilde{S_{\rho\Xi}} = \frac{1}{3}\sum_i S^i_{\rho\Xi}, i \in \{x, y, z\} \\ \tilde{K_{\tau\Xi}} = \frac{1}{3}\sum_i K^i_{\tau\Xi}, i \in \{x, y, z\} \end{cases} \tag{6.13}$$

The rest of the section describes how the elements in $\mathbf{\Xi}_s$ are derived from $\mathbf{\Delta}_s$.

The first element of $\mathbf{\Xi}_s$, $N_{\%\Xi}$, is given the value of the candidate match, $N_{\%cand}$ (see Fig. 6.2 on the structure of $\mathbf{\Pi}_s$) if the candidate match row is found in $\mathbf{\Delta}_s$; if

not, the largest $N_{\%\Delta}$ in $\mathbf{N}_{\%\Delta}$ is used:

$$N_{\%\Xi} = \begin{cases} N_{\%cand} \text{ if } \mathbb{G}_{cand} \in \mathbf{\Delta}_s \\ \\ \max\left(\mathbf{N}_{\%\Delta}\right) \text{ otherwise} \end{cases} \tag{6.14}$$

The reason for this assignment rule is very simple - if $\mathbb{G}_{cand} \notin \mathbf{\Delta}_s$, it is likely to be unreliable and should be rejected. Using the largest $N_{\%\Delta}$ will give us a D_t that is likely to be *larger* than G_{cand} since $N_{\%\Xi}$ determines partially the value of D_t (6.12). Invoking (6.10) allows the proposed SRS to effectively reject the unrelible $\mathbb{G}_{cand}$.

The rest of the elements in $\mathbf{\Xi}_s$ are determined in a three step process:

1. Collect the rank correlations over the three spatial directions together to form a *composite rank correlation matrix*, denoted as $(\mathbf{\Sigma}_{\rho\Delta}, \mathbf{\Lambda}_{\tau\Delta})$:

$$\begin{cases} \mathbf{\Sigma}_{\rho\Delta} = \left[\mathbf{S}^x_{\rho\Delta}\ \mathbf{S}^y_{\rho\Delta}\ \mathbf{S}^z_{\rho\Delta}\right] \\ \\ \mathbf{\Lambda}_{\tau\Delta} = \left[\mathbf{K}^x_{\tau\Delta}\ \mathbf{K}^y_{\tau\Delta}\ \mathbf{K}^z_{\tau\Delta}\right] \end{cases} \tag{6.15}$$

2. Compute the *median* value among the elements of $(\mathbf{\Sigma}_{\rho\Delta}, \mathbf{\Lambda}_{\tau\Delta})$, denoted as

$\text{med}(\mathbf{\Sigma}_{\rho\Delta}), \text{med}(\mathbf{\Lambda}_{\tau\Delta})$ and take the *minimum* among the two values to determine a threshold for significant rank correlations, t_{rank}:

$$t_{rank} = \min\left(\text{med}(\mathbf{\Sigma}_{\rho\Delta}), \text{med}(\mathbf{\Lambda}_{\tau\Delta})\right) \tag{6.16}$$

The value of t_{rank} is limited to a maximum value so that a sufficient number of rank correlations can be used to estimate D_t from $\mathbf{\Xi}_s$ (see the next step). A t_{rank} that is too large yields too few rank correlations for the subsequent computations to be reliable. In this work, t_{rank} is limited to 0.6.

3. Using t_{rank}, the statistics of the rank correlation elements in $\mathbf{\Xi}_s$ are determined by computing once again the median of these rank correlation entries that are *larger* than t_{rank} in $\mathbf{\Delta}_s$.

$$\begin{cases} S^i_{\rho\Xi} = \text{med}\left\{\mathbf{S}^i_{\rho\Delta} | \mathbf{S}^i_{\rho\Delta} \triangleright t_{rank}\right\}, i \in \{x, y, z\} \\ K^i_{\tau\Xi} = \text{med}\left\{\mathbf{K}^i_{\tau\Delta} | \mathbf{K}^i_{\tau\Delta} \triangleright t_{rank}\right\}, i \in \{x, y, z\} \end{cases} \tag{6.17}$$

where the $\triangleright$ operator represents a '>' comparison between the *elements* of a vector on the LHS with a scalar on the RHS. Using t_{rank} ensures that only the most significant rank correlations that contribute to the best matches in $\mathbf{\Delta}_s$ are used in the computation of D_t in (6.12).

With D_t determined, the proposed SRS arrives at the final decision D_f, to

accept or reject the input test scene by comparing D_t and the candidate match score G_{cand} (6.10). The next section explains briefly how D_t actually works in providing a reasonable adaptive threshold for the scene decision module.

6.2.3 How D_t works

The design of D_t, estimated from the best few matches in $\mathbf{\Delta}_s$, is useful as a reasonable threshold as D_t represents what one can term as an *average* match that sets the benchmark for reliability. If G_{cand} is indeed a *reliable* match, most if not all of its elements should have higher values than the corresponding entries in $\mathbf{\Delta}_s$.

The intuitive idea of using the median of the rank correlations of the N_{best} matches in $\mathbf{\Delta}_s$ (6.17) is illustrated in Fig. 6.3. The six median rank correlation components are shown as boxes - black boxes have values that differ significantly from G_{cand}. Since the candidate match must be in the N_{best} matches in the case of a true *positive match*, the majority of the rank correlations for D_t used in (6.13) will be smaller than the components of the candidate match (black boxes with '<' sign). This will make D_t *smaller* than G_{cand} for the match to be accepted (6.10).

For the *negative* case, since the percentage matches are likely to be small, the number of N_{best} matches is expected to be fewer. As was highlighted earlier in section 6.1.2, the reliability of the rank correlations degrades significantly with fewer matches. This makes the rank correlations contributing to D_t to be *varied*,

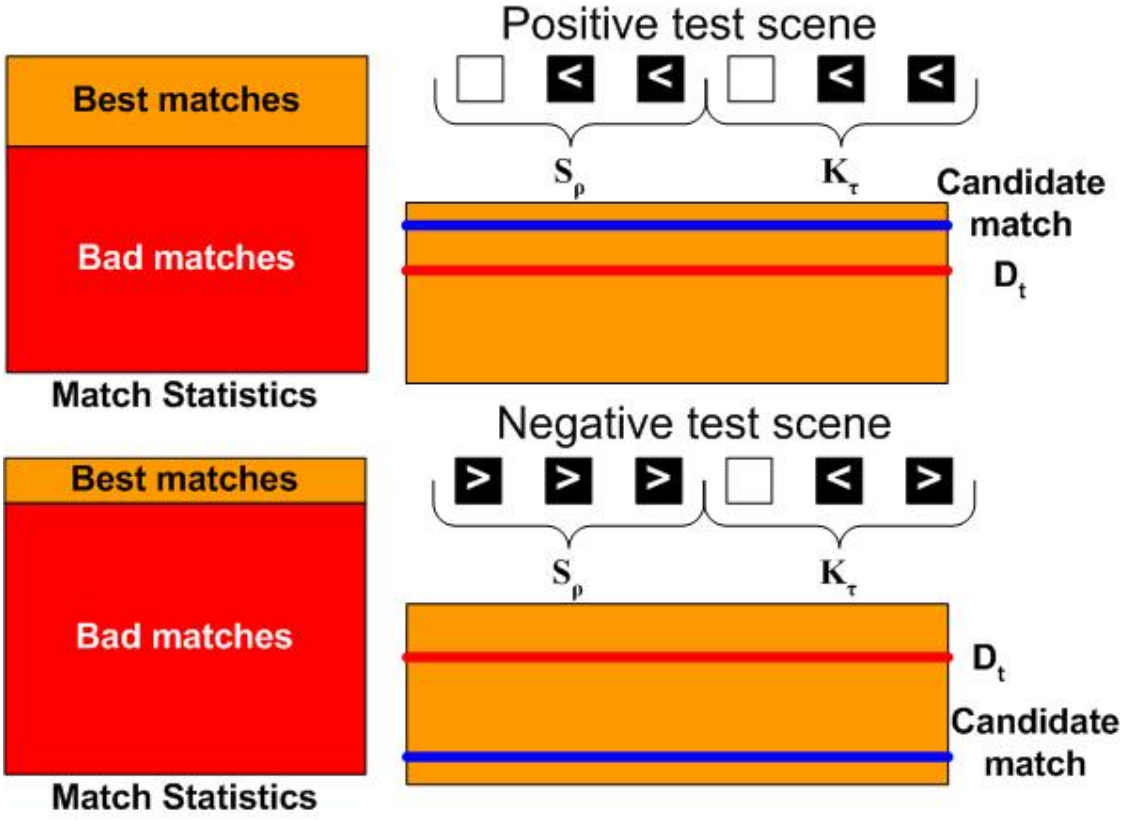

Figure 6.3: An illustration of how D_t, computed from the best few matches in $\boldsymbol{\Delta}_s$, arrives at providing a reasonable threshold in the case of a positive scene (top) and a negative scene (bottom), see text for details.

with more contribution from reference scenes having rank correlations that are likely to be larger (black boxes with '>' sign) than that of the candidate match. The result is a D_t that is *larger* than G_{cand} which rejects the negative match by invoking (6.10).

Furthermore, as D_t is estimated from the N_{best} matches in $\boldsymbol{\Delta}_s$, it *evolves* together with G_{cand}. Hence, D_t is *adaptive* and varies for different test scenes and

reference databases. This adaptability of D_t is a critical component for the proposed SRS to perform in vastly different environments under different image conditions. A complex outdoor natural test scene with numerous occlusions (especially in enclosed or highly cluttered forest scenes) and illumination changes (especially those dimly lit with few distinct features) will generally give lower *absolute* values of G_c even for positive matches. An indoor scene typically yields higher values of G_c as the features are more well defined and there are less image distortions due to occlusions and illumination changes. Such global changes in the absolute values of G_c does not affect the performance of the proposed SRS since G_{cand} is obtained from the *best* match (6.9) and D_t is estimated from the statistics of the best few matches in $\mathbf{\Delta}_s$ (6.12). This results in an adaptive threshold that reflects the current environment of the test scene so that a reliable decision can be made on G_{cand}.

However, there must be an absolute *minimum* acceptable threshold for D_t to be effective. The candidate match cannot be accepted if G_{cand} is very small. A minimum threshold, D_{min} represents the minimum value that we can accept G_{cand} before it is considered as unreliable and is rejected immediately. The final decision,

D_f must take this into account:

$$D_f = \begin{cases} \text{ACCEPT} & \text{if } (G_{cand} \geq D_{min}) \cap (G_{cand} \geq D_t) \\ \\ \text{REJECT} & \text{otherwise} \end{cases} \tag{6.18}$$

The choice of D_{min} is however application dependent. A small value of D_{min} increases the robustness of the algorithm to large image distortions that tend to degrade G_c rapidly, at the expense of losing discriminatory power for ambiguous scenes. This will result in more *false* positives in the SRS. Once again, this is an illustration of the antagonism between robustness and discriminatory power of any practical SRS, highlighted in section 1.4. In this work, D_{min} is set between 0.01 to 0.03 so as to tolerate a larger amount of image distortions that the proposed SRS encounters.

Three examples are illustrated in Appendix C for three cases of $\mathbb{G}_{cand}$ - A positive match (C.1), a negative match (C.2) and finally an *ambiguous match* (C.3). These examples show how D_t adapts itself to various situations for a reliable decision to be made. The case of an ambiguous match is detailed in the next section.

6.2.4 Scene decision for ambiguous cases

The above procedure of estimating D_t will fail if the Decision matrix $\mathbf{\Delta}_s$ is *empty*. Since $\mathbf{\Delta}_s$ is constructed from the best few matches in $\mathbf{\Pi}_s$ based on G_c and $N_{\%test}$

(section 6.2.2), an empty $\mathbf{\Delta}_s$ will occur if there are no matches that concurrently satisfy the two criteria - some matches may have a high G_c but a small $N_{\%test}$ and vice versa. In this case, D_t cannot be estimated from the procedure described in section 6.2.2 above and the test scene in this case is deemed to be *ambiguous*.

It is highly likely that $\mathbb{G}_{cand}$ is an unreliable match since the pairwise matching with the entire database does not produce a single good match in terms of G_c and $N_{\%test}$ and is thus inconclusive. The solution proposed is based on the assumption that a *good positive* match is not likely to result in such an ambiguous case and hence this match should be *rejected* as unreliable if possible.

Once an ambiguous case is detected, the scene decision module will set $D_t = G_{cand}$ directly. From (6.10), all matches will be accepted if no more modifications are made. Instead, D_{min} is further modified to a *higher* value, denoted as D^*_{min} which makes it more likely to reject unreliable matches (6.18). An illustration of how changing the value of D_{min} to D^*_{min} helps in rejecting an unreliable $\mathbb{G}_{cand}$ is shown in Fig. 6.4

This procedure is justified on the basis that since D_t is effectively unable to decide if $\mathbb{G}_{cand}$ should be accepted, one can only *heuristically* decrease the tolerance for false matches in this unreliable case. This is done by directly manipulating the value of D_{min}, making it larger to D^*_{min} so that it is unlikely the $\mathbb{G}_{cand}$ is accepted. However, this does not rule out the possibility that the scene is a true positive and

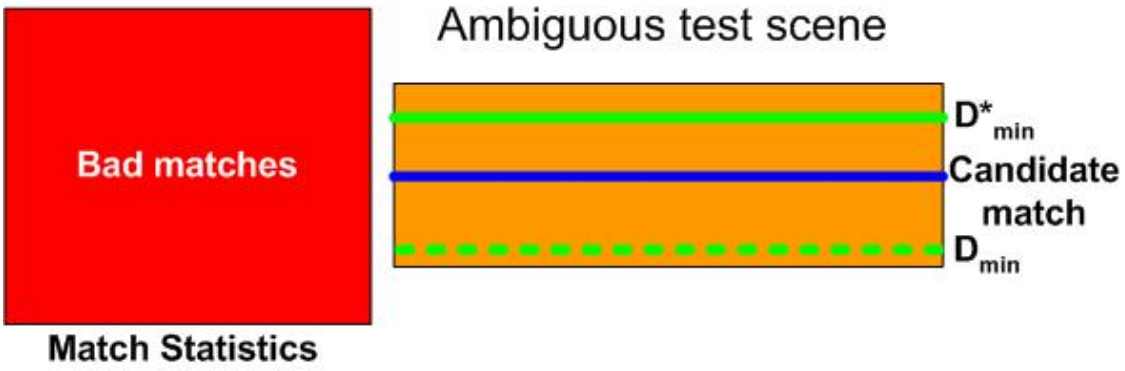

Figure 6.4: For the case of ambiguous scenes, D_t cannot be computed for scene decision. Instead, modifying the value of D_{min} to a higher D^*_{min} value allows such ambiguous scenes to be rejected.

should be accepted. G_{cand} must then be larger than D^*_{min} for acceptance which is still possible if the true positive has $N_{\%cand}$ or G_{cand} that are very near (but lower) than the threshold criteria for constructing $\mathbf{\Delta}_s$. In this work, D^*_{min} is set to 0.05. An example of a *positive* ambiguous test scene is shown in Appendix C.3.2 that demonstrates how this procedure works.

6.3 Final remarks

This chapter completes the description of the proposed SRS with the scene decision module. A novel similarity measure is introduced, known as the Global Configuration Coefficient, G_c which combines both 2D pixel correlation information ($N_{\%test}$) as well as rank correlation measures of the spatial configuration in ($\overline{S_\rho}, \overline{K_\tau}$). G_c is then used in an extended framework for determining scene equivalance between an

input test scene and a reference image database. This framework describes how the initial candidate match, $\mathbb{G}_{cand}$ is extracted and validated by estimating an adaptive decision threshold, D_t (6.10). Modifications to the procedure for ambiguous scenes are also considered in (6.18). Finally, examples which show how D_t can produce a reasonable threshold are illustrated in Appendix C for a variety of common cases of $\mathbb{G}_{cand}$.

The next chapter describes the experimental setup and tests that are used to validate the proposed SRS's performance and effectiveness for a variety of environments under various image distortions. A detailed discussion of the experimental results follows thereafter, and attempts to highlight the contribution of the various components to the recognition accuracy of the proposed SRS.

Chapter 7

Experimental Results and Discussion

This chapter presents the experiments conducted to verify the proposed SRS. The experimental setup is first introduced in section 7.1 where the four different image databases used are described. The experimental procedure is subsequently presented in section 7.2 where various measures of recognition accuracy are introduced so as to evaluate the performance of the proposed SRS. In order to highlight the performance of the proposed SRS, several comparative studies with various similarly designed SRSs are described in section 7.3. The results of the experiments are summarised in section 7.4, and various interesting examples from the image databases are highlighted. Finally, an analysis and discussion of the experimental results are presented in section 7.5.

7.1 Experimental setup

In this section, the four image databases used in all the experiments are described. These databases contain images taken from four different environments (their referenced name in this thesis is denoted in **bold**) - indoors(**IND**), a sandy shore(**UBIN**), a tropical rainforest(**NS**) and a mangrove forest(**SBWR**). For this thesis, the distinction between *reference* and *test* scenes is based on how the image scenes are used in the experiments - scenes that make up the *reference database*, $\mathbb{D}_{ref}$, are reference scenes while scenes used for testing the recognition accuracy of the proposed SRS are test scenes. In order to validate the robustness and discriminatory power of the proposed SRS (section 1.3), the scenes in the database often contain significant image distortions. A summary of the four databases is shown in Table 7.1 where the number of scenes used in each environment is shown as a triplet (N_{ref} N_{pos} N_{neg}) which are respectively the number of reference, positive and negative scenes used in the particular database. Some typical example scenes from the four databases are shown in Fig. 7.1. The following sections describe the databases in greater detail. More examples of the reference and test scenes used in the experiments are shown in Appendix D.3.

Table 7.1: The four databases used in the experiments.

Database	$(N_{ref}, N_{pos}, N_{neg})$	Type
IND	$(18, 25, 21)$	Indoor
UBIN	$(20, 63, 69)$	Outdoor coastal
NS	$(20, 41, 52)$	Outdoor varied
SBWR	$(15, 15, 16)$	Outdoor enclosed

7.1.1 Database IND

This database consists of indoor scenes taken under typical lighting conditions. Included is a set of artificial scenes with simple features that are configured differently in space, so as to test the usefulness of rank correlations in detecting changes in the ordinal configuration of ambiguous scenes sharing the same features (Fig. 7.1(**IND**: top)). Another set of images contains scenes from a typical office/factory with significant clutter and people moving around (Fig. 7.1(**IND**: bottom)). This database verifies the robustness of the proposed SRS against various image distortions due to viewpoint changes and human movements. The database also tests the proposed SRS's ability to discriminate ambiguous scenes containing numerous similar features that confuse other methods (*e.g.* [3, 76, 87])

7.1.2 Database UBIN

This database consists of outdoor images taken predominantly along a sandy shore and among the surrounding vegetation of an island. It is the nesting habitat of many species of tropical sand-digging wasps (section 3.5.2 and Fig. 3.20) where

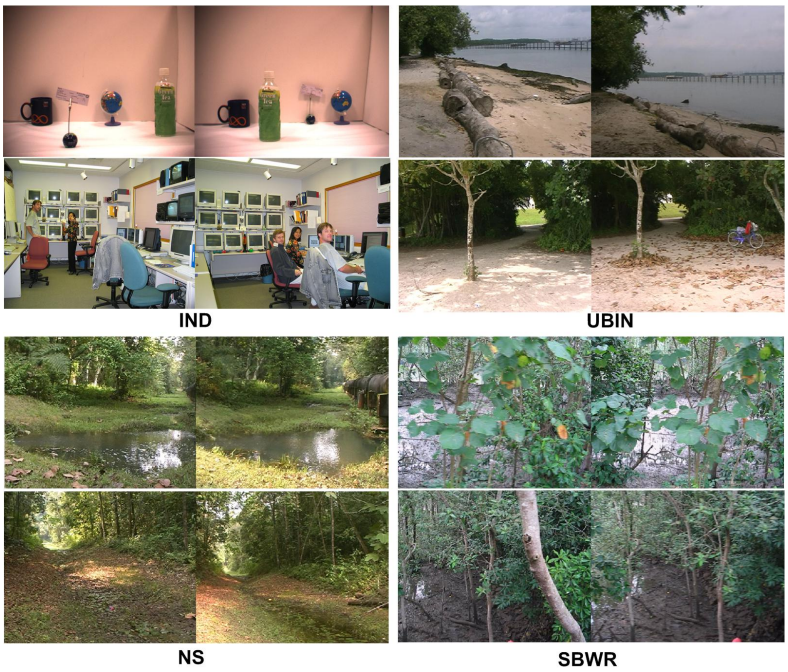

Figure 7.1: Various challenging test (left) and reference scenes (right) of the four databases, two rows ((t)op,(b)ottom) shown per scene. **IND**: ambiguous scenes(t) and viewpoint changes with significant clutter(b), **UBIN**: clear *vs.*hazy overcast sky(t) with differences in tides and shadows *vs.*leaves swept up(b), **NS**: non-uniform illumination(t) and changes in scene content due to rain and tree fall(b) and **SBWR**: numerous occlusions due to dense vegetation. See text for a detailed description of each database.

one can see them making foraging trips to and fro their nests in an unerring manner. The scenes are taken on two different days a month apart from each other at around the same time but under very different weather conditions. The reference scenes are taken on a clear sunny day while a portion of the test scenes are taken under very hazy (dim) conditions. Furthermore, the test scenes have also suffered from significant changes due to natural erosion and the dynamic nature of a coastal environment. For example, the reference scenes are taken at low tides while the

test scenes are taken at high tides which make this database very challenging (Fig. 7.1(**UBIN**: top)). Human intervention can also cause scenes taken from similar places to appear very different - leaves being swept up as well as the addition/removal of man-made structures in the scene (Fig. 7.1(**UBIN**: bottom)) further makes the recognition of this database difficult. Using this database will verify the robustness of the proposed SRS against such changes in a simple and open coastal environment with relatively sparse vegetation. The skyline is also particularly evident in such an environment which is exploited to aid in scene recognition.

7.1.3 Database NS

The **NS** database consists of scenes with lush green vegetation taken at a primary swamp forest in a nature reserve. The test scenes are varied in structure, from enclosed forests to semi-open clearings such as streams and ponds (Fig. 7.2).

Figure 7.2: The **NS** database consists of three environments: Enclosed forest (left), streams and ponds (middle) and semi-open clearings (right).

There are three sets of test scenes. The first set is taken from the morning till

noon time on a clear day, the second set is taken three weeks later from the period between the late afternoon and the evening, also on a clear day while the third set is taken at around noontime on a hazy, cloudy day one week after the second set. As the first two sets are taken on clear days at very different times, changes in illumination caused by the movement of the sun are particularly evident. The effects of shadows and the non-uniform lighting in the environment due mainly to the foilage can be quite drastic and are particularly challenging (Fig. 7.1(**NS**: top)). Finally, because of the separation in time between the three sets of test scenes, changes due to the dynamic nature of the environment add to the difficulty in recognising the scenes (Fig. 7.1(**NS**: bottom)).

7.1.4 Database SBWR

In contrast to the 'openness' of the **UBIN** database, **SBWR** contains relatively complex scenes taken from an enclosed tropical mangrove forest. As the mangrove environment is dominated by a few plant species, this database contains many similar-looking vegetation, and is characterised by dense foliage and numerous occlusions (Fig. 7.1(**SBWR**)). The difficulty in recognition is compounded as the reference scenes are taken purposely at random points in the forest, with no distinct landmarks that could be used by human observers, unlike the other two databases of natural scenes.

The design of the reference database is also slightly different than the other three databases. Many of the reference scenes are represented by two or three *snapshots* of the same scene at the beginning, middle and end of a TBL arc. This is motivated by the increased complexity of the environment which requires for its representation several slightly displaced snapshots of the same scene as they indeed look remarkably different (Fig. 5.8)! Furthermore, several authors have hypothesised that the view at the *endpoints* of the TBL arcs are remembered by insects as they contain useful information for scene recognition (see [25]'s conclusion on the purpose of TBL flights and section 5.2.1). The reference database constructed in this case thus models this statement.

This database tests the proposed SRS's tolerance to such natural scenes with many occlusions and clutter, common in an enclosed forest.

7.2 Experimental procedure

The experimental procedure evaluates the performance of the proposed SRS by computing the recognition accuracy in terms of *positive acceptance*, P_{acc} and *positive rejection*, P_{rej} rates (in %) when positive and negative test scenes from the four databases are presented respectively to the proposed SRS. The entire procedure mimicks a typical scene recognition situation (section 6.2) - a reference database $\mathbb{D}_{ref}$ of N_{ref} reference scenes is constructued and scene recognition is performed

with N_{test} test scenes with the database. Obviously $N_{test} = N_{pos} + N_{neg}$. The entire evaluation procedure is summarised in the following steps:

1. The Scene matrix cells, $\mathbf{M}_s$ (section 5.3), of *all* of the images (both reference and test scenes) in the database are first extracted from the raw input images and saved.

2. The reference image database, $\mathbb{D}_{ref}$, is constructed. Ideally this step should be performed automatically by a separate algorithm that decides which scenes are *distinct* enough to be used as reference images. This can be achieved in a practical navigation system during the learning phase when the agent explores its environment for the first time. For this work, one assumes that this has been done and the N_{ref} reference images are chosen manually to produce $\mathbb{D}_{ref}$.

3. The rest of the $\mathbf{M}_s$ are then grouped into two test sets containing N_{pos} positive or N_{neg} negative scenes depending if $\mathbb{D}_{ref}$ contains a known positive match for the test scenes or not.

4. The two test sets containing N_{pos} and N_{neg} scenes are then used to obtain the positive acceptance and positive rejection accuracies (P_{acc}, P_{rej}) respectively. This is done by presenting a test scene matrix cell, $\mathbf{M}_s^{test}$ to the scene decision module as described in section 6.2 so as to obtain the final decision, D_f on the test scene. Different runs of the scene decision module for the *same* scene

may yield different results due to the fact that RANSAC is implemented in matching the salient-SURF keypoints between the test and reference scenes similar to the method used to extract ordinal depth by TBL (section 6.1.2). The accuracy of the proposed SRS for the same test scene may thus vary over several trials. In order to arrive at a reasonable estimation of the recognition accuracy, the scene decision with the same test scene is repeated for N_{iter} times. At each iteration, a *correct* decision is given one point while an *incorrect* decision is given zero point. For positive scenes, a correct decision occurs when the SRS correctly matches the reference scene in the database. For negative scenes, a correct decision effectively rejects the scene since no reliable matches can be found. This is done by comparing D_f with a known database of correct response the SRS should give if there are no errors. In this work, N_{iter} is fixed at 20 for all of the experiments.

The recognition accuracies (P_{acc}, P_{neg}) is given as the percentage of correct decision over all the (N_{pos}, N_{neg}) test scenes with each scene iterated over N_{iter} times:

$$P_i = \frac{\sum B}{N_j \times N_{iter}} \times 100, j \in [(acc, pos), (rej, neg)]$$
$$B = \begin{cases} 1 & \text{if } D_f \text{ is correct} \\ 0 & \text{if } D_f \text{ is incorrect} \end{cases} \tag{7.1}$$

The *overall* recognition accuracy, $P_{overall}$ is obtained by combining (P_{acc}, P_{rej}) together as a weighted average:

$$P_{overall} = \frac{P_{acc}N_{pos} + P_{rej}N_{neg}}{N_{test}} \tag{7.2}$$

The above procedure is repeated for the four databases described in section 7.1 so as to validate the SRS's robustness against various image distortions imposed by these databases. The results of the recognition accuracy of the proposed SRS is given in section 7.4.

7.3 Comparative studies with similarly designed SRSs

In order to have a better understanding of how the proposed SRS functions, a series of five comparative studies are conducted over several variants of the proposed SRS. The objective is to find out how the different components of the proposed SRS contribute to the recognition rates presented in section 7.4. This is done by enabling/disabling the two main components in the scene decision module (section 6.2) - the spatial configuration (x, y, z) and the HSV (hue, saturation and value (grayscale)) colour space. Throughout this chapter, the short-forms of the colour spaces are used - *hue* for hue, *sat* for saturation and *gs* for grayscale. The variants

of the proposed SRS (with their short-forms in **bold**) are described in the following paragraphs.

'Simple' SRS (**SimpSRS**): This variant uses *only* the percentage matches, $\mathbf{N}_{\%test}$ over the three colour spaces in the match statistic matrix $\mathbf{\Pi}_s$ (6.8) to determine if a match is found in the reference database, $\mathbb{D}_{ref}$. The ordinal configuration is entirely ignored. The aim is to show the inadequecies of using this naive method for scene recognition. As was explained in section 6.1.1, one of the problems faced by this method is the determination of a reliable threshold for accepting a candidate match as reliable or not. In this comparative study, two *fixed* percentage thresholds (@10%, @5%) are used for scene decision by comparing it with the percentage matches in the candidate match, $N_{\%cand}$. If $N_{\%cand}$ is greater than the threshold, $\mathbb{G}_{cand}$ is accepted. If this is not the case, the candidate match is then rejected.

Disable one spatial component (**DIS_1spatial**_i): One of the three ordinal measures of the spatial components $i \in \{x, y, z\}$ is disabled over all the three colour spaces. The aim of this comparative study is to observe how the recognition rates are affected when only two of the spatial components are preserved.

Enable one spatial component (**EN_1spatial**_i): Similar to the previous variant, only one out of the three ordinal measures of the spatial components $i \in \{x, y, z\}$ is enabled over the three colour spaces.

Disable one colour component (**DIS_1col_**j): In this variant, the three ordinal measures are preserved while one of the colour space $j \in \{hue, sat, gs\}$ is disabled - only two colour spaces are used in this comparative study.

Enable one colour component (**EN_1col_**j): This final variant uses only one colour component in $j \in \{hue, sat, gs\}$ while preserving the full spatial configuration of the matches found.

In all the five comparative tests, the three recognition accuracies: P_{acc}, P_{rej} and $P_{overall}$ (7.1,7.2) are computed using the procedure described in section 7.2. All of the results are compared to the proposed SRS that serves as the baseline. The next section presents the results of all the comparative studies as well as the recognition accuracies of the proposed SRS for the four databases.

7.4 Experimental results

The recognition accuracy in terms of P_{acc}, P_{rej} and $P_{overall}$ defined in (7.1, 7.2) of the proposed SRS and the five variants in the comparative studies are summarised in tabular form for easy comparison over the four databases. All the tables containing the results are listed as follows: **Proposed SRS**: Table 7.2; **SimpSRS**: Table 7.3; **DIS_1spatial**: Tables (7.4,7.5,7.6); **EN_1spatial**: Tables (7.7,7.8,7.9); **DIS_1col**: Tables (7.10,7.11,7.12); **EN_1col**: Tables (7.13,7.14,7.15).

The tables referenced are presented in the next few pages.

Table 7.2: Proposed SRS

Database	P_{acc}	P_{rej}	$P_{overall}$
IND	96.6	94.52	95.65
UBIN	94.84	100	97.54
NS	100	99.42	99.67
SBWR	99.33	100	99.68

Table 7.3: SimpSRS@10% and 5% threshold

	P_{acc}		P_{rej}		$P_{overall}$	
Database	10%	5%	10%	5%	10%	5%
IND	87.2	93.8	80.95	38.1	84.35	68.37
UBIN	40.71	71.27	94.2	43.48	68.67	56.74
NS	46.46	66.59	96.15	73.08	74.25	70.22
SBWR	42	80	100	87.5	71.94	83.87

Table 7.4: DIS_1spatial_x

Database	P_{acc}	P_{rej}	$P_{overall}$
IND	86.8	90.24	88.37
UBIN	88.57	76.81	82.42
NS	92.44	86.63	89.19
SBWR	91.67	100	95.97

Table 7.5: DIS_1spatial_y

Database	P_{acc}	P_{rej}	$P_{overall}$
IND	83.6	86.19	84.78
UBIN	79.84	72.46	75.98
NS	80.12	90.38	85.86
SBWR	80	93.75	87.1

Table 7.6: DIS_1spatial_z

Database	P_{acc}	P_{rej}	$P_{overall}$
IND	85.8	90	87.72
UBIN	88.41	76.81	82.35
NS	92.44	86.54	89.14
SBWR	90	100	95.16

Table 7.7: EN_1spatial_x

Database	P_{acc}	P_{rej}	$P_{overall}$
IND	83.6	86.19	84.78
UBIN	80.24	72.46	76.17
NS	80.24	90.38	85.91
SBWR	80	93.75	87.1

Table 7.8: EN_1spatial_y

Database	P_{acc}	P_{rej}	$P_{overall}$
IND	95.8	73.1	85.43
UBIN	84.92	66.67	75.38
NS	90.24	88.17	89.09
SBWR	99	100	99.52

Table 7.9: EN_1spatial_z

Database	P_{acc}	P_{rej}	$P_{overall}$
IND	83.8	87.14	85.33
UBIN	79.92	72.46	76.02
NS	80.12	90.38	85.86
SBWR	80	93.75	87.1

Table 7.10: DIS_1col_*gs*

Database	P_{acc}	P_{rej}	$P_{overall}$
IND	75.4	91.43	82.72
UBIN	46.83	69.57	58.71
NS	70.73	86.54	79.57
SBWR	73.33	75	74.19

Table 7.11: DIS_1col_*sat*

Database	P_{acc}	P_{rej}	$P_{overall}$
IND	85.4	90.48	87.72
UBIN	83.89	75.36	79.43
NS	87.8	88.46	88.17
SBWR	93.33	100	96.77

Table 7.12: DIS_1col_*hue*

Database	P_{acc}	P_{rej}	$P_{overall}$
IND	96.2	81.19	89.35
UBIN	81.75	84.06	82.95
NS	80.49	94.23	88.17
SBWR	86.67	93.75	90.32

Table 7.13: EN_1col_*gs*

Database	P_{acc}	P_{rej}	$P_{overall}$
IND	84	80.95	82.61
UBIN	79.37	71.01	75
NS	78.05	90.38	84.95
SBWR	86.67	93.75	90.32

Table 7.14: EN_1col_*sat*

Database	P_{acc}	P_{rej}	$P_{overall}$
IND	76	85.71	80.43
UBIN	38.1	57.97	48.48
NS	65.85	75	70.97
SBWR	60	62.5	61.29

Table 7.15: EN_1col_*hue*

Database	P_{acc}	P_{rej}	$P_{overall}$
IND	44	80.95	60.87
UBIN	28.57	44.93	37.12
NS	48.78	53.85	51.61
SBWR	46.67	43.75	45.16

A few illustrative examples of the resulting matches are highlighted in the sections that follow. They aim to demonstrate the robustness and generality of the proposed SRS against various image distortions in both indoor and outdoor natural environments. The figures show the matched reference scene on the left and the test scene on the right. The correspondences are shown as cyan lines connecting the matched keypoints across the two images.

7.4.1 Database IND results

The **IND** scenes contain a few artificial scenes that test the usefulness of ordinal measures in encoding the landmark configuration for robust scene recognition. Fig. 7.3 (top) shows a positive test scene where a viewpoint change to the right by $\sim 20^\circ$ is correctly recognised. This does not always work when the landmarks are not *correctly* matched, as is shown in Fig. 7.3 (bottom). The negative scene high-

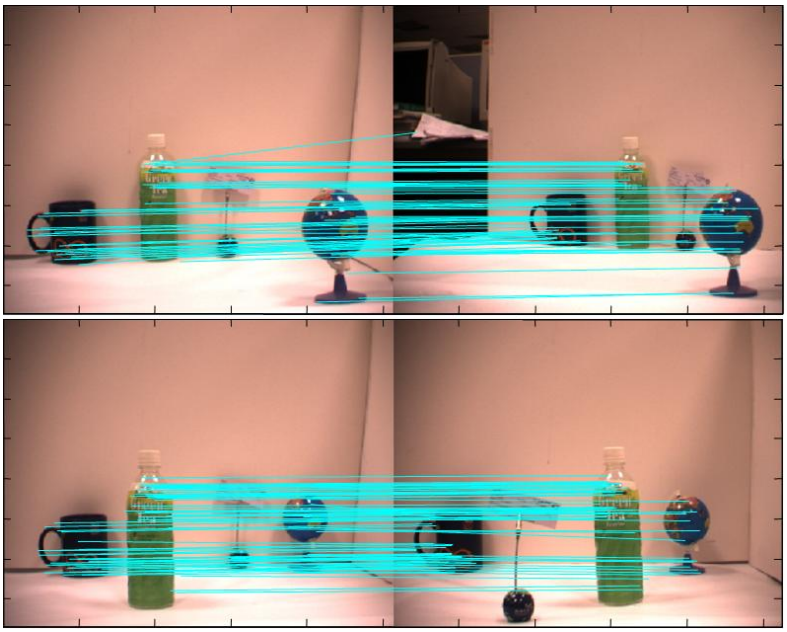

Figure 7.3: Correctly matched **IND** positive scene (top). Incorrectly matched **IND** negative scene (bottom).

lights clearly the limitations of using ordinal measures. If a landmark is improperly matched or not matched at all, the robustness that ordinal measures *should* give becomes an impediment to effective discrimination of ambiguous scenes. The unmatched landmark shown in the example is unfortunately the *crucial* landmark

that has changed in its x and z position. Because of its relatively small size, the number of SURF keypoints it contains is equally small in number and this makes the change statistically *insignificant*. Although this is unlikely to happen in real scenes where numerous features of various sizes exist, it highlights one of the many possible future improvements for the proposed SRS (section 8.3).

Another set of test scenes with clutter and people working in a typical office highlights the tolerance of the proposed SRS to such dynamic changes in the scene (Fig. 7.4).

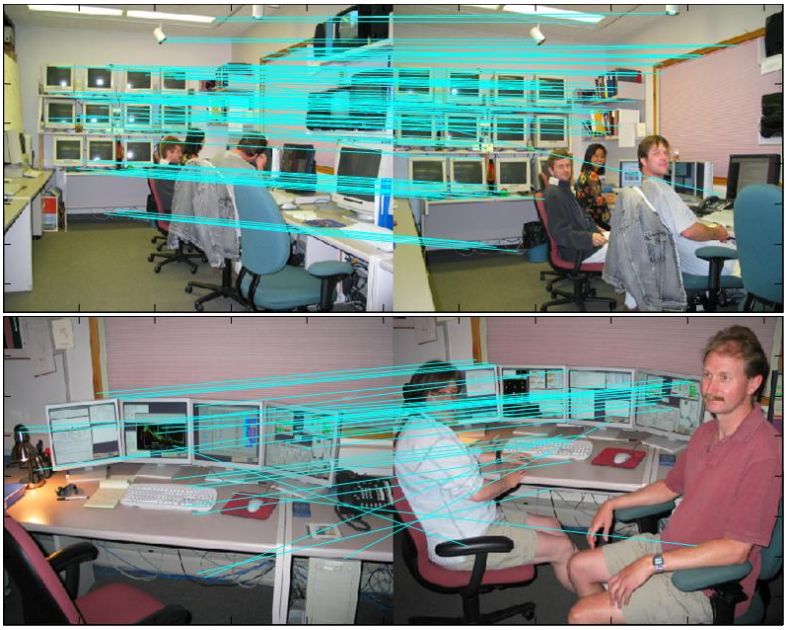

Figure 7.4: Tolerance to clutter and people in the **IND** database: small changes (top), large changes (bottom).

This is due to the accuracy of the salient-SURF keypoints in finding reliable matches between the test and reference images in the database and confirms the

formulation of G_c (section 6.1.2) and adaptive decision threshold, D_t (section 6.2.2) in detecting scene equivalence for difficult scenes with significant dynamic content. For example, in Fig. 7.4 (bottom), the test scene (left) contains significant occlusion caused by people sitting in front of the table which would have made scene recognition particular difficult. However, the encoded salient-SURF keypoints are matched with high accuracy on regions that are not occluded. This preserves the spatial configuration of the test scene with the true reference scene in the database, which enables the scene decision module to ignore these dynamic changes for reliable scene recognition.

7.4.2 Database UBIN results

Varying weather conditions as well as natural and man-made interventions contribute to the abundant dynamic changes in this database. As was highlighted in section 7.1.2, the main challenge of this database is the significant image distortion due to the long time difference between certain test scenes and the reference scenes. The proposed SRS is however tolerant to an extent to these changes and produces excellent results as illustrated in Fig. 7.5. The main reason for the proposed SRS's tolerance is that the skyline is exploited to provide invariant and stable features for reliable recognition even under such extreme image distortions. This robustness from using the skyline may however lead to a loss of discriminatory power of

Figure 7.5: Challenging **UBIN** test scenes with significant image distortion - viewpoint, weather and illumination changes.

the proposed SRS when different scenes may share the *same* skyline (sections 4.3 and 3.5.1). This is especially common in open places such as the coastal areas in the **UBIN** database. An example of a mismatched scene[1] with similar skyline is shown in Fig. 7.6 (which is the same as in Fig. 3.19). This is clearly an important limitation of using the skyline as a composite feature for scene recognition. This particular test scene is difficult as the changes in the foreground are too significant for reliable matching and only the skyline is reliable. Since there are a few reference scenes with similar skylines (as they were taken along the same coastal stretch), the best match using the skyline information is incorrect for this particular test

[1]the mismatched scene is around 50m from the correct reference scene

scene.

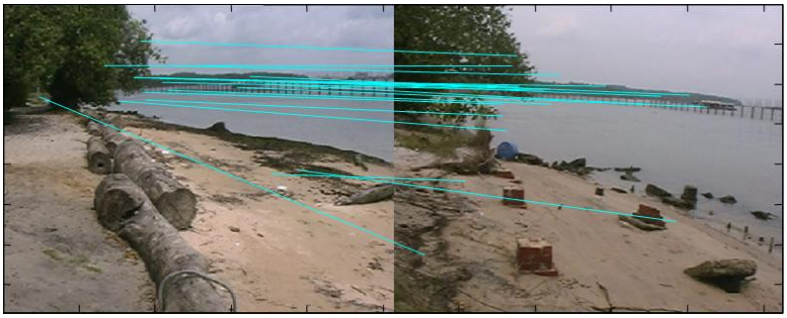

Figure 7.6: A mismatched **UBIN** scene with similar skyline.

7.4.3 Database NS results

The **NS** scenes are characterised by a mix of enclosed and semi-enclosed forest scenes with significant weather and illumination changes (section 7.1.3). This is due to the fact that the three sets of test scenes are collected at different times of the day, on three different days under different weather conditions. Nonetheless, the proposed SRS achieves a very impressive $P_{overall} = 99.67\%$ which is almost the same as the **SBWR** database. Fig. 7.7 shows three positive test scenes that are recognised despite significant non-uniform illumination, viewpoint changes and image distortions due to natural erosion. Some of the changes are so significant that even this author was initially unaware that they were taken at similar locations! The strength of the proposed SRS is clearly evident - even scenes that may confuse human observers could be reliably detected by the algorithm, making it a useful

tool for reliable and robust navigation in complex outdoor environments performed by future autonomous artificial agents.

Figure 7.7: Difficult **NS** test scenes that are correctly recognised.

7.4.4 Database SBWR results

The **SBWR** database contains enclosed mangrove forest scenes with an augmented reference database as described in section 7.1.4. This reference database models

after the TBL arc by having several reference snapshots of the same scene. Using this augmented database greatly *improves* the recognition accuracy of the proposed SRS despite the fact that the skyline is unusable. From Table 7.2, one can see that the **SBWR** test scenes reported the highest $P_{overall} = 99.68\%$ compared to the other databases. Since the proposed SRS tends to match the closest reference snapshot, an additional information about the approximate position of where the test scene is located on the TBL arc can be obtained. Fig. 7.8 illustrates two input test scenes that are correctly recognised in spite of the high complexity of the scenes with numerous occlusions.

Figure 7.8: Matched **SBWR** test scenes with significant occlusions and viewpoint distortions.

The few examples given in this section serve only to provide an illustrative idea

of the potential of the proposed SRS in recognising difficult scenes under various image distortions in both indoor and outdoor natural environments. A better understanding of how the components of the proposed SRS contribute to correct recognition is discussed next. This is done by analysing the recognition accuracy of the proposed SRS and its variants presented in this section.

7.5 Analysis and discussion of experimental results

A detailed analysis of the experimental results is presented in this section. The analysis is separated into several subsections that focus on the contribution of a particular component to the recognition accuracies,$(P_{acc}, P_{rej}, P_{overall})$, of the proposed SRS. Section 7.5.1 analyses the contribution of ordinal measures in general by comparing the **Proposed SRS** with **SimpSRS**. The contributions of the ordinal measures in the three spatial directions (x, y, z), denoted as (x_{om}, y_{om}, z_{om}), are analysed in sections 7.5.2–7.5.4. An evaluation of the relative importance of the three ordinal measures follows in section 7.5.5. The contributions of the three colour (gs, sat, hue) components, denoted as (gs_c, sat_c, hue_c), are analysed in sections 7.5.6–7.5.8. Section 7.5.9 then evaluates the relative importance of the three colour spaces. Finally, section 7.5.10 concludes this section based on the results of

the analysis.

7.5.1 Proposed SRS *vs.*SimpSRS

The results presented in Tables 7.2 and 7.3 are compared. The superiority of the **Proposed SRS** over **SimpSRS** which uses only the percentage matches as a decision criterion is evident. The proposed SRS achieves generally *better* recognition accuracy, $(P_{acc}, P_{rej}, P_{overall})$ in all four databases. Table 7.3 also highlights an important shortcoming of using a fixed rejection threshold - different rejection thresholds affect the accuracy of the databases differently and there is no simple method to determine in advance a good threshold that will achieve a high P_{acc} without compromising P_{rej}. As one can observe, decreasing the threshold from @10% to @5% improves P_{acc} but degrades P_{rej} which is a clear illustration of the conflicting requirements posed by the robustness (measured by P_{acc}) and the discriminatory power (measured by P_{rej}) of any SRS.

7.5.2 Contribution of x_{om}

The results in Tables 7.5 and 7.9 are compared where the only parameter that is changed is the addition of x_{om}. In terms of P_{acc}, the four databases report virtually no significant change in positive recognition accuracy when x_{om} is used. This is not a surprising result for positive scenes that preserves x_{om} in general. Similarly

for P_{rej}, the difference between the results when x_{om} is added is extremely small. Hence the addition of x_{om} plays only a small role in improving the discriminatory power of the SRS which is already high when z_{om} is used alone. This is an indication that other spatial/colour components contain more *distinctive* features for better discrimination.

7.5.3 Contribution of y_{om}

Tables 7.4 and 7.9 are compared as the only parameter that is changed is y_{om}. In terms of P_{acc}, the addition of y_{om} generally improves the positive recognition accuracy from 3%(**IND**) to 12%(**NS**). This is an indication that a small amount of invariance is encoded in y_{om} that aids in positive recognition. In terms of P_{rej}, three databases (**UBIN**, **SBWR** and **IND**) reported small improvements of varying degrees from 3.1%(**IND**) to 6.25%(**SBWR**) while there is a slight degradation of in P_{reg} for **NS** scenes (-3.75%). This shows that with the exception for the **NS** database, y_{om} does encode a small amount of distintiveness that allows for effective discrimination. The slight degradation observed in the **NS** scenes, however, indicates that the distintiveness encoded in y_{om} for this database is not very significant.

7.5.4 Contribution of z_{om}

Tables 7.5 and 7.7 are compared separately from Tables 7.4 and 7.8 where the only parameter that changes is the addition of z_{om}. Databases **NS** and **UBIN** reported similar results where the addition of z_{om} preserves P_{acc}. This is again due to the fact that ordinal measures are preserved in general for positive scenes. The other two databases, **SBWR** and **IND**, however reported a degradation in P_{acc} when z_{om} is added to y_{om} (Table 7.4): -7.33%(**SBWR**) and -9%(**IND**). Detailed analysis of the results attributes the degradation in database **SBWR** to its greater complexity with numerous occlusions that severely affected the estimation of Z_{ord} (section 5.2.2). For database **IND**, the degradation is due entirely to the ambiguous test scenes (Fig. 7.3 (top)) with numerous similar features that caused many mismatches to occur. The addition of z_{om} probably *amplified* the effect of such errors that is reflected in $\mathbf{\Pi}_s$. This in turn affects the reliablity of the decision threshold, D_t. $\mathbb{G}_{cand}$ is likely to be seen as highly unreliable and is falsely rejected by D_t, lowering P_{acc}.

In terms of P_{rej}, the **IND** and **UBIN** databases display a significant improvement in discriminatory power with the addition of z_{om} to y_{om}: 17.14%(**IND**) and 10.16%(**UBIN**) while it remains constant and high when z_{om} is added to x_{om} in both databases. This indicates that for these two databases, z_{om} encodes a significant amount of distinctiveness among the test scenes for effective discrimination,

which is more than y_{om} and is of equal importance to x_{om}. For the remaining two cases (**NS**, **SBWR**), the addition of z_{om} does not contribute to any significant changes in P_{rej}. This indicates that z_{om} does not contain enough distinct features for discrimination in the two databases. Comparing the results of the two previous sections reveals that for the **NS** database, P_{rej} remains largely unaffected by the addition of ordinal measures on the whole. This indicates that all three spatial ordinal measures, (x_{om}, y_{om}, z_{om}) are *equally* significant in determining the P_{rej} of the proposed SRS. This initial observation is confirmed in section 7.5.5.

7.5.5 Relative importance of (x_{om}, y_{om}, z_{om})

The relative importance of (x_{om}, y_{om}, z_{om}) are determined in terms of P_{acc} and P_{rej} separately in this section. The *relative rankings* of the components are then denoted as $(\cdots)$ for P_{acc} and $[\cdots]$ for P_{rej}. A slight abuse in notations is used in this section: the symbols $(>, =)$ are used to denote a greater and equal contribution respectively of a particular component to the particular recognition accuracy (P_{acc}, P_{rej}) concerned. The optional symbol $\sim$ denotes a slight difference in contribution of the components. Three comparisons are made between (x_{om}, y_{om}), (z_{om}, x_{om}) and (z_{om}, y_{om}). From the results of the pairwise comparisons, one can conclude on the overall relative importance of the three ordinal measures.

Between x_{om} and y_{om}

Tables 7.6, 7.7 and 7.8 are compared. In terms of P_{acc}, database **NS** and **UBIN** display a common trend where y_{om} is slightly more important than x_{om} with a generally higher, and hence *complementary*, combined P_{acc} when both ordinal measures are used together (Table 7.6). ($y_{om} \sim> x_{om}$)

For databases **SBWR** and **IND**, y_{om} is more significantly important than x_{om} but the combined P_{acc} actually degrades (but remains high) when both ordinal measures are used together: 90%(**SBWR**, **IND**). A closer analysis of the two databases shows that most of the viewpoint changes occur in the x spatial direction, resulting in more mismatches when x_{om} is used with y_{om}. ($y_{om} > x_{om}$)

In terms of P_{rej}, x_{om} is more dominant than y_{om} by various amounts for the three databases: 2.21%(**NS**), 5.79%(**UBIN**), 13.09%(**IND**) which is also complementary when both ordinal measures are used together. This follows the general observation that because most of the viewpoint distortion occurs in the x direction, x_{om} thus encodes more distinctive information that aids in discrimination. [$x_{om} > y_{om}$]

For the **SBWR** database, the situation is interestingly *reversed* - y_{om} is more dominant than x_{om} by 6.25% but the combined P_{rej} remains constant @100%. This shows that the two ordinal measures are complementary. The reason for the dominence of y_{om} over x_{om} is attributed once again to the complexity of the

SBWR test scenes - the numerous mismatches in the x spatial direction degrade P_{rej}. $[y_{om} > x_{om}]$

Between z_{om} and x_{om}

Tables 7.5, 7.7 and 7.9 are compared. In terms of P_{acc}, all the four databases report virtually *equal* importance between z_{om} and x_{om} with no significant variations when both ordinal measures are used. This confirms that an equal amount of invariance is encoded in these two ordinal measures for positive recognition. $(z_{om} = x_{om})$

Similarly for P_{rej}, all the databases display almost *equal* dominance between z_{om} and x_{om} with no significant variation when both measures are used. This is not a surprising result as both the x and z spatial directions capture significant distinctiveness that aids in discrimination. $[z_{om} = x_{om}]$

Between z_{om} and y_{om}

Tables 7.4, 7.8 and 7.9 are compared. In terms of P_{acc}, all the four databases report similar results with y_{om} being *more* significant than z_{om}. The variation in the combined accuracy when both ordinal measures are used is however small. $(y_{om} > z_{om})$

In terms of P_{rej}, databases **NS**, **UBIN** and **IND** display a dominance of z_{om} over y_{om}: 2.21%(**NS**), 5.79%(**UBIN**) and 14.04%(**IND**). This is once again due to the fact that most of the distinctiveness of the test scenes occur in the z spatial

direction while the y spatial direction remains largely invarient. The combined P_{rej} for the three databases are in general complementary. $[z_{om} > y_{om}]$

Database **SBWR** interestingly reports a *reversed* trend, with y_{om} being more dominant (@100%) than the z_{om} (@93.75%) with a combined accuracy that is maintained @100%. This can be explained by the errors in estimating Z_{ord} in the complex **SBWR** test scenes that reduces P_{rej} slightly. $[y_{om} > z_{om}]$

Relative importance

From the results of the comparisons in terms of P_{acc}, the relative importance is given as:

$$P_{acc} \Rightarrow \begin{cases} (y_{om} \sim> x_{om} = z_{om}) & \text{for } \mathbf{NS} \text{ and } \mathbf{UBIN} \\ (y_{om} > x_{om} = z_{om}) & \text{for } \mathbf{IND} \text{ and } \mathbf{SBWR} \end{cases} \tag{7.3}$$

For **NS** and **UBIN**, the difference between all three ordinal measures is *very* small. This confirms the fact that positive scenes *preserve* the ordinal configuration such that the use of just one ordinal measure is usually sufficient. For **IND** and **SBWR** scenes, however, the distortions in x_{om} and z_{om} emphasise the importance of y_{om} in preserving the invariance needed for positive recognition.

In terms of P_{rej}, the relative importance is given as:

$$P_{rej} \Rightarrow \begin{cases} [x_{om} = z_{om} > y_{om}] & \text{for } \mathbf{NS}, \mathbf{UBIN} \text{ and } \mathbf{IND} \\ [y_{om} > x_{om} = z_{om}] & \text{for } \mathbf{SBWR} \end{cases} \tag{7.4}$$

This is an *important* result that justifies the use of Z_{ord} in the proposed SRS. From the three databases (**NS**, **UBIN** and **IND**), z_{om} is equally dominant as x_{om} and is more important than y_{om} in encoding the distinctiveness of the scenes for effective discrimination. For the **NS** database, the difference in dominance is very small and this confirms the initial suggestion in section 7.5.4 that the distinctiveness of the test scenes is captured equally over the three ordinal measures. This is attributed to the variability of the scene structure in **NS** (section 7.1.3). For **UBIN** and **IND**, the main reason for the dominance of z_{om} and x_{om} over y_{om} is due to the fact that more distinctiveness is encoded in these two directions that aids in discrimination. Overall, the test scenes tend to maintain a general 'orderliness' in the y direction for the same environment. For example, in **UBIN**, the structure of sky (top), vegetation (middle) and ground (bottom) is maintained throughout and this tends to lower the usefulness of y_{om} in discrimination.

The interesting results come from **SBWR** that has a greater *emphasis* on y_{om} then x_{om} and z_{om} which are equally dominant. This is attributed to the complex scene structure of **SBWR**. The degradation in P_{rej} when x_{om} and z_{om}

are used makes y_{om} relatively more important. This highlights a *crucial* factor that determines the discriminating power of the proposed SRS - only the most stable and reliable ordinal measures are useful (y_{om} in this database) for discrimination.

7.5.6 Contribution of gs_c

Tables (7.12, 7.14) and Tables (7.11, 7.15) are compared. In terms of P_{acc}, the inclusion of gs_c generally improves P_{acc}: by 14.64%(**NS**) to 43.65%(**UBIN**)($sat_c + gs_c$) and by 39.1%(**NS**) to 55.32%(**UBIN**)($hue_c + gs_c$), with the + sign denoting the use of two colour components. This is an indication that the invariance encoded by the keypoints in gs_c is greater than the other two colour components. This is due to the fact that there are *more* keypoints detected in gs for reliable positive recognition.

In terms of P_{rej}, the three outdoor databases: **NS**, **UBIN** and **SBWR** display a general trend where the addition of gs_c improves P_{rej}: by 19.23%(**NS**) to 31.25(**SBWR**)($sat_c + gs_c$) and by 34.61%(**NS**) to 56.25%(**SBWR**)($hue_c + gs_c$). Once again, this improvement can be attributed to the fact that the keypoints encoded in gs are more distinctive and abundent for better discrimination. Furthermore the *monochromatic* nature of the outdoor databases, that contains significant natural vegetation, reduces the uniqueness of hue_c and sat_c keypoints for discrimination (section 3.4.2).

For the **IND** scenes, there are two surprising observations. The first observation is that hue_c and sat_c encode a surprisingly *significant* amount of information. The P_{rej} of 80.95% when hue_c is used alone and 85.71% when sat_c is used alone are surprisingly good (Tables 7.14, 7.15). This highlights the fact that for indoor scenes with many man-made features of various colours, hue_c and sat_c are clearly more discriminatory than the outdoor databases. The second surprising observation is in the addition of gs_c that enhances P_{rej} of hue_c to 90.48% while combining with sat_c causes a slight degradation to 81.19%. It is possible that gs_c is less reliable for indoor scenes compared to sat_c, due to a large number of ambiguous features which becomes virtually *indistinguishable* in grayscale, a case that was stated in the design of 'CSIFT' that uses colour information for better discrimination [1]. This ambiguity is less pronounced when used in conjunction with hue_c, since the hue_c maps are usually more different (and hence complementary) from the gs_c maps as compared to the sat_c maps (Fig. 7.9).

7.5.7 Contribution of sat_c

Tables(7.12, 7.13) and Tables(7.10, 7.15) are compared. In terms of P_{acc}, the addition of sat_c improves P_{acc} in general: by 0%(**SBWR**) to 12.2%(**IND**)($gs_c + sat_c$) and by 18.26%(**UBIN**) to 31.4%(**IND**)($hue_c + sat_c$). It is interesting to see that the improvement is the largest for the **IND**. This is an indication that for

Figure 7.9: Two **IND** scenes with their HSV components. Comparing the grayscale images *vs.*the hue images reveals the strong *dissimilarity* between the two colour spaces. On the other hand, saturation images bear a stronger resemblance to grayscale.

indoor scenes, sat_c encodes more useful information for positive recognition than in outdoor scenes.

The same trend is seen in the case of P_{rej}. sat_c provides more distinct keypoints for better discrimination when added to hue_c and complements the use of gs_c in all the four databases: by 10.48%(**IND**) to 32.7%(**NS**)($hue_c + sat_c$) and 0%(**SBWR**) to 13.05%(**UBIN**)($hue_c + gs_c$). This highlights the importance of using sat_c to enhance the discriminatory power of the proposed SRS.

7.5.8 Contribution of hue_c

Tables(7.10, 7.14) and Tables(7.11, 7.13) are compared. In terms of P_{acc}, the addition of hue_c in general improves slightly the combined P_{acc} in all the four databases. The small improvments in P_{acc} with the addition of hue_c, however, is an indication that it only plays a minor but complementary role in positive recognition.

In terms of P_{rej}, the addition of hue_c to the other colour components improves in general P_{rej} over the three outdoor databases. The small improvement with sat_c and slight degradation/improvement with gs_c is a hint that the distinctiveness of the features encoded in hue_c is rather small. This is true as there are usually less hue_c keypoints detected than in the other colour spaces. This reduces their contribution to the discriminatory power of the proposed SRS.

The *interesting* result comes from **IND** where the improvement for $(hue_c + gs_c)$ is the *largest* among the four databases (9.35%). This is attributed to the indoor nature of the scenes where hue_c provides more discriminating information than gs_c.

7.5.9 Relative importance of (gs_c, sat_c, hue_c)

Similar to the evaluation of the relative importance of the three spatial components in section 7.5.5, the relative importance of the three colour components

(gs_c, sat_c, hue_c) in determining P_{acc} and P_{rej} are evaluated by making three pairwise comparisons between (gs_c, sat_c), (gs_c, hue_c) and (sat_c, hue_c). The results of these comparisons are similarly enclosed in $(\cdots)$ for P_{acc} and $[\cdots]$ for P_{rej} with the symbols $(>, =, \sim)$ retaining the same meanings as defined in section 7.5.5. The overall relative importance can then be inferred.

Between gs_c and sat_c

Tables 7.12, 7.13 and 7.14 are compared. In terms of P_{acc}, gs_c is more significant and complementary with sat_c in all the four databases.$(gs_c > sat_c)$

In terms of P_{rej}, the three outdoor databases (**NS**, **UBIN** and **SBWR**) report the same dominance for gs_c over sat_c which are also complementary. This shows that gs_c contains more distinctive features for reliable discrimination.$[gs > sat]$

The most *interesting* results come from **IND**, with sat_c being slightly more important than gs_c by 4.76% and a combined P_{rej} that has a slight degradation (-4.52%). This is an indication that for indoor scenes, sat_c is more discriminatory than gs_c. The small degradation in the combined P_{rej} may be due to the unreliablity of the gs_c keypoints in **IND** (section 7.5.6). $[sat_c \sim> gs_c]$

Between gs_c and hue_c

Tables 7.11, 7.13 and 7.15 are compared. In terms of P_{acc}, gs_c plays a more significant and complementary role with hue_c over the four databases. The reason for

this dominance can be explained by the monochromatic (green) nature of the outdoor databases. Although hue_c is somewhat invariant to lighting and illumination changes, the keypoints are *unreliable* due to their small number and this in turn degrades P_{acc}. The smaller number of hue_c keypoints even in **IND** affects similarly its usefulness for positive recognition. $(gs_c > hue_c)$

In terms of P_{rej}, the three outdoor databases display the same general trend - gs_c is clearly more dominant and complementary with hue_c for discrimination. The weakness of the hue_c keypoints in discrimination is due to the monochromatic environment that results in many mismatches, hence degrading its discriminatory power. $[gs_c > hue_c]$

Once again, **IND** displays a different trend from the outdoor databases, with hue_c being of *equal* importance as gs_c with a high P_{rej} of 80.95% and a combined P_{rej} that is complementary @90.48%. As was explained in section 7.5.6, the hue_c maps are more different than gs_c maps in an indoor environment compared to sat_c maps. The test scenes are thus more distinctive in $hue_c + gs_c$. The equidominance of the two colour spaces clearly shows the usefulness of hue_c for discrimination in indoor scenes. $[hue_c = gs_c]$

Between sat_c and hue_c

Tables 7.10, 7.14 and 7.15 are compared. In terms of P_{acc}, sat_c is clearly more significant than hue_c with a generally improved P_{acc} when the two components

are used together. This shows that sat_c, when compared to hue_c, encodes more invariant information for positive recognition. ($sat_c > hue_c$)

In terms of P_{rej}, all the four databases display the same general trend with sat_c being more dominant and complementary with hue_c. This shows that sat_c does encode in general a greater amount of distinctiveness than hue_c. The interesting observation is in **IND** as sat_c is only *slightly* more important than hue_c by 4.76% while the other outdoor databases reported significantly larger improvements: 21.15%(**NS**), 13.04%(**UBIN**) and 18.75%(**SBWR**). The combined P_{rej} is also interesting for **IND** that reported the *highest* P_{rej} of 91.43%. These two observations indicate that the distinctiveness encoded by hue_c and sat_c for **IND** scenes is clearly more significant compared to other outdoor databases which follows the results in the previous subsections. [$sat_c > hue_c$] for (**NS**, **UBIN** and **SBWR**) and [$sat_c \sim> hue_c$] for **IND**.

Relative importance

The relative importance of the three colour components in terms of P_{acc} is given as:

$$P_{acc} \Rightarrow (gs_c > sat_c > hue_c) \tag{7.5}$$

This result confirms that the use of gs_c is often sufficient for reliable positive recognition in the literature (*e.g.* [68, 98]). What is interesting from the above analysis is that the addition of the other two colour components are in general *complementary* as they improve the combined P_{acc} when used with gs_c. This shows that when used appropriately, hue_c and sat_c do contribute to P_{rej} for outdoor scenes with illumination changes. Furthermore, the additional information contributed by hue_c and sat_c also *improves* P_{acc} for **IND** scenes. This contribution is, however, smaller than gs_c because less keypoints are detected, especially in monochromatic outdoor scenes. With less keypoints, the reliability in determining a positive match is called into question as a single mismatch can degrade G_c, making positive recognition difficult.

In terms of P_{rej}, the relative importance of the three colour components is given as:

$$P_{rej} \Rightarrow \begin{cases} [gs_c > sat_c > hue_c] & \text{for } \mathbf{NS}, \mathbf{UBIN} \text{ and } \mathbf{SBWR} \\ [sat_c \sim> gs_c = hue_c] & \text{for } \mathbf{IND} \end{cases} \tag{7.6}$$

For the three outdoor databases (**NS**, **UBIN** and **SBWR**), the usefulness of gs_c in encoding the distinctiveness of the scenes for discrimination is evident. Similar to the P_{acc} case, the addition of hue_c and sat_c information in general further

improves P_{rej} and are thus *complementary*. The monochromatic nature of the outdoor scenes, however, contains a lot of invariant information captured in hue_c (and to a lesser extent in sat_c) which makes hue_c and sat_c less reliable for discrimination. The reliability is further reduced as there are less of hue_c and sat_c keypoints detected than the gs_c keypoints.

The most important and *surprising* result comes from **IND** that justifies the use of hue_c and sat_c for improved discrimination. From (7.6), one can see that the distinctiveness of **IND** scenes are more *evenly* spread apart in the different colour components. sat_c is the *most* important colour component while hue_c is *equally* important as gs_c in discriminating indoor scenes. This result clearly deviates from the conclusion drawn from the outdoor databases. For indoor scenes containing numerous man-made structures, the use of sat_c and hue_c even equals or surpasses the discriminatory role that gs_c originally possesses.

7.5.10 Conclusion and discussion of the analysis

For the spatial ordinal measures (x_{om}, y_{om}, z_{om}), the addition of just one ordinal measure often improves the recognition accuracies (P_{acc}, P_{rej}) compared to **Simp-SRS** (Table 7.3). This highlights the importance of ordinal measures in improving the performance of the SRS. In particular, the *importance* of z_{om}, ignored in the majority of literatures, is clearly shown for effective scene recognition. In terms of

P_{acc}, z_{om} is equally important as x_{om} (7.3) while in the case of P_{rej}, z_{om} is even more significant than y_{om} for **NS**, **UBIN** and **IND** databases (7.4). The only exception is in **SBWR** where the complexity of the scene structure reduces the improvement rendered by x_{om} and z_{om}.

For the colour components, (gs_c, sat_c, hue_c), the analysis confirms the usefulness of gs_c for effective recognition. The main reason for gs_c's dominance is the fact that a larger number of gs_c keypoints is detected and this increases the reliability of gs_c for positive recognition as well as effective discrimination of negative scenes. The use of the other two colour components is also shown to be *complementary* in improving P_{acc} and P_{rej}. Most importantly, the contribution of hue_c and sat_c in improving the P_{rej} of **IND** scenes by encoding more distinctive information is highlighted (7.6). Comparing **Proposed SRS** *vs.***EN_1col**$\_gs$ (Tables 7.2 and 7.13) shows that using *all three* colour components produce significant improvements in the recognition accuracies $(P_{acc}, P_{rej}, P_{overall})$ compared to using gs_c alone. The results in terms of $P_{overall}$ are summarised below, denoted as ($P_{overall}$ of **EN_1col**$\_gs$ $\rightarrow$ (to) the $P_{overall}$ of **Proposed SRS**) +(percentage change):

- **IND**(82.61%→95.65%) +13.04%
- **UBIN**(75%→97.54%) +22.54%
- **SBWR**(90.32%→99.68%) +9.36%
- **NS**(84.95%→99.67%) +14.72%

From the above results, the use of more colour components thus comes at only a

slight increase in computational time and storage (due to the extraction, matching and storage of extra salient-SURF keypoints) and is justified by the significant improvement (from 9.36% to 22.54%) in $P_{overall}$.

In conclusion, for a particular component to contribute effectively to *positive acceptance* of matching scenes (measured by P_{acc}) and *positive rejection* (measured by P_{rej}), it must have the following characteristics:

Invariance. The greater the invariance encoded by a particular component, spatial or colour, the larger is the contribution of that particular component in determining the P_{acc} for two positive scenes. An invariant component that is robust and tolerant to all forms of image distortions (viewpoint, illumination as well as natural erosion) is highly *valuable* for positive recognition under various viewing conditions.

Distinctiveness. In contrast with invariance, the greater the amount of distinctiveness encoded by a particular component, the larger is its contribution in determining P_{rej} between two very similar (but different) scenes for discrimination. Such a component is able to detect significant changes in the scene structure (for spatial measures) or distinct properties in one of the colour spaces (for colour measures) that *enhances* the discriminatory power of the SRS. For example, in the **IND** database, the sat_c component is more significant for discriminating indoor

scenes than gs_c (7.6). This is because man-made objects become more distinguishable under the *sat* colour space compared to *gs* which improves the discriminatory power of the proposed SRS.

Proportion. As the proposed SRS uses the *percentage matches*, $N_{\%test}$, in the partial computation of G_c (6.3) and in the determination of the adaptive threshold D_t (section 6.2.2) and (6.12), the proportion of a particular *colour* component determines the importance of the component's contribution to P_{acc} and/or P_{rej}. This is clearly shown by the use of hue_c to improve the recognition of outdoor scenes under varying illumination. Although hue_c encodes invariance that *should* aid in positive recognition, the relatively small number of hue_c keypoints compared to gs_c or sat_c keypoints reduces its contribution.

Reliability. For a component to contribute significantly to P_{acc} and/or P_{rej}, not only must it encode enough invariance and/or distinctiveness in sufficient numbers, it must also be reliable and tolerates potential mismatches in the keypoints. A component that is unreliable with numerous mismatches severely degrades the performance of the proposed SRS. A good example is seen in the **NS** database where the addition of one or two ordinal measures (Tables 7.4–7.9) compared to the P_{rej}@10% threshold of **SimpSRS** (Table 7.3) shows a degradation of between 5.77% to 9.61%. A similar (but smaller) degradation is observed for **SBWR** too. The usefulness of ordinal measures is in fact a *double-edged sword* - when the scenes are not too complex and the salient-SURF keypoints are mostly reliable, ordinal

measures often maintain and even enhance the overall recognition accuracy of the SRS (section 3.3.4). However when large amounts of distortions (lighting, viewpoint and occlusions) occuring in the scenes (*e.g.* **NS**, **SBWR**) lead to mismatches in the keypoints, the ordinal measures are affected and P_{rej} is reduced significantly.

Note that the *same* component may possess multiple characteristics that contribute to *both* P_{acc} and P_{rej}. For example, a certain component may contain sufficient distinct features for discrimination of negative scenes, and it may also encode sufficient invariant information that aids in the recognition of positive scenes. This depends on whether or not the scene is a positive or negative test scene that contributes to P_{acc} or P_{rej} respectively (section 7.2).

7.6 Final remarks

This chapter has presented the details of how the proposed SRS is validated using four challenging image databases. Comparative studies with variants of the proposed SRS with a detailed analysis of the results confirms that the inclusion of Z_{ord} and the HSV colour space improves in general the performance of the proposed SRS.

From the analysis of the recognition results, the proposed SRS using all three ordinal measures and colour components reports the *best* overall recognition accuracy in terms of (P_{acc}, P_{rej} and $P_{overall}$) compared to all variants of the SRS over

the four databases. The results presented thus shows that the proposed SRS is *general*, *robust* and is able to *discriminate* a variety of challenging and difficult scenes under various image distortions. All of these are desirable characteristics of a good SRS discussed in section 1.3.

Chapter 8

Conclusions

In this concluding chapter, the main characteristics of the proposed SRS are reviewed in section 8.1. The major concepts introduced in this work are also reviewed in section 8.2. Finally, the thesis closes in section 8.3 with some concluding remarks on potential future research directions that could be undertaken as an extension to the work presented in this thesis.

8.1 Characteristics of the proposed SRS

This thesis has introduced a novel SRS that to the best of this author's knowledge surpasses all other competing scene recognition algorithms that exist in the literature (chapter 2). The novel SRS has the following characteristics that make it stand out from other algorithms:

Generality. The proposed SRS reports very good recognition rates for a wide variety of scenes, from cluttered indoor scenes with people moving around, dynamically changing open natural scenes to enclosed forested areas. This generality is assured because the proposed SRS extracts landmarks from salient ROIs with no prior assumptions on the type of environment that is encountered (section 7.4).

Robustness. The various image distortions - viewpoint changes, changes in illumination as well as changes in the scene structure due to natural erosion, climatic changes and human intervention, exist in all the test scenes. The proposed SRS is able to tolerate these changes and find a positive match, if it exists, in the reference scene. The ability to have such a good degree of robustness is built upon three features in the design of the proposed SRS. Firstly, the use of ordinal measures of spatial correlation in the three directions (x, y, z) (section 3.3) enhances the robustness of the SRS to viewpoint changes. Secondly, the use of the three colour components (H, S, V) (section 3.4) improves the robustness of the SRS to illumination changes for outdoor scenes and provides more discriminatory information for indoor scenes. Lastly, a reliable final decision can be made by the scene decision module thanks to the design of the adaptive decision threshold, D_t, that evolves with the changes in the scene content (section 6.2.2).

Discriminatory power. The proposed SRS is able to discriminate ambiguous scenes that contain similar features and effectively reject them using the scene decision module. The discriminatory power of the proposed SRS comes from the

formulation of G_c (6.3) that includes the spatial configuration information of the matched keypoints that provides a greater amount of information on the scene structure. An ambiguous scene with similar features may have numerous matches with a certain reference scene, but it is *unlikely* to possess a similar scene structure (section 3.3.1). This will degrade G_c and is a clear indication that the test scene is unreliable. Additional discriminatory information is included as the salient-SURF keypoints are matched over three colour spaces and this was shown to enhance the P_{rej} of the proposed SRS compared to when grayscale information alone is used (section 7.5.10).

Accuracy. The recognition results of the proposed SRS are highly reliable and accurate. These results are also repeatable when the same test images are used again. A certain amount of instability may occur for extremely ambiguous scenes in which numerical instabilities in the scene decision module may cause varying decisions to arise from repeated iterations of same test scene (section 7.2). For scenes that are obviously matching or are obviously different, however, the proposed SRS is able to reliably accept or reject these scenes with little difficulty.

8.2 Review of important concepts introduced

Several new concepts introduced by this thesis contribute to the excellent performance of the proposed SRS which are summarised in the next few paragraphs.

A modified **saliency map**. This thesis extends the original saliency map model of [51] by including long edges and the skyline as new composite features to detect stable and salient regions in a scene (sections 4.2 and 4.3). The saliency map is subsequently weighted by a dense ordinal proximity map, $\hat{\mathbf{D}}_{\mathbf{prox}}$, obtained from optic flow (4.5) in order to enhance the saliency of regions in the immediate surroundings for better discrimination of ambiguous scenes. For example, two scenes may share many common features in the background but it is the foreground that discriminates them from one another.

Efficient **SURF keypoint matching**. A fast and efficient keypoint matching algorithm inspired from Lowe's work [68] is introduced. The matching algorithm imposes a uniqueness constraint that results in more correct correspondences (section 5.1.3). Furthermore, the improved efficiency in the matching algorithm allows a large number of features (>1000) to be encoded in the Scene matrix, $\mathbf{m}_s$, of each colour space for effective recognition.

The **Scene matrix** and **Scene matrix cell**. A novel and compact representation of the scene structure is proposed using the Scene matrix cell, $\mathbf{M}_s$, that encodes the augmented salient-SURF keypoints over the three (H, S, V) colour spaces, with each colour space represented by an individual Scene matrix, $\mathbf{m}_s^j, j \in \{H, S, V\}$. The structure of $\mathbf{m}_s$ allows for easy computation of the rank correlations of the spatial configuration in order to determine the G_c between two Scene matrix cells.

The **Global Configuration Coefficient**, G_c. In order to determine a measure of similarity between two scenes, a novel measure of scene similarity termed the Global Configuration Coefficient, G_c, is introduced. This similarity metric exploits the spatial configuration information of the matched salient-SURF keypoints between the scenes. The inherent robustness of the proposed SRS is therefore derived partially from the design of G_c, which is computed in part from the rank correlations of ordinal measures in the three spatial directions (x, y, z) (section 6.1.2).

Computation of an **adaptive decision threshold**, D_t. In order to come to a decision to accept the query scene as a positive match or to reject it as a negative match, a certain decision threshold must be determined beforehand. A novel and intuitive method is introduced that uses the best few matches of the query scene with the database to estimate a reasonable adaptive threshold, D_t (section 6.2.2). The premise is based on the observation that for a true positive scene, the rank correlation components that compute G_c tend to agree with one another while in the opposite situation, a negative scene will cause a lot of fluctuations in these components (Fig. 6.3). The construction of D_t thus attempts to use only the most stable of these components among the best few matches so as to discriminate between difficult positive and negative test scenes (section 6.2.3).

Various **comparative studies** on the **recognition accuracy** of the proposed SRS are also conducted so as to determine which components contribute to positive

recognition and/or positive rejection (section 7.3). The main components tested are the three ordinal measures of spatial configuration (x_{om}, y_{om}, z_{om}) as well as the three colour components (hue_c, sat_c, gs_c) from the HSV colour space. From the experimental results and subsequent analysis presented, the relative importance of the components are determined (section 7.5). A discussion of the results in section 7.5.10 also reveals certain characteristics that a component must possess so as to contribute to the recognition accuracy.

8.3 Future research directions

Although the proposed SRS achieves *excellent* recognition accuracy, there remain several aspects of the algorithm that can be improved in future projects. The following paragraphs briefly describe these aspects that future research should focus on.

Combined saliency-SURF detector/descriptor. The current version of the proposed SRS separates the detection of salient ROIs and the detection of SURF keypoints into two operations (section 1.6). The main reason is that SURF was created to be a stand-alone application which is independent of the saliency algorithm presented. Potential future work should attempt to *integrate* the two components together so that a combined descriptor of the detected salient ROIs that retains the accuracy of the original SURF keypoints can be used. A combined

keypoint descriptor presents several advantages. Firstly, the descriptor can be designed to encode more information concerning the *size* and *global position* of the salient ROIs. This allows for better detection of important missing regions when the two scenes are compared (see Fig. 7.3(bottom)). Secondly, the descriptor can encode directly the *same* keypoint position over the three colour spaces which should lead to a better discrimination of the keypoints for even more accurate matching. Currently, different keypoints are detected in different colour spaces.

Improved segmentation of the scene structure for ordinal depth adjustment. In this work, a simple procedure using AHC is proposed to 'smooth out' the inconsistencies in the ordinal depth estimation so that the computation of the rank correlations will be more accurate (section 5.2.3). The current procedure suffers from two major problems. Firstly, the determination of a *prior* number of depth layers in the scene is often difficult and is compounded by the fact that the scenes used are often complex natural scenes with large depth variations. Secondly, the large number of occlusions and depth variations at the boundaries and within the vegetation itself (for *e.g.* trees and bushes) imply that AHC will often give erroneous depth estimates at such regions since the algorithm takes the mode of the depth in the cluster formed. A possible solution is to use advanced segmentation techniques that use for example the texture of a region [95] or colours (see [20, 30] for a good survey) as a preprocessing step to delimit homogeneous regions in the scene. AHC can then be applied separately to these regions to prevent the depth

estimate from spreading across region boundaries.

Improving the reliability of keypoint matching. Errors in keypoint matching do affect the recognition accuracy of the proposed SRS as was discussed in section 7.5.10. The main problem is that $N_{\%test}$ in (6.3) assumes that the matches are *correct* but this is not always the case. The current version of the SRS tolerates the errors by computing an adaptive threshold, D_t, that reflects the current state of the matching by the heuristic that there should be *more* bad matches for a negative test scene compared to a positive test scene (section 6.2.3). However, this assumption could be wrong, especially for positive test scenes that suffer from significant image distortions that make rejecting the wrong matches even more crucial. The wrong matches may overwhelm the correct matches and this may cause D_t to wrongly reject the positive match. This is however a challenging task as the discriminatory power of the SRS will be affected. This is because a negative match will only have false matches that degrade the rank correlations. This degradation is therefore an important indicator to reject the test scene. Hence, if *all* the bad matches are removed, the rank correlations would become insignificant altogether. A better solution would be to match not only the keypoints but also *groups* of keypoint together, similar to the technique of *semi-local contrained* matching introduced in [98]. This can be easily achieved in the proposed SRS as the keypoints are already grouped into salient ROIs.

Differential weighting of salient regions. A problem highlighted in the

experimental results with the **UBIN** database (Fig. 7.6) showed that false recognition can occur for natural scenes that share a similar skyline. The problem occurs because the landmarks in the scene are *equally* important. This is clearly not always true. Background features in the skyline alone are insufficient to determine a positive match as many reference scenes from the same environment may possess the same skyline. The definition of saliency must therefore be extended to a larger scale, over the whole database if possible. Features that are salient over the whole database will be weighed less than *unique* features that identify a particular scene. The computation of the G_c must also be modified. Highly weighted keypoint matches will increase G_c while the matching of common (or non-globally salient) keypoints alone will only yield a weak G_c as there are insufficient unique matches to conclude a positive match. This results in a new definition of match similarity, denoted as G_{cw} modified from (6.3), where the subscript w stands for 'weighted':

$$G_{cw}(\dot{\mathbf{M}}_{kp}) = \frac{N_{\%test}}{200} \times (\overline{S_{\rho w}} + \overline{K_{\tau w}}) \tag{8.1}$$

where $\overline{S_{\rho w}}, \overline{K_{\tau w}}$ are the *weighted* means of the rank correlations computed by assigning different weights to keypoint matches between the two scenes. Another solution would be to modify the saliency algorithm such that the dense ordinal proximity map, $\hat{\mathbf{D}}_{prox}$ (4.5), is used directly as one of the composite feature maps

(Fig. 4.1). This will enable features that are in the immediate vicinity to be extracted as salient ROIs.

Semi-local computations of $N_{\%test}$. G_c is partially computed from $N_{\%test}$ which is the percentage matches between the salient-SURF keypoints of two scenes. This $N_{\%test}$ is however computed *globally*, with respect to the total number of keypoints in the test scene. The problem with this approach is that the salient region information from which the keypoints originate is completely ignored. Hence an important and critical change in the scene may not be detected if it does not significantly degrade the computed $N_{\%test}$. This could be caused by the small size of the landmark involved as was highlighted in one of the experimental results (Fig. 7.3(bottom)). Future revisions of the proposed SRS should modify the computation of $N_{\%test}$ such that it is computed with respect to the *individual* salient ROIs. If a particular region is completely ignored, a potential mismatch may have occured. The downside of this method is that the SRS's robustness to changes in the scene content is degraded, which is especially true for dynamic natural environments. A possible solution would be to incorporate a certain size information that scales with the current scene structure (*e.g.* a close up shot with large objects or a scene with numerous small objects and clutter) and determine a *size threshold* that allows the proposed SRS to ignore or retain the degradation caused by missing landmarks.

Improved efficiency in scene decision. The pairwise comparisons made with the reference database of N_{ref} scenes have a complexity of $O(dN_{ref})$ where $d = 64$

is the dimension of the SURF keypoints. This naive method is obviously *not* going to work when N_{ref} increases with d remaining constant. A large database containing $N_{ref} > 1000$ means that the SRS must use an efficient search algorithm to determine the best N_{top} matches with the database as fast as possible (section 6.2). The problem of effective database searching has been well researched in the CBIR literature. A solution presented in [44] uses an optimised *Approximated Nearest Neighbour* (ANN) matching algorithm [4] by constructing a *Kd-tree* of the detected features in the reference scenes which are subsequently used for fast matching. Future work should explore using similar methods by modifying the structure of the Scene matrix, $\mathbf{m}_s$, and the Scene matrix cell, $\mathbf{M}_s$, for more efficient matching.

Automatic selection of reference scenes. The reference scenes used in all the experiments are selected *manually* from the complete database. Furthermore, the majority of the reference databases are created using scenes that this author thinks are distinct enough for navigation. Future work should focus on how such scenes can be detected using a global measure of saliency in the database, discussed earlier in this section. Ideally, reference scenes should contain certain distinctive and unique features that make them stand out from the whole database so that recognition is facilitated (section 3.1.1). Modelling how humans organise and choose salient objects from the database remains a difficult and open problem. Such models are more complex than the model of human attention proposed in

[51] which is only relevant for a single image. They will likely entail modelling how salient landmarks are organised and compared before these landmarks are selected as *globally-salient* landmarks.

When to update the reference database? Another fundamental problem that this thesis does not address is how the reference database, $\mathbb{D}_{ref}$, should be updated. A good updating algorithm optimises memory usage and reduces the time needed for database searching as the database size should be kept small. This problem is linked to the automatic selection of reference scenes discussed above. Reference scenes should be updated whenever novel test scenes are encountered by the agent. Such novel scenes are usually rejected with a very low G_c and are analysed for globally salient features in the database. Old and redundant reference scenes are periodically removed from the database by computing a certain threshold of global-salience which changes whenever new scenes are added. Old scenes that fall below this threshold are then rejected.

Future extensions of the proposed SRS include the incorporation of the algorithm with a working practical visual SLAM system for extensive testing of the scene recognition accuracy on a real mobile agent. This agent should be able navigate in various environments - indoors and outdoors, as well as under different illumination and weather conditions so as to validate and improve the proposed SRS further. A parametric learning module, using neural networks for example, can be added to the proposed SRS so that the various parameters and global

thresholds that control the SRS can be calibrated for a particular environment. The algorithm should be ported to a high-level programming language (C++ or Java) for further improvements in the speed and efficiency of the algorithm.

8.4 Closure

This chapter has summarised the major work of this thesis - the introduction of a novel SRS that is general and effective for a variety of challenging environments. The robustness and discriminatory power of the proposed SRS are achieved by exploiting the novel use of ordinal measures on the spatial configuration (section 3.3.4) of the salient ROIs, extracted from a novel depth-weighted saliency map (section 4.4). The extracted ROIs are encoded by salient-SURF keypoints augmented with ordinal depth (section 5.2) that are useful for determining good correspondences (section 5.1.4). These keypoints are extracted from the HSV colour space so as to provide an illumination invariant representation of the scene (section 3.4). A simple decision module, using an adaptive decision threshold, is proposed to effectively accept or reject positive and negative test scenes (section 6.2). The proposed SRS is also validated using extensive tests on challenging image databases. The superior performance of the proposed SRS, as well as the contribution of its various components to the recognition accuracy, are highlighted when it is compared with several similar variants (section 7.5).

The chapter has also presented the potential future research directions that should be undertaken to improve the current SRS. These improvements are aimed at increasing the *speed* and *accuracy* of the proposed SRS so that even ambiguous scenes can be recognised. The integration of the proposed SRS in a practical visual SLAM framework is the ultimate objective of any future work.

The potential applications of the proposed SRS are not only limited to bio-mimetic navigation. The same strategy can be modified for applications such as CBIR or visual SLAM loop closing that were reviewed in chapter 2. Since the proposed SRS is shown to perform remarkably well for complex outdoor natural environments, its potential use as a navigational aid for soldiers or hikers in the field holds interesting possibilities. Humans tend to get lost in unfamiliar and often confusing environments such as in an enclosed forest. The proposed SRS, however, has been shown to be remarkably accurate for such confusing environments (*e.g.* **SBWR**). Furthermore, using the proposed SRS as a navigational aid is useful in environments where current navigational technologies (*e.g.* GPS) remain unusable due to the thick forest foliage. Finally, by incorporating different saliency algorithms or different decision thresholds, the proposed SRS provides the fundamental framework to develop a future SRS that will one day approach or even surpass the scene recognition capabilities of insects or even humans.

Bibliography

[1] Alaa E. Abdel-Hakim and Aly A. Farag. Csift: A sift descriptor with color invariant characteristics. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, pages 1978–1983, 2006.

[2] G. J. Andersen and A. F. Kramer. Limits of focused attention in three-dimensional space. *Perception and Psychophysics*, 53(6):658–667, June 1993.

[3] A. Angeli, D. Fillat, S. Doncieux, and J-A. Meyer. 2d simultaneous localization and mapping for micro air vehicles. In *European Micro Air Vehicle Conference (EMAV)*, July 2006.

[4] Sunil Arya, David M. Mount, Nathan S. Netanyahu, Ruth Silverman, and Angela Y. Wu. An optimal algorithm for approximate nearest neighbor searching fixed dimensions. *J. ACM*, 45(6):891–923, 1998.

[5] Daniel C. Asmar, John S. Zelek, and Samer M. Abdallah. Seeing the trees before the forest. In *Proc. of CRV*, pages 587–593, Washington, DC, USA, 2005. IEEE Computer Society.

[6] Daniel C. Asmar, John S. Zelek, and Samer M. Abdallah. Tree trunks as landmarks for outdoor vision slam. In *CVPRW '06: Proceedings of the 2006 Conference on Computer Vision and Pattern Recognition Workshop*, page 196, Washington, DC, USA, 2006. IEEE Computer Society.

[7] A. Averbuch and A. Schclar. A real-time algorithm for vision-based localization. In *46th International Symposium Electronics in Marine*, pages 125–130, June 2004.

[8] C. Balkenius and L. Kopp. Elastic template matching as a basis for visual landmark recognition and spatial navigation. In *AISB workshop on Spatial Reasoning in Mobile Robots and Animals*, Manchester, 1997.

[9] T. D. Barfoot. Online visual motion estimation using fastslam with sift features. In *Proc. of the IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, 2005.

[10] H. Bay, T. Tuytelaars, and L. Van Gool. Surf: Speeded up robust features. In *Proc. of the ECCV*, May 2006.

[11] Dinkar N. Bhat. An evolutionary measure for image matching. In *Proc. of ICPR*, volume 1, pages 850–852, 1998.

[12] Dinkar N. Bhat and Shree K. Nayar. Ordinal measures for image correspondence. *IEEE Trans. Pattern Anal. Mach. Intell.*, 20(4):415–423, 1998.

[13] G. Bianco and A Zelinsky. Biologically-inspired visual landmark learning and navigation for mobile robots. In *Proc. of the IEEE/RSJ IROS'99*, volume 2, pages 671–676, October 1999.

[14] G. Bianco, A. Zelinsky, and M. Lehrer. Visual landmark learning. In *Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, volume 1, pages 227–232, Takamatsu, Japan, Oct 2000.

[15] Giovanni Bianco and Riccardo Cassinis. Biologically-inspired visual landmark learning for mobile robots. In *EWLR-8: Proceedings of the 8th European Workshop on Learning Robots*, pages 138–164, London, UK, 2000. Springer-Verlag.

[16] Myron Z. Brown, Darius Burschka, and Gregory D. Hager. Advances in computational stereo. *IEEE Trans. Pattern Anal. Mach. Intell.*, 25(8):993–1008, 2003.

[17] R. Burge, J. Mulligan, and P.D. Lawrence. Tree trunks as landmarks for outdoor vision slam. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, volume 1, pages 539–544, October 1998.

[18] J. Canny. A computational approach to edge detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 8(6):679–698, 1986.

[19] B. A. Carwright and T. S. Collet. Landmark learning in bees: Experiments and models. *J. of Comp. Physio.*, 151:521–543, 1983.

[20] H. D. Cheng, X. H. Jiang, Y. Sun, and Jingli Wang. Color image segmentation: advances and prospects. *Pattern Recognition*, 34:2259–2281, 2001.

[21] K. Cheng, T. S. Collett, A. Pickhard, and R. Wehner. The use of visual landmarks by honeybees: Bees weight landmarks according to their distance from the goal. *J. of Comp. Physio. A*, 161(3):469–475, May 1987.

[22] L. F. Cheong and T. Xiang. Characterizing depth distortion under different generic motion. *Int. J. Computer Vision*, 44(3):199–217, 2001.

[23] T. S. Collett. Landmark learning and guidance in insects. *Phil. Trans. Royal Society of London B*, 337:295–303, 1992.

[24] T. S. Collett and M. Collett. Memory use in insect navigation. *Nature Reviews Neuroscience*, 3:542–552, 2002.

[25] T. S. Collett and M. Lehrer. Orientation flights of wasps. In *Proc. R. Soc. of Lond. B,*, volume 252, pages 129–134, 1993.

[26] F. Cozman and E. Krotkov. Automatic mountain detection and pose estimation for teleoperation of lunar rovers. In *Proc. of ICRA*, volume 3, pages 2452–2457, 1997.

[27] Fabio Gagliardi Cozman, Eric Krotkov, and Carlos Guestrin. Outdoor visual position estimation for planetary rovers. *Autonomous Robots*, 9(2):135–150, 2000.

[28] J. W Dauben. *Georg Cantor : his mathematics and philosophy of the infinite.* Princeton University Press, Princeton, N.J., 1990.

[29] Jesse Davis and Mark Goadrich. The relationship between precision-recall and roc curves. In *ICML '06: Proceedings of the 23rd international conference on Machine learning*, pages 233–240, New York, NY, USA, 2006. ACM Press.

[30] Yining Deng, B.S. Manjunath, and Hyundoo Shin. Color image segmentation. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, volume 02, pages 446–451, 1999.

[31] Guilherme N. DeSouza and Avinash C. Kak. Vision for mobile robot navigation: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(2):237–267, 2002.

[32] M. Drew, G. Finlayson, and S. Hordley. Recovery of chromaticity image free from shadows via illumination invariance. In *Proc. of ICCV*, pages 32–39, 2003.

[33] Todt E. and Torras C. Detecting salient cues through illumination-invariant color ratios. *Robotics and Autonomous Systems*, 48(2–3):111–130, September 2004.

[34] David Filliat and Jean-Arcady Meyer. Map-based navigation in mobile robots. i. a review of localization strategies. *J. of Cognitive Systems Research*, 4(4):243–283, 2003.

[35] G. Finlayson and G. Schaefer. Hue that is invariant to brightness and gamma. In *Proc. of the BMVC*, September 2001.

[36] G. D. Finlayson and S. D. Hordley. Color constancy at a pixel. *J. Opt. Soc. Am. A*, 18:253–264, 2001.

[37] Graham D. Finlayson, Steven D. Hordley, and Mark S. Drew. Removing shadows from images. In *Proc. of ECCV*, pages 823–836, London, UK, 2002. Springer-Verlag.

[38] Graham D. Finlayson, Steven D. Hordley, and Mark S. Drew. Removing shadows from images using retinex. In *Colour Imaging Conference*, pages 73–79, 2002.

[39] M. O. Franz and H. A. Mallot. Biomimetic robot navigation. *Robotics and Autonomous Systems*, 30:133–153, 2000.

[40] Friedrich Fraundorfer and Horst Bischof. A novel performance evaluation method of local detectors on non-planar scenes. In *Proc. of CVPR*, page 33, 2005.

[41] S. Frintrop. *VOCUS: A Visual Attention System for Object Detection and Goal-directed Search.* Number 3899 in Lecture Notes in Artificial Intelligence. Springer, 2006.

[42] Simone Frintrop, Patric Jensfelt, and Henrik I. Christensen. Attentional landmark selection for visual slam. In *Proc. of the IEEE/RSJ IROS'06*, October 2006.

[43] Simone Frintrop, Patric Jensfelt, and Henrik I. Christensen. Pay attention when selecting features. In *Proc. of the ICPR*, pages 163–166, 2006.

[44] Toon Goedeme, Tinne Tuytelaars, and Luc Van Gool. Fast wide baseline matching for visual navigation. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, volume 01, pages 24–29, 2004.

[45] E. B. Goldstein. *Sensation and Perception.* Thomson Wadsworth, 7th edition, 2007.

[46] R. I. Hartley. In defence of the 8-point algorithm. In *ICCV '95: Proceedings of the Fifth International Conference on Computer Vision*, page 1064, 1995.

[47] G. Heidemann. The long-range saliency of edge- and corner-based salient points. *IEEE Trans. Pattern Anal. Mach. Intell.*, 14(11):1701–1706, November 2005.

[48] John M. Henderson. Human gaze control during real-world scene perception. *Trends in Cognitive Sciences*, 7(11):498–504, November 2003.

[49] B. K. P. Horn and B. G. Schunck. 'determining optical flow': a retrospective. *Artificial Intelligence*, 59(1–2):81–87, February 1993.

[50] L. Itti and C. Koch. Comparison of feature combination strategies for saliency-based visual attention systems. In B. E. Rogowitz and T. N. Pappas, editors, *Proc. SPIE Vol. 3644, p. 473-482, Human Vision and Electronic Imaging IV*, pages 473–482, May 1999.

[51] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. Pattern Anal. Mach. Intell.*, 20(11):1254–1259, 1998.

[52] P. Corke J. Banks. Quantitative evaluation of matching methods and validity measures for stereo vision. *Int. J. of Robotics Research*, 20:512–532, July 2001.

[53] Derek Johns and Gregory Dudek. Urban position estimation from one dimensional visual cues. In *Proc. of 3rd Canadian Conf. on Computer and Robot Vision (CRV'06)*, page 22, 2006.

[54] Timor Kadir and Michael Brady. Saliency, scale and image description. *Int. J. Computer Vision*, 45(2):83–105, 2001.

[55] Y. Ke and R. Sukthankar. Pca-sift: A more distinctive representation for local image descriptors. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, 2004.

[56] M. Kendall and J.D. Gibbons. *Rank Correlation Methods*. Edward Arnold, 5th edition, 1990.

[57] C. Koch and S. Ullman. Shifts in selective visual attention: Towards the underlying neural circuitry. *Human Neurobiology*, 4:219–227, 1985.

[58] D. Lambrinos, R. Moller, T. Labhart, R. Pfeifer, and R. Wehner. A mobile robot employing insect strategies for navigation. *Robotics and Autonomous Systems*, 30:39–64, 2000.

[59] Edwin H. Land. Recent advances in retinex theory. *Vision Research*, 26(1):7–21, 1986.

[60] L. Ledwich and S. Williams. Reduced sift features for image retrieval and indoor localization. In *Aust. Conf. on Robotics and Automation (ACRA)*, 2004.

[61] M. Lehrer. Why do bees turn back and look? *J. of Comp. Physio. A*, 173:549–563, 1993.

[62] M. Lehrer. Looking all around: Honeybees use different cues in different eye regions. *J. of Experimental Biology*, 201:3275–3292, 1998.

[63] M. Lehrer and G. Bianco. The turn-back-and-look behaviour: bee versus robot. *Biological Cybernetics*, 83(3):211–229, Aug 2000.

[64] M. Lehrer and T. S. Collett. Approaching and departing bees learn different cues to the distance of a landmark. *J. of Comp. Physio. A*, 175:171–177, 1994.

[65] T. Lemaire, C. Berger, I.-K. Jung, and S. Lacroix. Vision-based slam: Stereo and monocular approaches. *Submitted to IJCV/IJRR special joint issue*, 2006.

[66] T. Lemaire and S. Lacroix. Long term slam with panoramic vision. *Submitted to J. of Field Robotics*, 2006.

[67] H. C. Longuet-Higgins. A computer algorithm for reconstruction of a scene from two projections. *Nature*, 293:133–135, 1981.

[68] D. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. of Computer Vision*, 20:91–110, 2003.

[69] Jiebo Luo, Stephen P. Etz, and Robert T. Gray. Normalized kemeny and snell distance: A novel metric for quantitative evaluation of rank-order similarity of images. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(8):1147–1151, 2002.

[70] J. I. Marden. *Analyzing and Modeling Rank data*. Chapman and Hall, 1995.

[71] J Matas, O Chum, U Martin, and T Pajdla. Robust wide baseline stereo from maximally stable extremal regions. In *Proc. of the BMVC*, volume 1, pages 384–393, London, 2002.

[72] Jean-Arcady Meyer and David Filliat. Map-based navigation in mobile robots - ii. a review of map-learning and path-planing strategies. *J. of Cognitive Systems Research*, 4(4):283–317, 2003.

[73] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. Van Gool. A comparison of affine region detectors. *Int. J. Comput. Vision*, 65(1-2):43–72, 2005.

[74] Krystian Mikolajczyk and Cordelia Schmid. A performance evaluation of local descriptors. In *Proc. of ICPR*, volume 2, pages 257–263, June 2003.

[75] Krystian Mikolajczyk and Cordelia Schmid. A performance evaluation of local descriptors. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27(10):1615–1630, 2005.

[76] J. V. Miro, W. Zhou, and G. Dissanayaje. Towards vision based navigation in large indoor environments. In *Proc. of the IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, pages 2096–2102, 2006.

[77] T. M. Mitchell. *Machine Learning*. New York, NY:McGraw-Hill, 1997.

[78] R. Moller. Insect visual homing strategies in a robot with analog processing. *Biological Cybernetics*, 83(3):231–243, 2000.

[79] R. Moller. Insects could exploit uv-green contrast for landmark navigation. *J. of Theoretical Biology*, 214(4):619–631, February 2002.

[80] R. Moller, D. Lambrinos, T. Roggendorf, R. Pfeifer, and R. Wehner. Insect strategies of visual homing in mobile robots. *Biorobotics - Methods and Applications*, 2001.

[81] P. Moreels and P. Perona. Evaluation of features detectors and descriptors based on 3d objects. In *Proc. of ICCV*, volume 1, pages 800–807, 2005.

[82] P. Moreels and P. Perona. Evaluation of features detectors and descriptors base on 3d objects. *International J. of Computer Vision (IJCV)*, 2006.

[83] T. Mori, Y. Matsumoto, T. Shibata, M. Inaba, and H. Inoue. Trackable attention point generation based on classification of correlation value distribution. In *ROBOMEC*, pages 1076–1079, Kavasaki, Japan, 1995.

[84] K. Nakayama and G. H. Silverman. Serial and parallel processing of visual feature conjunctions. *Nature*, 320:264–265, March 1986.

[85] O. Nestares, R. Navarro, J. Portilla, and A. Tabernero. Efficient spatial-domain implementation of a multiscale image representation based on gabor functions. *J. of Electronic Imaging*, 7:166–173, 1996.

[86] P. Newman, D. Cole, and K. Ho. Outdoor slam using visual appearance and laser ranging. In *Proc. of ICRA 2006*, pages 1180–1187, 2006.

[87] P. Newman and K. Ho. Slam-loop closing with visually salient features. In *Proc. of ICRA 05*, pages 644–651, 2005.

[88] D. Parkhurst, K. Law, and E. Neibur. Modelling the role of salience in the allocation of overt visual attention. *Vision research*, 42(1):107–123, January 2002.

[89] R. J. Peters, A. Iyer, L. Itti, and C. Koch. Components of bottom-up gaze allocation in natural images. *Vision research*, 45(18):2397–2416, August 2005.

[90] M. Pilu. A direct method for stereo correspondence based on singular value decomposition. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, page 261, 1997.

[91] M. I. Posner and S. E. Petersen. The attention system in the human brain. *Annual Review of neuroscience*, 13:24–42, 1990.

[92] A. J. Pritchard, R.E.N. Horne, and S.J. Sangwine. Achieving brightness-insensitive measurements of colour saturation for use in object recognition. *IEE Conference Publications*, pages 791–795, 1995.

[93] A. Rosenfeld. *Multiresolution Image Processing and Analysis*. Springer Series in Information Sciences. Springer-Verlag, New York, 1984.

[94] J. E. Rubin. *Set theory for the mathematician*. Holden-Day, San Francisco, 1967.

[95] F. Schaffalitzky and A. Zisserman. Viewpoint invariant texture matching and wide baseline stereo. In *Proc. 8th International Conference on Computer Vision*, July 2001.

[96] Daniel Scharstein and Richard Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *Int. J. Comput. Vision*, 47(1-3):7–42, 2002.

[97] S. Scherer, A. Pinz, and P. Werth. The discriminatory power of ordinal measures - towards a new coefficient. In *Proc. of CVPR*, volume 01, pages 1076–1081, 1999.

[98] Cordelia Schmid and Roger Mohr. Local grayvalue invariants for image retrieval. *IEEE Trans. Pattern Anal. Mach. Intell.*, 19(5):530–535, 1997.

[99] G. L. Scott and H. C. Longuet-Higgins. An algorithm for associating the features of two images. *Proceedings: Biological Science, Royal Society London*, 244(1309):21–26, April 1991.

[100] S. Se, D. Lowe, and J. Little. Vision-based mobile robot localization and mapping using scale-invariant features. In *Proc. of the Int. Conf. on Robotics and Automation (ICRA)*, pages 2051–2058, Seoul, Korea, May 2001.

[101] R. Sekuler and R. Blake. *Perception*. McGraw-Hil, 5th edition, 2005.

[102] Larry S. Shapiro and J. Michael Brady. Feature-based correspondence: an eigenvector approach. *Image Vision Comput.*, 10(5):283–288, 1992.

[103] I. Shmuulevich, B. Cramariuc, and M. Gabbouj. A framework for ordinal-based image correspondence. In *Proc. of EUSIPCO 2000*, pages 1389–1392, Tampere, Finland, September 2000.

[104] S. Siggelkow and H. Burkhardt. Image retrieval based on colour and nonlinear texture invariants. In *Proceedings of the Noblesse Workshop on Non-Linear Model Based Image Analysis*, pages 217–224, July 1998.

[105] A. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrival at the end of the early years. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(12):1349–1380, 2000.

[106] M. V. Srinivasan, M. Lehrer, S.W. Zhang, and G.A. Horridge. How honeybees measure their distance from objects of unknown size. *J. of Comp. Physio. A*, 165:605–613, 1989.

[107] M. V. Srinivasan, S. W. Zhang, J. S. Chahl, E. Barth, and S. Venkatesh. How honeybees make grazing landings on at surfaces. *Biological Cybernetics*, 83(3):171–183, September 2000.

[108] F. Stein and G. Medioni. Map-based localization using the panoramic horizon. In *Proc. of ICRA*, volume 3, pages 2631–2637, May 1992.

[109] S. Todorovic, M. C. Nechyba, and P. G. Ifju. Sky/ground modeling for autonomous mav flight. In *Proc. of ICRA*, volume 1, pages 1422–1427, September 2003.

[110] S. Todorovic and M.C. Nechyba. A vision system for intelligent mission profiles of micro air vehicles. *IEEE Trans. on vehicular technology*, 53(6):1713–1725, Nov 2004.

[111] E. Todt and C. Torras. Color constancy for landmark detection in outdoor environments. In *Proc. 4th. European Workshop on Advanced Mobile Robots (Eurobot)*, pages 75–82, 2001.

[112] A. Triesman. *The perception of features and objects, Attention: Selection Awareness and Control*. Oxford University Press, 1995.

[113] O. Trullier, S. Wiener, A. Berthoz, and J. Meyer. Biologically-based artificial navigation systems: Review and prospects. *Progress in Neurobiology*, 51:483–544, 1997.

[114] S. Ullman. *The interpretation of Visual Motion*. MIT Press, Cambridge, MA, 1979.

[115] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, 2001.

[116] R. Voss and J. Zeil. Active vision in insects: an analysis of object-directed zigzag flights in wasps (*Odynerus spinipes*, Eumenidae). *J. of Comp. Physio. A*, 182(3):377–387, February 1998.

[117] W. Weaver and C. E. Shannon. *The Mathematical Theory of Communication*. University of Illinois Press, Urbana, Illinois, 1963.

[118] J. Wolf, W. Burgard, and H. Burkhardt. Robust vision-based localization for mobile robots using an image retrieval system based on invariant features. In *Proc. of the Int. Conf. on Robotics and Automation (ICRA)*, pages 359–365, May 2002.

[119] J. M. Wolfe. *Visual Search*. Attention. Pyschology Press, 1998.

[120] Ramin Zabih and John Woodfill. Non-parametric local transforms for computing visual correspondence. In *Proc. of the ECCV*, pages 151–158, 1994.

[121] J. Zeil, A. Kelber, and R. Voss. Structure and function of learning flights. *J. of Experimental Biology*, 199:245–252, 1996.

Appendix A

Demonstration of rank correlation measures

In this appendix, the computations of Spearman's ρ (3.1) and Kendall's τ (3.2) are demonstrated using two numerical examples.

Example A.2. *Computation of S_ρ* Consider two rankings of 10 elements $(\mathbf{A}, \mathbf{B})$, the difference, d of two rankings and the square of the differences d^2 are shown below.

Table A.1: Computation of S_ρ

A :	7	4	3	10	6	2	9	8	1	5
B :	5	7	3	10	1	9	6	2	8	4
d	2	-3	0	0	5	-7	3	6	-7	1
d^2	4	9	0	0	25	49	36	9	49	1

Summing the bottom row of d^2 gives the component $\sum d^2$ in (3.1). With $n = 10$

and $\sum d^2 = 182$, one obtains $S_\rho = -0.103$. □

Example A.3. *Computation of* K_τ Consider the same two rankings of 10 elements $(\mathbf{A}, \mathbf{B})$ together with the *natural* order ranking $\mathbb{N}_r$ as shown below.

Table A.2: Computation of K_τ

$\mathbb{N}_r$:	1	2	3	4	5	6	7	8	9	10
A :	7	4	3	10	6	2	9	8	1	5
B :	5	7	3	10	1	9	6	2	8	4

The number 1 ranking in **B** has 6 above in ranking **A**. In $\mathbb{N}_r$, ranking 6 has four members to its right, so the current score now is 4 and delete 6 from $\mathbb{N}$. Moving on to the number 2 ranking in **B**, it has a 8 in **A**. Since 8 has two members to its right in $\mathbb{N}_r$, the current score is $4 + 2$ and 8 is deleted from $\mathbb{N}_r$. Continuing this way until all the members in **Y** are compared leads to the full score, P, given as $P = 4 + 2 + 5 + 3 + 2 + 1 + 1 + 2 + 1 + 0 = 21$. Applying (3.2), one obtains $K_\tau = \frac{4(21)}{10(9)} - 1 = -0.07$. □

Appendix B

Derivation of Z_{ord} from optical flow

The optical flow (u, v) of an image location p at (x, y) by a projection of a scene point P at (X, Y, Z) in the world is given by [67]:

$$\begin{aligned} u &= u_{trans} + u_{rot} = \frac{W}{Z}(x - \frac{fU}{W}) + u_{rot} \\ &= \frac{W}{Z}(x - \frac{fU}{W}) + \frac{\omega_x xy}{f} - \omega_y(\frac{x^2}{f} + f) + \omega_z y \\ v &= v_{trans} + v_{rot} = \frac{W}{Z}(y - \frac{fV}{W}) + v_{rot} \\ &= \frac{W}{Z}(y - \frac{fV}{W}) - \frac{\omega_y xy}{f} + \omega_x(\frac{y^2}{f} + f) - \omega_z x \end{aligned} \tag{B.1}$$

where (U, V, W) and $(\omega_x, \omega_y, \omega_z)$ are the translation and rotation components respectively. f is the focal length in pixels. (u_{trans}, v_{trans}) are the horizontal and vertical components of the flow due to translation and (u_{rot}, v_{rot}) are the horizontal and vertical components of the flow due to rotation. Z is the depth of scene point P that corresponds to the imaged point p. The focus of expansion (FOE)

(x_0, y_0) is given by $(\frac{fU}{W}, \frac{fV}{W})$. For the case of TBL [61] motion, $W \to 0$ and the rotation $(\omega_x, \omega_y, \omega_z) \to (0, \omega_y, 0)$, one can thus simplify (B.1) to:

$$\begin{aligned} u &= -\frac{fU}{Z} + u_{rot} = -\frac{fU}{Z} - \omega_y(\frac{x^2}{f} + f) \Rightarrow Z = \frac{-fU}{u - u_{rot}} \\ v &= -\frac{fV}{Z} + v_{rot} = -\frac{fV}{Z} - \frac{\omega_y xy}{f} \Rightarrow Z = \frac{-fV}{v - v_{rot}} \end{aligned} \tag{B.2}$$

Combining the results in (B.2), one can write:

$$Z = \frac{-(fU, fV) \cdot (n_x, n_y)}{(u - u_{rot}, v - v_{rot}) \cdot (n_x, n_y)} \tag{B.3}$$

where (n_x, n_y) is an unit vector specifying a direction from which Z is recovered. As the TBL motion can be approximated by a lateral translation, one can approximate (U, V) as $(U, 0)$ and set $(n_x, n_y) \to (1, 0)$ to obtain the *scaled depth*, $\frac{Z}{U}$ from (B.3):

$$\frac{Z}{U} = \frac{-f}{u + \omega_y f} \tag{B.4}$$

where the second order flow in the term that contains ω_y is ignored. From [22], it is proven that even if the estimates for ω_y and f are in error, the recovered scaled depths are related to the true scaled depths by a relief transformation and thus the order of the depths will be *preserved*. Since the exact value of ω_y is not crucial, one can approximate it as zero. Henceforth the scale depth, $\frac{Z}{U}$ recovered in (B.4) is denoted as Z_{ord}, the ordinal depth which is shown in (5.5). □

Appendix C

Demonstration of scene decision using D_t

The effectiveness of D_t in three common scenarios that the proposed SRS encounters are demonstrated using the **NS** and **UBIN** databases (section 7.1). The same parameters used to obtain the results in section 7.4 are similarly used. For each case, how the *Threshold vector*, $\mathbf{\Xi}_s$ (6.11), is obtained from the *Decision matrix*, $\mathbf{\Delta}_s$ (definition 6.12), and the *Match statistic matrix*, $\mathbf{\Pi}_s$ (6.8), detailed in section 6.2.2, is shown. The three cases considered are: a typical *positive* case, a *negative* case and finally an *ambiguous* case.

C.1 Positive case

The test scene used for this example is **Nat2_14** shown in Fig. C.1 (left). After running the test scene with the reference database, $\mathbb{D}_{ref}$, one obtains $G_{cand} = 0.40$

(6.9) from the 19^{th} row of $\mathbf{\Pi}_s$, denoted as $\mathbf{\Pi}_s(19)$:

$$\mathbf{\Pi}_s(19) = \begin{bmatrix} 45.0 & 1.0 & 1.0 & 0.68 & 0.98 & 0.98 & 0.67 & 0.40 \end{bmatrix}$$

Following the procedure in section 6.2.2, $\mathbf{\Delta}_s$ is obtained as a row vector:

$$\mathbf{\Delta}_s = \begin{bmatrix} 45.0 & 1.0 & 1.0 & 0.68 & 0.98 & 0.98 & 0.67 & 0.40 \end{bmatrix}$$

$\mathbf{\Xi}_s$ is then obtained as:

$$\mathbf{\Xi}_s = \begin{bmatrix} 45.0 & 1.0 & 1.0 & 0.68 & 0.98 & 0.98 & 0.67 \end{bmatrix}$$

and $D_t = 0.40 = G_{cand}$ in this case. Applying (6.10) yields:

$$D_f = \text{ACCEPT}$$

$\mathbb{G}_{cand}$ is thus accepted as a *positive* match (Fig. C.1). This is an extremely simple case as the test scene has only suffered a slight viewpoint change from the reference scene in $\mathbb{D}_{ref}$. This explains why $\mathbf{\Delta}_s$ only contains a single row, and this row is the correctly matched $\mathbb{G}_{cand}$. This highlights the importance of the equality in (6.10) for accepting $\mathbb{G}_{cand}$ as a positive match since $G_{cand} = D_t$ for such cases. □

Figure C.1: Matched reference scene (right) with the test scene (left).

C.2 Negative case

The input test scene is **Nat2_66a** shown in Fig. C.2. This is a *negative* test scene

Figure C.2: Input negative test scene.

as there are no reference scenes in $\mathbb{D}_{ref}$ that correspond to this particular location. Nonetheless, because the general environment is almost the same, many similar features exist that may confuse the salient-SURF correspondences. D_t is however designed to detect such ambiguities to reject this scene. After comparing the test

scene with $\mathbb{D}_{ref}$, one obtains $G_{cand} = 0.035$ from the 4^{th} row of $\mathbf{\Pi}_s$:

$$\mathbf{\Pi}_s(4) = \begin{bmatrix} 12.0 & 0.57 & 0.31 & 0.11 & 0.39 & 0.25 & 0.10 & 0.035 \end{bmatrix}$$

Following the procedure in section 6.2.2, $\mathbf{\Delta}_s$ is obtained as:

$$\mathbf{\Delta}_s = \begin{bmatrix} 12.0 & 0.57 & 0.31 & 0.11 & 0.39 & 0.25 & 0.10 & 0.035 \\ 11.0 & -0.18 & 0.19 & 0.60 & -0.19 & 0.10 & 0.58 & 0.021 \end{bmatrix}$$

$\mathbf{\Delta}_s$ shows that there are two competing matches, including that of $\mathbb{G}_{cand}$. These two matches are then used to obtain $\mathbf{\Xi}_s$:

$$\mathbf{\Xi}_s = \begin{bmatrix} 12.0 & 0.57 & 0.31 & 0.60 & 0.39 & 0.25 & 0.58 \end{bmatrix}$$

Comparing $\mathbf{\Xi}_s$ to $\mathbf{\Pi}_s$, one can see that *most* of the elements in $\mathbf{\Xi}_s$ comes from the match that yields $\mathbb{G}_{cand}$. This is not a surprise, as $\mathbb{G}_{cand}$ is obtained from the *best* performing match in the whole database. However, a reliable match should give more or less *consistent* rank correlations that will be reflected in $\mathbf{\Xi}_s$. Notice that the 4^{th} element, $S^z_{\rho\Xi}$, takes on the value of the 2^{nd} row in $\mathbf{\Delta}_s$ which is larger than that found in the row belonging to $\mathbb{G}_{cand}$. This is an indication that $\mathbb{G}_{cand}$ does not respect the general configuration for a true match with a significant reduction

in its own $S^z_{\rho\Xi}$. Using $\mathbf{\Xi}_s$ to compute D_t gives $D_t = 0.054$. Applying (6.10) yields:

$$D_f = \text{REJECT}$$

as $G_{cand} < D_t$. This example highlights the importance of the rank correlations in rejecting scenes with similar features and different scene structures. □

C.3 Ambiguous case

A match is defined as *ambiguous* when $\mathbf{\Delta}_s$ is empty (section 6.2.4). Two test cases are considered: **Nat2_38a** (**NS**) as a negative test scene and **Nat27a** (**UBIN**) as a positive test scene (Fig. C.3). Both of these test scenes yield an empty $\mathbf{\Delta}_s$ but the correct D_f is made by the proposed SRS.

Figure C.3: Nat2_38a (left) and **Nat27a** (right) input test scenes.

C.3.1 Ambiguous rejection

Running the image **Nat_38a** through the pairwise comparisons with $\mathbb{D}_{ref}$, one obtains $G_{cand} = 0.018$ from the 2^{nd} row of $\mathbf{\Pi}_s$:

$$\mathbf{\Pi}_s(2) = \begin{bmatrix} 3.8 & 0.37 & 0.70 & 0.44 & 0.28 & 0.61 & 0.41 & 0.018 \end{bmatrix}$$

Following the procedure in section 6.2.2 yields an empty $\mathbf{\Delta}_s$ as the number of matched keypoints are small. This means that this match is likely to be unreliable and should be rejected. The alternate procedure described in section 6.2.4 is followed where $D_t = G_{cand} = 0.018$ immediately and D_{min} is modified to a larger value $D^*_{min} = 0.05$. Applying (6.18) for the ambiguous case, one obtains the final decision as:

$$D_f = \text{REJECT}$$

since $G_{cand} < D_{min} = 0.05$. The importance of manipulating D_{min} is thus highlighted as ambiguous scenes require a higher threshold to prevent false matches from occurring. □

C.3.2 Ambiguous acceptance

The likelihood of an ambiguous test scene that has a positive match in $\mathbb{D}_{ref}$ is *low* as most positive scenes usually have sufficient matches with their true matching scenes to prevent ambiguity. This condition does arise, however, when the positive test scene has undergone significant distortions. A positive ambiguous match is different from the negative case as G_{cand} tends to be *larger* due to the better preservation of the rank correlations. Running the image **Nat27a** through the pairwise comparisons with $\mathbb{D}_{ref}$ gives $G_{cand} = 0.036$ from the 19^{th} row of $\mathbf{\Pi}_s$:

$$\mathbf{\Pi}_s(19) = \begin{bmatrix} 7.0 & 0.61 & 0.68 & 0.45 & 0.41 & 0.54 & 0.39 & 0.036 \end{bmatrix}$$

Since the number of matches are small, $\mathbf{\Delta}_s$ is empty. This is caused by the large viewpoint change between the test image and the reference image that reduces the number of common features between the two scenes available for an *unambiguous* recognition. The result is that *all* of the matches, including the true positive match has a percentage match below $t_{\%} = 10\%$ (section 6.2.2). The procedure described in section 6.2.4 is followed, by adjusting the value of D_{min} from 0.01 to 0.03 for the **SBWR** database. Fixing $D_t = G_{cand} = 0.036$, D_f is determined as:

$$D_f = \text{ACCEPT}$$

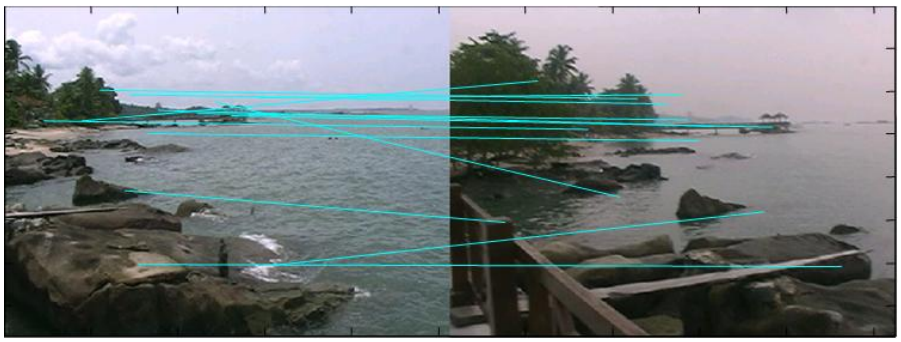

Figure C.4: Ambiguous positive scene: Reference scene (left) with the test scene (right). The correspondences are shown as cyan lines.

since $G_{cand} > D_{min}$ (6.18). This test scene is thus accepted in spite of its initial ambiguity (Fig. C.4). □

These two examples highlight the effectiveness of the proposed SRS in handling difficult ambiguous cases. The crucial idea is to adjust the value of D_{min} as soon as the ambiguity is detected. The main problem is the value to set for D_{min}, which is entirely left to the discretion of the user. A very large value is almost certain to reject all ambiguous matches and is not likely to identify difficult *positive* scenes which may not be desirable. On the other hand, a value that is too small will run the risk of false positives that *should* be rejected. As was mentioned in section 8.3, since these parameters are entirely application dependent, a learning algorithm can be implemented to determine the optimum range of values to obtain the highest recognition accuracy.

Appendix D

Reference Database and Test scenes

This appendix shows two examples of the MATLAB® console output of the proposed SRS for two test cases: a positive scene (section D.1) and a negative scene (section D.2). Examples from the four databases (section 7.1) of reference scenes in $\mathbb{D}_{ref}$ together with their corresponding test scenes are shown in section D.3.

D.1 MATLAB® output for a positive scene

The Console output 1 comes from a positively matched scene in the **NS** database (Fig. D.1 (left)).

As can be seen from this positive example, because the scene is correctly recognised, the rank correlations are preserved with a significant number of matches

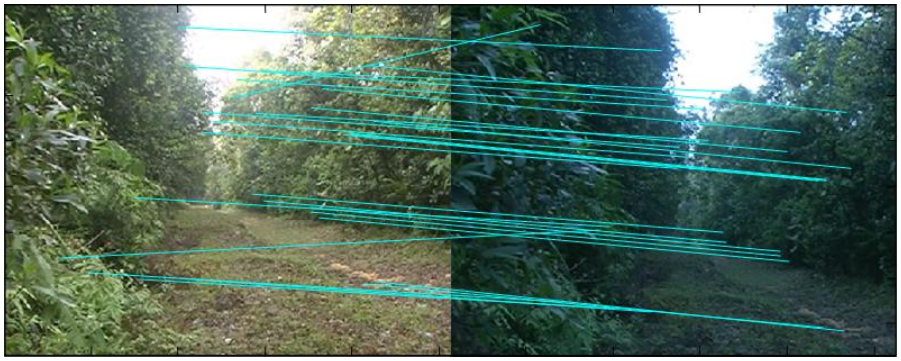

Figure D.1: Sample matched positive reference scene (left) with the input test scene (right). Correspondences are shown as cyan lines.

```
****** Scene Matching Algorithm version 5.0 ******

Running scene matching routines with 20 reference scenes...
.....................done.
***** Results of the Scene Matching Algorithm version 5.0 ********

********* Global Statistics *******
Threshold for SURF matching: 0.82,
Threshold (adaptive) for decision metric: 0.0904533
Threshold for statistical significance: 0.6,
Absolute min for decision metric: 0.01
Threshold for RANSAC tolerance: 0.02,
Number of RANSAC iterations performed: 5
Threshold for min RANSAC points: 20, Points per RANSAC trials used: 10
Number of matches found: 28
Percentage of matches found: 12.1739 percent
******** Detailed results *******
Spearman's rho for x coordinate: 0.963875
Kendall's tau for x coordinate: 0.899471
Spearman's rho for y coordinate: 0.970443
Kendall's tau for y coordinate: 0.878307
Spearman's rho for depth: 0.384551
Kendall's tau for depth: 0.36141
matched weighted Global correlation using Spearman's rho: 0.0468149
matched weighted Global correlation using Kendall's tau: 0.0439978
Mean of the Global correlation: 0.0904533
********* Match Decision *********
Match is found with:Nat2_4_f2191_2197
displaying the matched scenes...
displaying the correspondences...
Saving results and matlab figures...
Done.
Elapsed time is 15.313000 seconds.
```

Console output 1: Positive recognition.

found. This results in a G_{cand} (`Global correlation`) that is equal to the decision threshold, D_t(`Threshold (adaptive)`) computed, and the test scene is thus accepted. The matched reference scene together with the correspondences found are shown in Fig. D.1.

D.2 MATLAB® output for a negative scene

The Console output 2 comes from a rejected (negative) scene in the **NS** database (Fig. D.2)

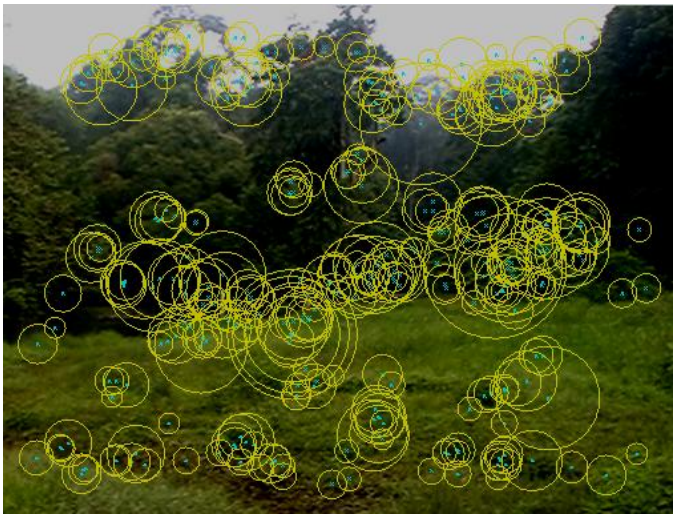

Figure D.2: The negative sample test scene shown with extracted salient-SURF keypoints.

As can be seen, because this is a negative test scene, all of the rank correlations are low with only a small number of matches found. This results in a small G_{cand} which is threshold rejected.

```
****** Scene Matching Algorithm version 5.0 ******

Running scene matching routines with 20 reference scenes...
.....................done.
***** Results of the Scene Matching Algorithm version 5.0 ********

********* Global Statistics *******
Threshold for SURF matching: 0.82,
Threshold (adaptive) for decision metric: 0.05
Threshold for statistical significance: 0.6,
Absolute min for decision metric: 0.01
Threshold for RANSAC tolerance: 0.02,
Number of RANSAC iterations performed: 5
Threshold for min RANSAC points: 20, Points per RANSAC trials used: 10
Number of matches found: 17
Percentage of matches found: 5.43131 percent
******** Detailed results *******
Spearman's rho for x coordinate: 0.0367647
Kendall's tau for x coordinate: 0
Spearman's rho for y coordinate: 0.25
Kendall's tau for y coordinate: 0.132353
Spearman's rho for depth: 0.2125
Kendall's tau for depth: 0.204545
matched weighted Global correlation using Spearman's rho: 0.00199681
matched weighted Global correlation using Kendall's tau: 0
Mean of the Global correlation: 0.00909393
********* Match Decision *********
No reliable matches were found in the database.
Saving results and matlab figures...
Done.
Elapsed time is 15.938000 seconds.
```

Console output 2: Rejection of a negative test scene.

D.3 Sample positive results from the four databases

In the following few figures, the recognition results for various challenging *positive* test scenes are presented that illustrate the performance of the proposed SRS. The results for the four databases are shown separately in each figure. The matched reference scene is shown on the left and the input test scene is shown on the right with the correspondences shown as cyan lines. The complete image database is submitted electronically with the attached CD-ROM containing this thesis.

Figure D.3: IND database matches.

Figure D.4: UBIN database matches.

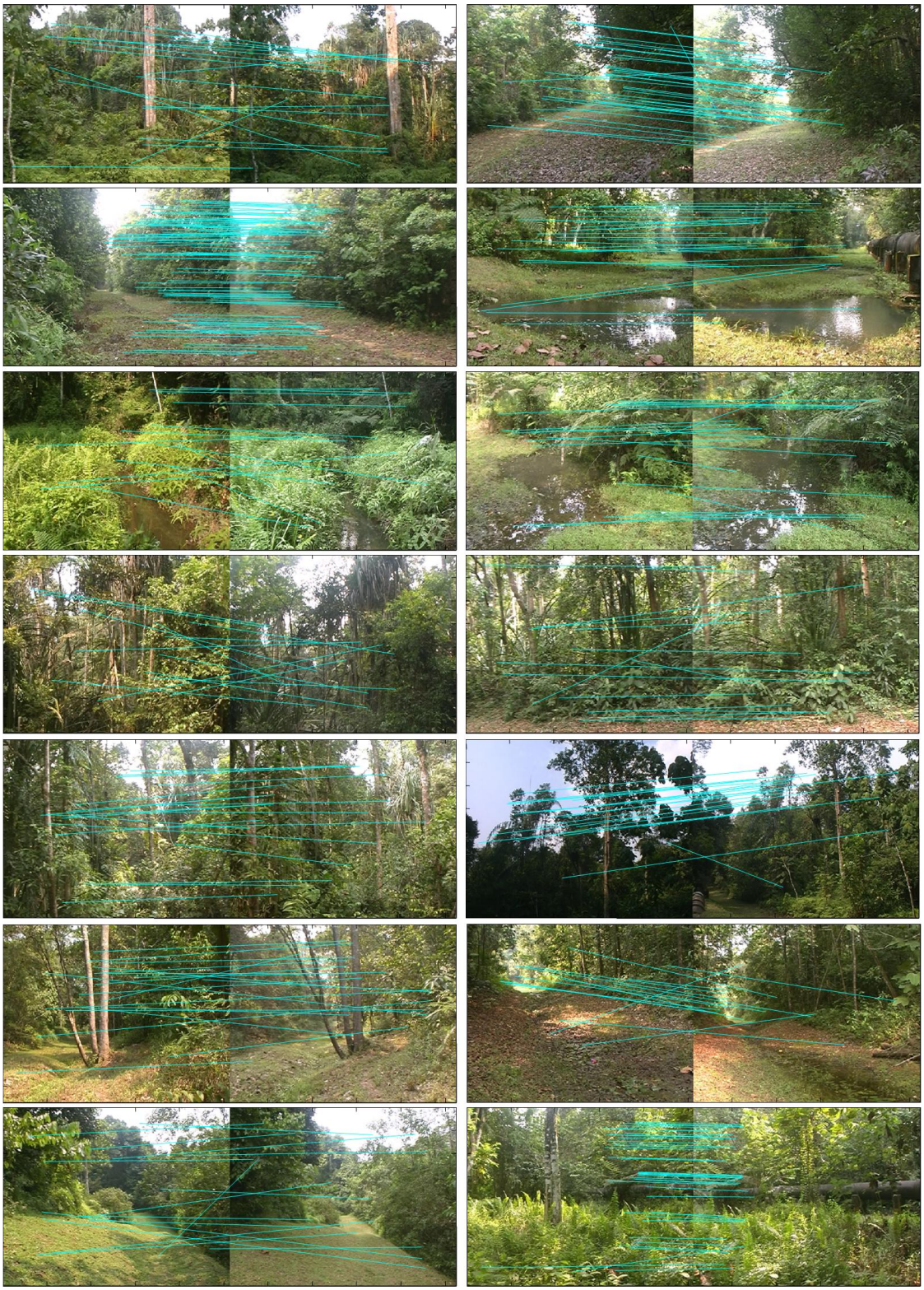

Figure D.5: NS database matches.

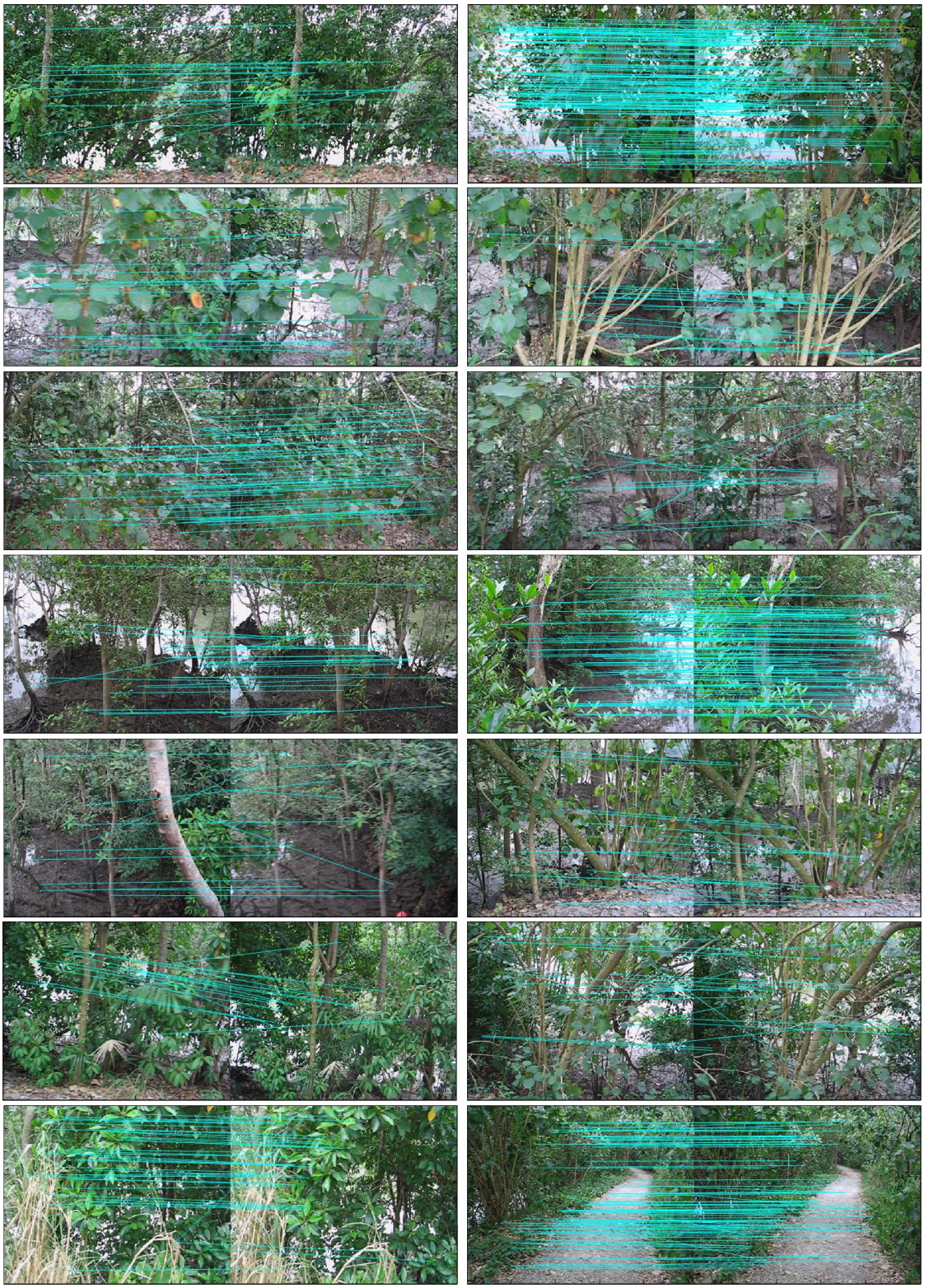

Figure D.6: SBWR database matches.