

ORDINAL DEPTH FROM SFM AND ITS APPLICATION IN
ROBUST SCENE RECOGNITION

Li Shimiao

NATIONAL UNIVERSITY OF SINGAPORE

2009

ORDINAL DEPTH FROM SFM AND ITS APPLICATION IN
ROBUST SCENE RECOGNITION

Li Shimiao

(B.Eng. Dalian University of Technology)

A THESIS SUBMITTED
FOR THE DEGREE OF DOCTOR OF PHILOSOPHY
DEPT. ELECTRICAL AND COMPUTER ENGINEERING
NATIONAL UNIVERSITY OF SINGAPORE

2009

Acknowledgments

First of all I would like to express my sincere gratitude to my thesis supervisor, Professor Cheong Loong Fah for his valuable advices, constant support and encouragement through out the years.

I would also like to thank Mr. Teo Ching Lik for our good collaboration.

I am grateful to Professor Tan Chew Lim for his understanding and support during the last one year.

Thanks to all my colleagues in Vision and Image Processing Lab for their sharing of ideas, help and friendship. Many thanks to Mr. Francis Hoon, our lab technician, for providing me with all the technical facilities during the years.

Finally, my special thanks to my parents and Dat, for their encouragement, support, love and sacrifices in making this thesis possible.

Abstract

Ordinal Depth from SFM and Its Application in Robust Scene Recognition

Li Shimiao

Under the purposive vision paradigm, visual data sensing, space representation and visual processing are task driven. Visual information in this paradigm can be weak or qualitative as long as it successfully subserves some vision task, but it should be easy and robust to recover.

In this thesis, we propose the qualitative structure information - *ordinal depth* as a computationally robust way to represent 3D geometry obtained from motion cues and in particular, advocate it as an informative and powerful component in the task of robust scene recognition.

The first part of this thesis analyzes the computational property of ordinal depth when being recovered from the motion cues and proposes an active camera control method - the biomimetic *TBL motion* as a strategy to robustly recover ordinal depth. This strategy is inspired by the behavior of insects from the order hymenoptera (bees and wasps). Specifically, we investigate the resolution of the ordinal depth extracted via motion cues when facing errors in 3D motion estimates. It is found that although metric depth estimates are inaccurate, ordinal depth can still be discerned reliably if the physical depth difference is beyond a certain discrimination threshold. Findings in this part of our work suggest that accurate knowledge of qualitative 3D structure can be ensured in a relatively small local image neighborhood and that resolution of ordinal depth decreases as the visual angle between points increases. Findings

also advocate camera lateral motion as a robust way to recovery ordinal depth.

The second part of this thesis proposes a scene recognition strategy that integrates the appearance-based local SURF features and the geometry-based *3D ordinal constraint* to recognize different views of a scene, possibly under different illumination and subject to various dynamic changes common in natural scenes.

Ordinal depth information provides the crucial 3D information when dealing with outdoor scenes with large depth relief, and helps to distinguish ambiguous scenes with repeated local image features. In our investigation, geometrical ordinal relations of landmark feature points in each of the three dimensions are found to complement each other under different types of camera movements and with different types of scene structures. Based on these insights, we propose the *3D ordinal space representation* and put forth a scheme to measure similarities among two scenes represented in this way. This leads us to a novel scene recognition algorithm which combines appearance information and geometrical information together.

We carried out extensive scene recognition testing over four sets of scene databases, consisting mainly of outdoor natural images with significant view-point changes, illumination changes and moderate changes in scene content over time. The results show that our scene recognition strategy outperforms other algorithms that are based purely on visual appearance or exploit global or semi-local geometrical transformations such as epipolar constraint or affine constraint.

Table of Contents

Acknowledgments	i
Abstract	ii
List of Tables	ix
List of Figures	xii
1 Introduction	1
1.1 What is This Thesis About?	1
1.2 Space Representation and Computational Limitation of Shape from X	4
1.3 What Can Human Visual System Tell Us?	5
1.4 Purposive Paradigm, Active Vision and Qualitative Vision . .	6
1.5 Ordinal Depth	8
1.6 Turn-Back-and-Look(TBL) Motion	8
1.7 Scene Recognition	9
1.8 Contribution of the Thesis	10
1.9 Thesis Organization	14

2	Resolving Ordinal Depth in SFM	16
2.1	Overview	16
2.2	Related Works	19
2.2.1	The Structure from Motion (SFM) Problem	19
2.2.2	Error Analysis of 3D Motion Estimation in SFM	20
2.2.3	Analysis of 3D Structure Distortion in SFM	21
2.2.4	Ordinal Depth Information: Psychophysical Insights	23
2.3	Depth from Motion and its Distortion : A General Model	24
2.4	Estimation of Ordinal Depth Relation	27
2.4.1	Ordinal Depth Estimator	27
2.4.2	Valid Ordinal Depth (VOD) Condition and VOD Inequality	28
2.5	Resolving Ordinal Depth under Weak-perspective Projection	30
2.5.1	Depth Recovery and Its Distortion under Orthographic or Weak-perspective Projection	30
2.5.2	VOD Inequality under Weak-perspective Projection	32
2.5.3	Ordinal Depth Resolution and Discrimination Threshold(DT)	33
2.5.4	VOD Function and VOD Region	33
2.5.5	Ordinal Depth Resolution and Visual Angle	34
2.5.6	VOD Reliability	36
2.6	Resolving Ordinal Depth under Perspective Projection	38
2.6.1	The Pure Lateral Motion Case	39
2.6.2	Adding Forward Motion: The Influence of FOE	41
2.7	Discussion	43
2.7.1	Practical Implications	43

2.7.2	Psychophysical and Biological Implication	44
2.8	Summary	45
3	Robust Acquisition of Ordinal Depth using Turn-Back-and- Look (TBL) Motion	47
3.1	Background	47
3.1.1	Turn-Back-and-Look (TBL) Behavior and Zig-Zag Flight	47
3.1.2	Why TBL Motion Is Performed?	49
3.1.3	Active Camera Control and TBL Motion	49
3.2	Recovery of Ordinal Depth using TBL Motion	51
3.2.1	Camera TBL motion	51
3.2.2	Gross Ego-motion Estimation and Ordinal Depth Recovery	52
3.3	Dealing With Negative Depth Value	54
3.4	Experimental Results	55
3.5	Summary	56
4	Robust Scene Recognition Using 3D Ordinal Constraint	58
4.1	Background	58
4.1.1	2D vs 3D Scene Recognition	60
4.1.2	Revisiting 3D Representation	64
4.1.3	Organization of this Chapter	65
4.2	3D Ordinal Space Representation	65
4.3	Robustness of Ordinal Depth Recovery	67
4.4	Stability of Pairwise Ordinal Relations under Viewpoint Change	68
4.4.1	Changes to Pairwise Ordinal Depth Relations	68
4.4.2	Changes to Pairwise Ordinal x and y Relations	72
4.4.3	Summary of Effects of Viewpoint Changes	75

4.5	Geometrical Similarity between Two 3D Ordinal Spaces	77
4.5.1	Kendall's τ and Rank Correlation Coefficient	77
4.5.2	Weighting of Individual Pairs	82
4.6	Robust Scene Recognition	85
4.6.1	Salient Point Selection	86
4.6.2	Encoding the Appearance and Geometry of the Salient Points	89
4.6.3	Measuring Scene Similarity and Recognition Decision	91
4.7	Summary	93
5	Robust Scene Recognition: the Experiment	95
5.1	Experimental Setup	95
5.1.1	Database IND	96
5.1.2	Database UBIN	97
5.1.3	Database NS	101
5.1.4	Database SBWR	101
5.2	Experimental Results	103
5.2.1	Recognition Performance and Comparison	103
5.2.2	Component Evaluation and Discussions	104
5.3	Summary	115
6	Future Work and Conclusion	118
6.1	Future Work Directions	118
6.1.1	Space Representation: Further Studies	118
6.1.2	Scene Recognition and SLAM	119
6.1.3	Ordinal Distance Information for 3D Object Classification	119
6.2	Conclusion	124

A Acronyms	127
B Author's Publications	128
Bibliography	129

List of Tables

2.1	DT values for different visual angles under different translation-to-rotation ratio h . $\bar{Z} = 100m$ and $p_e = 5\%$	44
4.1	Invariant properties of ordinal relations in x , y , and Z dimensions to different types of camera movements and in different types of scenes. It can be seen that different dimensions complement each other.	76
5.1	Description of the four databases used in the experiments . . .	96
5.2	Rank correlation coefficient in the x , y , and Z dimensions for two types of scenes. 1 and 2: locally planar or largely fronto-parallel scenes. 3 and 4: in-depth scenes.	109

- 5.3 The comparison between local appearance matching and overall geometrical consistency: positive test example 1. The top left image pair represents the correspondences between the test and its correct reference scene; the middle left image pair represents the correspondences between the test and the best of the remaining reference scenes (wrong reference scene); the top right pair and middle right pair represent the correspondences left after pruning by the epipolar constraint (RANSAC is used); the bottom table shows the detailed values of $\frac{N_{match}}{N_{tot}}$, τ_{3D} , G , and $\frac{N_{match}}{N_{tot}}$ after pruning by the epipolar constraint (RANSAC). . . 110
- 5.4 The comparison between local appearance matching and overall geometrical consistency: positive test example 2. The top left image pair represents the correspondences between the test and its correct reference scene; the middle left image pair represents the correspondences between the test and the best of the remaining reference scenes (wrong reference scene); the top right pair and middle right pair represent the correspondences left after pruning by the epipolar constraint (RANSAC is used); the bottom table shows the detailed values of $\frac{N_{match}}{N_{tot}}$, τ_{3D} , G , and $\frac{N_{match}}{N_{tot}}$ after pruning by the epipolar constraint (RANSAC). . . 111

- 5.5 The comparison between local appearance matching and overall geometrical consistency: positive test example 3. The top left image pair represents the correspondences between the test and its correct reference scene; the middle left image pair represents the correspondences between the test and the best of the remaining reference scenes (wrong reference scene); the top right pair and middle right pair represent the correspondences left after pruning by the epipolar constraint (RANSAC is used); the bottom table shows the detailed values of $\frac{N_{match}}{N_{tot}}$, τ_{3D} , G , and $\frac{N_{match}}{N_{tot}}$ after pruning by the epipolar constraint (RANSAC). 112
- 5.6 Some negative test examples which have high $\frac{N_{match}}{N_{tot}}$ value with some reference scenes. The correspondences and the actual values of $\frac{N_{match}}{N_{tot}}$, τ_{3D} , G between the test and the reference scene are shown. 117

List of Figures

- 2.1 Realization of the distortion maps A , B under perspective projection, iso-a contour, iso-b contour are shown. Motion parameters are: focus of expansion (FOE) $(x_0, y_0) = (26, 30.5)$, rotation velocity $\alpha = 0.005, \beta = 0.004, \gamma = 0.0002$. Error in FOE estimates: $(x_{0e}, y_{0e}) = (8, 9)$, error in rotation: $\alpha_e = 0.001, \beta_e = 0.001, \gamma_e = 0.00005$. Focal length: 50 pixels, FOV= 90° , epipolar reconstruction scheme was adopted ($\mathbf{n} = \frac{\hat{\mathbf{d}}}{\|\hat{\mathbf{d}}\|}$), blue * indicates the true FOE, red * indicates the estimated FOE. . 27
- 2.2 Realization of VOD region of $\mathbf{p}_0 = (\mathbf{0}, \mathbf{0})^T$ (\mathbf{p}_0 is denoted by the red asterisk) for different DT under weak-perspective projection. VOD region is bounded by black lines. The big red circles show the width of the region bands. τ is the visual angle between points on the circle and \mathbf{p}_0 . The rainbow at the background shows the change of distortion factor b . . Motion parameters and errors: $\mathbf{T} = (0.81, 0.2, 0.15)^T$, $\Omega = (0.008, 0.009, 0.0001)$, $\bar{Z} = 35000$, $\delta = -4.2857e - 006$, $\phi_e = 28.6^\circ$, $\delta_e = 1.0e - 006$, $\gamma_e = 1.0e - 006$, $\dot{\mathbf{p}}_n = 0$, $f = 250$. 35

- 2.3 Top: VOD Reliability of image points w.r.t. the image center for $DT = 100$ at $\bar{Z} = 35000$. Bottom: VOD Reliability of image points w.r.t. the image center for different DT at $\bar{Z} = 35000$ as visual angle ($^\circ$) between the point pair changes. $(U, V, W) = (0.001, 0.002, 0.001)$, $(\alpha, \beta, \gamma) = (0.004, 0.002, 0.003)$ 37
- 2.4 Realization of VOD region of $\mathbf{p}_0 = (\mathbf{0}, \mathbf{0})^T$ (denoted by red cross) for different DT under perspective projection and pure lateral motion. Top: second order flow ignored. Bottom: second-order flow considered. The VOD region is bounded by black lines. The background rainbow shows the change of distortion factor b . Motion parameters and errors are: $\mathbf{T} = (18, 22, 0)^T$, $\mathbf{T}_e = (15.3, 24.5, 0)^T$, (translation direction estimation error is -7.3°), $\Omega_e = (0.00002, 0.00002, 0.00005)$, $\bar{Z} = 20000$, $\dot{\mathbf{p}}_n = 0$, $f = 250$ 40
- 2.5 Realization of VOD region of $\mathbf{p}_0 = (\mathbf{0}, \mathbf{0})^T$ (denoted by red cross) for different DT under perspective projection with forward translation added to the motion configuration shown in Figure 2.4. Top: $\mu = 15^\circ$. Bottom: $\mu = 25^\circ$. $\mu_e = 0$ in both cases. Only first-order optical flow is considered for the illustration. 42
- 3.1 The Zig-Zag flight of a wasp directed towards a target (large circle) as seen from above [112]. Notice the significant translational motion that is almost perpendicular to the target at each arc formed. The complete path is shown on the right. 48
- 3.2 Simple camera TBL motion 52

3.3	Recovered ordinal depth of feature points in indoor and outdoor scenes, depicted using the rainbow color coding scheme (red stands for near depth; violet for far depth). Gross 3D motion estimates $(\hat{\phi}, \hat{\alpha}, \hat{\beta}, \hat{\gamma}, \hat{f})$ are shown under each image.	57
4.1	SIFT matching in the natural environment. Top: SIFT matches between two different views of a scene. Bottom: SIFT matches between images of two different scenes. The same matching threshold is used for both examples.	61
4.2	Examples of scenes with large depth discontinuities on which 2D-geometry-enhanced approach may fail and 3D method is required.	63
4.3	Landmark rank (based on the x -coordinate) remains unchanged under small viewpoint change.	67
4.4	Pairwise ordinal depth relation varies as optical axis direction changes.	69
4.5	Pairwise ordinal depth relation under camera rotation around the Y -axis. The figure is the projection of the scene in Figure 4.4 onto the XCZ plane. Forbidden orientation zone for $C'Z'$ is indicated by the shaded region that passes through C' . For feature pair that are almost fronto-parallel, like \mathbf{P}_i and \mathbf{P}_j when viewed from C , small camera rotation around the Y axis may cause the line of sight $C'Z'$ at the new viewpoint to cross into the forbidden zone.	70

4.6	Pairwise x relation under camera translation in the XCZ plane is preserved as long as C' does not enter the forbidden zone, which is the half space indicated by the shaded region. $Dist_X$ is the shortest camera translation that will bring about the crossing of this half space.	72
4.7	The range image of a forest scene from the Brown range image database. Intensity represents distance values, with distant object looking brighter.	80
4.8	Computing RCC on the Brown range data (forest scene): different RCCs across the views are shown when the camera undergoes different types of movements. The top, middle and bottom left figures correspond to translations in the X , Y , and the Z direction respectively, whereas the top, middle and bottom right figures correspond to rotations around the X , Y , and the Z direction respectively. The horizontal axis in each plot represents the various view positions (view 0-9) as the camera moves away from the original position.	81
4.9	Grayscale(top row) and saturation(bottom row) for the same scene taken under different illumination conditions.	87
4.10	An example outdoor scene with its sky-ground segmentation (top right), detected skyline (bottom left) and the resulting saliency map (bottom right).	88
4.11	Steps that describe the various stages of extracting the salient ROIs using various image morphological operations. The initial saliency map is extracted based on a down-sampled image. The final salient ROIs are boxed in white and highlighted in green.	90

4.12	Ordinal depths extracted from gross depth estimates under TBL motion, depicted using the rainbow color coding scheme (red stands for near depth; violet for far depth).	92
5.1	Reference scenes in the IND database.	97
5.2	Reference scenes in the UBIN database.	98
5.3	Reference scenes in the NS database.	99
5.4	Various challenging positive test scenes and reference scenes from the four databases.	100
5.5	Reference scenes in the SBWR database.	102
5.6	Comparison of the proposed SRS(SURF + 3D ordinal constraint), the SURF only method, the SURF + Epipolar Constraint (RANSAC) method, and the SURF + affine constraint method.	105
5.7	Successfully recognized positive test scenes (right image in each subfigure) and their respective reference matches (left image in each subfigure), despite substantial viewpoint changes, natural dynamic scene changes or illumination changes.	106
5.8	More successfully recognized positive test scenes (right image in each subfigure) and their respective reference matches (left image in each subfigure), despite substantial viewpoint changes, natural dynamic scene changes or illumination changes.	107
5.9	Component evaluation: comparing the 3D weighted scheme, the 2D weighted scheme, and the 3D unweighted scheme over the four databases.	108

5.10	Separation of the positive test set and the negative test set in IND and NS databases (respectively the four rows). Left column: histogram of the SURF matching percentage ($P = \frac{N_{match}}{N_{tot}}$) for both the positive and the negative test set. Right column: histogram of the global scene correlation coefficient (G) for both the positive and the negative test set. For positive test scene, P and G are the values between the test scene and its correct reference scene; for negative test scene, P and G are the biggest values obtained when the test scene is compared with all the reference scenes.	114
5.11	Separation of the positive test set and the negative test set in UBIN and SBWR databases (respectively the four rows). Left column: histogram of the SURF matching percentage ($P = \frac{N_{match}}{N_{tot}}$) for both the positive and the negative test set. Right column: histogram of the global scene correlation coefficient (G) for both the positive and the negative test set. For positive test scene, P and G are the values between the test scene and its correct reference scene; for negative test scene, P and G are the biggest values obtained when the test scene is compared with all the reference scenes.	115
6.1	Images of models of tables and planes	121
6.2	Sampling with different number of vertices.	121
6.3	Rank proximity matrices of table models, computed from 343 sampled vertices.	122

6.4	Rank proximity matrices of plane models, computed from 343 sampled vertices.	123
6.5	Rank proximity matrices with different number of sampled vertices. Upper row: table class, lower row: plane class. Sample number increases from left to right (as shown in Figure 6.2). .	124

Chapter 1

Introduction

1.1 What is This Thesis About?

3D reconstruction has been a key problem since the emergence of the computer vision field. Marr and Poggio founded the theory of computational vision [65, 66, 64]. According to this theory, 3D representation of the physical world can be built through three description levels from 2D images [64]. This was applied to the shape from X problems, which aim to reconstruct full 3D structure from its projections on 2D images using various visual cues such as texture, shading, stereo, motion etc. It is believed by the Marr school that reconstructing an internal representation of the physical world is a prerequisite for carrying out any vision tasks [32]. However, despite the many ensuing efforts on 3D reconstruction since the 1980s, it was found that the shape from X(SFX) problems are ill-posed or very difficult to solve computationally [115, 29]. Accurate and robust 3D reconstruction from 2D images seems to be infeasible in practice. Even low-level or mid-level representation is

very difficult to construct accurately. Thus Marr’s paradigm does not lead to many successful robotic vision applications such as recognition and navigation.

Probably due to the difficulty of 3D reconstruction, researchers have been seeking alternative approaches to fulfil vision tasks without geometrical reconstruction. In image based object recognition task, 2D local feature descriptors, which encode the local visual appearance information, have been the mainstay since the late 1990s and have been proven to be successful, especially with the recent development of locally invariant descriptors [62, 9]. In spite of the success, the visual appearance information encoded by the descriptors may change significantly when camera has large viewpoint change or when the lighting condition changes. This limits the power of these 2D local descriptor methods. To overcome the limitation, the visual appearance information is often combined with geometrical constraints so as to enhance the discriminating power of the local feature descriptors [6, 18, 34, 73, 91, 88]. 2D geometrical constraints [6, 18, 34, 73, 91], due to the assumption on the scene structure they are based on, are always restricted to certain types of objects or scenes. Therefore, 3D geometrical information is again required in robust recognition tasks. However, to combine 3D geometrical information with 2D visual appearance information, we again face the difficulties encountered in the 3D reconstruction problem.

Contrasting Marr’s paradigm of general-purpose reconstruction is the purposive vision paradigm [3, 98]. Its main tenet is that if we consider the specific vision task we are dealing with, e.g. the recognition task, the situation can be simplified [86]. Instead of seeking a solution for full 3D reconstruction following Marr’s paradigm, we may look for some weak or qualitative 3D geometrical information useful for the recognition task and at the same time, can

be recovered in an easy and robust way from some visual cues.

This thesis aims at finding such robust and useful geometrical information for vision tasks. We aim to answer the following questions.

- Although the reconstructed 3D structure may in general be inaccurate due to the computational difficulties in shape from X, can we still extract some valid and useful geometrical information from the inaccurate structures?
- How to acquire such geometrical information in a simple and robust way?
- How to use such geometrical information in practical vision tasks?

Specifically, in this thesis, we propose the qualitative structure information - *ordinal depth*¹ as a computationally robust way to represent 3D geometry in shape/structure from motion problem and advocate it as a powerful component in the robust scene recognition task.

The first part of this thesis answers the question "How to recover ", specifically, we analyze ordinal depth's computational properties when being recovered from the motion cues. Based on these properties, we propose a simple way called *TBL motion*, which is inspired from the behavior of biological insects, to recover ordinal depth robustly. The second part answers the question "How to use". The invariance properties of ordinal depth w.r.t. camera viewpoint change are analyzed. Based on these insights, we propose the *3D ordinal space representation*. Finally, we design a strategy to exploit the 3D ordinal space

¹By ordinal depth, we mean the order of the distances of points in the physical world to the observer or camera along the optical axis direction.

representation successfully in the robust scene recognition task, especially in the outdoor natural scene environment.

The remainder of this chapter is organized as follows. In Section 1.2 to Section 1.7, we give brief accounts to the various background topics relevant to this thesis. Section 1.8 presents a summary of the key contributions of the thesis. Finally, Section 1.9 presents the organization of the thesis.

1.2 Space Representation and Computational Limitation of Shape from X

Marr’s paradigm aims at recovering metric representation of the space. However, techniques of shape from X for this purpose suffer from noise in image measurements and errors in the computation stages. Taking the structure from motion problem for example, small noise in image velocity measurements can lead the algorithm to very different solutions. In spite of the many algorithms proposed for structure from motion, we still lack methods robust to noise in image velocities, and errors in motion estimates or calibration parameters. Error analysis of this problem shows that there are inherent ambiguities in the motion estimation and calibration stage which may cause severe 3D structure distortions [29, 21, 119]. Similar problems exist in shape recovery from other visual cues [31, 67].

In a vision system, the geometrical information conveyed in a *3D space representation*² is usually computed by some 3D reconstruction technique.

²By space representation, we mean the way geometrical information of the physical world structure is described in any vision system.

However, due to the ill-conditioned and noise sensitive nature of shape from X, the robustness of this computation should be given a careful evaluation, especially for vision tasks requiring robust performance.

In this thesis, we present a comprehensive analysis on the computational robustness of structure from motion algorithms to recover the ordinal depth information. The insights obtained from this analysis serve as guidelines for ordinal depth to be exploited in the robust scene recognition task.

1.3 What Can Human Visual System Tell Us?

To find a proper space representation suitable for a wide range of vision tasks, researchers in cognition and psychophysics have been referring to one of the most powerful vision systems present in nature - the human vision system. Many studies were carried out exploring the properties of space representation in human visual system. It is believed by most researchers that the representation is anything but Euclidean [106, 50, 38]. This may indicate that human perception of space is metrically imprecise.

Studies have also been carried out on how humans measure distances in space. Some psychophysical experiments were designed to test observers' judgement on interval and ordinal depth measurements [100, 107, 76, 28]. Results show that human are good at judging the weaker measurements such as the ordinal measurement. It was suggested that human vision might only perceive ordinal distance information from sparse points in the space, and as the number of points increases, metric information could be recovered from dense ordinal measurements using methods like multi-dimensional scaling [28]. Therefore, it seems that qualitative geometry information might be a key step

towards finding a proper space representation.

Studies also show that human visual attention changes as subjects are asked to perform different vision tasks [120, 117]. This shows that the visual data acquisition process is purposively and actively controlled, rather than being a passively general process. It implies that vision might be a task-driven process and thus, geometrical information recovered by SFX could also vary with different tasks.

Inspired by the above findings from human visual system, this thesis focuses on understanding the qualitative geometry information, that is, the computational properties and practical application of ordinal depth. We also propose a bio-inspired strategy for active acquisition of such geometrical information.

1.4 Purposive Paradigm, Active Vision and Qualitative Vision

The fundamental difference between Marr’s reconstruction paradigm and the purposive paradigm [3, 103] lies in the way they see the final goal of vision. According to the panel discussion in [13], from the view of the *reconstruction paradigm*, the goal of vision is:

- “*The description of three dimensional world in terms of the surfaces and objects present and their physical properties and spatial relationships.*”

while from the view of the *purposive paradigm*, the goal of vision is:

- “*The development of fast visual abilities which are tied to specific behaviors and which access the scene directly without intervening representations.*”

In traditional reconstruction paradigm, reconstruction is a task-independent process and is therefore general-purpose. On the other hand, the purposive paradigm is task-driven. In the purposive paradigm, data acquisition, space representation, and the 3D geometry information needed all become task-oriented.

Data acquisition often becomes an active process in the purposive paradigm. For example, the eye (or camera) movement can be actively controlled depending on the information the agent needs for performing the current task and the status of current scene interpretation. Such data acquisition strategy is known as the *active vision* paradigm [4, 3, 8].

In another aspect, space representation and 3D reconstruction in the purposive paradigm are used to subserve specific task performing. Only geometry information needed for the robust performance of the current task is to be represented and constructed. Such geometry information can be imprecise or even qualitative in nature; this is in contrast to metric 3D reconstruction in the reconstruction paradigm. If only the qualitative description of the physical world is needed for some specific task, the system is said to subscribe to the *qualitative vision* paradigm [5, 36]. Qualitative information exhibits greater invariance to the various factors in vision system such as viewpoint or illumination changes, and noise in data acquisition. It is hoped that qualitative vision, if proven to be adequate for some specific task, would have more robust performance than the traditional quantitative system.

In this thesis, we develop a recognition system for individual scene identification. Our system subscribes to the active vision and qualitative vision paradigms. We use controlled camera movement though not requiring precise camera motion to robustly recover the qualitative ordinal depth information.

Using ordinal depth, we develop the *3D ordinal space representation* which only encodes the ordinal spatial information and couple it successfully to the task of scene recognition.

1.5 Ordinal Depth

Being the simplest qualitative description of the third dimension of the physical world, ordinal depth measures the order of the distances of 3D points to the observer along the viewing direction. Due to its qualitative nature, ordinal depth information is robust to noise and errors in shape from X [23]. It was proposed as one of the qualitative structures that can be used in active vision [36]. However firstly, the computational capability of shape from X algorithms to judge ordinal depth under different resolutions of depth variation has not been well analyzed. Secondly, the power of the ordinal depth information has not been well demonstrated in practical vision tasks.

Ordinal depth is the focus of this thesis. In this thesis, we will gain more understanding towards this qualitative geometry information, specifically, its computational properties and practical application. This thesis put ordinal depth into the proposed *3D ordinal space representation* and show how ordinal depth complements spatial information in the other two dimensions under different types of camera viewpoint changes.

1.6 Turn-Back-and-Look(TBL) Motion

In this thesis, we adopt an active data acquisition scheme which can acquire the ordinal depths in a simple and robust manner. For this purpose, we pro-

pose the use of motion cues, motion being an omnipresent cue for a mobile agent navigating in the environment. As is well-known, structure from motion analysis is sensitive to noise [29]. However, Cheong and Xiang [23] showed that for a certain kind of generic motions, the recovered depths preserve their depth relief, despite the gross egomotion estimates. Such motion consisting of a lateral translation plus a rotation is referred to as a lateral motion. The analysis and experiments in this thesis will further advocate lateral motion as a robust way to recover ordinal depth.

The ecological relevance of lateral motion is underlined by the prevalence of lateral motion used by different animals in nature to appreciate distances [112]. In the case of bees and wasps, this type of motion is known as zig-zag flights in Turn-Back-and-Look (TBL) behavior. In this thesis, we call such flight the *Turn-Back-and-Look (TBL) motion*. It was believed [24, 122] that TBL is important for the bees to recognize these scenes on their return trip. In our proposed scheme, camera performs a roughly controlled TBL motion to actively recover the ordinal depths.

1.7 Scene Recognition

Scene recognition is to recognize a specific location that has been previously visited. This is in contrast to the problem of scene classification or scene categorization (e.g. [81]) which recognizes scene class. Knowing where I am is important to visual navigation [7, 16, 27, 45, 57, 85, 92, 97, 110, 116], for instance, in relation to the SLAM loop closing problem, or to various emerging applications stemming from large scale image databases of the world [40, 90]. In the domain of biomimetic navigation, it also forms an integral component of what

is known as the place recognition-triggered response [110] — the biological agent has a set of places in memory that is linked with a learnt set of actions that it must take once it recognizes that it has returned to the same place again. Compared to object recognition, robust scene recognition (especially outdoor natural scene recognition) requires algorithms that are able to deal with large viewpoint change, illumination change, and natural dynamic change of the scene itself.

This thesis tackles indoor and outdoor scene recognition problem and shows that the proposed *3D ordinal space representation* is a robust geometry descriptor adequate for this vision task. We have also built up indoor and outdoor databases, which contain extensive sets of scenes with complex changing effects between reference scene and test scene.

1.8 Contribution of the Thesis

The major contributions of this thesis are summarized as follows:

Computational properties of ordinal depth in structure from motion:

We investigate the resolution of the ordinal depth extracted via motion cues in the perceived visual space, which is distorted from the physical space due to errors in the motion estimates. It is found that although metric depth estimates are inaccurate, ordinal depth can still be discerned reliably if physical metric depth difference is beyond a certain discrimination threshold. Moreover, the resolution level of discernible ordinal depth decreases as the image distance or visual angle between the point pairs increases. Ordinal depth resolution also decreases as points receding from the camera or as the speed of the motion com-

ponent carrying depth information decreases. Ordinal depth resolution also decreases as image region approaching the focus of expansion (FOE), which indicates the resolution can be high under camera lateral motion and provides theoretical support for using TBL motion to extract ordinal depth. Findings in this part of work suggest that accurate knowledge of qualitative 3D structure is ensured in a relatively small local image neighborhood. By fleshing out the computational properties of the qualitative visual space perception under estimation uncertainty, we hope to inspire future computational and psychophysical ventures into the study of visual space representation.

Scene recognition strategy: We put forth a scene recognition algorithm that is able to deal with both indoor and outdoor environments. In the current state of the art, outdoor natural environments without any man-made structures are deemed to be very challenging. Such scenes remain largely untouched by robotics and vision researchers due to the lack of distinguishable landmarks. Our scene recognition strategy is tested on four databases, consisting of one set for indoor environment and three for outdoor natural environments without man-made structures. As far as we are aware, they constitute the most extensive sets of outdoor scenes for specific scene recognition, covering a spatial extent much more extensive than those typically encountered in SLAM experiments, and containing much more complex illumination changes and viewpoint effects than those found in typical object recognition database. These changes will degrade the performance in methods using 2D local feature matching, even when enhanced with the epipolar or affine constraint [61], as

we show in the experiment. Nevertheless, our proposed algorithm exhibits good performance on all the four databases, demonstrating its accuracy and generality. While the visual appearance aspect of SLAM loop-closing [45,57,85,97] has common grounds with the work described here, given the large spatial extent encountered in our work, internal maps and vehicle estimates are apt to be in gross errors and hence not useful.

TBL motion for active ordinal depth acquisition: By using TBL motion scheme, ordinal depths can be obtained robustly in an active way, solely from a gross estimate of the motion parameters. It is thus stripped of the excess baggage of strict egomotion recovery, much faster, and more relevant for biological organisms in rapid motions without ample computational resources. TBL motion scheme also raises interesting questions about the actual role of TBL in insects during navigation. Some authors [56] have proposed the use of TBL to extract landmarks only, whereas others [56,99] suggested distance learning from such flights. However, in the latter works, either no computational details are forthcoming or restrictive conditions are required of the insect flights (e.g. translation only, in which case the recovery of relative depths is trivial). These works have overlooked the robustly obtainable ordinal depths, even in the presence of camera rotational perturbation.

3D ordinal space representation: We propose the use of weak 3D geometrical constraint based on an 3D ordinal representation of space. This constraint is combined with local feature descriptor for robust scene recognition. Compared to some recent works that exploit global rigid-

ity for 3D object recognition [15, 88] and scene recognition [57, 92, 97], we exploit the qualitative geometrical information for scene recognition. Computing 3D rigid transformation (or tensors among multiple views) is difficult because, as discussed previously, image appearance changes substantially under different illumination and different viewpoint especially in outdoor natural scenes, as well as due to the non-static nature of natural scenery over time, making local feature matching unreliable. Instead, we propose the *3D ordinal constraint* which uses correlation to verify the geometrical consistency between the test scene and reference scene, thus avoid the difficulty of computing transformation with numbers of outliers. Our weak geometrical characterization is similar in spirit to those works in 3D object problem [46, 52, 89], because both have to deal with variability in appearance. However, our task of specific scene recognition requires a much more powerful geometrical constraint than the qualitative constraints typically used in these works. For scene categorization and classification, [54, 96] exploit the 2D geometrical configuration of the image sub-regions characterized by their image feature statistics. Our proposed method not only adopt the 3D geometrical information, but we are also able to offer a robustness analysis of the 3D ordinal geometrical consistency with respect to viewpoint change and errors in the 3D reconstruction stage.

Invariance properties of ordinal depth w.r.t. viewpoint changes: The use of ordinal depths for vision tasks have been proposed by [36, 44, 114]. However, its invariance property with respect to viewpoint change have not been investigated. We carry out such analysis and show clearly that

3D ordinal measurements provide complementary information to those provided by 2D ordinal measurements in the image dimension, and are especially important for certain types of scenes and viewpoint changes. The analysis also furnishes a scheme which weighs the different pairwise ordinal relationships appropriately, depending on various factors such as image separation and separation in depth, so that they can be combined in a more optimal way.

1.9 Thesis Organization

The remainder of this thesis is organized as follows.

Chapter 2 gives an analytic analysis of the resolution of ordinal depth recovered from motion cues when facing errors in 3D motion estimates. Detailed analysis is carried out under orthographic/weak-perspective camera and perspective camera. In particular, lateral motion and forward motion cases are discussed.

In Chapter 3, an active camera control method - TBL motion is proposed for fast and robust acquisition of ordinal depth. A simple yet effective algorithm is designed and tested.

Chapter 4 presents a strategy to use ordinal depth in performing scene recognition task. Firstly, we propose the *3D ordinal space representation*. Secondly, invariance properties of geometrical entities in this space w.r.t. camera viewpoint changes are analyzed; a similarity measure based on these properties is developed. Thirdly, we develop a scene recognition scheme which successfully combines the geometrical information in 3D ordinal space with the appearance information encoded by SURF feature descriptors.

Chapter 5 gives extensive experimental testing results on the proposed scene recognition strategy. These testings are carried out on databases of indoor and outdoor natural scenes, with various changing effects. The proposed method is compared with methods based on global and semi-local transformations. Evaluation of various components of the proposed system is also provided.

Chapter 6 gives some brief proposals of future work directions and the conclusion of this thesis.

Chapter 2

Resolving Ordinal Depth in SFM

2.1 Overview

The shape/structure from motion (SFM) problem, which is to recover 3D structure from motion cues in 2D images, has attracted many concerns in the last two decades from researchers in the computer vision community and many other disciplines. Despite the large amount of algorithms proposed, the estimation of motion and structure is beset by the noise sensitivity problem. This has led to many error analyses trying to understand the behavior of the SFM algorithms in the presence of noise [2] [115] [29] [79]. These works have shown that some motion ambiguities are inherent and errors in the motion estimates are inevitable.

Since motion errors are inevitable, it is important to understand how the errors and noise may affect the recovered 3D structure information. A few works investigating this problem can be found in the literature [101] [21] [23]. It was shown that errors in motion estimates may cause severe systematic

distortion in the estimated depth and metrically accurate depth estimate is difficult to obtain [21].

However, despite the above works, there is still little understanding about the nature of the distorted perceived visual space. Are there any systematic laws governing the uncertainty of the recovered structure? Specifically, although the estimated metric depth might differ significantly from the physical value, can we still extract some valid and useful information of depth from these inaccurate estimates? Moreover, instead of recovering the depth of individual points, robustly recovering some information about the relative positions among points might be of more importance. Such information extracted may be of a less precise form, such as ordinal or interval depth measurement [100]. It may be qualitative rather than quantitative. It could be more robustly achieved than metric depth estimates and might suffice for many vision tasks such as navigation and recognition. Exploring such geometry information and its possible applications is important for developing vision systems that subscribe to the purposive vision paradigm [3].

In the computer vision literature, a qualitative description of depth is given in [5,36]. Qualitative depth representation such as ordinal depth map has been adopted for visual motion segmentation and independent motion detection tasks [36,60,77,78]. In the area of visual psychophysics, some psychophysical experiments were designed to test observers' judgement on interval and ordinal depth relations [107,76,51]. However, in spite of these works, the computational property of shape from X algorithms to resolve qualitative depth information from inaccurate metric depth estimates is as yet unknown. Such an understanding might provide us with better insight about the nature of the perceived visual space and shed light on a proper space representation whereby

structure information could be obtained robustly and applied to vision tasks.

In this chapter, we aim to investigate the resolution of the ordinal depth extracted via motion cues in the perceived visual space, which is distorted from the physical space due to errors in the motion estimates. Based on a general model describing how recovered depth is distorted by errors in the motion estimates, we derive a sufficient condition under which ordinal depth can be estimated validly. Then the condition is explored under orthographic/weak-perspective and perspective projection. Image regions that have valid ordinal depth estimates up to certain levels of resolution are delineated. By studying the geometry and statistics of these regions, we found that although metric depth estimates are inaccurate, ordinal depth can still be discerned reliably if the physical metric depth difference is beyond a certain discrimination threshold. Moreover, the resolution level of discernible ordinal depth decreases as the image distance or visual angle between the point pairs increases. Ordinal depth resolution also decreases as points recede from the camera (as average depth increases) or as the speed of the motion component carrying depth information decreases. These findings suggest that accurate knowledge of qualitative 3D structure is ensured in a relatively small local image neighborhood, which might account for biological foveated vision. By fleshing out the computational properties of on the qualitative visual space perception under estimation uncertainty, we hope to inspire future computational and psychophysical ventures into the study of visual space representations and their practical applications in vision systems. The findings in this chapter will be used as guidelines in developing ordinal depth recovery strategy and applying ordinal depth information in the scene recognition task in later chapters.

The remainder of this chapter is organized as follows. In Section 2.2, we

give a review of the relevant works. Section 2.3 describes depth recovery via motion and the associated distortion model. Section 2.4 presents the ordinal depth estimator and conditions for its validity (valid ordinal depth(VOD) condition). Section 2.5 investigates VOD condition under orthographic/weak-perspective projection and presents analytical results and delineated how various factors affect the resolution of discernible ordinal depth. Section 2.6 investigates VOD condition under perspective projection. Section 2.7 discusses possible implications. Section 2.8 presents a summary.

2.2 Related Works

2.2.1 The Structure from Motion (SFM) Problem

In computer vision, structure from motion(SFM) refers to the process of recovering 3D structure of object/scene from analyzing the image projection of the 3D relative motion between object/scene and the camera. Following Marr's reconstruction paradigm and shape from X studies, SFM became one of the central problems in computer vision since the early 1980s and has attracted much attention in the ensuing decades. The problem is normally divided into three subproblems: 1. the measurement of 2D displacement in the image; 2. the recovery of the 3D relative motion; 3. the reconstruction of the 3D structure. These three subproblems are usually solved in sequence. SFM algorithms can be categorized into two different approaches according to the two different ways of measuring the image displacement. The differential approach measures the 2D image velocities (optical flow), while the discrete approach measures the feature correspondences between the views. The discrete ap-

proach for SFM is also known as the *shape from stereo* problem. Our analysis in this chapter adopts the differential approach.

Early studies in SFM focused on proving that a unique solution exists. Algorithms were proposed for both the differential case [59, 113, 1, 42] and the discrete case [58, 111]. Most of these algorithms are based on the epipolar constraint, which relates 2D image displacement to 3D motion parameters based on the rigidity assumption, eliminating the unknown 3D structure from the computation. These early algorithms using the epipolar constraint have closed-form solution and can be solved linearly. Thus these algorithms are simple and easy to implement. Other methods include the factorization approach [109], the pattern recognition approach [35] etc. Reviews for SFM algorithms can be found in [68, 33, 79].

However, in practice, SFM algorithms face two problems: ambiguity and noise sensitivity. Firstly, ambiguity problem refers to the fact that for some special scene or motion configuration, more than one solution may exist, e.g. the camera is viewing a planar scene [2, 67, 101]. Secondly, because measuring 2D displacements from image intensities is an ill-conditioned problem, noise is inevitable in this process. These noisy measurements are taken as input in the 3D motion estimation stage. Error analysis of SFM studies the effects of such noise on the final 3D motion estimates and the recovered 3D structure.

2.2.2 Error Analysis of 3D Motion Estimation in SFM

To design robust practical SFM algorithms, error analysis of SFM has been carried out to understand the behavior of the algorithms with noisy input [42, 115, 68, 29, 37, 80, 119]. One major approach is to express the errors in 3D motion

estimates as bias or variance through statistical analysis [121, 115, 29]. Another approach is to characterize the topology of the cost function being minimized in solving SFM. [37, 80, 119]. Error analysis of 3D motion estimation has led to new optimization criteria and numerical methods in SFM [115, 48, 29, 123, 63]. Many of these works share common results regarding the noise sensitivity properties in 3D motion estimation. We briefly summarize these results below:

1. **Translation-rotation confounding:** When the field of view is small or depth variation in the scene is not sufficient, a rotation about an axis parallel to the image plane may easily be confounded with a lateral translation perpendicular to the axis.
2. **Bias towards the viewing direction:** The estimated translation tend to be biased towards the viewing direction if the cost function is not normalized properly.
3. **Bas-relief valley:** The plane defined by the true translation and the viewing direction can be estimated reliably by most algorithms. However, analysis on the topology of error surface found that there is a valley lying trough the image centroid and the true focus of expansion (FOE: the intersection of the translation vector with the image plane). The estimated FOE is likely to fall into this valley, especially for scenes with small depth variation. This is related to the well known bas-relief ambiguity in 3D structure reconstruction.

2.2.3 Analysis of 3D Structure Distortion in SFM

Error analysis in SFM show that small perturbations in the image measurement input may lead to erroneous solutions in 3D motion estimation. There-

fore, understanding the effects of errors in 3D motion estimates on 3D structure reconstruction is important.

The most well-known 3D structure distortion in SFM is the *bas-relief ambiguity* [50, 51, 101]. It refers to the confusion between the relative depth of objects and the amount of camera rotation (in orthographic cameras) or translation (in perspective camera). Bas-relief ambiguity causes shape distortion on the 3D structure with a bas-relief effect. Early studies on bas-relief ambiguity were carried out under orthographic cameras. It was later shown by Szeliski and Kang [101] that bas-relief ambiguity is significant even with many images under perspective projection. Besides, bas-relief ambiguity is also observed in shape from shading problem [12]. However, in spite of the shape distortion, it is worth noting that recovered depth is monotonically invariant to the true depth under bas-relief ambiguity.

To obtain further understanding towards how the perceived 3D structure is distorted by errors in motion estimates, Cheong developed the iso-distortion framework [21], which is a tool to systematically characterize the depth distortion. The framework was used to explain various phenomena in psychophysical vision, develop new strategy of independent motion detection, and analyze the robustness of shape recovery in SFM [20, 118]. Based on the iso-distortion framework, Cheong and Xiang [23] analyzed depth distortion under generic motions in both calibrated and uncalibrated cases. It was found that depth information is difficult to recover under forward motion case; while for lateral motion, although depth is difficult to recover, ordinal depth information is obtainable.

2.2.4 Ordinal Depth Information: Psychophysical Insights

Due to the distortion, perceived 3D structure is far from being accurate metrically. However, we may still extract some reliable qualitative information from this inaccurate quantitative structure. On the other hand, it has been suggested that some vision tasks can be performed on the basis of some qualitative structure information such as ordinal or topological relations among objects; thus explicit knowledge of Euclidean metric structure may not be required. To build real vision systems based on ordinal geometrical information, it is important to understand the role of such information in perceiving 3D space and performing vision tasks.

In Norman and Todd's study [76], psychophysical experiments were carried out to test human's judgement of relative depths of small probe dots using stereo cue. Discriminations of ordinal depth were found to be more precise than discriminations of depth intervals. Moreover, performance was higher when observers evaluated the depth relationships between nearby points in the projected images, and lower when the points were more widely separated. These findings indicate that the human visual system is good at measuring qualitative information (ordinal information), and that accurate knowledge of 3D structure is limited to small local neighborhood. In [107, 75], shading and texture cues were used. It was shown that relative depth judgments on smoothly curved surfaces were influenced by a monotonic depth change. This result suggests that our visual knowledge of smoothly curved surfaces can be defined in terms of local, nonmetric order relations. All the above works demonstrate the importance of ordinal depth information in human visual perception.

In [28], Cutting compared among human’s judgement of ordinal depth using nine different visual cues. Cutting suggested that the perceived spaces are really ordinal. However, as denser ordinal measurements are obtained, these spaces converge to a metric space. He further suggested that such convergence could possibly be realized using multidimensional scaling techniques. This hypothesis places ordinal geometrical information right in the center of 3D space perception.

2.3 Depth from Motion and its Distortion : A General Model

We introduce in this section a novel distortion model, which relates recovered depth to errors in 3D motion estimates. Compared to the iso-distortion framework [21] [23], the model proposed in this section is more general and can be adapted to other camera models besides the perspective model used in the iso-distortion framework.

In this chapter, we denote the estimated parameters with the hat symbol $\hat{\cdot}$ and errors in the estimated parameters with the subscript e . The error of any estimated parameter l is defined as $l_e = l - \hat{l}$. \mathbf{p}^\perp is the vector perpendicular to vector \mathbf{p} .

Generally, 2D image velocities $\dot{\mathbf{p}} = (\dot{p}_x, \dot{p}_y)^\mathbf{T}$ due to 3D rigid motion (translation $\mathbf{T} = (U, V, W)^\mathbf{T}$ and rotation $\mathbf{\Omega} = (\alpha, \beta, \gamma)^\mathbf{T}$) between the camera and the scene under any camera projection model can be written as

$$\dot{\mathbf{p}} = \mathbf{d}g(Z) + \dot{\mathbf{p}}_{indep} \tag{2.1}$$

where \mathbf{d} is the direction in which image velocity carries the depth information; we call this direction the *epipolar direction* since it is the direction of epipolar line in the differential case. $g(Z)$ is a monotonic function of depth $Z(Z > 0)$ and $\dot{\mathbf{p}}_{indep}$ is the component in image velocity independent of depth. Then the depth information of a scene point can be recovered up to a scale factor as

$$g(\hat{Z}) = \frac{(\tilde{\mathbf{p}} - \hat{\mathbf{p}}_{indep}) \cdot \mathbf{n}}{\hat{\mathbf{d}} \cdot \mathbf{n}} \quad (2.2)$$

where $\tilde{\mathbf{p}} = \dot{\mathbf{p}} + \dot{\mathbf{p}}_n$ is the measured image velocity. $\dot{\mathbf{p}}_n$ is the noise term in the optical flow measurement which is random and its distribution is up to the image formation process and the optical flow computation process. \mathbf{n} is a unit vector which specifies a direction. \mathbf{n} 's value depends on the approach we use to recover depth. For example, the epipolar reconstruction approach uses $\mathbf{n} = \hat{\mathbf{d}}$; while reconstruction from normal flow uses local image gradient direction as \mathbf{n} . Concrete forms of Equation (2.1) and (2.2) under weak-perspective and perspective camera projection models will be given in Section 2.5 and Section 2.6 respectively.

Due to errors in the motion estimates and noise in the optical flow measurements, actual estimate of depth information $g(\hat{Z})(\hat{Z} > 0)$ would be erroneous and can be readily shown to be related to the true $g(Z)$ as

$$g(\hat{Z}) = ag(Z) + b + c \quad (2.3)$$

where a , b and c are the distortion factors:

$$a = \frac{\mathbf{d} \cdot \mathbf{n}}{\hat{\mathbf{d}} \cdot \mathbf{n}} = 1 + \frac{\mathbf{d}_e \cdot \mathbf{n}}{\hat{\mathbf{d}} \cdot \mathbf{n}}, \quad b = \frac{\dot{\mathbf{p}}_{indep} \cdot \mathbf{n}}{\hat{\mathbf{d}} \cdot \mathbf{n}}, \quad c = \frac{\dot{\mathbf{p}}_n \cdot \mathbf{n}}{\hat{\mathbf{d}} \cdot \mathbf{n}} \quad (2.4)$$

\mathbf{d}_e and $\dot{\mathbf{p}}_{indep\ e}$ are both functions of the image coordinates and motion errors. The latter can be regarded as random variables whose distribution functions rely on the motion estimation process and motion-scene configurations. a , b and c are undefined when $\hat{\mathbf{d}} = \mathbf{0}$.

Equation (2.3) shows how the errors in the motion estimates and noise in the image measurements may distort the actual recovered depth. The error in the estimates of the epipolar direction \mathbf{d}_e causes a multiplicative distortion in $g(\hat{Z})$, while error in the estimates of the depth independent component $\dot{\mathbf{p}}_{indep\ e}$ and noise in the optical flow measurement $\dot{\mathbf{p}}_n$ result in additive distortions.

Note that a , b and c are functions of image coordinates. We denote them as $a_{i,j}$, $b_{i,j}$, $c_{i,j}$, where i, j are the indices of image pixels. Let matrices $\mathbf{A} = [a_{i,j}]$, $\mathbf{B} = [b_{i,j}]$, $\mathbf{C} = [c_{i,j}]$. We call \mathbf{A} , \mathbf{B} and \mathbf{C} the *distortion maps*, whose entries are random variables. In each depth recovery process from motion cues, there exist certain realizations of the distortion maps A , B and C . Figure 2.1 illustrates realizations of the distortion maps \mathbf{A} and \mathbf{B} given specific motion configurations and errors in the image velocities under perspective projection.

The depth distortion model described above can be applied into any camera projection model, such as orthographic, weak-perspective, perspective and catadioptric cameras. Under perspective projection, the distortion factors a , b and c in the model are related to the distortion factor D in [21] [23] by

$$D = \frac{1}{a+(b+c)Z}.$$

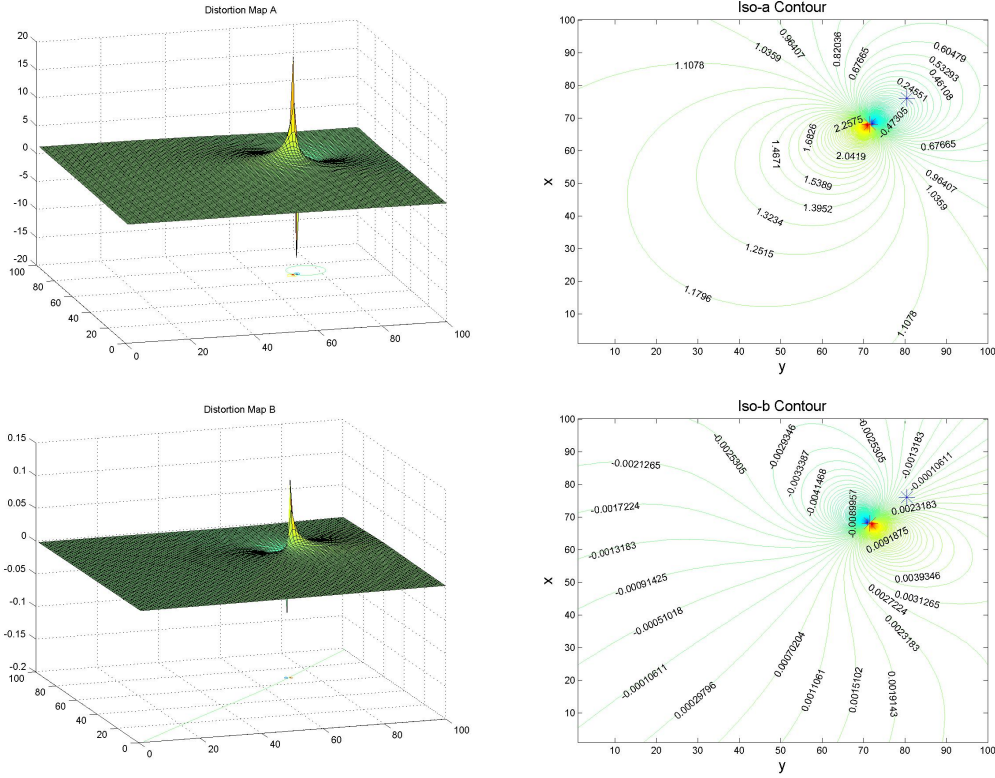


Figure 2.1: Realization of the distortion maps A , B under perspective projection, iso-a contour, iso-b contour are shown. Motion parameters are: focus of expansion (FOE) $(x_0, y_0) = (26, 30.5)$, rotation velocity $\alpha = 0.005, \beta = 0.004, \gamma = 0.0002$. Error in FOE estimates: $(x_{0e}, y_{0e}) = (8, 9)$, error in rotation: $\alpha_e = 0.001, \beta_e = 0.001, \gamma_e = 0.00005$. Focal length: 50 pixels, FOV= 90° , epipolar reconstruction scheme was adopted ($\mathbf{n} = \frac{\hat{\mathbf{a}}}{\|\hat{\mathbf{a}}\|}$), blue * indicates the true FOE, red * indicates the estimated FOE.

2.4 Estimation of Ordinal Depth Relation

2.4.1 Ordinal Depth Estimator

Suppose Z_0 and Z_1 are the depths of two scene points \mathbf{P}_0 and \mathbf{P}_1 , whose image points are \mathbf{p}_0 and \mathbf{p}_1 . We denote $g(Z_0)$ as g_0 , $g(Z_1)$ as g_1 . The function $\text{sgn}(g_0 - g_1)$, i.e. the sign of $(g_0 - g_1)$, reveals the ordinal depth relationship between points \mathbf{P}_0 and \mathbf{P}_1 , since $g(Z)$ is a monotonic function of depth

$Z(Z > 0)$. For example, given $g(Z) = \frac{1}{Z}$ which is the case under perspective projection, we have

$$\begin{cases} \text{sgn}(g_0 - g_1) = 1 & Z_0 < Z_1 \\ \text{sgn}(g_0 - g_1) = -1 & Z_0 > Z_1 \end{cases} \quad (2.5)$$

2.4.2 Valid Ordinal Depth (VOD) Condition and VOD Inequality

From Equation (2.3) we only have \hat{g}_0 and \hat{g}_1 at our disposal. Unfortunately, $\text{sgn}(\hat{g}_0 - \hat{g}_1)$ may not reveal the correct ordinal relation information, because $g(\hat{Z})$ may not be a monotonic function of Z due to the distortion. We now derive the general condition under which $\text{sgn}(\hat{g}_0 - \hat{g}_1)$ is a valid estimator for ordinal depth relation.

$\text{sgn}(\hat{g}_0 - \hat{g}_1)$ is a valid estimator for ordinal depth relation if and only if

$$\text{sgn}(\hat{g}_0 - \hat{g}_1) \text{sgn}(g_0 - g_1) > 0 \quad (2.6)$$

Referring to (2.3), the above is the same as

$$((a_0 g_0 - a_1 g_1) + (b_0 - b_1) + (c_0 - c_1))(g_0 - g_1) > 0 \quad (2.7)$$

where (a_0, b_0, c_0) and (a_1, b_1, c_1) are realizations of the distortion factors associated with points \mathbf{p}_0 and \mathbf{p}_1 . Equation (2.6) or (2.7) is a sufficient and necessary condition for $\text{sgn}(\hat{g}_0 - \hat{g}_1)$ to be a valid estimator for ordinal depth. We call it the *Valid Ordinal Depth (VOD) Condition*. It reveals how the distortion factors may affect the judgement of the ordinal depth relation.

To obtain more insight into the condition, we define $\bar{g} = \frac{g_0+g_1}{2}$, $\bar{a} = \frac{a_0+a_1}{2}$, $\Delta a = a_0 - a_1$, $\Delta b = b_0 - b_1$, $\Delta c = c_0 - c_1$, $\Delta g = g_0 - g_1$. Then it is clear that $a_0g_0 - a_1g_1 = \bar{a}\Delta g + \Delta a\bar{g}$. The VOD Condition (2.7) becomes

$$(\bar{a}\Delta g + (\Delta a\bar{g} + \Delta b + \Delta c)) \Delta g > 0 \quad (2.8)$$

Generally, given $\bar{a} > 0$, it can be shown that a sufficient condition (but not necessary) for (2.8) to be satisfied is

$$|\Delta g| > \left| \frac{\Delta a\bar{g} + \Delta b + \Delta c}{\bar{a}} \right| \quad (2.9)$$

We call (2.9) the *VOD Inequality*. It is a sufficient condition for $sgn(\hat{g}_0 - \hat{g}_1)$ to be a valid ordinal depth relation estimator given $\bar{a} > 0$. If $\bar{a} < 0$, depth order between the two points is ensured to be estimated reversely by the VOD Inequality.

Equations (2.8) and (2.9) show that when the average of depth function \bar{g} , depth function difference Δg , and the difference of the distortion factors of the two points Δa , Δb , and Δc satisfy certain conditions defined by the inequality, ordinal depth can be validly discerned up to a certain resolution even in the presence of motion errors and image measurement noise. To understand the VOD Condition and VOD Inequality better, we will look into specific projection models, reconstruction schemes, and motion configurations in Section 2.5 and Section 2.6 to see how various factors may affect the judgement of ordinal depth.

2.5 Resolving Ordinal Depth under Weak-perspective Projection

We begin our investigation with the orthographic/weak-perspective cameras, which are good approximations of the perspective camera model under small FOV (small FOV is usually the case which gives rise to some typical errors in 3D motion estimates). An orthographic/weak-perspective camera belongs to the *affine camera model* [95, 41]. Equations associated with these models are relatively simple and easy to handle. The concepts introduced however can be applied to the perspective camera model in the next section.

2.5.1 Depth Recovery and Its Distortion under Orthographic or Weak-perspective Projection

Motion field equations under orthographic and weak-perspective cameras can be written as follows [22]:

$$\dot{p}_x = -sZ\beta - sU + \gamma y + \delta x, \quad \dot{p}_y = sZ\alpha - sV - \gamma x + \delta y \quad (2.10)$$

where $\delta = \frac{1}{s} \frac{ds}{dt}$ is the relative changing rate of the scaling factor s ($s = 1$ for orthographic camera; $s = \frac{f}{Z}$ for weak-perspective camera, where f is the focal length and \bar{Z} is the average depth of scene points). Here we have $\mathbf{d} = (-\beta, \alpha)^T$, $g(Z) = sZ$. Like under perspective projection, depth can only be recovered up to a scale factor. The magnitude of frontal rotation is unsolvable. We set $\|\hat{\mathbf{d}}\| = 1$, which means we will recover depth information up to a scale factor $k = \sqrt{(\alpha^2 + \beta^2)}$.

It is known that in the 2-frame motion estimation process under affine camera, translation parallel to the image plane can only be estimated in the direction perpendicular to the epipolar direction [22], thus $\dot{\mathbf{p}}_{indep}$ can only be partially estimated. Depth can be recovered as a scaled and offset version of \hat{Z} .

$$g'(\hat{Z}) = ks\hat{Z} + Z_c = \frac{(\tilde{\mathbf{p}} - \hat{\mathbf{p}}_{indep-known}) \cdot \mathbf{n}}{\hat{\mathbf{d}} \cdot \mathbf{n}} \quad (2.11)$$

where $\hat{\mathbf{p}}_{indep-known} = (\hat{\gamma}y + \hat{\delta}x, -\hat{\gamma}x + \hat{\delta}y)^T$ and $Z_c = \frac{(-sU, -sV)^T \cdot \mathbf{n}}{\hat{\mathbf{d}} \cdot \mathbf{n}}$ which is unknown. Depth distortion due to motion errors and noise can be written as

$$g'(\hat{Z}) = ks\hat{Z} + Z_c = a(ksZ) + Z_c + b + c \quad (2.12)$$

where $a = \frac{\mathbf{d} \cdot \mathbf{n}}{\hat{\mathbf{d}} \cdot \mathbf{n}}$, $b = \frac{\dot{\mathbf{p}}_{indep-known} \cdot \mathbf{n}}{\hat{\mathbf{d}} \cdot \mathbf{n}}$, $\dot{\mathbf{p}}_{indep-known} = (\gamma_e y + \delta_e x, -\gamma_e x + \delta_e y)^T$, $c = \frac{\dot{\mathbf{p}}_n \cdot \mathbf{n}}{\hat{\mathbf{d}} \cdot \mathbf{n}}$.

In the following discussion in this section, we assume the unit vector \mathbf{n} in Equation (2.2) is the same for every feature point; thus Z_c is a constant. This allows relative depth between any two points to be recovered up to a scale factor. The scaled relative depth between points \mathbf{p}_0 and \mathbf{p}_1 can be recovered as

$$g(\hat{Z}_0) - g(\hat{Z}_1) = ks\Delta\hat{Z} = g'(\hat{Z}_0) - g'(\hat{Z}_1) = a(ks\Delta Z) + \Delta b + \Delta c \quad (2.13)$$

where $\Delta\hat{Z} = \hat{Z}_0 - \hat{Z}_1$ and $\Delta Z = Z_0 - Z_1$, $a = a_0 = a_1$, $\Delta b = b_0 - b_1$, $\Delta c = c_0 - c_1$.

Specifically, if the epipolar reconstruction scheme ($\mathbf{n} = \frac{\hat{\mathbf{d}}}{\|\hat{\mathbf{d}}\|}$) is adopted, we have $a = \cos \phi_e$ ($\phi = \tan^{-1} \frac{\beta}{\alpha}$ and ϕ_e is the angle between \mathbf{d} and $\hat{\mathbf{d}}$), $b = (\gamma_e \mathbf{p}^\perp + \delta_e \mathbf{p}) \cdot \hat{\mathbf{d}}$ and $c = \dot{\mathbf{p}}_n \cdot \hat{\mathbf{d}}$. Note that ϕ can only be recovered up to

a 180° ambiguity in the model and thus a may be negative. If $a < 0$, all the relative depths will be recovered reversely, and the whole scene structure will be recovered up to a mirror transformation.

2.5.2 VOD Inequality under Weak-perspective Projection

Here we will consider the VOD Inequality (2.9) under orthographic/weak-perspective projection. We adopt the epipolar reconstruction scheme where $\mathbf{n} = \hat{\mathbf{d}}$ (the derivation can be modified using other reconstruction schemes). We then have $\Delta a = 0$, $\Delta b = (\gamma_e \Delta \mathbf{p}^\perp + \delta_e \Delta \mathbf{p}) \cdot \hat{\mathbf{d}}$, $\Delta c = \Delta \dot{\mathbf{p}}_n \cdot \hat{\mathbf{d}}$ where $\Delta \mathbf{p} = \mathbf{p}_0 - \mathbf{p}_1$, $\Delta \dot{\mathbf{p}}_n = \dot{\mathbf{p}}_{n0} - \dot{\mathbf{p}}_{n1}$. We write $\Delta \mathbf{p} = r(\sin \theta, \cos \theta)$, where r indicates the image distance between \mathbf{p}_0 and \mathbf{p}_1 . After some manipulation, the VOD Inequality under orthographic/weak-perspective cameras takes the form

$$\frac{r}{|\Delta Z|} < ks\varepsilon \quad (2.14)$$

where $\varepsilon = \left| \frac{\cos \phi_e}{\gamma_e \cos(\hat{\phi} - \theta) + \delta_e \sin(\hat{\phi} - \theta) + \Delta \dot{\mathbf{p}}'_n} \right|$, where $\Delta \dot{\mathbf{p}}'_n = \frac{\Delta \dot{\mathbf{p}}_n \cdot \hat{\mathbf{d}}}{r}$. $\varepsilon = \infty$ in the error-free and noise-free ideal case, which implies that VOD inequality is satisfied in the entire image plane. Equation (2.14) shows that for two points, if the ratio between the image distance r and depth variation $|\Delta Z|$ is less than a certain value $ks\varepsilon$ defined by a particular realization of motion errors and noise in the optical flow measurements, the SFM system can still get a valid ordinal depth relation judgement even in the presence of errors and noise.

2.5.3 Ordinal Depth Resolution and Discrimination Threshold(DT)

Equation (2.14) can be written as :

$$|\Delta Z| > DT, \quad DT = \frac{r}{ks\varepsilon} \quad (2.15)$$

Equation (2.15) indicates that when depth variation is larger than a *discrimination threshold(DT)*, ordinal depth relation can be judged correctly by the SFM system. *DT* is an indication of the *ordinal depth resolution*. It gives us the smallest depth difference that ensures ordinal depth can be resolved correctly by VOD Inequality. The bigger *DT* is, the poorer the ordinal depth resolution.

It is noted that *DT* is a function of *r* – the distance between \mathbf{p}_0 and \mathbf{p}_1 in the image. Generally, for a certain realization of errors in the motion estimates and noise in the image velocity, *DT* increases as *r* increases. This means that ordinal depth resolution decreases as image distance increases. Equation (2.15) further shows that ordinal depth resolution decreases as motion errors (ε) increase and as the magnitude of the motion component carrying depth information(frontal rotation here) *k* decreases.

2.5.4 VOD Function and VOD Region

To have an intuitive understanding of Equation (2.15), we define *VOD function* and *VOD region* as follows:

- ***VOD function***: *Given certain realization of errors in the motion estimates and noise in the optical flow measurements, for an image point*

\mathbf{p}_0 , if image point \mathbf{p}_i satisfies the VOD Inequality (2.9) with respect to \mathbf{p}_0 for depth variation $|\Delta Z| = DT$ and average depth \bar{Z} , function $VOD(\mathbf{p}_0, \mathbf{p}_i, DT, \bar{Z}) = 1$; otherwise function $VOD(\mathbf{p}_0, \mathbf{p}_i, DT, \bar{Z}) = 0$.

- **VOD region:** VOD region \mathbb{R} of image point \mathbf{p}_0 for DT at \bar{Z} is a set of image points: $\mathbb{R}_{(\mathbf{p}_0, DT, \bar{Z})} = \{\mathbf{p}_i | VOD(\mathbf{p}_0, \mathbf{p}_i, DT, \bar{Z}) = 1\}$.

The VOD region contains all the image points that satisfy the VOD inequality with respect to \mathbf{p}_0 given particular DT and \bar{Z} . Since motion errors and noise are random, the VOD region is a random region in the image plane.

Figure 2.2 illustrates the realizations of VOD regions for different DT under certain motion error realizations when the effect of image noise is ignored. As can be seen, the VOD region of an image point \mathbf{p}_0 has a band shape under orthographic/weak-perspective projection. The width of the band increases as DT increases. The band stretches along the direction of the estimated frontal rotation axis ($\hat{\phi}$). This anisotropic property is due to the dependence of ε on θ . We indicate the width of the band region by the biggest circle that can be drawn inside the region and centered at the investigated point.

2.5.5 Ordinal Depth Resolution and Visual Angle

We now investigate the relationship between the ordinal depth resolution and the visual angle. Define the visual angle subtended by two image points as $\tau = 2 \tan^{-1} \frac{r}{2f}$. The VOD inequality can be written in terms of visual angle as

$$|\Delta Z| > \frac{2}{\varepsilon k} \tan \frac{\tau}{2} |\bar{Z}| \quad (2.16)$$

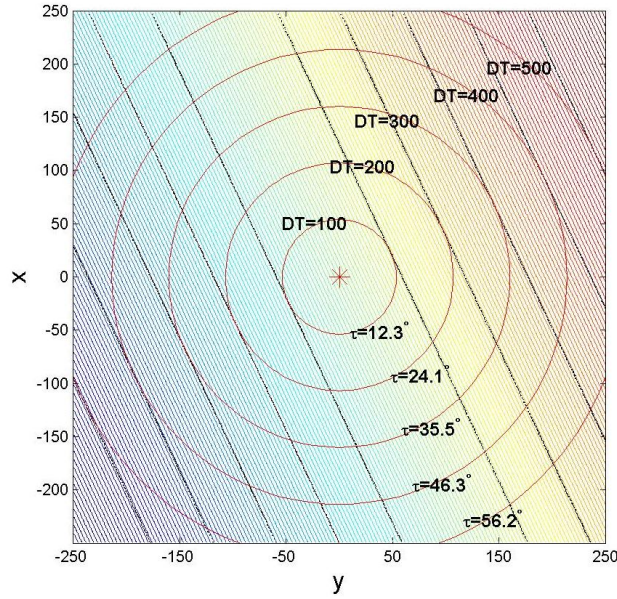


Figure 2.2: Realization of VOD region of $\mathbf{p}_0 = (\mathbf{0}, \mathbf{0})^T$ (\mathbf{p}_0 is denoted by the red asterisk) for different DT under weak-perspective projection. VOD region is bounded by black lines. The big red circles show the width of the region bands. τ is the visual angle between points on the circle and \mathbf{p}_0 . The rainbow at the background shows the change of distortion factor b . Motion parameters and errors: $\mathbf{T} = (0.81, 0.2, 0.15)^T$, $\Omega = (0.008, 0.009, 0.0001)$, $\bar{Z} = 35000$, $\delta = -4.2857e - 006$, $\phi_e = 28.6^\circ$, $\delta_e = 1.0e - 006$, $\gamma_e = 1.0e - 006$, $\dot{\mathbf{p}}_n = 0$, $f = 250$.

This shows that given ε and k , for two image points subtending a visual angle of τ , the ordinal depth relation between points in this region can be correctly resolved when the depth variation $|\Delta Z|$ is greater than $DT = \frac{2}{\varepsilon k} \tan \frac{\tau}{2} |\bar{Z}|$. The bigger the visual angle, the higher the DT. Therefore, ordinal depth resolution decreases as visual angle increases. Moreover, ordinal depth resolution also decreases as average depth \bar{Z} increases. Figure 2.2 also shows the increase of DT in the direction perpendicular to the band as the visual angle τ increases.

2.5.6 VOD Reliability

Practically, the VOD region is stochastic due to the random nature of errors and noise. To deal with the statistical issue, we define the *VOD reliability* of image point \mathbf{p}_i with respect to the investigated point \mathbf{p}_0 as

$$P_{VOD(\mathbf{p}_0, \mathbf{p}_i, DT, \bar{Z})} = P(VOD(\mathbf{p}_0, \mathbf{p}_i, DT, \bar{Z}) = 1) = P(\mathbf{p}_i \in \mathbb{R}_{(\mathbf{p}_0, DT, \bar{Z})}) \quad (2.17)$$

where $P(\cdot)$ is the probability of certain event. VOD reliability gives us the probability that image point \mathbf{p}_i falls inside \mathbf{p}_0 's VOD region for DT at \bar{Z} . It gives us the lower bound of the probability of correct judgement of the depth order relationship between point \mathbf{p}_0 and \mathbf{p}_i (for depth variation bigger than DT at average depth \bar{Z}). Particularly, under orthographic/weak-perspective projection, we have:

$$P_{VOD(\mathbf{p}_0, \mathbf{p}_i, DT, \bar{Z})} = P(r < |\Delta Z|ks\varepsilon) = P(\tau < 2\tan^{-1}\left(\left|\frac{\Delta Z}{\bar{Z}}\right|\frac{\varepsilon k}{2}\right)) \quad (2.18)$$

It is clear that, generally, under certain error and noise level, the VOD reliability decreases as the distance r between the points and the visual angle τ subtended by the points increase.

Figure 2.3 (Top) shows the VOD reliability of image points w.r.t. the image center \mathbf{p}_0 for $DT = 100$ at average depth $\bar{Z} = 35000$. This figure is the result of repeating the SFM process described in [22] 500 times on 1000 randomly generated points when the level of isotropic gaussian noise in the optical flow is 10%. Figure 2.3 (Bottom) shows the result for different DT as visual angle increases. It is shown that VOD reliability drops down significantly as distance between \mathbf{p}_i and \mathbf{p}_0 increases. This indicates that for

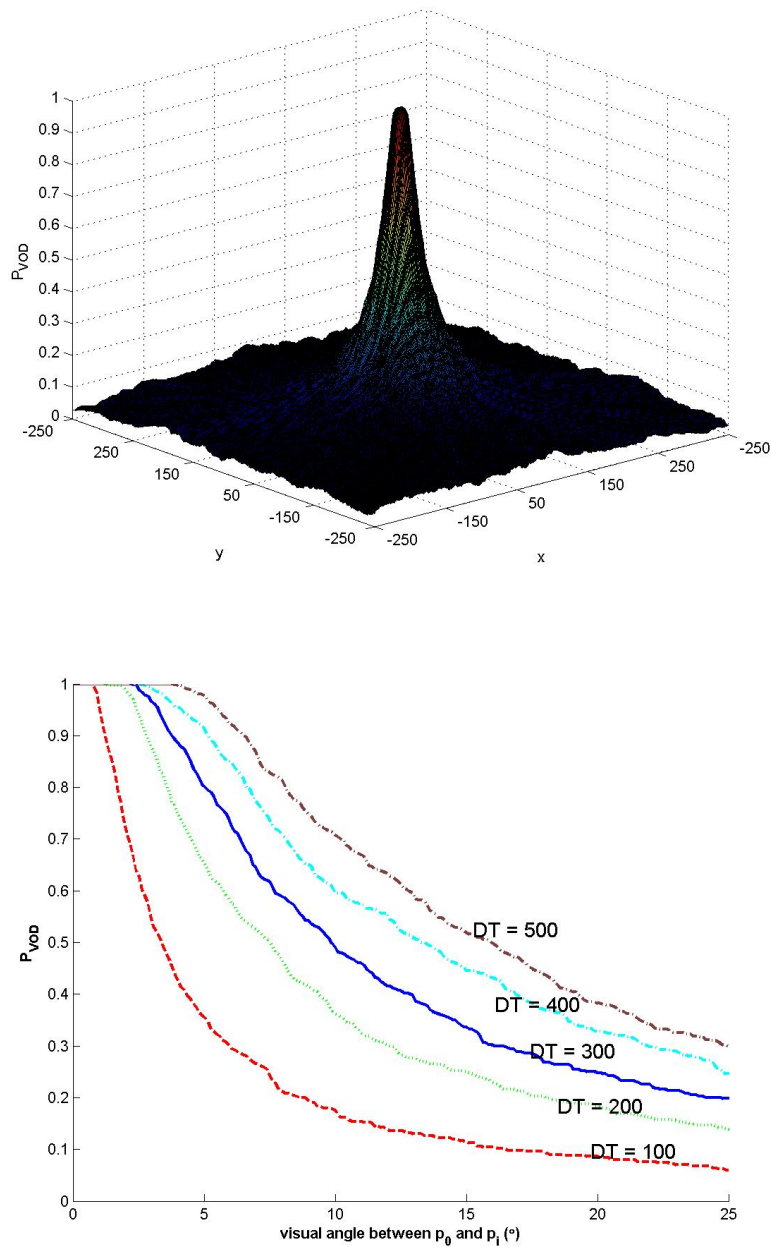


Figure 2.3: Top: VOD Reliability of image points w.r.t. the image center for $DT = 100$ at $\bar{Z} = 35000$. Bottom: VOD Reliability of image points w.r.t. the image center for different DT at $\bar{Z} = 35000$ as visual angle ($^\circ$) between the point pair changes. $(U, V, W) = (0.001, 0.002, 0.001)$, $(\alpha, \beta, \gamma) = (0.004, 0.002, 0.003)$.

the same depth variation, ordinal depth judgement by SFM systems for a pair of closer points can be considered as more reliable and trustworthy. Pairs of points subtending smaller visual angle have more reliable ordinal depth judgement. Ordinal depth information is strong in the local image areas within small visual angle despite the motion uncertainties and noise.

2.6 Resolving Ordinal Depth under Perspective Projection

In this section, ordinal depth resolution is investigated under perspective projection. Our analysis is first carried out under pure lateral motion configuration (Section 2.6.1). Lateral motion is a camera 3D motion without any forward translation component; note that 3D camera rotation can present in the lateral motion case. The analysis under lateral motion is very similar to the orthographic/weak-perspective projection analysis above, in the sense that all points have epipolar lines lying in the same direction. Then the effect of adding forward motion is analyzed in Section 2.6.2. Results in this Section will serve as guidelines for developing robust ordinal depth acquisition method and applying ordinal depth information into scene recognition algorithm in Chapter 3 and Chapter 4.

2.6.1 The Pure Lateral Motion Case

We assume that the SFM system knows that a pure lateral motion is executed.

Therefore $\hat{W} = W = 0$. The image velocity equation in this case is

$$\dot{\mathbf{p}}_x = \frac{-Uf}{Z} - \beta f + \gamma y + \frac{\alpha xy}{f} - \frac{\beta x^2}{f}, \quad \dot{\mathbf{p}}_y = \frac{-Vf}{Z} + \alpha f - \gamma x - \frac{\beta xy}{f} + \frac{\alpha y^2}{f} \quad (2.19)$$

We denote the direction of lateral translation (which is also the epipolar direction in this case) $\mathbf{d} = \frac{(\mathbf{U}, \mathbf{V})^T}{\sqrt{U^2 + V^2}}$ as $(\cos \phi, \sin \phi)^T$. The distortion factors can be written as $a = \cos \phi_e$, $b = f(\alpha_e \sin \hat{\phi} - \beta_e \cos \hat{\phi}) + \gamma_e(\cos \hat{\phi} y - \sin \hat{\phi} x) + O^2(x, y)$, where $O^2(x, y) = \cos \hat{\phi}(\frac{\alpha_e xy}{f} - \frac{\beta_e x^2}{f}) + \sin \hat{\phi}(\frac{-\beta_e xy}{f} + \frac{\alpha_e y^2}{f})$ is the second order term and it only exists under errors in the frontal rotation estimates (This second order term can be ignored when field of view is small). The VOD inequality can be written as

$$|\Delta Z| > DT, \quad DT = \frac{r}{ks\varepsilon} \quad (2.20)$$

which takes the same form as (2.15) but with the meaning of the parameters slightly different here. $k = \sqrt{U^2 + V^2}$ is the magnitude of lateral translation. $s = \frac{f}{\bar{Z}}$, where $\bar{Z} = \sqrt{Z_0 Z_1}$ is the geometric mean of the depths of the two points. $\varepsilon = \left| \frac{\cos \phi_e}{\gamma_e \sin(\hat{\phi} - \theta) + \Delta O^{2'} + \Delta \mathbf{p}'_n} \right|$, where $\Delta O^{2'} = \frac{O^2(x_0, y_0) - O^2(x_1, y_1)}{r}$.

Figure 2.4 shows the realization of VOD regions of the image center point under pure lateral motion. When the second-order flow is ignored, the shape of the region is the same as that under orthographic/weak-perspective projection (Top). With the second-order flow considered, the lines change to hyperbolae and the band shapes are distorted (Bottom), though the general topology remains. Note that the direction of the VOD region band in this case is the

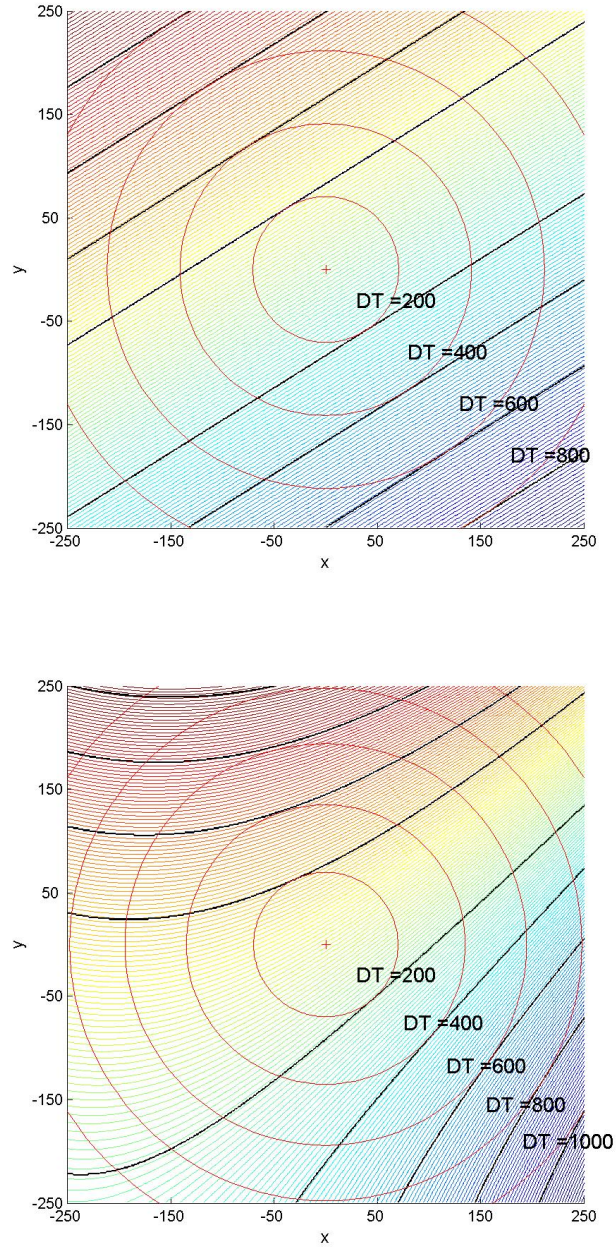


Figure 2.4: Realization of VOD region of $\mathbf{p}_0 = (\mathbf{0}, \mathbf{0})^T$ (denoted by red cross) for different DT under perspective projection and pure lateral motion. Top: second order flow ignored. Bottom: second-order flow considered. The VOD region is bounded by black lines. The background rainbow shows the change of distortion factor b . Motion parameters and errors are: $\mathbf{T} = (18, 22, 0)^T$, $\mathbf{T}_e = (15.3, 24.5, 0)^T$, (translation direction estimation error is -7.3°), $\Omega_e = (0.00002, 0.00002, 0.00005)$, $\bar{Z} = 20000$, $\dot{\mathbf{p}}_n = 0$, $f = 250$.

estimated lateral translation direction ($\hat{\phi}$).

If $\gamma_e = 0$ and second order term is ignored, Equation 2.20 becomes $|\Delta Z| > 0$. This means any ordinal depth relation is ensured to be recovered correctly. Such case corresponds to the discussion result in [23]. Compared to the result in [23], Equation 2.20 further shows that the effect of γ_e is to reduce the ordinal depth resolution as image distance between the points increases. This effect is expressed in an analytic way here.

2.6.2 Adding Forward Motion: The Influence of FOE

Now we add the forward translation component. It is well known that when the focus of expansion (FOE) is near the image center, the recovered depth is highly unreliable. This phenomenon is also shown in Figure 2.1, from which it can be seen that the values of the distortion factors change rapidly near the estimated FOE. This is in contrast to the lateral motion case, in which FOE can be regarded as lying at infinity and the distortion factors change slowly over the image.

Ordinal depth recovery in this forward motion case is of little practicability. Therefore, our investigation here is restricted to the case that FOE is far away from the boundary of the image. We use the angle $\mu = \arctan \frac{|W|}{\sqrt{U^2+V^2}}$ to indicate the relative amount of forward translation component compared to the lateral translation. The bigger μ is, the bigger the forward translation executed, and the nearer the FOE is to the image center. The VOD region with the forward translation added are shown in Figure 2.5. Several observations can be noted below:

1. Adding forward translation narrows the width of the VOD region and

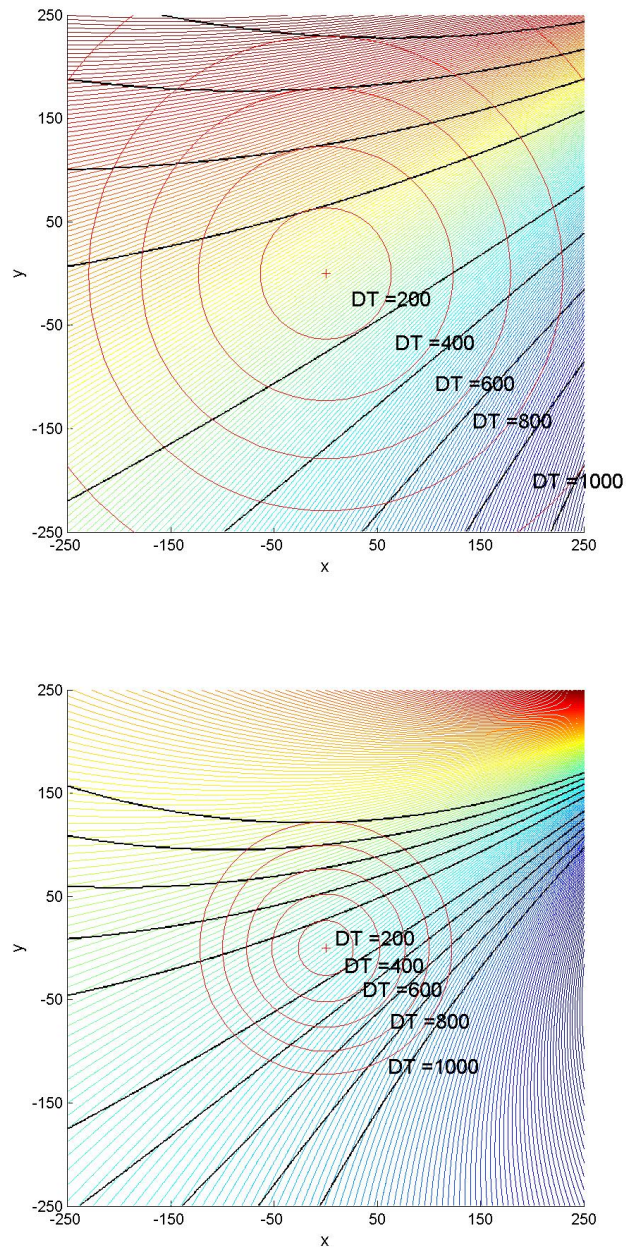


Figure 2.5: Realization of VOD region of $\mathbf{p}_0 = (\mathbf{0}, \mathbf{0})^T$ (denoted by red cross) for different DT under perspective projection with forward translation added to the motion configuration shown in Figure 2.4. Top: $\mu = 15^\circ$. Bottom: $\mu = 25^\circ$. $\mu_e = 0$ in both cases. Only first-order optical flow is considered for the illustration.

distorts the band shape. The VOD region is narrower in the image region nearer to the estimated FOE. However, the topology remains the same as that under pure lateral motion.

2. The VOD region is bounded by curves which can be shown to pass through the estimated FOE.
3. The bigger the forward translation component, the more the VOD region shrinks. Therefore, ordinal depth resolution decreases as image points approach the FOE. In other words, camera lateral motion (of which FOE is at infinity) is a good motion configuration for robust ordinal depth recovery.

2.7 Discussion

2.7.1 Practical Implications

To show the practical implication of the above results, we now compute the values of DT under lateral motion given values of \bar{Z} and certain practical error rate. We use h to denote the ratio between the average translational flow magnitude and average rotational flow magnitude¹. We assume error in the rotational estimate is $p_e(\%)$ of the magnitude of the rotational parameter and $\phi_e = 0$ (due to the bas-relief valley as we have discussed in Section 2.2). Then if we ignore the second order distortion effect, it can be readily shown

¹Here rotational flow is computed using first order term only with the assumption that $\sqrt{\alpha^2 + \beta^2} \gg \gamma$.

	$\tau = 10^\circ$	$\tau = 20^\circ$	$\tau = 30^\circ$	$\tau = 40^\circ$	$\tau = 50^\circ$	$\tau = 60^\circ$	$\tau = 70^\circ$
$h = 1$	0.8816m	1.8199m	2.8868m	4.1955m	5.9588m	8.6603m	13.7374m
$h = 5$	0.1763m	0.3640m	0.5774m	0.8391m	1.1918m	1.7321m	2.7475m
$h = 20$	0.0441m	0.0910m	0.1443m	0.2098m	0.2979m	0.4330m	0.6869m

Table 2.1: DT values for different visual angles under different translation-to-rotation ratio h . $\bar{Z} = 100m$ and $p_e = 5\%$.

from Equation (2.20) that an upper bound of DT is:

$$DT \leq \frac{|(\tan \tau)\bar{Z}p_e|}{h} \quad (2.21)$$

where τ is the visual angle subtended by the point pair. Table 2.1 shows the DT values computed from the above equation for different visual angles under different *translation-to-rotation ratio* h , given $\bar{Z} = 100m$ and $p_e = 5\%$.

2.7.2 Psychophysical and Biological Implication

Our results show that SFM algorithms can obtain reliable ordinal depth resolution within small visual angles despite the motion uncertainties. This is especially so for lateral motion. Ordinal depth resolution decreases as visual angle increases. This agrees with the intuition that in human vision, the depth order of two objects close together can be determined with much greater ease than that of objects far apart. Moreover, our result is also consistent with the experimental findings in psychophysics [107] [76] which showed that human vision gives better judgement of ordinal depth relation and depth intervals for pairs of close points using stereo or texture depth cue.

From an evolutionary perspective, foveated vision is adopted for many biological vision systems. For example, humans have a sharp foveated vision. The spatial resolution of the human eye decreases by more than an order

of magnitude within a few degrees from the optical axis and at least two orders at ten degrees from the optical axis. One possible explanation for this phenomenon may be that depth cues such as motion can only resolve various levels of depth information precisely in a small visual angle due to errors in ego-motion estimation, as shown by our results. Therefore, foveated vision might be an adaptive result of natural selection in response to the computational capability and limitation of Shape from X modules.

2.8 Summary

In this chapter, the resolution of ordinal depth from the inaccurate metric depth estimates in SFM was investigated theoretically based on a novel, general depth distortion model. It was shown that:

1. In SFM algorithms, although accurate metric depth may be difficult to obtain due to motion errors, ordinal depth can still be discerned locally if the actual depth difference is beyond a certain discrimination threshold.
2. The reliable ordinal depth resolution was found to decrease as the visual angle between point pair increases, as the speed of the motion component carrying depth information decreases, as scene points recede from the camera, and as the image points approach the estimated FOE.

These findings are important since they suggest that accurate qualitative 3D structure information is ensured in small local image neighborhood. Therefore, it also follows that qualitative structure information, when being used for performing vision tasks, should be carefully weighted according to its resolution under system errors, as we will do in Chapter 4. Besides, the result in this

chapter consolidates the conclusion in [23], which advocates lateral motion as a good active vision strategy to obtain ordinal structure information.

Chapter 3

Robust Acquisition of Ordinal Depth using Turn-Back-and-Look (TBL) Motion¹

3.1 Background

3.1.1 Turn-Back-and-Look (TBL) Behavior and Zig-Zag Flight

In behavioral physiology, researchers studied the learning behavior of bees around food sources. A special behavior called the *Turn-Back-and-Look behavior* (TBL) was observed and studied [56]: when bees depart from a new food source, they turn around to view it at a short distance, before departing

¹The work presented in this chapter was carried out in collaboration with Mr. Ching Lik Teo.

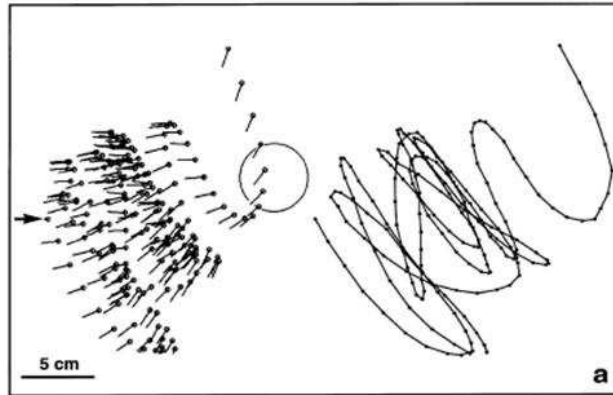


Figure 3.1: The Zig-Zag flight of a wasp directed towards a target (large circle) as seen from above [112]. Notice the significant translational motion that is almost perpendicular to the target at each arc formed. The complete path is shown on the right.

for the hive; this behavior was repeated on several successive visits.

In Voss and Zeil’s study, the zig-zag flight of wasps was observed: The insects repeatedly approach objects and start a series of rapidly swaying sideways movements, roughly perpendicular to their line of sight (see Figure 3.1) [112]. Studies on the occurrence, cessation and re-emergence of zig-zag flights suggested that zig-zag flight happens in ‘Turn-Back-and-Look’ behavior. In this thesis, we call zig-zag flight in turn-back-and-look behavior the *Turn-Back-and-Look (TBL) Motion*.

It was believed [24,122] that TBL behavior and zig-zag flight are important for the bees to recognize food source scenes on their return trip. It was shown that zig-zag flights were triggered only by 3D objects [112]. Experiments [56, 99] have also shown that bees do in fact obtain 3D distance information from the selected features so as to perform precise landing at feeding sites and at their nests. Furthermore, Voss and Zeil [112] suggested that zig-zag flights seem to be a ‘depth from motion’ procedure for the extraction of object-related

depth information.

3.1.2 Why TBL Motion Is Performed?

By examining the path of zig-zag flight in Figure 3.1, it can be seen that each arc in the TBL motion is actually a lateral motion without any forward translation component. Recalling the result in Cheong and Xiang’s work [23], we know that such lateral motion is especially conducive to depth recovery. If we further consider the observation that TBL motion is only triggered by 3D objects [112], it is reasonable to say that TBL motion might be performed for the purpose of obtaining 3D structure information of scenes.

However, to obtain reliable depth information, biological vision system needs to have mechanism for accurate ego-motion estimation. As we have discussed in Section 2.2, due to noise in the 2D optical flow measurements, accurate estimation of 3D motion and depth is difficult. Even under camera lateral motion, there may be systematic distortion of depth due to errors in the 3D motion estimates [23]. Therefore, qualitative depth seems to be an appropriate source of information to exploit. As we have shown in Section 2.6, ordinal depth can be recovered up to certain resolutions under camera lateral motion, even under various errors. It is reasonable to suggest that ordinal depth could be the 3D structure information obtained by TBL motion and used in high-level recognition tasks.

3.1.3 Active Camera Control and TBL Motion

The central tenet of the active vision paradigm is that a vision system is not solitary but one part of the perception-and-action system. In such systems,

sensing, processing and action components work in a cooperative way. Camera parameters and visual processing are actively controlled in response to the activity and task context (such as navigation, recognizing persons, obstacle avoidance) [17].

Active control of camera parameters such as orientation, focus, zoom, aperture, and vergence has been well studied. Through camera parameter controls, efficiency of visual data sensing can be improved, and the computation involved in early vision can be simplified dramatically. For example, by camera focus control (gaze control) and resolution control, areas of interest can be examined at the desired resolution without the high cost of uniform resolution sensing [17, 87]. Active camera control has been applied to many vision tasks such as measuring time-to-contact [93], 3D perception from static scenes [82], face detection and tracking [84] etc.

One way of active camera control is to control camera motion. For example, in the segmentation of an object from the background, some controlled camera motion can disambiguate solutions that are otherwise underconstrained [17]. In [25], a camera calibration procedure was developed for an outdoor active camera system with pan, tilt and zoom control. These active camera systems make use of the fact that computation of vision tasks can be simplified or made more robust under some special type of camera motion.

In this chapter, we use TBL motion as a strategy of active camera control for obtaining robust ordinal depth information. This strategy is inspired by the zig-zag flight of wasps in TBL behavior. It is based on the insight that ordinal depth recovery is robust under TBL motion, as we have discussed in the last subsection.

3.2 Recovery of Ordinal Depth using TBL Motion

3.2.1 Camera TBL motion

In our proposed approach, the camera TBL motion is just a relatively simple lateral motion (comprising of lateral translations and some rotations) facing the scene, without necessarily carrying out the whole suite of trajectories as found in the zig-zag flight of some wasps [112]. This strategy can also be applied to small baseline stereo systems where the two cameras or eyes are frontally placed, because such configuration is equivalent to a lateral monocular translation. Since stereo can be treated as a special case of motion, our formulation will be dealing with the case of motion.

The simple camera TBL motion is a pure translation in the horizontal direction without any rotation (see Figure 3.2). Under such simple motion, ordinal depth can be perfectly recovered if the optical flow can be measured properly. However, carrying out such controlled camera motion needs special mechanical device. Since ordinal depth recovery can tolerate some errors in the 3D motion estimation, the precise execution of such an ideal simple camera TBL motion is not required. We allow some 3D camera rotation and some camera translation in the vertical direction. These components can be estimated with a relatively gross ego-motion estimation step. Even some small forward translation component is also allowed, as long as it is relatively small compared to the lateral translation. As we have discussed in Chapter 2, despite some errors in the estimates of these motion components, ordinal depth can still be recovered up to certain resolution for points, especially in a local

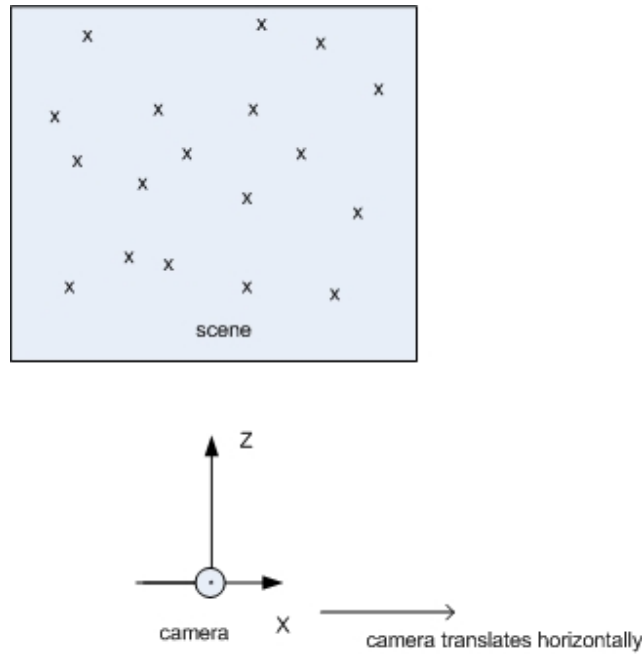


Figure 3.2: Simple camera TBL motion

image neighborhood. Thus, a common video camera held by hand performing a roughly lateral motion, probably with some camera rotation, suffices for our purpose.

3.2.2 Gross Ego-motion Estimation and Ordinal Depth Recovery

Now we develop a simple procedure to estimate the camera 3D ego-motion and the gross depth values which contain robust ordinal depth information, under camera TBL motion. Here we use the same notations as those used in Chapter 2.

Optical flow measurement: We measure the optical flow from SURF feature correspondences [9] between two consecutive video frames ($T = 0.8$

as the matching threshold; for details, see [9]). This gives us the optical flow of salient feature points in the scene (we will give a more comprehensive discussion on what saliency means in our work in Chapter 4). The reason we use the local feature matching method instead of the dense optical flow computation [11] is that the extracted features can be readily used in the scene recognition task later as we will show in Chapter 4.

Gross ego-motion estimation: The image velocity equations under lateral motion are:

$$\dot{\mathbf{p}}_x = \frac{-Uf}{Z} - \beta f + \gamma y + \frac{\alpha xy}{f} - \frac{\beta x^2}{f}, \quad \dot{\mathbf{p}}_y = \frac{-Vf}{Z} + \alpha f - \gamma x - \frac{\beta xy}{f} + \frac{\alpha y^2}{f} \quad (3.1)$$

We denote the direction of lateral translation $\mathbf{d} = \frac{(\mathbf{U}, \mathbf{V})^T}{\sqrt{\mathbf{U}^2 + \mathbf{V}^2}}$ as $(\cos \phi, \sin \phi)^T$ (this is also the epipolar direction). Eliminating depth Z in the above equations gives us the epipolar constraint:

$$\begin{aligned} a \cos \phi + b \sin \phi + c(-y \sin \phi - x \cos \phi) + d(-xy \sin \phi + y^2 \cos \phi) \\ + e(x^2 \sin \phi - xy \cos \phi) = (-\dot{\mathbf{p}}_x \sin \phi + \dot{\mathbf{p}}_y \cos \phi) \end{aligned} \quad (3.2)$$

where $a = \alpha f$, $b = \beta f$, $c = \gamma$, $d = \frac{\alpha}{f}$, $e = \frac{\beta}{f}$. Suppose the lateral translation direction ϕ is known (see below), if we have now more than five feature points and their optical flow measurements, we can solve for a , b , c , d and e by linear least square fitting. Then we have $f = \sqrt{\frac{a}{e}}$, $\alpha = \frac{a}{f}$, $\beta = \frac{b}{f}$ and $\gamma = c$. In simple camera TBL motion case, $\phi = 0$.

Searching for the translation direction: Since we do not require precise

control of the camera movement, the camera’s TBL motion has unknown ϕ , although it is largely horizontal. We take $[-30^\circ, 30^\circ]$ as the search range for ϕ . For each possible ϕ value within the range, we do a linear least square fitting to solve for α , β , γ and f . The residual error in the fitting is computed and the ϕ value with the smallest residual error is taken as the final estimate $\hat{\phi}$ for the lateral translation direction.

Extracting ordinal depth: Having obtained the gross estimates of $\hat{\phi}$, $\hat{\alpha}$, $\hat{\beta}$, $\hat{\gamma}$ and \hat{f} , we can compute the depth value of each feature point as:

$$\hat{Z} = \frac{\hat{f} \cos \hat{\phi}}{-\dot{\mathbf{p}}_x - \hat{\beta}\hat{f} + \hat{\gamma}y + \frac{\hat{\alpha}xy}{\hat{f}} - \frac{\hat{\beta}x^2}{\hat{f}}} \quad (3.3)$$

The \hat{Z} recovered here is a depth scaled by $s = \frac{1}{\sqrt{U^2+V^2}}$. Smoothing is applied locally to remove noise. The size of this local smoothing neighborhood is chosen empirically so that the noise effect can be removed without disturbing the fine structure of the scene. Due to the noise sensitive nature of SFM and the simple 3D motion estimation scheme used, the metric depth value \hat{Z} is unreliable. However, we can extract the more reliable ordinal depth information between each pair of feature points \mathbf{p}_i and \mathbf{p}_j by:

$$\text{sgn}(\hat{Z}_i - \hat{Z}_j) \quad (3.4)$$

3.3 Dealing With Negative Depth Value

We note that any image noise in the optical flow will add a stochastic component to the uncertainty in the ordinal depth recovery, over and above the

systematic errors caused by the errors in the egomotion estimates. Thus we need a robust way of handling the depth recovery and in particular those negative depth estimates. We carry out the following adjustment with regards to a negative depth estimate \hat{Z}_{neg} recovered at the image location (x, y) :

1. If the number of negative depths in a small image neighborhood centered at (x, y) is less than half of the total number of feature points within this neighborhood, regard the negativity in \hat{Z}_{neg} as caused by noise, and set the depth value \hat{Z}_{neg} to the average depth value of those positive depths in the neighborhood.
2. Otherwise, the negative depth value is caused by the systematic distortion arising from errors in the egomotion estimates. Since it is only the far depths that yield negative depth estimates [23], set the depth value \hat{Z}_{neg} to Z_{max} , where Z_{max} is the maximum positive depth recovered from the current scene.

3.4 Experimental Results

We use a SONY hand held video camera to take image sequences of indoor and outdoor scenes. Approximate camera TBL motion is executed by hand movement during recording. There will be a more detailed description of the scene content of these videos in Chapter 5.

Figure 3.3 depicts the recovered ordinal depths, using a color coding scheme that follows the rainbow sequence; warm colors such as red mean that the feature points are close to the observer, while cool colors such as violet represent points that are far away from the observer. Gross 3D motion estimates are

also shown. It can be seen that depth orders among feature points are well recovered in both indoor and outdoor environments.

In some environment or under some lighting condition, SURF features can be sparse (e.g. Figure 3.3: left of second row). This may cause 3D motion estimation unreliable due to the numerical issues. However, it is shown that ordinal depth can still be well recovered. It also can be seen that fine structure like trunk or branch of the tree (Figure 3.3: left of third row) is well recovered.

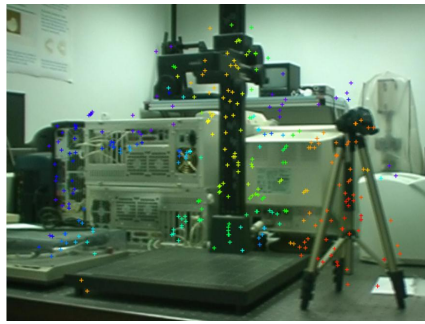
From the estimated 3D motion shown in Figure 3.3, we can see that the hand held camera's motion is not precisely the simple TBL motion. Some rotations and translation in vertical direction are involved in. It is difficult to accurately estimate these parameters. Our 3D motion estimation algorithm only gives gross estimates. However, the errors seem do not affect the robustness of ordinal depth recovery, as we see in Figure 3.3.

3.5 Summary

In this chapter, we develop an active camera motion strategy for robust acquisition of ordinal depth information. The controlled TBL motion used in the strategy is inspired by biological insect behaviors as well as the computational properties of ordinal depth from SFM. A simple yet effective algorithm to recover ordinal depth under the camera TBL motion is designed and tested.



(15.00°, 0.0007, -0.0065, 0.0030, 359.9(pixs))



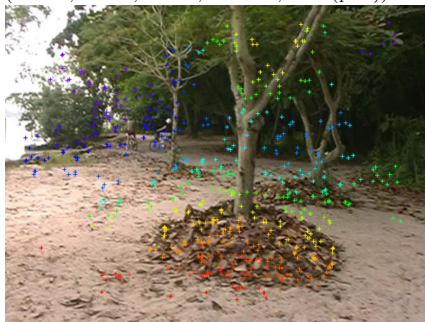
(0.00°, 0.0024, -0.0050, -0.0002, 567.2(pixs))



(10.00°, 0.0294, 0.0132, -0.0178, 558.8(pixs))



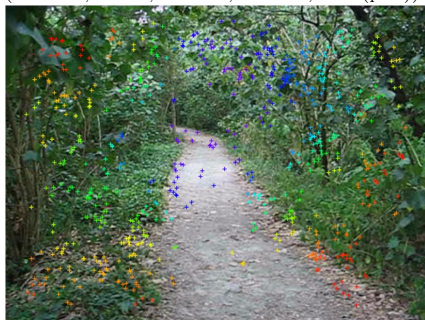
(-5.00°, -0.0068, 0.0151, 0.0024, 565.7(pixs))



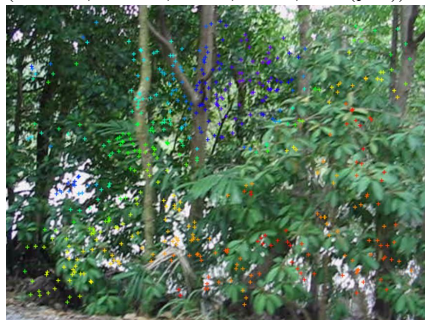
(-10.00°, 0.0042, -0.0002, -0.0057, 230.2(pixs))



(-10.00°, -0.0046, 0.0050, 0.0034, 464.5(pixs))



(-10.00°, -0.0009, -0.0003, 0.0003, 331.9(pixs))



(-30.00°, -0.0038, 0.0091, 0.0059, 310.6(pixs))

Figure 3.3: Recovered ordinal depth of feature points in indoor and outdoor scenes, depicted using the rainbow color coding scheme (red stands for near depth; violet for far depth). Gross 3D motion estimates $(\hat{\phi}, \hat{\alpha}, \hat{\beta}, \hat{\gamma}, \hat{f})$ are shown under each image.

Chapter 4

Robust Scene Recognition Using 3D Ordinal Constraint

4.1 Background

Due to the difficulty of 3D reconstruction, the reconstruction and the recognition problems are very much treated as separate problems, and 2D local descriptors have been the mainstay of object and scene recognition algorithms. Despite some success and the invariant properties of many 2D local descriptors, they cannot deal with large 3D viewpoint changes [72]. Illumination change is another challenge, especially in uncontrolled outdoor environment. Often, to accommodate these large changes, the threshold for local feature matching must be lowered, with the attendant sacrifice of its discriminating power.

To overcome the problem, the local feature approach is often combined with geometrical constraints to eliminate mismatches or as a verification stage so as to enhance the discriminating power of the local descriptors. 2D geomet-

rical constraints [6, 18, 34, 73, 91] have been proven helpful in object and scene recognition. However, these constraints are often based on the assumption that surfaces of objects are planar or at least can be approximated as planes locally, and that the object has no rotation in depth. In scenes with strong 3D effects, for instance, natural outdoor environment, such constraints become highly unstable and complex operations to group image into affine regions becomes necessary. Thus, despite the enhanced power of these 2D methods, they will always be restricted to certain types of objects or scenes.

A handful of work has indeed appeared and showed that it is possible to implement 3D object recognition, either at the single object level or at the category level [15, 39, 46, 52, 61, 88, 89, 105]. In these approaches, geometrical constraints such as epipolar or multi-view constraint are used. However, the use of these geometrical constraints presupposes the ability to compute the transformation between the reference view and the test view, which in turn requires a sufficiently accurate set of point correspondences. The latter is very much limited by the repeatability of the feature extraction and the difficulty of matching itself, especially when there are large amounts of clutter and significant changes in viewing conditions. Due to the large change in image appearance, one often has to lower the threshold for matching, as mentioned previously, with the resultant increase in large number of outliers. Under such situation, fitting a correct global or semi-local transformation is difficult; even robust methods like RANSAC may fail to work properly.

4.1.1 2D vs 3D Scene Recognition

The problems discussed above are exacerbated for the problem of outdoor scene recognition, which is the problem we wish to address in this chapter.

The problem of scene recognition has also been tackled at the 2D and 3D levels. In the biomimetic navigation community, the simple snapshot model has been proposed for insect navigation in an outdoor environment [19, 53, 70], whereby 2D snapshot of a reference scene is memorized and compared against the current scene. However, these works have only been validated in artificially manipulated outdoor environments with obvious landmarks [53, 70], or only computer simulations have been used to validate the method in theory [19]. Other approaches make domain specific assumption such as a relatively open terrain in which the skyline is assumed to be the most salient feature and contain all available information [26]. The problem is that real animals such as insects navigate in complex outdoor environments in which the selection of landmarks from these scenes is much harder. More importantly, these snapshot-based models are too simple to withstand viewpoint or illumination distortion.

An additional difficulty of recognizing outdoor scenes using purely 2D descriptors is that the 2D features returned by algorithms such as SIFT [61, 62] or SURF [10] are not informative or discriminative enough in such environment. Many features are alike and repeated in natural surroundings and there are often no distinctive colors in such environment either. Figure 4.1 illustrates such difficulty. In the top pair of images, the SIFT matches between two views of the same scene are shown. Due to the large viewpoint change, it can be seen that the SIFT matcher can only return a few correct matches,

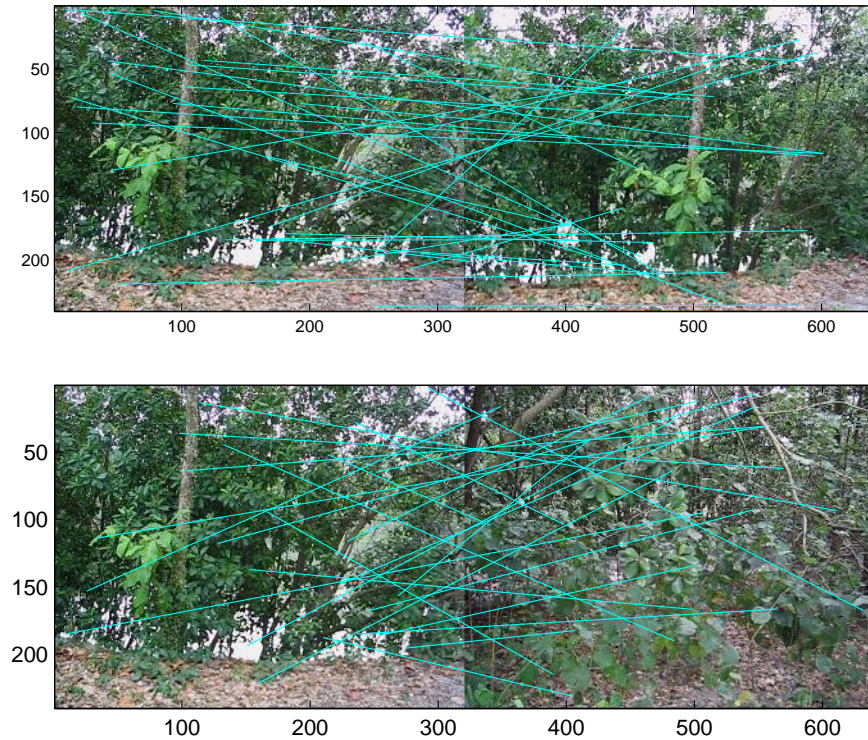


Figure 4.1: SIFT matching in the natural environment. Top: SIFT matches between two different views of a scene. Bottom: SIFT matches between images of two different scenes. The same matching threshold is used for both examples.

together with quite a number of mismatches. On the other hand, the bottom pair of images of Figure 4.1 show the SIFT matches obtained between images of two different scenes, using the same matching threshold as used for the top pair. Even though the scenes are different, a large number of matches are still obtained due to the presence of locally highly similar landmarks. This presence of highly similar features on the one hand, coupled with potentially large viewpoint change on the other, will pose severe difficulties for any 2D scene recognition schemes based entirely on local features. Lighting changes in an outdoor environment such as a wooded area further compound the issue as they are much more complex than the relatively “flat” variety experienced in

an indoor scene. Complex effects like inter-reflection, vignetting, reflexes, etc. mean that simple contrast normalization will not work. Under such complex illumination changes, pure 2D local feature approaches like [19, 27, 53, 70, 85] are simply ineffective.

While geometry-based information such as the 2D spatial configuration of the feature points could potentially relieve the problems highlighted in the preceding paragraph, the presence of large depth discontinuities in the environment would challenge and limit the effectiveness of these 2D-geometry-enhanced approaches. For such scenes, the ability to encode some 3D aspects of the feature landmarks will be crucial. Figure 4.2 shows examples of such scenes selected from our experimental database; the large depth discontinuities present means that some form of 3D information is required for successful recognition.

Unfortunately, 3D object recognition techniques such as those discussed in the opening paragraphs in this section have rarely been extended to the case of scene recognition due to various reasons. Firstly, the various difficulties discussed in the previous section, such as appearance variation caused by illumination and viewpoint changes, are multiplied manifold in the case of scene recognition, especially in the outdoor natural environment with its complex geometrical structure. Secondly, scene matching should allow mild changes and deformations as scene content changes somewhat over time (e.g., natural growth in a forest, erosion by the seashore, parked cars having moved in a city scene). A strict matching scheme based on global transformations such as homography or fundamental matrix [57, 92, 97] would fail in this situation. In addition, due to the potentially large spatial coverage (for instance, to recognize various places in a wood), it is also impractical to use a large number



Figure 4.2: Examples of scenes with large depth discontinuities on which 2D-geometry-enhanced approach may fail and 3D method is required.

of views for a specific location as a means to combat the large changes caused by various factors.

Most existing 3D model-based scene recognition works [26, 30, 45, 74, 102] either assume that 3D structure information is available from laser range finders [30, 74, 45] or a digital elevation map [102]. Such techniques are too slow to be useful in real time navigation and are limited by the physical properties of the objects being scanned which may not yield reliable results. Approaches such as [57, 92, 97] are based on a mix of appearance matching and epipolar geometry, as well as the use of non-visual information. However, they have not tested their approaches in extensive outdoor environments, especially with outdoor natural scenes under complex illumination changes. Clearly, due to the difficulty of recovering 3D scene structure from multiple views, such 3D

information has not been well exploited in the scene recognition problem, especially in scene recognition systems without recourse to active sensing device such as laser range sensors. Yet, due to the lack of highly distinguishing landmarks in natural outdoor scenes, a greater emphasis should be placed upon the 3D geometrical configuration of the landmarks for the successful recognition of such scenes. In our view, a proper use of 3D information in some form is crucial to the success of such a system.

4.1.2 Revisiting 3D Representation

Since exact 3D geometry is difficult to recover without recourse to sensing device, the question then becomes what sort of 3D spatial knowledge is required and how it is to be obtained. Central to our scene recognition approach is the usage of ordinal relationships between feature landmarks' spatial positions to represent the 3D geometry of the scene. The qualitative nature of ordinal representation is advantageous for the following reasons. Firstly, it makes the scene comparison robust to large viewpoint changes, as the spatial orders of the landmark points in the scene remains invariant to a significant extent to viewpoint changes. Our particular way of combining the spatial orders in all three dimensions also complements each other in terms of their stability under different types of viewpoint changes and for different scene types (as we show later). Secondly, the scene comparison strategy is also robust against the addition and deletion of new and old landmarks since the relational ordering of the remaining original landmarks is likely to remain unchanged. Thus our proposed similarity comparison based on ordinal ranks is robust against clutter and occlusion. Thirdly, unlike methods based on epipolar constraint, it does

not require fitting a rigid transformation on the landmark matches and can thus tolerate certain amount of deformation (e.g., trees have grown in size or fallen down). Fourthly, the proposed method is proved to be robust to feature matching outliers. Since unlike most of the existing geometrical constraint methods, the proposed method does not need to compute a geometrical transformation. Lastly, ordinal information along the third dimension (the depth dimension) can be obtained in a robust and simple manner without requiring full egomotion recovery. This has been discussed in Chapter 3.

4.1.3 Organization of this Chapter

The rest of this chapter is organized as follows. In Section 4.2 we propose a 3D ordinal space representation that qualitatively encodes the scene landmark's spatial configuration. We also analyze the robustness of the various ordinal relationships in different spatial dimensions under viewpoint changes. Section 4.5 addresses the issue of measuring the geometrical similarity between two scenes encoded in terms of 3D ordinal geometry. In Section 4.6, we present the full scene recognition system that combines the local appearance information together with the ordinal spatial constraint.

4.2 3D Ordinal Space Representation

Our proposed ordinal space representation is based on a weak characterization of the 3D geometric configuration. It is encoded in terms of the ordinal ranks of the landmark features based on their position along three dimensions: the two image dimension x and y , and the depth dimension Z . The reason for using the directly observable image coordinates x and y is that they will not be

affected by errors in depth recovery, unlike the unknown 3D space coordinates X and Y (Recall $X = \frac{xZ}{f}$ and $Y = \frac{yZ}{f}$, where f is the camera focal length) which depend on the estimated depth Z .

Figure 4.3 illustrates the idea of landmark rank. The numbers on the objects show the ranks of the landmark objects according to the x -coordinate of the landmarks' position in the image. Landmark rank has the property that it is more robust to viewpoint change than the metric coordinate. The arrows in Figure 4.3 show how the positions of the four landmarks have changed as the camera shifts to the right (see Figure 4.3, right). Although the landmarks' x coordinates change in the image when the camera undergoes some viewpoint change, the ranks remain unchanged.

Landmark rank can be regarded as the result that emerges from all pairwise comparisons between pairs of landmarks. It can be encoded implicitly as a matrix $\mathbf{A}_m = \{a_{mij}\}$ where m is the property upon which the landmarks are ranked, and the value of entry a_{mij} for two landmarks \mathbf{P}_i and \mathbf{P}_j is defined as:

$$a_{mij} = \begin{cases} 1 & \text{if } m_i > m_j \\ -1 & \text{if } m_i < m_j \\ 0 & \text{if } m_i = m_j \end{cases} \quad (4.1)$$

Clearly, the diagonal entries a_{mii} are 0 and $a_{mij} = -a_{mji}$. Since we represent the ordinal relationship along the three dimensions: the two image dimension x, y , and the depth dimension Z , we have $\mathbf{A}_x = \{a_{xij}\}$, $\mathbf{A}_y = \{a_{yij}\}$, and $\mathbf{A}_Z = \{a_{Zij}\}$. $\mathbf{A}_Z = \{a_{Zij}\}$ represents the ordinal depth information among landmark points.

In actual fact, not all ordinal relations can be obtained with the same degree of robustness nor used with the same confidence under viewpoint changes.

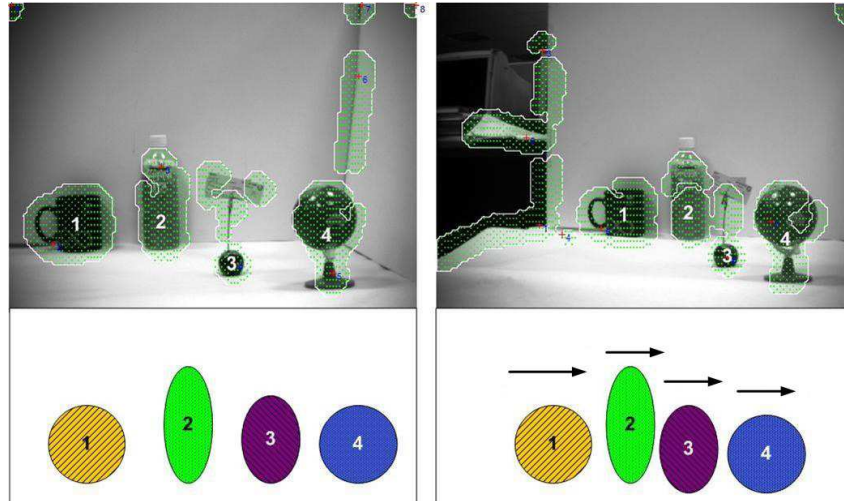


Figure 4.3: Landmark rank (based on the x -coordinate) remains unchanged under small viewpoint change.

Thus we also associate a confidence factor s_{mij} with each of these ordinal relations. Before formulating these s_{mij} , we will first analyze in the following the factors affecting the robustness of the various ordinal measurements, during both the depth recovery stage and during the scene recognition stage when there might be substantial viewpoint changes.

4.3 Robustness of Ordinal Depth Recovery

In Chapter 3, we have argued about the ecological relevance of lateral motion and highlighted the TBL motion of wasps in particular. We have also designed an algorithm to recover ordinal depth under controlled camera TBL motion. It has been shown in Section 2.6 and also in [23] that if we assume that the contribution of γ_e is small, and further assume that the quadratic terms in the rotational error flow is small relative to the other terms due to the limited field of view, ordinal depth can be recovered correctly in spite of errors in 3D mo-

tion estimates. Section 2.6 also shows that if the aforementioned assumptions are violated, global depth ordinality is no longer preserved. However, we can still obtain ordinality within a neighborhood where the VOD condition is satisfied. The size of this neighborhood depends on the size of the motion errors and the depth differences; smaller depth difference would need finer motion estimation for ordinal depth to be resolved. This property in ordinal depth recovery implies that the 3D spatial information should not be represented in a homogeneous manner but should have different levels of confidence for point pairs with different depth differences subtending different visual angles. This insight will be used for weighting the importance of ordinal depth of each individual point pair when we compute the global spatial similarity between two set of points in the 3D ordinal space.

4.4 Stability of Pairwise Ordinal Relations under Viewpoint Change

We now assess the stability property of the pairwise relations in each dimension (x , y , and Z) with respect to camera viewpoint change during the scene recognition stage. We denote the camera’s optical center before and after the viewpoint change as C and C' respectively.

4.4.1 Changes to Pairwise Ordinal Depth Relations

We first explore how the depth order of a pair of landmarks i and j may change as the camera viewpoint varies. We denote the position of the landmark pair in 3D space as: \mathbf{P}_i and \mathbf{P}_j . Referring to Figure 4.4, it is obvious that the

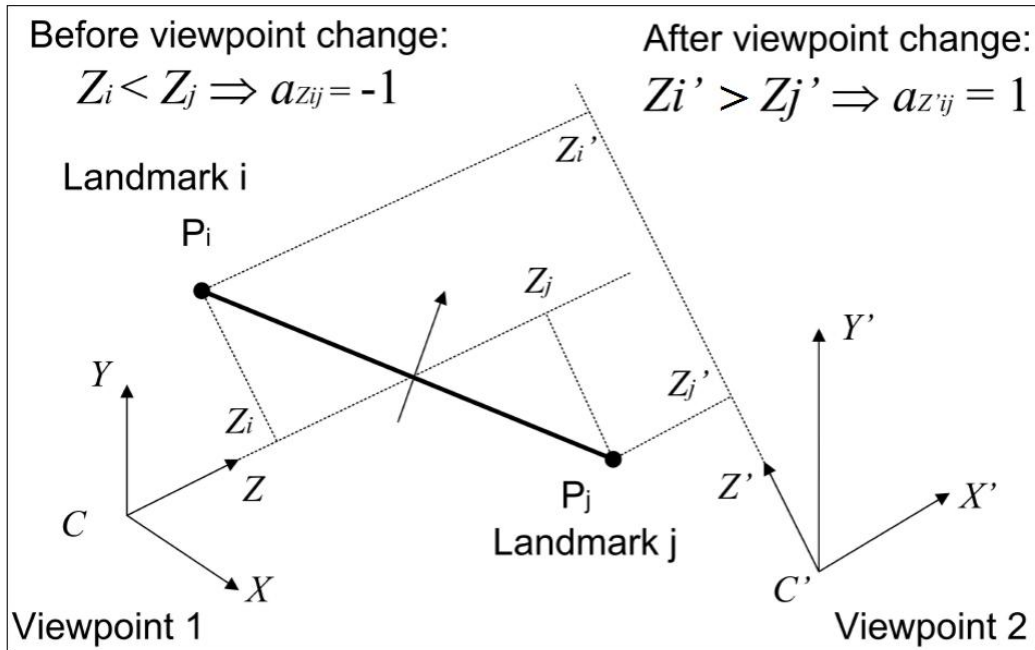


Figure 4.4: Pairwise ordinal depth relation varies as optical axis direction changes.

value of $a_{Z_{ij}}$ depends on the relative orientation between the camera's optical axis (CZ) and the line joining P_i and P_j , with the relative order $a_{Z_{ij}}$ flipping value when the viewpoint change brings the optical axis sweeping across the direction perpendicular to P_iP_j .

Since camera movements like translation and cyclotorsion (rotation around the optical axis) does not change the optical axis direction, we first conclude that the pairwise ordinal depth relation $a_{Z_{ij}}$ is invariant to camera translation and cyclotorsion.

The next question is: how does $a_{Z_{ij}}$ change as the camera rotates around the X - or Y -axis, resulting in changes in the optical axis direction? The critical point at which $a_{Z_{ij}}$ changes values can be obtained by simple algebra but it is easier to illustrate graphically through Figure 4.5. More specifically,

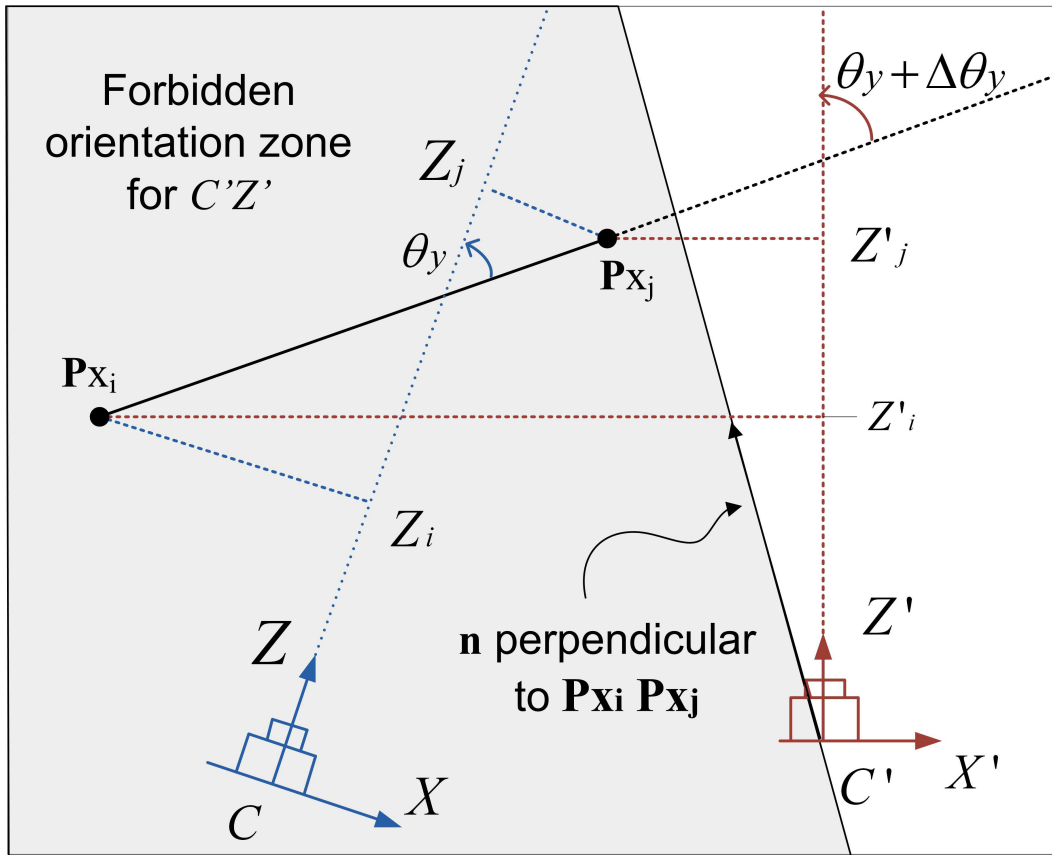


Figure 4.5: Pairwise ordinal depth relation under camera rotation around the Y -axis. The figure is the projection of the scene in Figure 4.4 onto the XCZ plane. Forbidden orientation zone for $C'Z'$ is indicated by the shaded region that passes through C' . For feature pair that are almost fronto-parallel, like P_i and P_j when viewed from C , small camera rotation around the Y axis may cause the line of sight $C'Z'$ at the new viewpoint to cross into the forbidden zone.

let \mathbf{P}_{X_i} and \mathbf{P}_{X_j} be the projections of \mathbf{P}_i and \mathbf{P}_j on the XCZ plane (see Figure 4.5). Suppose the line joining \mathbf{P}_{X_i} and \mathbf{P}_{X_j} forms an angle θ_Y with respect to the line of sight of the camera C at position 1, and suppose there is an orientation change of $\Delta\theta_Y$ around the Y -axis due to viewpoint change such that the line joining \mathbf{P}_{X_i} and \mathbf{P}_{X_j} forms an angle $\theta_Y + \Delta\theta_Y$ with respect to the line of sight of the camera C' at position 2, then it can be readily shown that $a_{Z_{ij}}$ is invariant to camera rotation around the Y -axis if and only if $\cos \theta_Y \cos(\theta_Y + \Delta\theta_Y) > 0$. In other words, the orientation of the line $\mathbf{P}_{X_i}\mathbf{P}_{X_j}$ defines a forbidden orientation zone for the forward direction of camera C' (see the shaded area in Figure 4.5). It is clear that the same arguments also apply to the rotation about the X -axis. Thus, we conclude that $a_{Z_{ij}}$ is invariant to the camera rotation around the X -axis if and only if $\cos \theta_X \cos(\theta_X + \Delta\theta_X) > 0$, where θ_X and $\Delta\theta_X$ are defined in a manner similar to θ_Y and $\Delta\theta_Y$ respectively.

We can summarize the above results as follows:

Proposition 4.4.1 *Pairwise ordinal depth relation $a_{Z_{ij}}$ is invariant to:*

1. camera translation and cyclotorsion;
2. camera rotation around the Y -axis iff $\cos \theta_Y \cos(\theta_Y + \Delta\theta_Y) > 0$;
3. camera rotation around the X -axis iff $\cos \theta_X \cos(\theta_X + \Delta\theta_X) > 0$.

Remark 4.4.1 *Given a orientation change of $\Delta\theta$, one can see that the ordinal relation is more readily satisfied when θ_X and θ_Y have small values, which corresponds to the case when the point pair \mathbf{P}_i and \mathbf{P}_j stretch along the Z -direction (we call it an in-depth pair).*

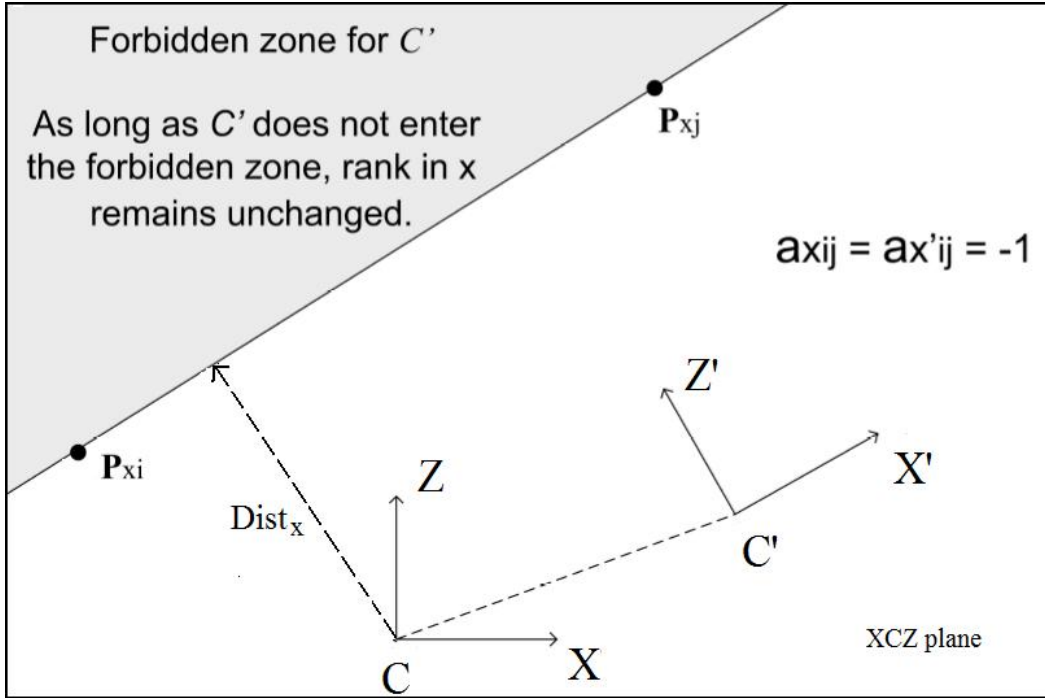


Figure 4.6: Pairwise x relation under camera translation in the XCZ plane is preserved as long as C' does not enter the forbidden zone, which is the half space indicated by the shaded region. $Dist_x$ is the shortest camera translation that will bring about the crossing of this half space.

4.4.2 Changes to Pairwise Ordinal x and y Relations

Now we look into how pairwise ordinal relations in the x or y dimension may change when the camera viewpoint changes.

From the geometry of a perspective camera, it is intuitively clear that as long as P_i and P_j remain visible, $a_{x_{ij}}$ and $a_{y_{ij}}$ are invariant to pure camera rotation around the X - or Y -axis (since the projection rays associated with P_i and P_j remain unchanged). The invariance of the relation can also be verified algebraically. For instance, for the ordinal relation in x , say $x_i - x_j$ of landmarks i and j , after a rotation $\Delta\theta_Y$ about the Y -axis, the new ordinal

relation in x can be readily obtained as

$$x'_i - x'_j = \frac{(x_i - x_j)(f + f \tan^2 \Delta\theta_Y)}{(f - x_i \tan \Delta\theta_Y)(f - x_j \tan \Delta\theta_Y)}.$$

where f is the focal length. As the numerator has the same sign as $x_i - x_j$, the sign of the new ordinal relation depends on the sign of the denominator terms $(f - x_i \tan \Delta\theta_Y)$ and $(f - x_j \tan \Delta\theta_Y)$. Clearly we only have to worry about sign change when $x_i \tan \Delta\theta_Y$ or $x_j \tan \Delta\theta_Y$ are positive. However it is clear from simple geometry that these two terms $(f - x_i \tan \Delta\theta_Y)$ and $(f - x_j \tan \Delta\theta_Y)$ are positive as long as both the landmarks remain visible, thus supporting our intuitive statement. Similar expression holds for a rotation $\Delta\theta_X$ about the X -axis. Thus, ordinal relations in x and y are invariant to any rotation about the X - and the Y -axis, as long as the points \mathbf{P}_i and \mathbf{P}_j remain visible.

Unfortunately, for a camera cyclotorsion $\Delta\theta_Z$ about the Z -axis, we have for the x relation the following:

$$x'_i - x'_j = (x_i - x_j) \cos \Delta\theta_Z - (y_i - y_j) \sin \Delta\theta_Z.$$

If the rotation $\Delta\theta_Z$ is substantial, then the ordinal relation in x can change. Clearly, the ordinal relation in x would be more likely to retain the same sign if $y_i \approx y_j$. Similar arguments can be applied, *mutatis mutandis* for the case of ordinal relation in y .

Next, for viewpoint change involving camera translation, we discuss the invariance condition for $a_{x_{ij}}$ (the extension to $a_{y_{ij}}$ is straightforward). Referring to Figure 4.6 which again depicts the projection onto the XCZ plane, it is clear that the plane passing through \mathbf{P}_{X_i} and \mathbf{P}_{X_j} and perpendicular to the

XCZ plane defines a forbidden half-space into which the camera should not tread. As long as this half-space is not crossed, $a_{x_{ij}}$ is unchanged. It also implies that the shortest camera translation that will bring about the crossing of this half space is to move along the direction of the shortest line segment from C to the line $\mathbf{P}_{X_i}\mathbf{P}_{X_j}$ (see $Dist_X$ in Figure 4.6).

We summarize the above results in the following proposition:

Proposition 4.4.2 *As long as \mathbf{P}_i and \mathbf{P}_j remain visible,*

1. *pairwise ordinal relation $a_{x_{ij}}$ and $a_{y_{ij}}$ are invariant to pure camera rotation around the X - or Y -axis.*
2. *$a_{x_{ij}}$ (respectively $a_{y_{ij}}$) is invariant to camera translation that does not cross the half space defined by the plane passing through \mathbf{P}_{X_i} and \mathbf{P}_{X_j} (respectively \mathbf{P}_{Y_i} and \mathbf{P}_{Y_j}) and perpendicular to the XCZ plane (respectively YCZ plane).*
3. *$a_{x_{ij}}$ (respectively $a_{y_{ij}}$) might be sensitive to camera cyclotorsion $\Delta\theta_Z$, with its sensitivity depending on $(y_i - y_j)$ (respectively $(x_i - x_j)$).*

Remark 4.4.2 *Other conditions being equal, this means that a favorable condition for the invariance of $a_{x_{ij}}$ or $a_{y_{ij}}$ with respect to camera translation is that the two points \mathbf{P}_i and \mathbf{P}_j are as fronto-parallel as possible (so that $a_{x_{ij}}$ or $a_{y_{ij}}$ is not easy to be disturbed by translation along X or Y direction); or as far away from the camera as possible (so that $a_{x_{ij}}$ or $a_{y_{ij}}$ is not easy to be disturbed by translation along Z direction), or the two image points are as widely separated as possible (so that $a_{x_{ij}}$ or $a_{y_{ij}}$ is not easy to be disturbed by cyclotorsion).*

4.4.3 Summary of Effects of Viewpoint Changes

To end this section, we summarize the kinds of changes in landmark ordinal relations that might arise due to various viewpoint changes.

- * Ordinal relation in the Z dimension is likely to be perturbed by viewpoint change arising from camera rotation around the X - or Y -axis. The more in-depth feature pairs there are in the scene, the lower the sensitivity; whereas scenes with more fronto-parallel pairs will have higher sensitivity. We call the former type of scene an *in-depth scene*, and the latter type a *fronto-parallel scene*. Ordinal relation in the Z dimension are invariant to any kind of pure camera translation.
- * Ordinal relation in the x (y) dimension is likely to be perturbed by viewpoint change arising from camera translation. Here, in contrast to the case for ordinal ranks in the Z dimension, the more in-depth feature pairs there are in the scene, the higher the sensitivity. On the other hand, these ordinal relations are invariant to pure camera rotation around the X - or Y -axis.
- * Ordinal relation in the x (y) dimension is not sensitive to camera forward translation in faraway scenes; while given the ordinal depth recovery property as we have discussed in Chapter 2, recovered ordinal relation in Z dimension is more reliable in the near scenes.
- * Viewpoint change arising from camera cyclotorsion may perturb the ordinal relation in the x and y dimension but not the ordinal relation in the Z dimension.

The above properties are briefly summarized in Table 4.4.3.

	$a_{x_{ij}}$	$a_{y_{ij}}$	$a_{Z_{ij}}$
horizontal translation	change	invariant	invariant
vertical translation	invariant	change	invariant
forward translation	change	change	invariant
rotation around horizontal axis	invariant	invariant	change
rotation around vertical axis	invariant	invariant	change
rotation around optical axis	change	change	invariant
favorable scene type	frontal-parallel scene, faraway scene	frontal-parallel scene, faraway scene	in-depth scene, near scene

Table 4.1: Invariant properties of ordinal relations in x , y , and Z dimensions to different types of camera movements and in different types of scenes. It can be seen that different dimensions complement each other.

Clearly, the ordinal relations in the Z dimension behave differently in terms of their response to types of viewpoint changes, compared to those in the x and y dimensions. In this sense, the ordinal depth relations capture a complementary aspect of the scene essence different from those in the x and y direction. We would expect that the ability to capture this aspect of the scene essence is especially critical for a scene recognition algorithm dealing primarily with in-depth scenes (e.g. forest scenes). The in-depth property of such scenes is more conducive toward preserving the invariance of the ordinal depth relations, while at the same time rendering ordinal x and y relationships highly unstable.

4.5 Geometrical Similarity between Two 3D Ordinal Spaces

Given correspondences between two groups of 3D scene points, we now propose a scheme to compare their geometrical similarity, based on the respective ordinal rankings in the x , y and Z dimensions. To do this, we have to first construct the global ordinal rankings from the pairwise ordinal relations, with their individual reliability properly taken into account.

4.5.1 Kendall's τ and Rank Correlation Coefficient

In the statistical literature, the similarity between two different rankings on a set of objects is measured by the *Rank Correlation Coefficient (RCC)* [49]. Normally, the coefficient is a variable within $[-1, 1]$, with 1 indicating perfect agreement between the two ranks, 0 indicating complete independence, and -1 indicating perfect disagreement. Many methods have been proposed to calculate the rank correlation coefficient. We briefly describe Kendall's τ , which will be used in the ensuing development.

Suppose we have a set of N objects $O = \{O_1, O_2, \dots, O_N\}$, which are being considered in relation to two properties represented by α and β . We may say that the objects exhibit values $\alpha_1, \alpha_2, \dots, \alpha_N$ according to α and $\beta_1, \beta_2, \dots, \beta_N$ according to β . These values may be variates or ranks, from which we can form matrices \mathbf{A}_α and \mathbf{A}_β based on the definition in (4.1). Entries in matrix \mathbf{A}_α are denoted by $a_{\alpha ij}$, while entries in matrix \mathbf{A}_β are denoted by $a_{\beta ij}$. Denoting by \sum as summation over all values of i and j from

1 to N , the Kendall's τ is defined by the equation:

$$\tau = \frac{\sum (a_{\alpha_{ij}} a_{\beta_{ij}})}{\sqrt{\sum a_{\alpha_{ij}}^2 \sum a_{\beta_{ij}}^2}}.$$

Kendall's τ as formulated treats every relation equally. However, as shown in Chapter 2 and Section 4.4, the ordinal relations of some pairs of objects may be more reliable or stable than others. Thus, we propose the *Weighted Kendall's τ* as follows:

$$\tau_w = \frac{\sum (s_{ij} a_{\alpha_{ij}} a_{\beta_{ij}})}{\frac{1}{N(N-1)} (\sum s_{ij}) \sqrt{\sum a_{\alpha_{ij}}^2 \sum a_{\beta_{ij}}^2}}$$

where $s_{ij} \in [0, 1]$ is the weighting factor of the relation between object pair O_i and O_j indicating its reliability or stability. The normalizing term is given by $N(N - 1)$ rather than N^2 because all s_{ii} are defined to be zero. In the context of this paper, the two rankings to be compared arise from that of a test scene and a reference scene over a particular property. The properties being measured include the x , y , and Z coordinates of the feature matches (the N objects), generating τ_x , τ_y and τ_Z respectively. The corresponding weighted version are denoted as τ_{x_w} , τ_{y_w} and τ_{Z_w} respectively; how the weighting factor s_{ij} is defined depends on which of the coordinates is being considered and will be discussed in the next subsection.

Finally, since the pairwise ordinal relations in the x , y and Z dimensions are differently sensitive to perturbations from various types of camera motions, we propose the following *3D Rank Correlation Coefficient (3D RCC)* that combines the rank correlation coefficients in all three dimensions so that each

dimension complements the strength of each other:

$$\tau_{3D} = w_Z\tau_Z + w_x\tau_x + w_y\tau_y; \quad (w_Z + w_x + w_y = 1). \quad (4.2)$$

The weights w_Z, w_x, w_y can be chosen based on our prior knowledge, if any, about the scene or about the system. For instance, if we have at our disposal a gravitational sensor, perturbations arising from cyclotorsion would not pose a serious problem (since it can be compensated) and thus τ_x and τ_y would be more reliable. In the case of the bees and wasps, they are able to use compass cues available from the sun or the Earth's magnetic field and take up a preferred compass orientation, thereby minimizing rotational perturbations about the X and the Y axis and enhancing the reliability of ordinal relations in depth. In the absence of such *a priori* information, we can simply average the three components to obtain ($\tau_{3D} = \frac{1}{3}(\tau_Z + \tau_x + \tau_y)$).

To illustrate how the various RCCs might vary as viewpoint changes, we choose a typical forest scene from the Brown range image database [55] and truncate it to a more conventional field of view of 45° (Figure 4.7). We execute different kinds of viewpoint changes, compute the various RCCs between the new and the original viewpoint, and plot them in Figure 4.8. It can be seen that firstly, as camera translation in horizontal (or vertical) direction increases (Figure 4.8, top left, middle left), the RCC in the x (or y) dimension τ_x (or τ_y) decreases, while that in the Z dimension τ_Z does not change. Secondly, as camera rotates around the horizontal (or vertical) axis (Figure 4.8, top right, middle right), the RCC in the Z dimension τ_Z decreases, while those in the x and y dimension do not change. Thirdly, the RCCs in the x and y dimensions decrease as camera forward translation increases, while that in the



Figure 4.7: The range image of a forest scene from the Brown range image database. Intensity represents distance values, with distant object looking brighter.

Z dimension does not change (Figure 4.8, bottom left). Fourthly, the RCCs in the x and y decrease as camera's rotation around the optical axis increases, while that in Z does not change its value (Figure 4.8, bottom right). The result is consistent with our summary in Section 4.4.3, showing that RCC in the depth dimension complements those in the two image dimensions. Finally, by virtue of capturing this complementary aspect of the scene geometry, the 3D RCC (τ_{3D}) is clearly superior to the 2D RCC (τ_{2D}) under different types of camera movements ($\tau_{2D} = \frac{1}{2}(\tau_x + \tau_y)$, $\tau_{3D} = \frac{1}{3}(\tau_x + \tau_y + \tau_Z)$, see Figure 4.8, all plots).

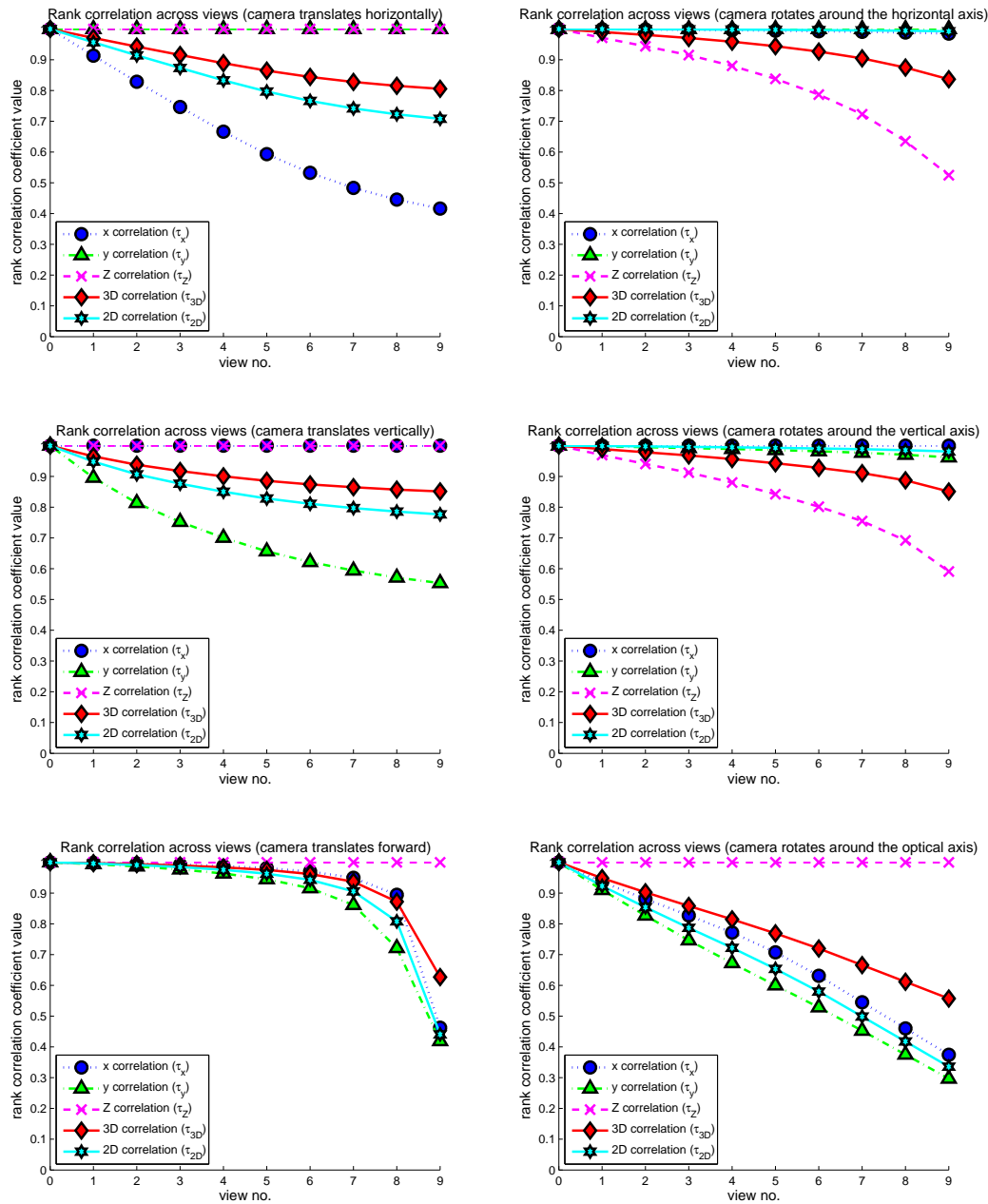


Figure 4.8: Computing RCC on the Brown range data (forest scene): different RCCs across the views are shown when the camera undergoes different types of movements. The top, middle and bottom left figures correspond to translations in the X, Y, and the Z direction respectively, whereas the top, middle and bottom right figures correspond to rotations around the X, Y, and the Z direction respectively. The horizontal axis in each plot represents the various view positions (view 0-9) as the camera moves away from the original position.

4.5.2 Weighting of Individual Pairs

We now discuss how the pairwise ordinal relations a_{ij} should be weighted by s_{ij} , in accordance with the individual pair’s reliability against errors in the TBL motion estimation and stability under viewpoint changes (as discussed in Chapter 2 and Section 4.4 in this chapter respectively). The details of these weighting factors described in this subsection can be skimmed without affecting the understanding of the rest of this thesis.

Weighting the Pairwise Ordinal Depth Relation

The usefulness of a particular pairwise ordinal depth relation between a pair of scene points \mathbf{P}_i and \mathbf{P}_j depends on two factors: 1) the accuracy of the ordinal depth estimates; 2) the amount of viewpoint change. From Proposition 4.4.1, we know that the robustness of the ordinal depth relation a_{Zij} with respect to viewpoint change depends on the angle θ_X and θ_Y (see Figure 4.5). The smaller these angles are, the more robust the ordinal depth relationship to viewing direction change. In Chapter 2, we have also seen that the reliability of the ordinal depth relation with respect to errors in the egomotion estimates is affected by the image separation of the two feature points, which in turn can be related to θ_X and θ_Y given a fixed depth difference. Thus, for both factors, we can characterize the usefulness of an ordinal depth relation by θ_X and θ_Y . In this thesis, we measure these angles from the test scene, but in principle, either the test scene or the reference scene involved can be used for the measurement.

For a more concise characterization of the usefulness, we first rotate the XYZ coordinate system to $X'Y'Z$ such that the new X' -axis is parallel to the

projection of \mathbf{P}_i and \mathbf{P}_j on the XY plane. We now only need to characterize the reliability with one angle θ'_Y , since it is indeed the rotation $\Delta\theta'_Y$ around the Y' -axis that has the greatest perturbation on the ordinal depth relation. In analogy with the angle θ_Y in Figure 4.5, we have in this rotated coordinate system:

$$|\tan \theta'_Y| = \left| \frac{X'_i - X'_j}{Z_i - Z_j} \right|$$

We define a weighting factor s_{Zij} based on θ'_Y to indicate the confidence we can attach to a particular pairwise ordinal depth relationship:

$$s_{Zij} = 1 - \frac{2}{\pi} \arctan \left| \frac{X'_i - X'_j}{Z_i - Z_j} \right|$$

where $s_{Zij} \in [0, 1]$. The above can be written in terms of the image coordinates as:

$$s_{Zij} = 1 - \frac{2}{\pi} \arctan \left(\frac{1}{f} \left| \bar{x}' + \Delta x' \frac{\bar{Z}}{\Delta Z} \right| \right)$$

where $\bar{x}' = \frac{x'_i + x'_j}{2}$, $\bar{Z} = \frac{Z_i + Z_j}{2}$, $\Delta x' = x'_i - x'_j$, and $\Delta Z = Z_i - Z_j$. Clearly, we do not have the true depths Z_i and Z_j at our disposal. Thus we have to replace all the depth quantities with their estimated version. We can also further introduce a factor k in the above equation inside the bracket to adjust between the discriminating power and the rate of acceptance:

$$s_{Zij} = 1 - \frac{2}{\pi} \arctan \left(\frac{k}{\hat{f}} \left| \bar{x}' + \Delta x' \frac{\bar{\hat{Z}}}{\Delta \hat{Z}} \right| \right).$$

For instance, if we want to tolerate large viewpoint change, we could raise the value of k to discount the more fronto-parallel landmark pairs. Only the more in-depth pairs, which are robust against large orientation changes, would

contribute towards computing the rank correlation. In this paper, $k = 1$.

Weighting the Pairwise x or y Relation

We use the shortest distance ($Dist_X$) from the camera center to the border of the forbidden zone in Figure 4.6 as an indicator of the robustness of the ordinal x relations a_{xij} against viewpoint change caused by camera translation in the XCZ plane. $Dist_X$ can be readily shown to be

$$Dist_X = \left| \frac{\bar{Z}(\Delta Z\bar{x} + \Delta x\bar{Z}) - \frac{\Delta Z}{2}(x_i Z_i + x_j Z_j)}{\sqrt{f^2 \Delta Z^2 + (\Delta Z\bar{x} + \Delta x\bar{Z})^2}} \right|$$

where we have used $\frac{X_i - X_j}{Z_i - Z_j} = \frac{1}{f} \left(\bar{x} + \Delta x \frac{\bar{Z}}{\Delta Z} \right)$. Again, we have to replace all the depth quantities with their estimates. Furthermore, to indicate the reliability of the ordinal x relation with respect to cyclotorsion (see Proposition 4.4.2), we further multiply the above with $\left(1 - \frac{|y_i - y_j|}{N_y}\right)$ where N_y is the image dimension in y :

$$Dist'_X = \left| \frac{\hat{Z}(\Delta \hat{Z}\hat{x} + \Delta x\hat{Z}) - \frac{\Delta \hat{Z}}{2}(x_i \hat{Z}_i + x_j \hat{Z}_j)}{\sqrt{\hat{f}^2 \Delta \hat{Z}^2 + (\Delta \hat{Z}\hat{x} + \Delta x\hat{Z})^2}} \right| \left(1 - \frac{|y_i - y_j|}{N_y}\right)$$

Normalizing the above to the interval between 0 and 1, we have

$$s_{xij} = 1 - \frac{2}{\pi} \arctan \frac{1}{Dist'_X}.$$

Similar expressions can be written for $Dist'_Y$ and s_{yij} for the case of a_{yij} . Again, like the preceding subsection, all the quantities are computed from the test scene.

Weighting by Matching Score

In addition to s_{xij} , s_{yij} , and s_{zij} , we should further weight all ij landmark pairs according to the quality of the feature matches with respect to the reference scene, in case one of the feature matches is bad. In particular, each of s_{xij} , s_{yij} , and s_{zij} is further multiplied by the following factor

$$s_{tij} = 1 - \max(t_i, t_j)/T \quad (4.3)$$

where $t_i \in [0, T]$ is the matching score when the correspondence for the i^{th} landmark is established with respect to the reference scene (respectively for j), and T is the upper threshold for the local feature matching score of an acceptable match (the smaller t_i is, the more reliable the matching).

4.6 Robust Scene Recognition¹

Having covered the novel aspect of our ordinal representation, we now briefly outline the other component steps in our scene recognition system (SRS). The first step in our approach acquires the landmark features that are salient and encodes them based on the SURF descriptor [10]. Ordinal depth information of the landmarks is acquired from TBL motion using a gross egomotion estimation step as described in Chapter 3. A scene recognition strategy is then proposed based on both the visual appearance of the SURF landmarks and their 3D geometrical configuration described in terms of the ordinal ranks.

¹The work presented in this section was carried out in collaboration with Mr. Ching Lik Teo [104].

4.6.1 Salient Point Selection

The first step of our proposed SRS is to extract salient regions in the scene for reliable recognition in the later stages. We use a modified version of the human visual saliency (HVS) model [47]. Apart from the intensity and orientation features in the original HVS model, we incorporate a modified opponent colour feature based on the Hue, Saturation and Intensity (HSI) color space to handle illumination changes. Lastly, we have also included two new salient features: long edges and skyline.

We use the HSI color space because it is more robust to a variety of illumination changes such as highlights, shading and shadowing [83] caused by the change in the position of the sun as well as changes in weather conditions. An example of such drastic change is shown in Figure 4.9. The top left image was taken under bright sunlight while the top right image was taken at the same location under diffused lighting from a hazy overcast sky. It can be seen that the appearance change is drastic. Transforming the original RGB image to the HSI space, this problem can be alleviated to some extent; this can be seen from the bottom row of Figure 4.9, which shows the saturation image of the same two scenes in the top row. Hue is known to be intrinsically unstable at low intensities; thus we use all three components of the HSI space to complement each other. Furthermore, for measuring hue similarity, we adopt the modulo distance operator suggested by [83] to handle the wrap-around nature of the angular measure hue.

Long edges are known in the literature as an extremely useful and viewpoint invariant salient feature [43] that are robust against illumination changes and occlusions. Such long edges were exploited in natural sceneries to reliably



Figure 4.9: Grayscale(top row) and saturation(bottom row) for the same scene taken under different illumination conditions.

detect tree trunks for outdoor visual SLAM [7, 16]. In this work, the edge map is extracted by applying the Canny edge detector on the intensity image.

The skyline is used by several authors in past works for scene recognition in navigation [26, 108]. The use of the skyline for scene recognition has also been hypothesized by behavioural scientists for certain species of bees and wasps [71]. Skyline offers the advantage that it remains unchanged for significant changes in the agent’s viewpoint. Motivated by these results, we make use of the skyline feature whenever it is visible in the image.

We propose a simple scheme where the skyline is detected from an image by assuming that: 1) the sky is in general at the top half of the image, 2) it is more luminous (brighter) than the ground, and 3) it contains a higher percentage of blue colour component. Furthermore, as the sky contains relatively few objects, it is relatively textureless compared to the ground that contains abundant vegetation.

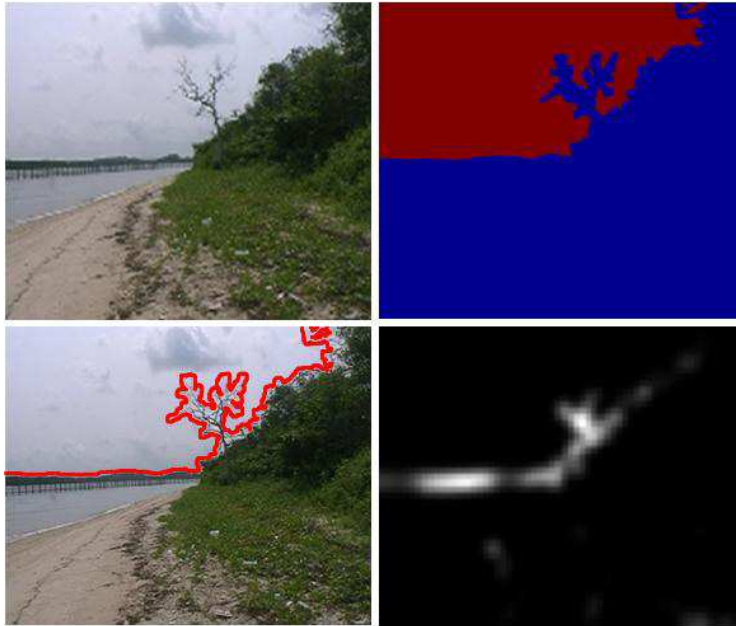


Figure 4.10: An example outdoor scene with its sky-ground segmentation (top right), detected skyline (bottom left) and the resulting saliency map (bottom right).

The first step of the skyline module detects edges using the Canny edge detector. The algorithm then performs several image morphological operations of dilation and filling to create a labeled image that should represent the segmented sky and 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. Figure 4.10 summarizes the various stages in the skyline detection algorithm.

With the five salient features extracted, the saliency algorithm then forms a saliency map [47] which 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. Various image morphological operations are

then applied on the saliency map to extract the salient regions (ROIs) which will serve as the initial landmarks for scene recognition (see Figure 4.11). This step can be summarized as follows:

1. Edges are detected from the input saliency map (Figure 4.11.1) using the Canny edge detector with an initial predefined threshold, t_{edge} to form an edge map (Figure 4.11.2).
2. Dilate the edges with a suitably chosen structuring element such that the broken edges are connected together (Figure 4.11.3). 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 (Figure 4.11.5). The number of ROIs detected is returned and if this number is not between the empirically predetermined minimum and maximum number of salient ROIs desired, the algorithm goes back to step 1 with a suitably adjusted t_{edge} .

The output is a binary labeled map that identifies salient ROIs (Figure 4.11.7). This map is used as a mask to indicate which regions of the image are salient for further encoding by SURF in the next subsection.

4.6.2 Encoding the Appearance and Geometry of the Salient Points

The SURF descriptor [10] attempts to improve the efficiency of the well known SIFT descriptor [62]. It has been shown in [10] to outperform the current state of the art (such as SIFT [62] and GLOH [69]) in terms of recognition accuracy and speed for recognition applications. This makes the SURF algorithm

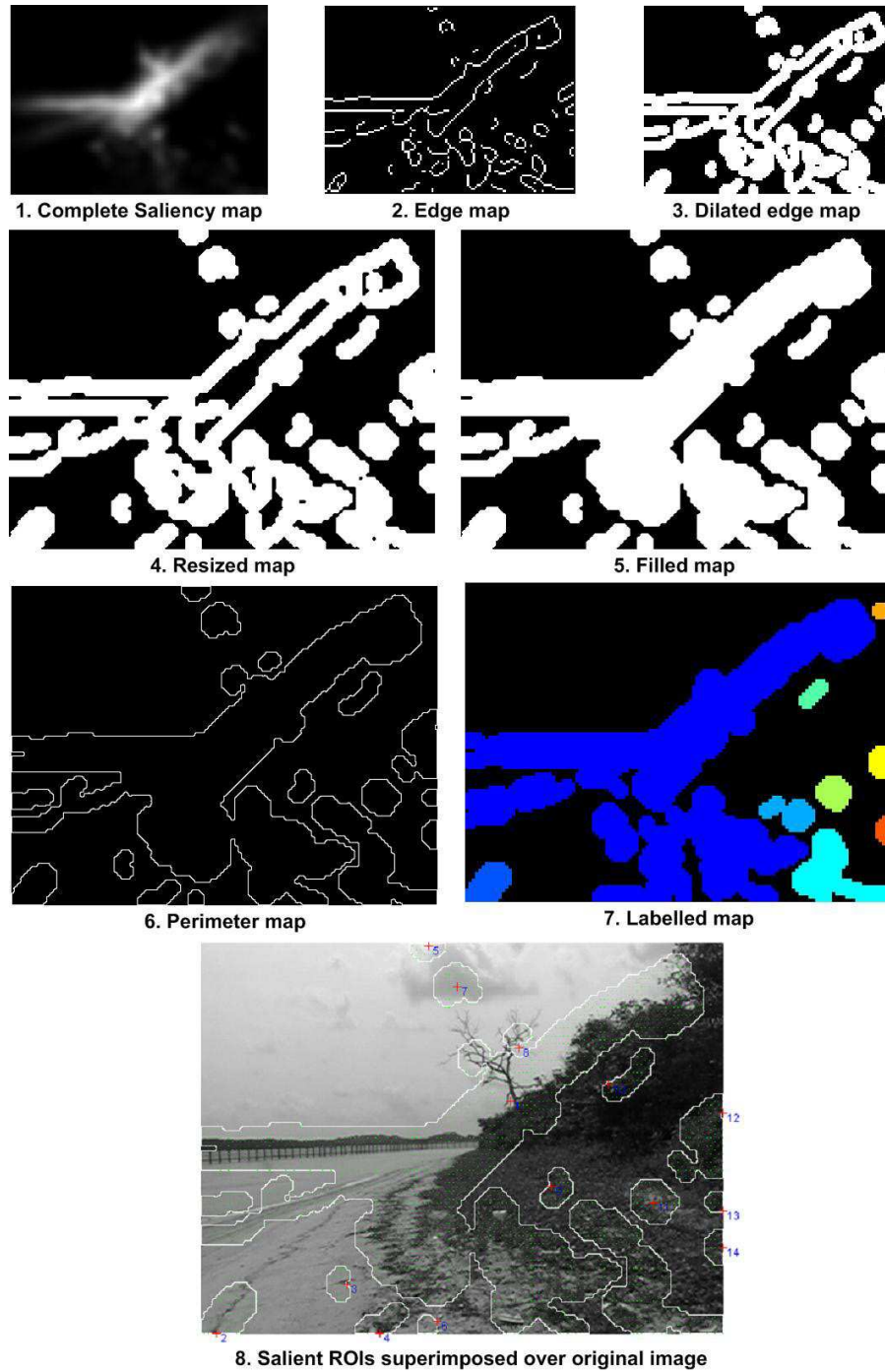


Figure 4.11: Steps that describe the various stages of extracting the salient ROIs using various image morphological operations. The initial saliency map is extracted based on a down-sampled image. The final salient ROIs are boxed in white and highlighted in green.

the descriptor of choice for our proposed SRS, in which we apply it to each component of the HSI space.

At the end of this step, the salient ROIs of the scene are encoded by the SURF descriptors over the three colour spaces. The SURF descriptor itself is made up of a 6D localisation component and a 64D description component (for details, see [10]). We also describe the depth aspect of the scene with the gross depth value \hat{Z} recovered from TBL motion (see Equation (3.3)). The gross depth values recovered under TBL motion are up to a relief transformation of the true depths, as we have shown in Chapter 2 and Chapter 3. Figure 4.12 depicts ordinal depths extracted from the recovered gross depth values. Similar results have been shown in Section 3.4. The same color coding scheme is used here. With these ordinal depths, the 3D RCC can be computed in the subsequent stage.

4.6.3 Measuring Scene Similarity and Recognition Decision

Given a test scene sc_1 and a reference scene sc_2 , both encoded by their respective salient SURF descriptors, we first obtain correspondences between the two sets of SURF features by using the matching algorithm for SIFT descriptors [62]. The threshold for accepting a feature match is purposely set at a less stringent level so that more matches can be obtained even for large illumination and viewpoint changes. We then measure the similarity of the two scenes in terms of both the local visual appearance and the overall geometrical configuration. We call this measure the *Global Scene Correlation Coefficient*



Figure 4.12: Ordinal depths extracted from gross depth estimates under TBL motion, depicted using the rainbow color coding scheme (red stands for near depth; violet for far depth).

G , which is defined as:

$$G(sc_1, sc_2) = \left(\frac{N_{match}}{N_{tot}} \right)^q \tau_{3D} \quad (4.4)$$

where N_{match} and N_{tot} are the number of matches and the total number of features with respect to the test scene respectively. The first term $\frac{N_{match}}{N_{tot}}$ measures the amount of appearance similarity in terms of the percentage of feature matches, whereas the second term τ_{3D} is the 3D RCC in Equation (4.2) that measures the 3D geometrical similarity of the matched feature points. It functions as geometrical constraint which requires high consistency between

the ordinal structures of two sets of matching feature landmarks. Here we name this constraint the *3D ordinal constraint*. The parameter q controls the relative importance attached to the appearance similarity $\frac{N_{match}}{N_{tot}}$ and the geometrical similarity τ_{3D} . In this thesis q is set to 1.

A database of N_{ref} reference images will require N_{ref} pairwise comparisons with the input test scene, each of which will use Equation (4.4) to compute a measure of scene similarity. The candidate match, s_{cand} , is the reference scene that yields the largest Global Scene Correlation Coefficient G_{cand} . G_{cand} thus represents the best match score that is produced by the pairwise comparisons. A decision threshold G_t is set such that a decision, D_f on the test scene

$$D_f = \begin{cases} \text{ACCEPT as positive test scene matching with } s_{cand} & \text{if } G_{cand} \geq G_t \\ \text{REJECT as negative scene} & \text{if } G_{cand} < G_t \end{cases} \quad (4.5)$$

can be made.

4.7 Summary

Now we give a brief summary of this chapter. In this chapter,

firstly, we propose the *3D ordinal space representation* which describes the qualitative structure of a set of landmark points in the scene. This novel space representation encodes the ordinal ranks of points in two image dimensions and the depth dimension. Ordinal depth information encoded in 3D ordinal space representation can be recovered robustly from the motion cue under camera TBl motion as having been discussed

in Chapter 3.

Secondly, the invariance properties of entities in the proposed 3D ordinal space under camera viewpoint change have been well studied. It is found that information in different dimensions compensates each other for different types of camera movements and in different types of scenes.

Thirdly, we propose the use of rank correlation coefficient - ‘*Kendall’s τ* ’ in measuring the global similarity between two sets of points represented in 3D ordinal space. 3D similarity measure τ_{3D} is proposed. We also propose a weighted version of Kendall’s τ , which gives different importance to individual point pairs in τ ’s computation. Proper weighting schemes are developed according to the computational robustness properties and invariance properties.

Lastly, a new scene recognition strategy has been developed. The strategy combines the appearance information encoded by local feature descriptors and the geometrical information encoded by 3D ordinal space representation. The ordinal structure similarity measure τ_{3D} forms a *3D ordinal constraint* for robust scene recognition. The proposed strategy also makes use of the human visual saliency (HVS) information while extracting local features from images. Experiments will be carried out to test the proposed scene recognition strategy in the next chapter.

Chapter 5

Robust Scene Recognition: the Experiment

5.1 Experimental Setup¹

In order to validate our proposed SRS, four challenging databases, often with significant image distortions between the test and reference scenes, are created. They contain image sequences taken from four different environments – indoor(**IND**), a sandy shore(**UBIN**), a tropical rainforest(**NS**) and a mangrove forest(**SBWR**). The sequences were taken with a standard semi-professional camcorder that was moved with hand to simulate a TBL motion. About one third of the scenes are chosen to make up the reference database; these scenes are termed the reference scenes, while the remaining scenes used for testing are called the test scenes . A summary of the four databases is shown in Ta-

¹The work presented in this section was carried out in collaboration with Mr. Ching Lik Teo [104].

Table 5.1: The four databases used in the experiments.

Database	$(N_{ref}, N_{pos}, N_{neg})$	Type
IND	(12, 18, 19)	Indoor
UBIN	(15, 34, 21)	Seashore, open
NS	(19, 32, 32)	Forest, mixed
SBWR	(15, 15, 16)	Mangrove, enclosed

Table 5.1 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 scenes, positive test scenes and negative test scenes used in the particular database. All the reference scenes of each database are shown in Figure 5.1, 5.2, 5.3, 5.5. Each positive test scene corresponds to one reference scene in the database. Negative test scenes are taken in the same type of environment as the reference scenes in each database. The full set of images can be downloaded online². Some challenging positive test examples from the four databases are shown in Figure 5.4.

5.1.1 Database IND

This database consists of indoor scenes taken under different lighting conditions. Test scenes are taken under significant viewpoint and illumination changes from the reference scenes (see Figure 5.1 and Figure 5.4(**IND**)). This database verifies the robustness of the proposed scene recognition strategy against various image distortions due to viewpoint changes and illumination changes in a structured man-made environment.

²http://www.ece.nus.edu.sg/stfpage/electf/robust_SRS.htm



Figure 5.1: Reference scenes in the IND database.

5.1.2 Database UBIN

This database consists of outdoor images taken predominantly along a sandy shore and among its sparse coastal vegetation (see Figure 5.2). It is the nesting habitat of many species of tropical sand-digging wasps where one can see them executing orientation flights and 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 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 and 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



Figure 5.2: Reference scenes in the UBIN database.

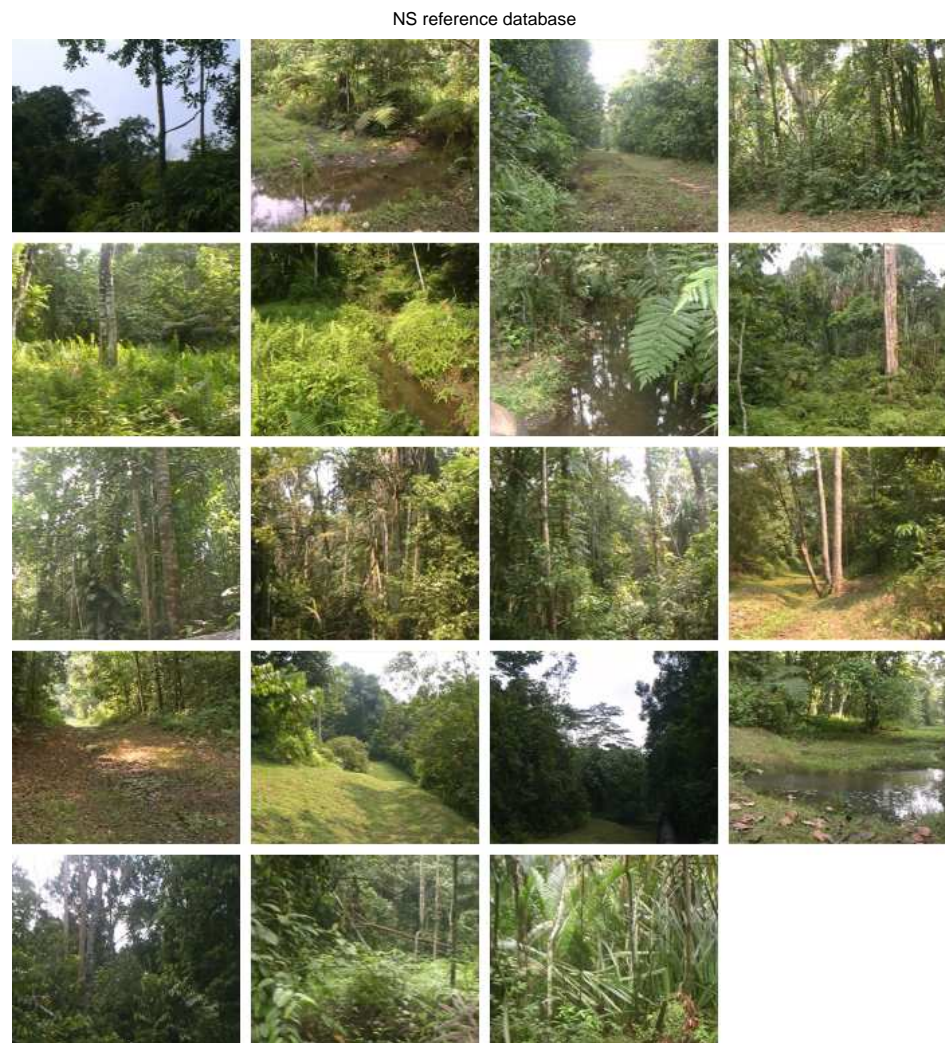


Figure 5.3: Reference scenes in the NS database.

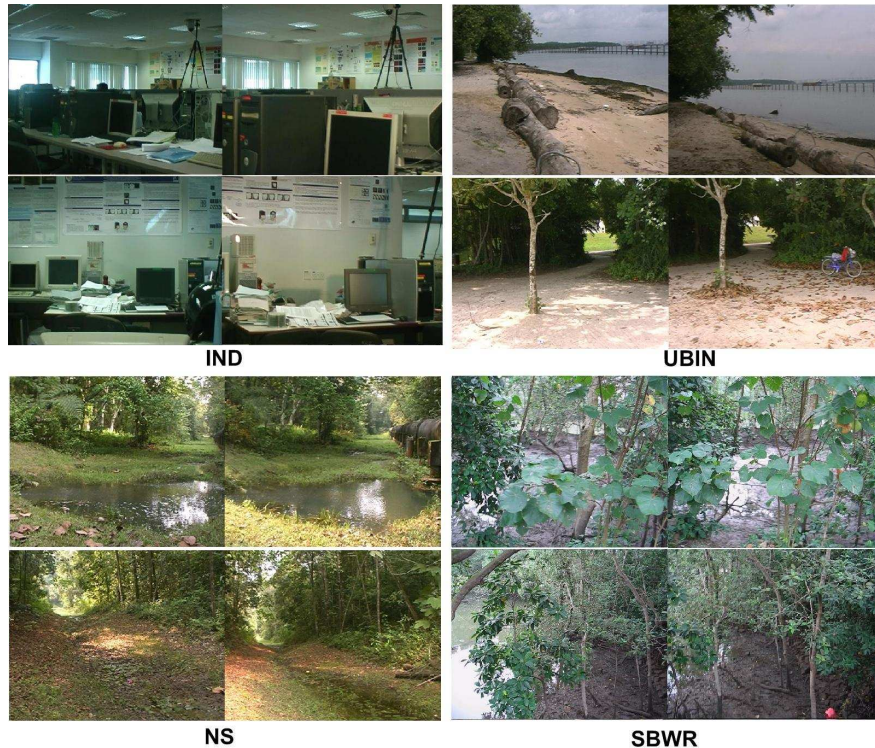


Figure 5.4: Various challenging test (left) and reference scenes (right) of the four databases, two rows ((t)op,(b)ottom) shown per database. **IND**: significant viewpoint and illumination changes ((t) and (b)), **UBIN**: clear versus hazy overcast sky (t) with difference in tide conditions(t), cast shadows and amount of leaf debris(b), **NS**: non-uniform illumination changes(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.

at low tides while the test scenes are taken at high tides which make this database challenging (Figure 5.4(**UBIN**: top)). Human intervention can also cause scenes taken from similar places to appear very different, e.g., leaf debris being swept up as well as the addition/removal of man-made objects in the scene (Figure 5.4(**UBIN**: bottom)). The skyline is also particularly evident in such an environment and has been exploited to aid in scene recognition.

5.1.3 Database NS

This and the next databases contain images with dense foliage and large depth discontinuities, the kind of images that originally motivate us to develop the proposed scene recognition scheme. The **NS** database consists of scenes with lush green vegetation taken at a primary swamp forest in a nature reserve. The scenes are varied in structure, from enclosed forests to semi-open clearings such as streams and ponds (see Figure 5.3). They consist of scenes taken over three occasions. 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 foliage can be quite drastic and are particularly challenging (Figure 5.4(**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 recognizing the scenes (Figure 5.4(**NS**: bottom)).

5.1.4 Database SBWR

In contrast to the ‘openness’ of the **UBIN** database, **SBWR** contains relatively complex scenes taken from an enclosed tropical mangrove forest. Due to the enclosed and dense foliage, there is a much higher percentage of scene points at near depth, resulting in large depth separation of many point pairs and the attendant violation of their ordinal relations in x and y as viewpoint



Figure 5.5: Reference scenes in the SBWR database.

changes. Furthermore, as the mangrove environment is dominated by a few plant species, this database contains many similar-looking vegetation (Figure 5.5). The difficulty in recognition is compounded as the reference scenes are taken purposely at random points in the mangrove forest, and are thus devoid of distinct landmarks that would have been used by human observers, unlike the other two databases of natural scenes. This database tests the proposed scene recognition strategy's tolerance to such natural scenes with many occlusions and clutter, common in an enclosed forest.

5.2 Experimental Results

5.2.1 Recognition Performance and Comparison

We evaluate the overall recognition performance of the proposed SRS in terms of its positive recognition rate P_{rec} , and negative rejection rate P_{rej} when positive and negative test scenes from the four databases are presented respectively. P_{rec} denotes the percentage of positive test scenes that are recognized correctly, whereas P_{rej} represents the percentage of negative test scenes that are correctly rejected. Figure 5.6 shows the recognition-rejection curves for our proposed scene recognition strategy (the outermost red curve in each plot) in the four databases. The curves are generated while the threshold G_t in Equation (4.5) changes. It can be seen that our proposed SRS algorithm achieved a consistently high level of performance over all four databases. It also outperforms the other three methods that we have implemented for comparison. The first method is that of a simple appearance-based scheme (Figure 5.6, the "SURF only" curve (in blue)). This scheme is based on the percentage of SURF feature matches between the test scenes and the reference scenes ($G = \frac{N_{match}}{N_{tot}}$). It is clear that our proposed SRS outperforms the simple appearance-based matching method significantly.

We also compare the performance of our proposed SRS with that of an appearance-based matching method augmented by the epipolar constraint to remove mismatches (Figure 5.6, the "SURF + epipolar constraint (RANSAC)" curve (in green)). RANSAC is used to compute the epipolar geometry in a robust manner. While the epipolar constraint improves the recognition performance over the "SURF only" appearance-based method, it is generally inferior to our proposed method, with the difference in performance much more pro-

nounced in environments with large depth discontinuities (**NS** and **SBWR**).

The third comparison is with the appearance-based matching method augmented with the affine constraint as described in [61] (Figure 5.6, the “SURF + affine constraint” curve (in black)). We divide the image into 12×8 blocks. Inside each block, an affine transformation is fitted between the reference block and the test block. Feature matches which are inconsistent with the affine model are discarded as mismatches. Figure 5.6 shows that the “SURF + affine constraint” method is only better than the “SURF only” method, especially in environments with planar objects where the affine assumption holds (such as in **IND** or **UBIN**). In all four databases, the proposed SRS is shown to outperform the affine method significantly.

Finally, to demonstrate the ability of the proposed SRS in handling substantial viewpoint change, natural dynamic scene change or illumination change, Figure 5.7 and 5.8 depict some specific examples of positive test images successfully recognized by the system (G_t is such that P_{rej} is at least 85%).

5.2.2 Component Evaluation and Discussions

This section compares the performance of the different constituent parts of the proposed SRS so as to understand their contribution to the overall performance. The three curves in each plot of Figure 5.9 respectively show the result of using the full 3D method with weighting scheme (in which $G = \frac{N_{match}}{N_{tot}} \times \tau_{3D}$, $\tau_{3D} = \frac{1}{3}(\tau_{xw} + \tau_{yw} + \tau_{Zw})$), that of the 2D method with weighting scheme (in which $G = \frac{N_{match}}{N_{tot}} \times \tau_{2D}$, $\tau_{2D} = \frac{1}{2}(\tau_{xw} + \tau_{yw})$), and finally that of the 3D method without weighting scheme (in which $G = \frac{N_{match}}{N_{tot}} \times \tau_{3D}$, $\tau_{3D} = \frac{1}{3}(\tau_x + \tau_y + \tau_Z)$). It can be seen that in general, the 3D method outperforms the

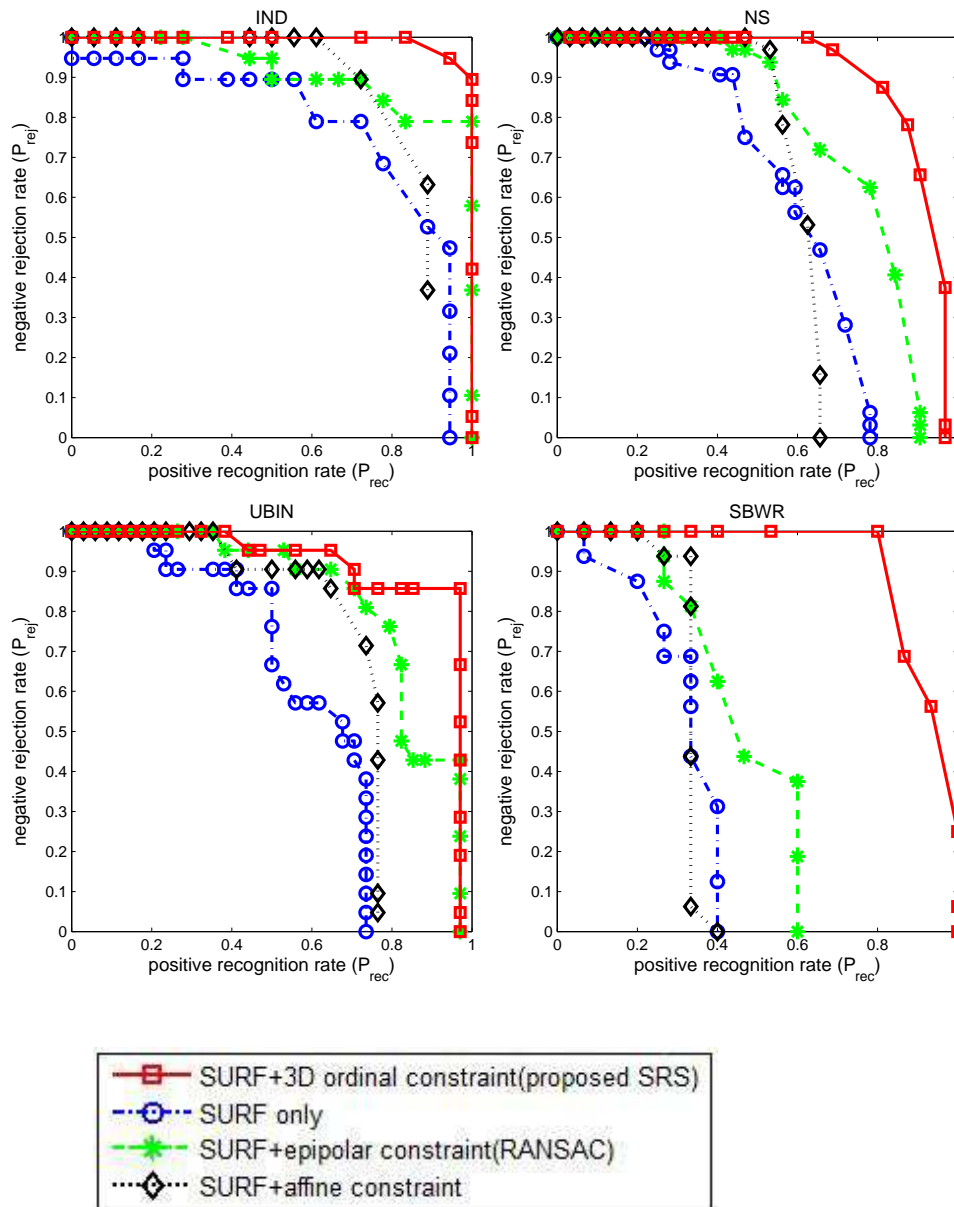


Figure 5.6: Comparison of the proposed SRS(SURF + 3D ordinal constraint), the SURF only method, the SURF + Epipolar Constraint (RANSAC) method, and the SURF + affine constraint method.

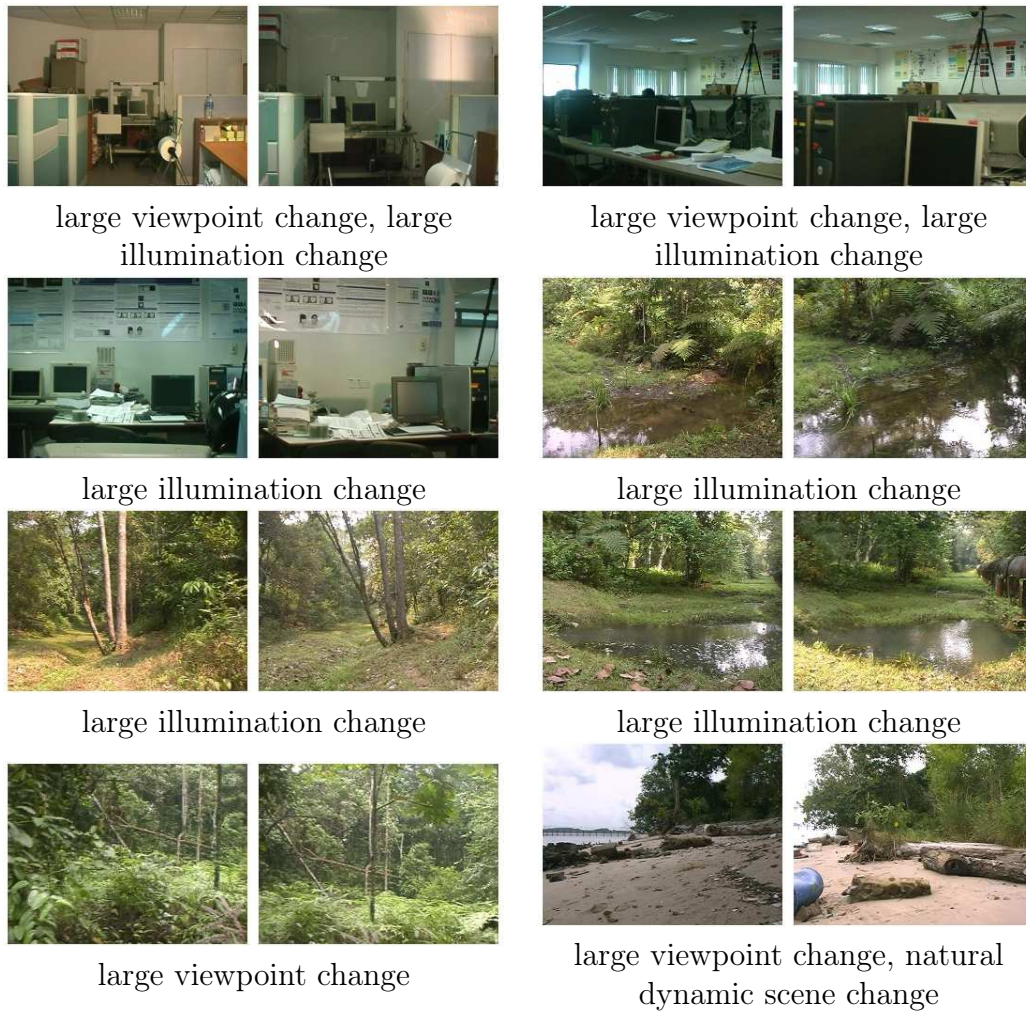


Figure 5.7: Successfully recognized positive test scenes (right image in each subfigure) and their respective reference matches (left image in each subfigure), despite substantial viewpoint changes, natural dynamic scene changes or illumination changes.

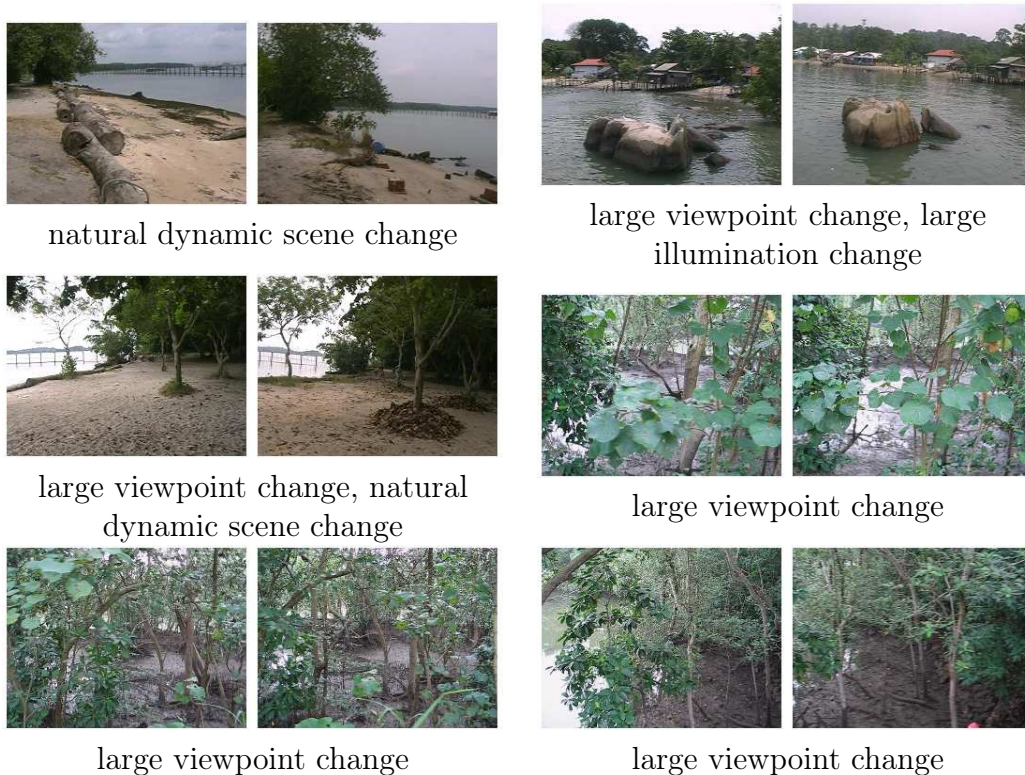


Figure 5.8: More successfully recognized positive test scenes (right image in each subfigure) and their respective reference matches (left image in each subfigure), despite substantial viewpoint changes, natural dynamic scene changes or illumination changes.

2D method, with the margin of improvement depending on the type of environment. In the outdoor forest environment, such as in the **NS** and especially the **SBWR** database, the improvement brought about by the depth dimension is especially significant. This is due to the presence of the large amount of depth discontinuities which destroy the 2D configuration of the features once viewpoint changes. In the indoor and the coastal environment, where scenes are mainly made up of large stretches of roughly planar surfaces, the improvement is not obvious, as 2D geometrical relationship already adequately describes the scene. Finally, it is also shown that the performance of the proposed SRS is significantly improved by adopting the proposed weighting scheme discussed

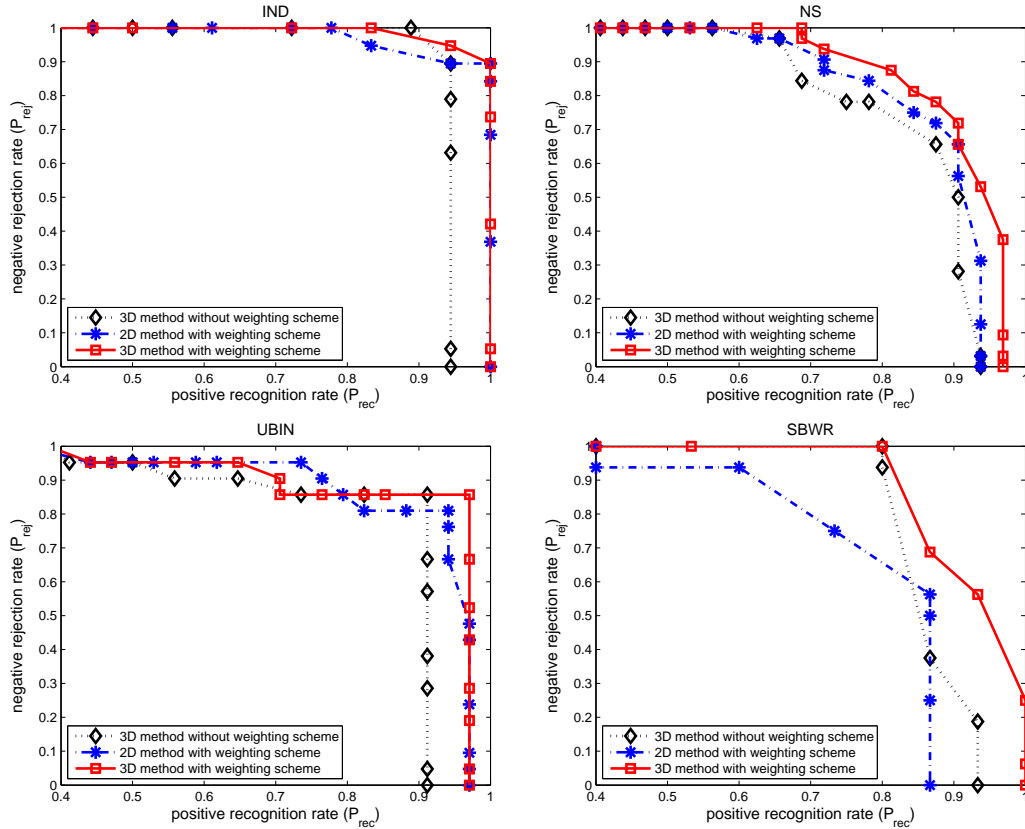


Figure 5.9: Component evaluation: comparing the 3D weighted scheme, the 2D weighted scheme, and the 3D unweighted scheme over the four databases.

in Section 4.5.2, as not all ordinal relations between feature pairs are equally reliable.

Next, we demonstrate with specific examples how the RCCs in the different directions (x , y , and Z) complement each other under different scene types. Table 5.2 shows the RCC values in the x , y , Z dimensions for two types of positive test scenes when being matched with their correct references. When the scene is locally planar or largely fronto-parallel, τ_{xw} registers a high value while τ_{Zw} is low (examples 1 and 2 in Table 5.2). On the contrary, when the scene has large depth variation within a local neighborhood (in-depth scene), τ_{Zw} is able to maintain a high value while τ_{xw} drops significantly (examples 3

and 4 in Table 5.2). This complementary nature of τ_{xw} and τ_{Zw} is consistent with our analysis in Section 4.2 regarding fronto-parallel and in-depth scenes. It is also noted that τ_{yw} maintains high values for both types of scenes. This is due to the fact that there is very little vertical translation between the different viewpoints in our database (most viewpoints are taken from a standing position).





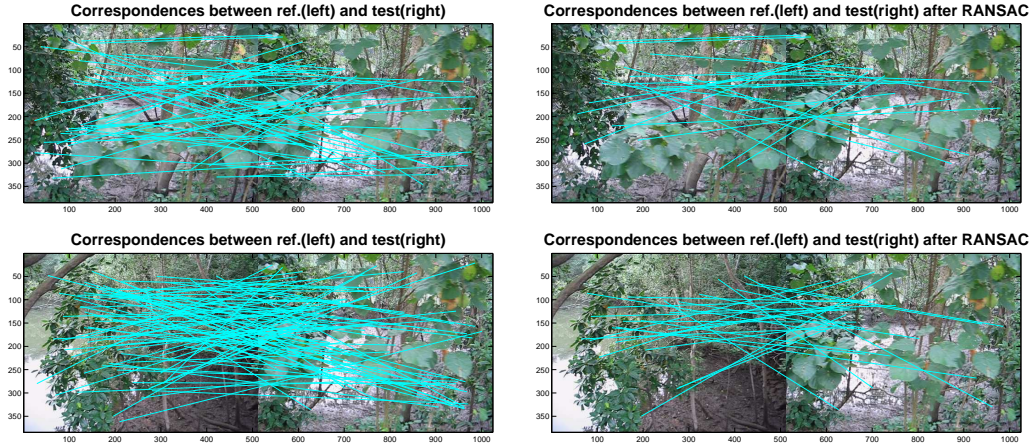
positive test scenes		
	1	2
$P = \frac{N_{match}}{N_{tot}}$	0.1408	0.1328
$(\tau_{xw}, \tau_{yw}, \tau_{Zw})$	(0.1978, 0.2283, 0.0139)	(0.6190, 0.4533, -0.0761)
positive test scenes		
	3	4
$P = \frac{N_{match}}{N_{tot}}$	0.1807	0.2211
$(\tau_{xw}, \tau_{yw}, \tau_{Zw})$	(-0.0613, 0.2325, 0.3617)	(0.0837, 0.1018, 0.2189)

Table 5.2: Rank correlation coefficient in the x , y , and Z dimensions for two types of scenes. 1 and 2: locally planar or largely fronto-parallel scenes. 3 and 4: in-depth scenes.

Finally, we would like to compare the role played by local appearance matching and overall geometric configuration. Three positive test examples are shown in Table 5.3, 5.4, 5.5. These tables show the actual values of $\frac{N_{match}}{N_{tot}}$, τ_{3D} , and G for the three positive test examples. In these examples, the test scenes did not have the largest number of feature matches with their correct

Positive test example 1



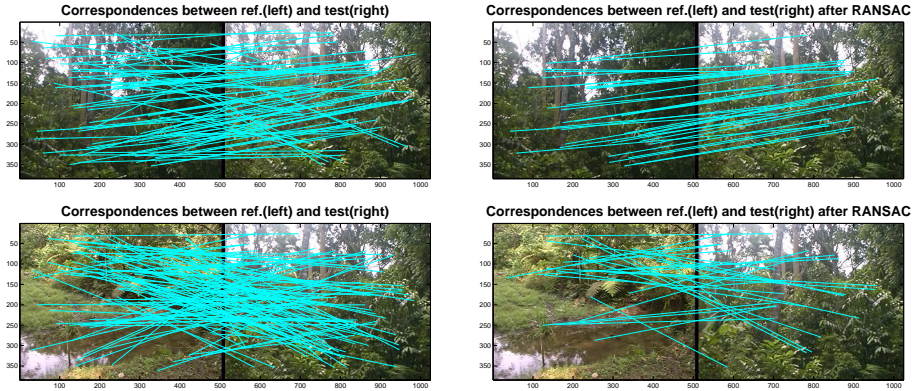
	matching with correct ref.	matching with wrong ref.
$\frac{N_{match}}{N_{tot}}$	0.1264 <	0.1929
$\tau_{3D} = \frac{\tau_{xw} + \tau_{yw} + \tau_{zw}}{3}$	0.2529 >	-0.0167
G	0.0320 >	-0.0032
$\frac{N_{match}}{N_{tot}}$ (after RANSAC)	0.0532 <	0.0621

Table 5.3: The comparison between local appearance matching and overall geometrical consistency: positive test example 1. The top left image pair represents the correspondences between the test and its correct reference scene; the middle left image pair represents the correspondences between the test and the best of the remaining reference scenes (wrong reference scene); the top right pair and middle right pair represent the correspondences left after pruning by the epipolar constraint (RANSAC is used); the bottom table shows the detailed values of $\frac{N_{match}}{N_{tot}}$, τ_{3D} , G , and $\frac{N_{match}}{N_{tot}}$ after pruning by the epipolar constraint (RANSAC).

reference scenes; thus simple appearance-based ("SURF only") method could not provide the correct scene match no matter what the value of the threshold G_t in Equation (4.5) is. This lack of feature matches between the test scenes and their correct reference scenes is due to the significant changes in the image arising from either large viewpoint change (Table 5.3), large illumination change (Table 5.4), or natural dynamic scene change (Table 5.5). Neverthe-

less, when the 3D geometrical information is taken into consideration through the proposed 3D rank correlation, the global scene correlation coefficient G between the correct pair of images registers the highest value (see the sub-table in the bottom of Table 5.3, 5.4 and 5.5).

Positive test example 2

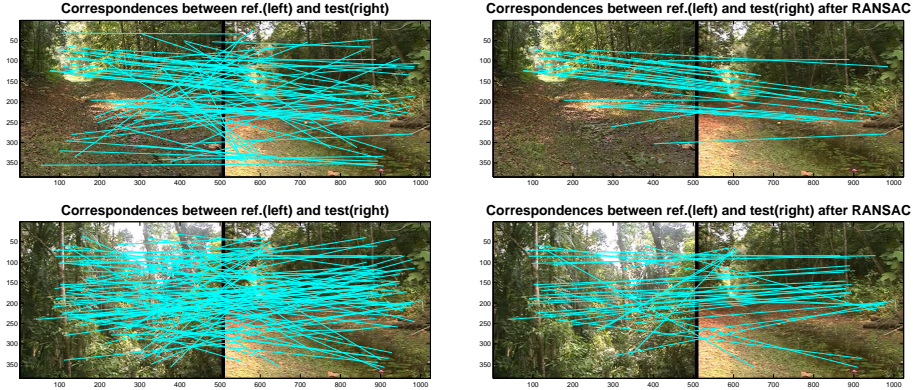


	matching with correct ref.	matching with wrong ref.
$\frac{N_{match}}{N_{tot}}$	0.1468 <	0.1962
$\tau_{3D} = \frac{\tau_{xw} + \tau_{yw} + \tau_{zw}}{3}$	0.3567 >	-0.0921
G	0.0524 >	-0.0181
$\frac{N_{match}}{N_{tot}}$ (after RANSAC)	0.0597 >	0.0529

Table 5.4: The comparison between local appearance matching and overall geometrical consistency: positive test example 2. The top left image pair represents the correspondences between the test and its correct reference scene; the middle left image pair represents the correspondences between the test and the best of the remaining reference scenes (wrong reference scene); the top right pair and middle right pair represent the correspondences left after pruning by the epipolar constraint (RANSAC is used); the bottom table shows the detailed values of $\frac{N_{match}}{N_{tot}}$, τ_{3D} , G , and $\frac{N_{match}}{N_{tot}}$ after pruning by the epipolar constraint (RANSAC).

It is also noteworthy that even when there is quite a number of mismatches between the test scene and its correct reference scene (Table 5.3, 5.4, 5.5: top left pair), the 3D RCC (τ_{3D}) still maintains a high value from the correct

Positive test example 3



	matching with correct ref.	matching with wrong ref.
$\frac{N_{match}}{N_{tot}}$	0.1609 <	0.1874
$\tau_{3D} = \frac{\tau_{xw} + \tau_{yw} + \tau_{zw}}{3}$	0.3119 >	0.0672
G	0.0502 >	0.0126
$\frac{N_{match}}{N_{tot}}$ (after RANSAC)	0.0611 <	0.0652

Table 5.5: The comparison between local appearance matching and overall geometrical consistency: positive test example 3. The top left image pair represents the correspondences between the test and its correct reference scene; the middle left image pair represents the correspondences between the test and the best of the remaining reference scenes (wrong reference scene); the top right pair and middle right pair represent the correspondences left after pruning by the epipolar constraint (RANSAC is used); the bottom table shows the detailed values of $\frac{N_{match}}{N_{tot}}$, τ_{3D} , G , and $\frac{N_{match}}{N_{tot}}$ after pruning by the epipolar constraint (RANSAC).

feature matches, thus indicating its high degree of robustness against outliers. In contrast, the last row in the sub-table of Table 5.3, 5.4 and 5.5 clearly shows that these outliers pose severe problems for methods that employ epipolar constraint (and generally methods that need to find a transformation). While the RANSAC-based epipolar constraint manages to disambiguate the second example (Table 5.4), it cannot successfully handle the first (Table 5.3) and the third examples (Table 5.5). Even in the second example, our proposed method produces a bigger difference between the score of the correct scene and those of the erroneous reference scenes. Clearly the large number of mismatches has fatally impacted on the ability of RANSAC to find the correct global minimum (be it the epipolar geometry or the affine transformation). Not only it cannot eliminate all the outliers between the positive test scene and its correct reference scene (Table 5.3, 5.4, and 5.5: top right pair in each example), it also fails to effectively eliminate the mismatches that occur between the test scene and the erroneous reference scene (Table 5.3, 5.4, and 5.5: middle right pair in each example). Yet such a large number of outliers is inevitable because one has to adopt a relatively lenient threshold for the SURF feature matching, so as to accommodate the potentially big changes between the test and its correct reference (even relatively small viewpoint and illumination change can induce significant local appearance change). This situation is exacerbated by the highly similar features present in natural environment (e.g. trees and foliage), thus generating many feature matches between a test scene (whether positive or negative) and a non-correct reference scene. In Table 5.6, some of the negative test examples that have high $P = \frac{N_{match}}{N_{tot}}$ value with some references are shown. However, these feature matches all have low geometrical consistency, as reflected by the low or negative τ_{3D} values.

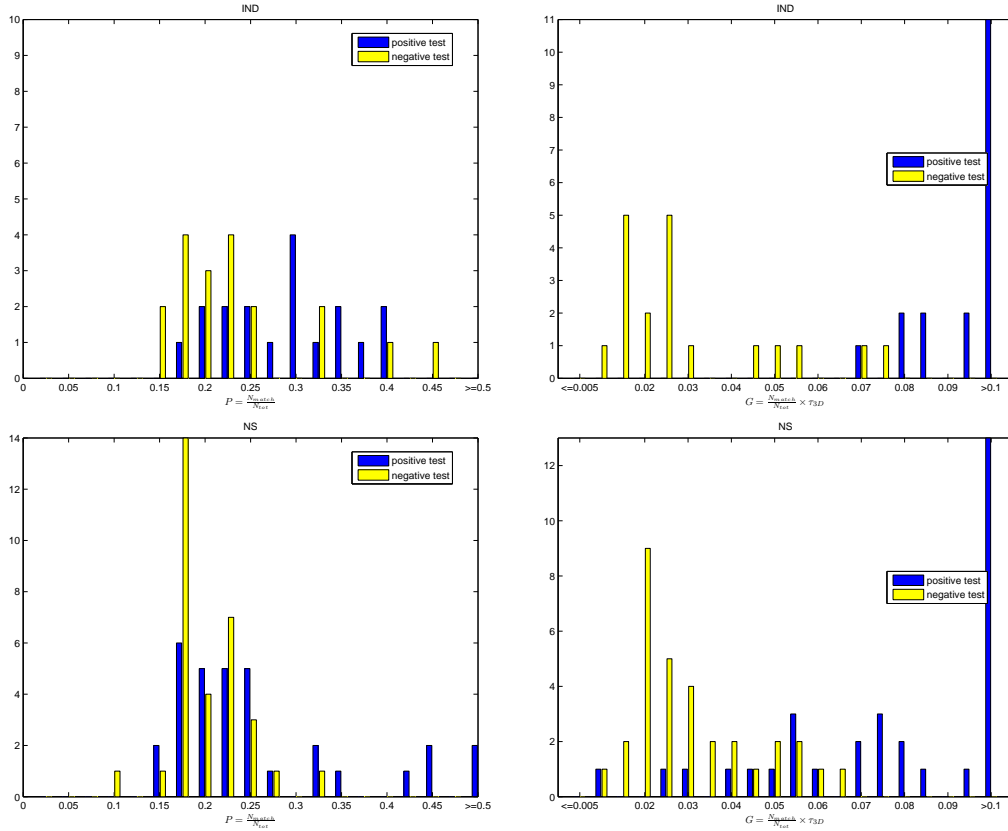


Figure 5.10: Separation of the positive test set and the negative test set in IND and NS databases (respectively the four rows). Left column: histogram of the SURF matching percentage ($P = \frac{N_{match}}{N_{tot}}$) for both the positive and the negative test set. Right column: histogram of the global scene correlation coefficient (G) for both the positive and the negative test set. For positive test scene, P and G are the values between the test scene and its correct reference scene; for negative test scene, P and G are the biggest values obtained when the test scene is compared with all the reference scenes.

To demonstrate the increase in discriminating power brought about by the 3D geometrical information in a more quantitative manner, we plot for both the positive and negative test set the distribution of the $P = \frac{N_{match}}{N_{tot}}$ value (percentage of SURF feature match) in the left column of Figure 5.10 and 5.11, and compare it with the distribution of the G value in the right column of Figure 5.10 and 5.11. As can be seen from the histograms, there is a significant degree of overlap in the values of P for the positive (in blue) and

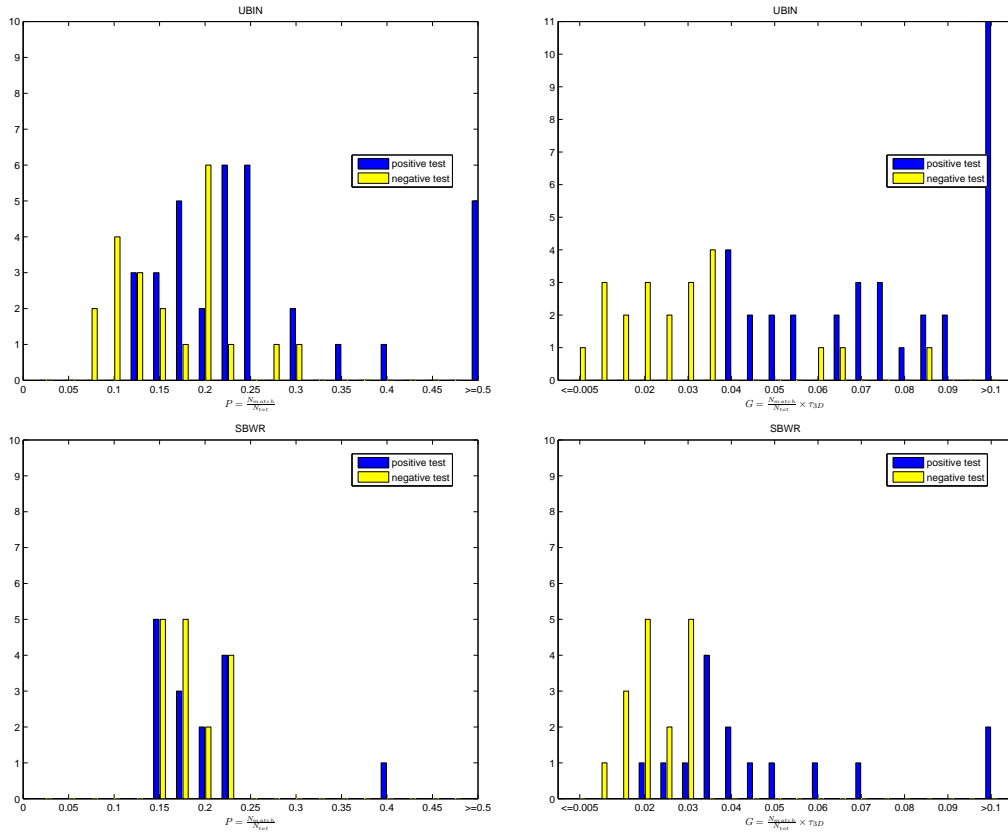


Figure 5.11: Separation of the positive test set and the negative test set in UBIN and SBWR databases (respectively the four rows). Left column: histogram of the SURF matching percentage ($P = \frac{N_{match}}{N_{tot}}$) for both the positive and the negative test set. Right column: histogram of the global scene correlation coefficient (G) for both the positive and the negative test set. For positive test scene, P and G are the values between the test scene and its correct reference scene; for negative test scene, P and G are the biggest values obtained when the test scene is compared with all the reference scenes.

negative (in yellow) test sets, making their separation impossible, whereas for the measure G , the distribution is in a much more obliging form for separation.

5.3 Summary

In this chapter, we have carried out a number of experiments to evaluate the performance of the proposed scene recognition strategy. Using 3D ordinal

constraints allow us to bypass the limitation imposed by global or semi-local transformation model such as the epipolar constraint or affine constraint; this is especially relevant in the context of outdoor natural scenes where local feature descriptors are not very informative and discriminative and the scene content might change over time. The result is that our scene recognition algorithm shows good performance on an extensive database of both indoor and outdoor natural scenes.

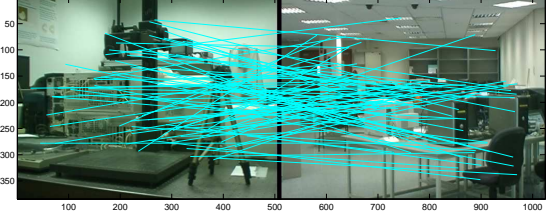
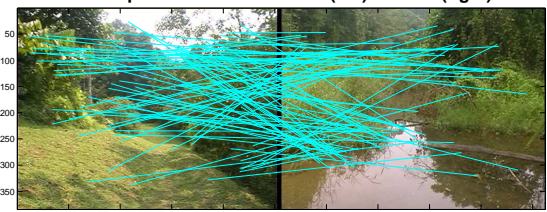
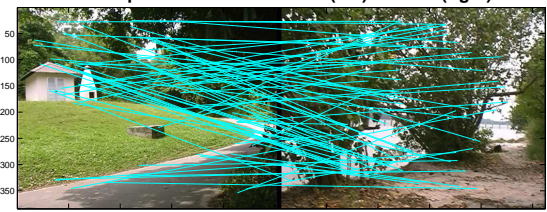

<p style="text-align: center;">Correspondences between ref.(left) and test(right)</p>  <p style="text-align: center;">$\frac{N_{match}}{N_{tot}} = 0.2069, \quad \tau_{3D} = 0.0227, \quad G = 0.0047$</p>
<p style="text-align: center;">Correspondences between ref.(left) and test(right)</p>  <p style="text-align: center;">$\frac{N_{match}}{N_{tot}} = 0.2231, \quad \tau_{3D} = -0.0309, \quad G = -0.0069$</p>
<p style="text-align: center;">Correspondences between ref.(left) and test(right)</p>  <p style="text-align: center;">$\frac{N_{match}}{N_{tot}} = 0.1825, \quad \tau_{3D} = -0.0473, \quad G = -0.0086$</p>
<p style="text-align: center;">Correspondences between ref.(left) and test(right)</p>  <p style="text-align: center;">$\frac{N_{match}}{N_{tot}} = 0.2009, \quad \tau_{3D} = 0.0181, \quad G = 0.0036$</p>

Table 5.6: Some negative test examples which have high $\frac{N_{match}}{N_{tot}}$ value with some reference scenes. The correspondences and the actual values of $\frac{N_{match}}{N_{tot}}$, τ_{3D} , G between the test and the reference scene are shown.

Chapter 6

Future Work and Conclusion

6.1 Future Work Directions

Now we sketch some brief proposals of possible future works.

6.1.1 Space Representation: Further Studies

Global vs. Local: The result that ordinal depth resolution decreases as visual angle increases suggests that accurate ordinal 3D structure recovery is ensured in small local image neighborhood. If a global space representation is desired, how to describe the global links between regions with locally accurate 3D ordinal structure is an important issue. Such an issue of mediating between the local and the global could perhaps be called the *glocalization problem*.

Ordinal vs. Metric: It was suggested by Cutting [28] that the perceptual space might be really ordinal and this space converges to a metric space when the feature points become dense. How to take into account of the

resolution of ordinal structures in the space (as we have discussed in this thesis) and construct metric information from these ordinal measurements could be interesting future research direction.

6.1.2 Scene Recognition and SLAM

Future work should explore the integration of the proposed scene recognition strategy to a visual SLAM application on a mobile agent capable of simulating TBL motion. This will be especially useful for localization in environments where current navigational technologies (e.g. GPS) remain unusable due to the thick forest foliage. Furthermore, future work should also focus on how the reference scenes can be automatically selected. Ideally, reference scenes should contain certain distinctive and unique features that make them stand out from the whole database so that place recognition is facilitated. Modeling how humans organize and choose salient scenes from the database remains a difficult and open problem. Finally, we hope that our proposed scene recognition system will spur more scene recognition research to focus on outdoor natural scenes, which remains a challenging problem. The availability of the online image databases serves this purpose.

6.1.3 Ordinal Distance Information for 3D Object Classification

Given the work in this thesis which has demonstrated that it is feasible to exploit 3D qualitative information for performing visual recognition task, one might be interested in developing other similar schemes and applying them to other recognition or classification tasks.

One possible future work could be using the pairwise ordinal distance on 3D objects to perform object classification. Here the ordinal distance could mean the rank of the distance value between each pair of points on the 3D object. If we have N feature points on the object, we will have $N(N - 1)$ distance values to rank. In ordinal multi-dimensional scaling, it is known that as the number of points is large, an Euclidean model can be embedded into the set of points of which only ordinal distances among points have been measured. Such Euclidean model becomes more and more precise as points become denser [14]. If we consider Cutting's hypothesis [28] that the human visual perceptual space is ordinal by its nature, we may suggest that some high level vision task such as recognition could be carried out based on the ordinal distances among the feature points in space. In the following, we give a preliminary test on this possibility using some 3D object models from the Princeton's repository of 3D models [94].

We select four models of tables and four models of airplanes from the repository (Figure 6.1). Our aim is to observe how the ordinal distances among points on the object may characterize the class of the object (table or plane), and how the number of feature points may affect such characterization. For this aim, the models are first aligned manually. We simply sample a number of vertices from each model. These sampled vertices are uniformly chosen from all the vertices ¹(see Figure 6.2). We order the sampled vertices according to certain prespecified direction of traversal. The 3D Euclidean distance between each pair of points on the object is computed. For an object with N sample

¹The Computational Geometry Algorithms Library (CGAL) is used. <http://www.cgal.org>.

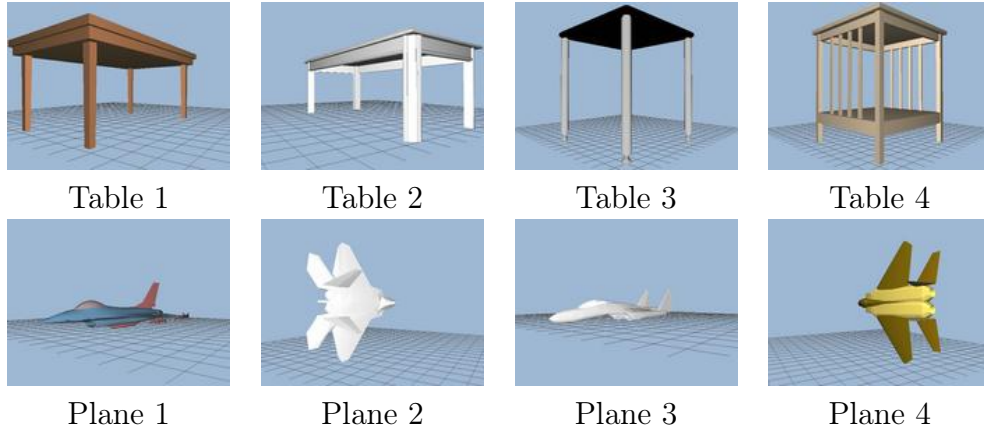


Figure 6.1: Images of models of tables and planes

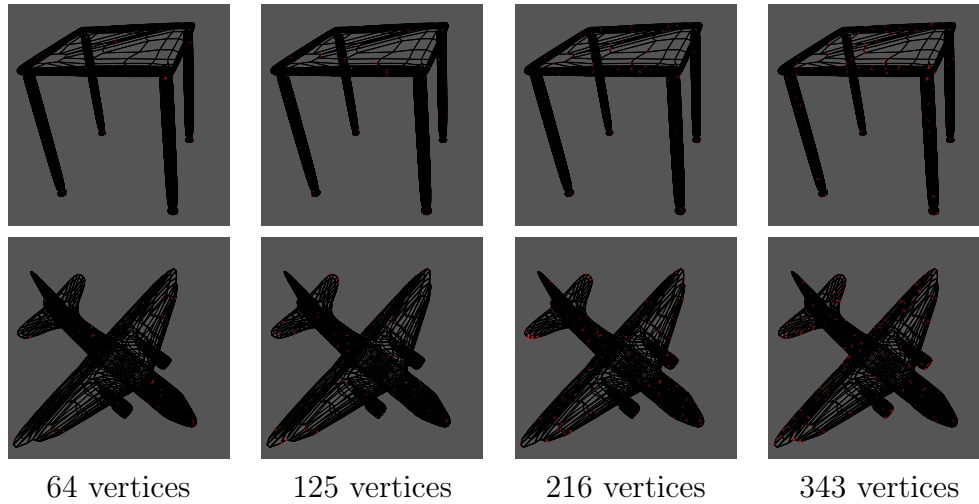


Figure 6.2: Sampling with different number of vertices.

points, the $N(N - 1)$ distance values are ranked. An $N \times N$ *rank proximity matrix* is formed with the entry (i, j) denoting the rank of the distance value between point i and point j ($i, j = 1, 2, \dots, N$).

The rank proximity matrices for the table objects are shown in Figure 6.3 and those for planes are shown in Figure 6.4. It can be seen that matrices within the same class exhibit similar patterns, whereas the patterns are different between classes. This indicates that the rank proximity matrix carries object shape information associated with the object class.

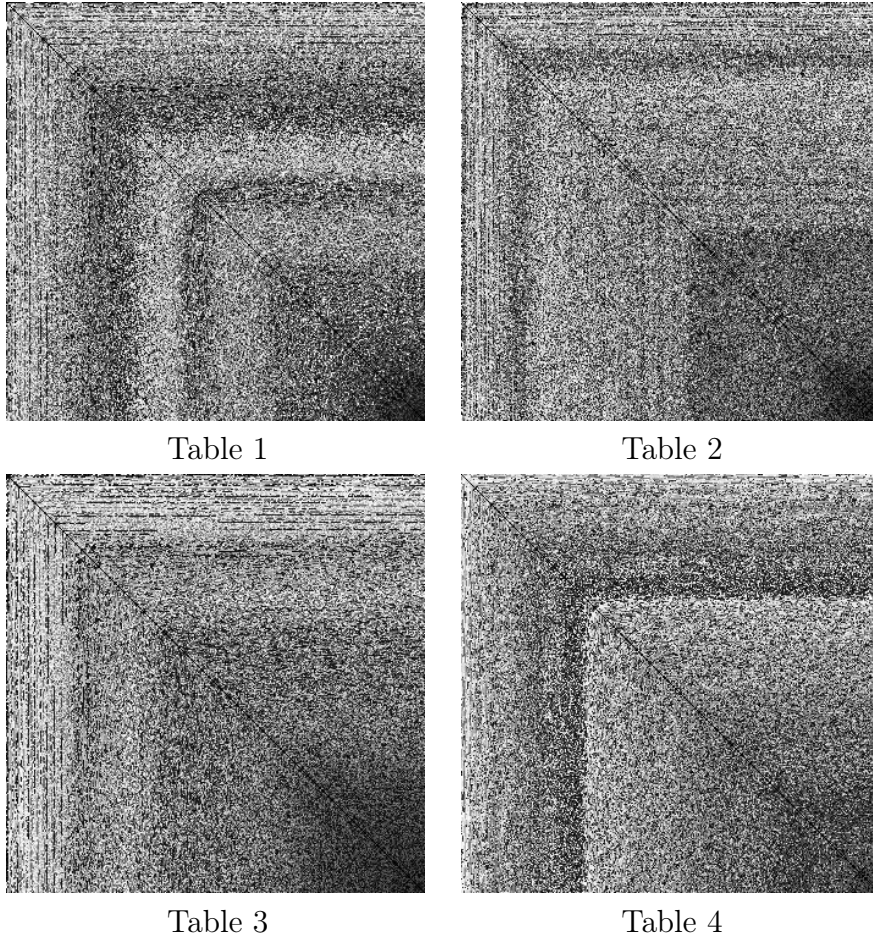


Figure 6.3: Rank proximity matrices of table models, computed from 343 sampled vertices.

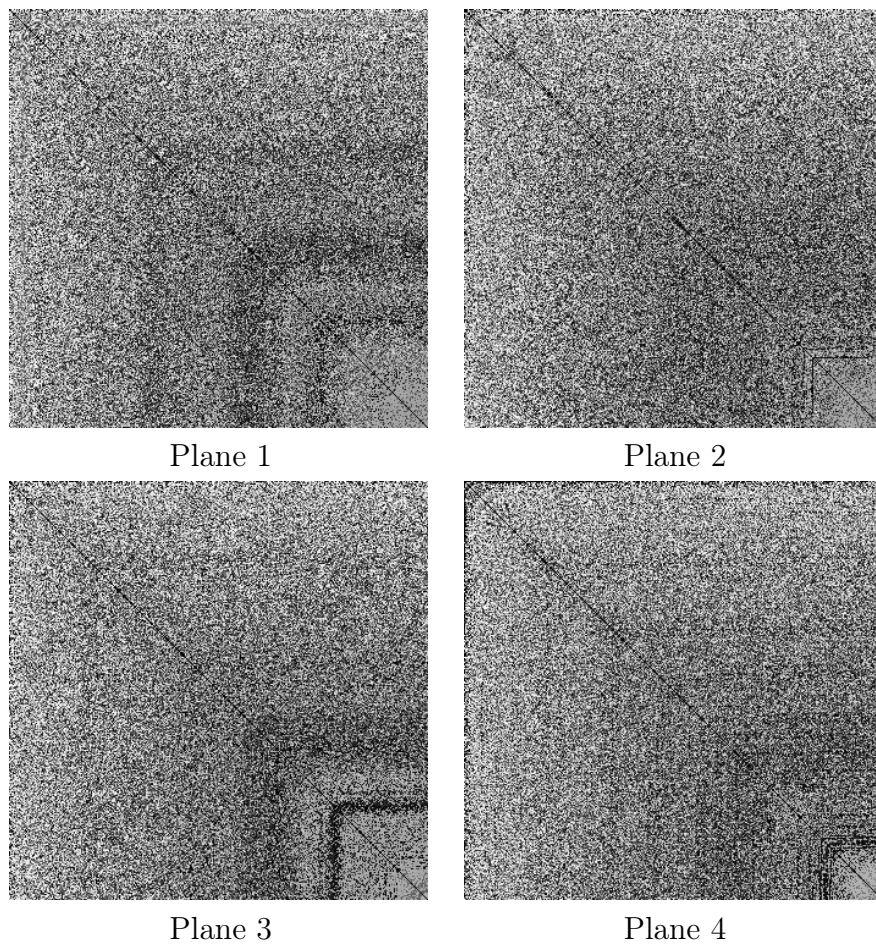


Figure 6.4: Rank proximity matrices of plane models, computed from 343 sampled vertices.

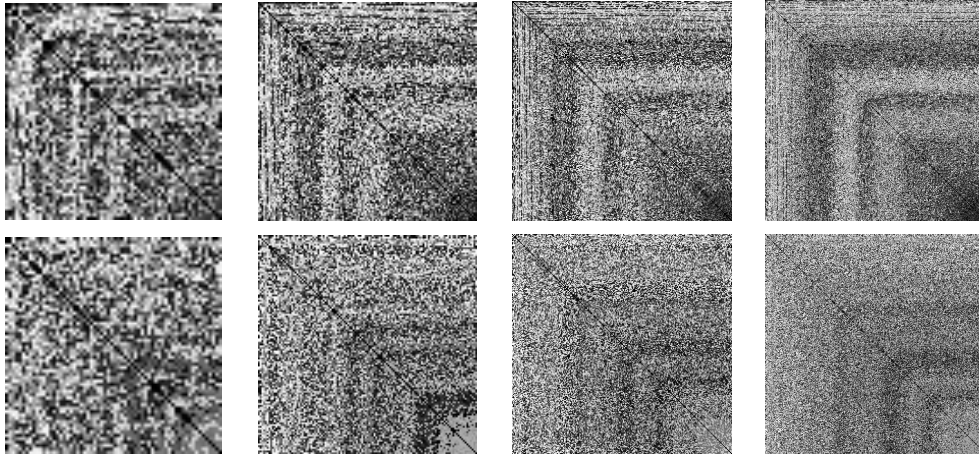


Figure 6.5: Rank proximity matrices with different number of sampled vertices. Upper row: table class, lower row: plane class. Sample number increases from left to right (as shown in Figure 6.2).

Another test is carried out to investigate the behavior of the rank proximity matrices under different number of sampled vertices (see Figure 6.5). It is shown here that the class pattern becomes more and more unclear as the number of sampled vertices decreases; however, the topology of the pattern can still be discerned. This might indicate that the rank proximity matrix still carries the class information under sparse feature points, though the information becomes weak in the case.

6.2 Conclusion

In this thesis, we have carried out extensive studies focusing on ordinal depth: from its computational properties from SFM and its robust acquisition from specific motion cue; to its application in scene recognition. Through these studies, new theories and techniques have been developed towards understanding such ordinal/qualitative geometrical information as well as its exploitation in practical vision systems.

Firstly, based on the proposed depth distortion model, we have analyzed the ability of SFM algorithms in judging ordinal depth. Analytic results have shown that in small image neighborhood, one can get ordinal depth up to certain resolution. The resolution decreases as the visual angle between the pair of image points increases. The results imply that a proper space representation might be non-uniform, with different resolutions varying according to different sizes of the neighborhoods. Future work can be carried out in developing such space representations, as we will discuss in more detail in the next section.

Secondly, we analyzed the ordinal depth properties and showed that the lateral motion is a good strategy for ordinal depth recovery. Based on this insight, together with the bio-inspired TBL motion, we developed an active camera control method to acquire robust ordinal depth and use it in our proposed scene recognition system. One feature of our proposed method is that precise camera control is not required.

Thirdly, we have shown that qualitative spatial information in the two image dimensions and the depth dimension complement each other in terms of their stability to camera viewpoint changes and in different types of scenes. Thus it is crucial to encode the *3D ordinal constraint* in our scene recognition system. Further studies on the invariance properties of various qualitative geometrical entities might lead us to more robust algorithms for performing various practical vision tasks.

Fourthly, a scene recognition strategy has been proposed and tested extensively under indoor and outdoor environments. The proposed strategy combines the local feature appearance information together with the 3D ordinal geometrical information. Results show that our proposed strategy outperforms the pure local feature based method as well as methods using global or

semi-local transformations. Our proposed scene recognition system provides a successful example of a system subscribing to the purposive and active vision paradigm. It also demonstrates the feasibility of exploiting 3D qualitative geometrical information in performing scene recognition.

Appendix A

Acronyms

FOE: Focus of Expansion

FOV: Field of View

RCC: Rank Correlation Coefficient

SFM: Structure from Motion

SRS: Scene Recognition System

TBL: Turn-Back-and-Look

VOD: Valid Ordinal Depth

Appendix B

Author's Publications

1. **Shimiao Li**, Loong-Fah Cheong: Behind the Depth Uncertainty: Resolving Ordinal Depth in SFM. European Conference on Computer Vision (3) 2008: 330-343.
2. Ching Lik Teo, **Shimiao Li**, Loong-Fah Cheong, Ju Sun: 3D Ordinal Constraint in Spatial Configuration for Robust Scene Recognition. International Conference on Pattern Recognition 2008: 1-5.
3. Loong-Fah Cheong, **Shimiao Li**: Error Analysis of SFM Under Weak-Perspective Projection. Asian Conference on Computer Vision (2) 2006: 862-871.
4. **Shimiao Li**, Loong-Fah Cheong, Ching Lik Teo: 3D Ordinal Geometry for Scene Recognition Using TBL Motion. submitted to International Journal of Computer Vision.

Bibliography

- [1] G. Adiv. Determining 3-D motion and structure from optical flow generated by several moving objects. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 7:384–401, 1985.
- [2] G. Adiv. Inherent ambiguities in recovering 3-D motion and structure from a noisy flow field. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 11:477–489, 1989.
- [3] J. Aloimonos. Purposive and qualitative active vision. In *Pattern Recognition, 1990. Proceedings., 10th International Conference on*, volume 1, pages 346–360, Atlantic City, NJ, USA, 1990.
- [4] J. Aloimonos, I. Weiss, and A. Bandyopadhyay. Active vision. *International Journal of Computer Vision*, 1:333–356, 1988.
- [5] Y. Aloimonos, C. Fermüller, and A. Rosenfeld. Seeing and understanding: Representing the visual world. *ACM Computing Surveys*, 27:307–309, 1995.
- [6] J. Amores, N. Sebe, and P. Radeva. Fast spatial pattern discovery integrating boosting with constellations of contextual descriptors. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2005.
- [7] D. C. Asmar, J. S. Zelek, and S. M. Abdallah. Tree trunks as landmarks for outdoor vision slam. In *IEEE Conference on Computer Vision and Pattern Recognition Workshop*, pages 196–196, 2006.
- [8] D. Ballard. Animate vision. *Artificial Intelligence*, 48:57–86, 1991.
- [9] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool. SURF: Speeded-up robust features. *Computer Vision and Image Understanding*, 110:346–359, 2008.
- [10] H. Bay, T. Tuytelaars, and L. Van Gool. SURF: Speeded up robust features. In *European Conference on Computer Vision*, May 2006.

-
- [11] S. S. Beauchemin and J. L. Barron. The computation of optical flow. *ACM Computing Surveys*, 27:433–467, 1996.
- [12] P. N. Belhumeur, D. J. Kriegman, and A. L. Yuille. The bas-relief ambiguity. *International Journal of Computer Vision*, 35:1040–1046, 2001.
- [13] M.J. Black, Y. Aloimonos, I. Horswill, G. Sandini, C.M. Brown, J. Malik, and M.J. Tarr. Action, representation, and purpose: Re-evaluating the foundations of computational vision. In *In Proceedings of the International Joint Conference on Artificial Intelligence*, 1993.
- [14] I. Borg and P. Groenen. *Modern Multidimensional Scaling: theory and applications*. Springer-Verlag, New York, 2005.
- [15] M. Brown and D. G. Lowe. Unsupervised 3D object recognition and reconstruction in unordered datasets. In *Proc. 5th Int'l Conf. 3-D Digital Imaging and Modeling (3DIM '05)*, pages 56–63, 2005.
- [16] R. Burge, J. Mulligan, and P.D. Lawrence. Using disparity gradients for robot navigation and registration. In *IEEE/RSJ Int'l Conf. on Intelligent Robots and Systems*, volume 1, pages 539–544, October 1998.
- [17] Edited by M. J. Swain and M. A. Stricker. Promising directions in active vision. *International Journal of Computer Vision*, 11:106–129, 1993.
- [18] G. Carneiro and A. D. Jepson. Flexible spatial configuration of local image features. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 29:2089–2104, 2007.
- [19] B. A. Cartwright and T. S. Collett. Landmark learning in bees: Experiments and models. *J. of Comparative Physiology A*, 151:521–543, 1983.
- [20] L-F. Cheong. *Geometry of the Interaction between 3D shape and Motion Perception*. PhD thesis, Dept. Computer Sci., University of Maryland, 1996.
- [21] L-F. Cheong, C. Fermüller, and Y. Aloimonos. Effects of errors in the viewing geometry on shape estimation. *Computer Vision and Image Understanding*, 71:356–372, 1998.
- [22] L-F. Cheong and S. Li. Error analysis of sfm under weak-perspective projection. In *Asian Conference on Computer Vision (ACCV'06)*, pages 862–871, 2006.

- [23] L-F. Cheong and T. Xiang. Characterizing depth distortion under different generic motions. *International Journal of Computer Vision*, 71:356–372, 1998.
- [24] T. S. Collett and M. Lehrer. Looking and learning: A spatial pattern in the orientation flight of the wasp *vespula vulgaris*. *Proceedings: Biological Sciences*, 252:129–134, May 1993.
- [25] Robert Collins and Yanghai Tsin. Calibration of an outdoor active camera system. In *IEEE Computer Vision and Pattern Recognition (CVPR '99)*, pages 528 – 534, June 1999.
- [26] F. G. Cozman, E. Krotkov, and C. E. Guestrin. Outdoor visual position estimation for planetary rovers. *Autonomous Robots*, 9:135–150, 2000.
- [27] M. Cummins and P. Newman. Fab-map: Probabilistic localization and mapping in the space of appearance. *Int'l J. Robotics Research*, 27:647–665, 2008.
- [28] J. E. Cutting. Reconceiving perceptual space. In H. Hecht, M. Atherton, and R. Schwartz, editors, *Looking into pictures : an interdisciplinary approach to pictorial space*. MIT Press, 2003.
- [29] K. Daniilidis and M. E. Spetsakis. Understanding noise sensitivity in structure from motion. In Y. Aloimonos, editor, *Visual Navigation*, pages 61–88. Academic Press, 1997.
- [30] M. Devy, R. Chatila, P. Fillatreau, S. Lacroix, and F. Nashashibi. On autonomous navigation in a natural environment. *Robotics and Autonomous Systems*, 16:5–16, 1995.
- [31] R. Dutta and M. A. Snyder. Robustness of structure from binocular known motion. In *Motion91*, pages 81–86, 1991.
- [32] S. Edelman. D. Marr. In N. J. Smelser and P. B. Baltes, editors, *International Encyclopaedia of Social and Behavioral Sciences*. ELSEVIER, 2001.
- [33] O. D. Faugeras. *Three-Dimensional Computer Vision*. MIT Press, 1993.
- [34] R. Fergus, P. Perona, and A. Zisserman. Weakly supervised scale-invariant learning of models for visual recognition. *International Journal of Computer Vision*, 71:273–303, 2007.
- [35] C. Fermüller. Passive navigation as a pattern recognition problem. *International Journal of Computer Vision*, 14:147–158, 1995.

- [36] C. Fermüller and Y. Aloimonos. Representations for active vision. In *Proc. IJCAI*, pages 20–26, 1995.
- [37] C. Fermüller, D. Shulman, and Y. Aloimonos. Observability of 3D motion. *International Journal of Computer Vision*, 37.
- [38] J. M. Fernandez and B. Farella. Is perceptual space inherently non-euclidean? *Journal of Mathematical Psychology*, 53:86–91, 2009.
- [39] V. Ferrari, T. Tuytelaars, and L. Van Gool. Simultaneous object recognition and segmentation from single or multiple model views. *Int'l J. Computer Vision*, 67:159–188, April 2006.
- [40] M. Goesele, N. Snavely, S.M. Seitz, B. Curless, and H. Hoppe. Multi-view stereo for community photo collections. In *Int'l Conf. on Computer Vision*, 2007.
- [41] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2000.
- [42] D. J. Heeger and A. D. Jepson. Subspace methods for recovering rigid motion i: Algorithm and implementation. *International Journal of Computer Vision*, 7:95–117, 1992.
- [43] G. Heidemann. The long-range saliency of edge- and corner-based salient points. *Perception*, 14:1701–1706, November 2005.
- [44] Y. Hel-Or and S. Edelman. A new approach to qualitative stereo. In *Int'l Conf. on Pattern Recognition*, volume 1, pages 316–320, 1994.
- [45] K.L. Ho and P. M. Newman. Detecting loop closure with scene sequences. *International Journal of Computer Vision*, 74:261–286, 2007.
- [46] D. Hoeim, C. Rother, and J. Winn. 3D layout for multi-view object class recognition and segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2007.
- [47] L. Itti, C. Koch, and E. Neibur. A model of saliency-based attention for rapid scene analysis. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 20:1254–1259, 1998.
- [48] K. Kanatani. 3-D interpretation of optical flow by renormalization. *International Journal of Computer Vision*, 11:267–282, 1993.
- [49] M. Kendall and J.D. Gibbons. *Rank Correlation Methods 5th edition*. Edward Arnold, 1990.

- [50] J. J. Koenderink and A. J. van Doorn. Relief: Pictorial and otherwise. *Image and Vision Computing*, 13.
- [51] J. J. Koenderink, A. J. van Doorn, and A. M. L. Kappers. Ambiguity and the 'mental eye' in pictorial relief. *Perception*, 30:431–448, 2001.
- [52] A. Kushal, C. Schmid, and J. Ponce. Flexible object models for category-level 3D object recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2007.
- [53] D. Lambrinos, R. Moller, T. Labhart, R. Pfeifer, and R. Wehner. A mobile robot employing insect strategies for navigation. *Robotics and Autonomous Systems*, 30.
- [54] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2006.
- [55] A.B. Lee and J. Huang. Brown range image database, 2000. [online] <http://www.dam.brown.edu/ptg/brid/index.html>.
- [56] M. Lehrer and G. Bianco. The turn-back-and-look behaviour: bee versus robot. *Biological Cybernetics*, 83:211–229, 2000.
- [57] A. Levin and R. Szeliski. Visual odometry and map correlation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2004.
- [58] H. C. Longuet-Higgins. A computer program for reconstructing a scene from two projections. *Nature*, 293.
- [59] H. C. Longuet-Higgins and K. Prazdny. The interpretation of a moving retinal image. *Proc. Royal Society of London B*, 208:385–397, 1980.
- [60] M. Lourakis. Non-metric depth representations: preliminary results. Technical Report TR-156, 1995.
- [61] D. G. Lowe. Object recognition from local scale-invariant features. In *IEEE International Conference on Computer Vision*, pages 1150–1157, 1999.
- [62] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *Int'l J. Computer Vision*, 60:91–110, 2004.
- [63] Y. Ma, J. Kořecká, and S. Sastry. Optimization criteria and geometric algorithms for motion and structure estimation. *International Journal of Computer Vision*, 44:219–249, 2001.

- [64] D. Marr. *Vision*. W. H. Freeman, 1982.
- [65] D. Marr and T. Poggio. From understanding computation to understanding neural circuitry. *Neurosciences Res. Prog. Bull.*, 15:470–488, 1977.
- [66] D. Marr and T. Poggio. A computational theory of human stereo vision. volume 204, pages 301–308, 1979.
- [67] S. J. Maybank. Ambiguity in reconstruction from image correspondences. In *Proc. European conference on computer vision*, pages 175–186, 1990.
- [68] S. J. Maybank. *Theory of Reconstruction from Image Motion*. Springer, Berlin, 1993.
- [69] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 27:1615–1630, 2005.
- [70] R. Moller. Insect visual homing strategies in a robot with analog processing. *Biological Cybernetics*, 83:231–243, 2000.
- [71] R. Moller. Insects could exploit uv-green contrast for landmark navigation. *J. of Theoretical Biology*, 214:619–631, 2002.
- [72] P. Moreels and P. Perona. Evaluation of features detectors and descriptors based on 3D objects. *Int'l J. Computer Vision*, 73:263–284, 2007.
- [73] E.N. Mortensen, H.Deng, and L.Shapiro. A sift descriptor with global context. In *IEEE International Conference on Computer Vision*, 2005.
- [74] R. Murrieta-Cid, C. Parra, and M. Devy. Visual navigation in natural environments: From range and color data to a landmark-based model. *Autonomous Robots*, 13:143–168, 2002.
- [75] J. F. Norman and J. T. Todd. The discriminability of local surface structure. *Perception*, 25:381–398, 1996.
- [76] J. F. Norman and J. T. Todd. Stereoscopic discrimination of interval and ordinal depth relations on smooth surfaces and in empty space. *Perception*, 27:257–272, 1998.
- [77] A. S. Ogale, C. Fermüller, and Y. Aloimonos. Occlusions in motion processing. In *Proc. BMVA symposium on Spatiotemporal Image Processing*, 2004.

- [78] A. S. Ogale, C. Fermüller, and Y. Aloimonos. Motion segmentation using occlusions. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 27:988–992, 2005.
- [79] J. Oliensis. A critique of structure from motion algorithms. *Computer Vision and Image Understanding*, 80:172–214, 2000.
- [80] J. Oliensis. A new structure-from-motion ambiguity. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 22:685–700, 2000.
- [81] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *Int'l J. Computer Vision*, 42:145–175, 2001.
- [82] D. Paulus and G. Schmidt. Approaches to depth estimation from active camera control. 1996.
- [83] F. Perez and C. Koch. Toward color image segmentation in analog vlsi: algorithm and hardware. *Int'l J. Computer Vision*, 12:17–42, 1994.
- [84] P. Petrov, O. Boumbarov, and K. Muratovski. Face detection and tracking with an active camera. In *Intelligent Systems, 2008. IS '08. 4th International IEEE Conference*, pages 1434–1439, 2008.
- [85] A. Ranganathan, E. Menegatti, and F. Dellaert. Bayesian inference in the space of topological maps. *IEEE Trans. Robotics*, 22:92–107, 2006.
- [86] E. Rivlin, Y. Aloimonos, , and A. Rosenfeld. Object recognition by a robotic agent: The purposive approach. In *IEEE Conference on Pattern Recognition*, pages 712–715, 1992.
- [87] E. Rivlin and H. Rotstein. Control of a camera for active vision: Foveal vision, smooth tracking and saccade. *International Journal of Computer Vision*, 39:81–96, 2000.
- [88] F. Rothganger, S. Lazebnik, C. Schmid, , and J. Ponce. 3D object modeling and recognition using local affine-invariant image descriptors and multi-view spatial constraints. *Int'l J. Computer Vision*, 66:231–259, 2006.
- [89] S. Savarese and Li Fei-Fei. 3D generic object categorization, localization and pose estimation. In *IEEE International Conference on Computer Vision*, pages 1–8, 2007.
- [90] G. Schindler, M. Brown, and R. Szeliski. City-scale location recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

- [91] C. Schmid. A structured probabilistic model for recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 485–490, 1999.
- [92] S. Se, D. Lowe, and J. Little. Vision-based global localization and mapping for mobile robots. *IEEE Trans. Robotics*, 21:364–375, 2005.
- [93] W. B. Seales. Measuring time-to-contact using active camera control. In *Computer Analysis of Images and Patterns*, volume 970/1995 of *Lecture Notes in Computer Science*, pages 944–949. Springer, 2006.
- [94] P. Shilane, P. Min, M. Kazhdan, and T. Funkhouser. The princeton shape benchmark. *Shape Modeling International, Genova, Italy*, June 2004.
- [95] I. Shimshoni, R. Basri, and E. Rivlin. A geometric interpretation of weak-perspective motion. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 21:252–257, 1999.
- [96] C. Siagian and L. Itti. Rapid biologically-inspired scene classification using features shared with visual attention. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 29:300–312, 2007.
- [97] C. Silpa-Anan and R. Hartley. Visual localization and loopback detection with a high resolution omnidirectional camera. In *Workshop on Omnidirectional Vision*, 2005.
- [98] M. Sonka, V. Hlavac, and R. Boyle. *Image Processing, Analysis, and Machine Vision*. PWS Publishing.
- [99] 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.
- [100] S. S. Stevens. On the theory of scales of measurement. *Science*, 103:677–680, 1946.
- [101] R. Szeliski and S. B. Kang. Shape ambiguities in structure from motion. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 19:506–512, 1997.
- [102] R. Talluri and J. K. Aggarwal. Position estimation for an autonomous mobile robot in an outdoor environment. *IEEE Trans. on Robotics and Automation*, 8:573–584, 1992.

- [103] M. J. Tarr and M. J. Black. A computational and evolutionary perspective on the role of representation in computer vision. Technical Report YALEU/DCS/RR-899, Yale University, 1991.
- [104] C. L. Teo. An effective scene recognition strategy for biomimetic robotic navigation. Master's thesis, Dept. Electrical and Computer Engineering, National University of Singapore, 2007.
- [105] A. Thomas, V. Ferrari, B. Leibe, T. Tuytelaars, B. Schiele, and L. Van Gool. Towards multi-view object class detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 1589–1596, 2006.
- [106] J. T. Todd and P. Bressan. The perception of 3-dimensional affine structure from minimal apparent motion sequences. *Perception and Psychophysics*, 48:419–430, 1990.
- [107] J. T. Todd and F. D. Reichel. Ordinal structure in the visual perception and cognition of smoothly curved surfaces. *Psychological Review*, 96:643–657, 1989.
- [108] S. Todorovic and M.C. Nechyba. A vision system for intelligent mission profiles of micro air vehicles. *IEEE Trans. vehicular technology*, 53:1713–1725, 2004.
- [109] C. Tomasi and T. Kanade. Shape and motion from image streams under orthography: a factorization method. *International Journal of Computer Vision*, 9:137–154, 1992.
- [110] O. Trullier, S. Wiener, A. Berthoz, and J. Meyer. Biologically-based artificial navigation systems: Review and prospects. *Progress in Neurobiology*, 51:483–544, 1997.
- [111] R. Y. Tsai and T. S. Huang. Uniqueness and estimation of 3-D motion parameters of rigid bodies with curved surfaces. *IEEE Trans. Pattern Analysis and machine Intelligence*, 6:13–27, 1984.
- [112] R. Voss and J. Zeil. Active vision in insects: An analysis of object-directed zig-zag flights in a ground-nesting wasp (*odynerus spinipes*, eumenidae). *J. of Comparative Physiology A*, 182:377–387, 1998.
- [113] A. M. Waxman and K. Wohn. Contour evolution, neighborhood deformation, and global image flow: planar surfaces in motion. *International Journal of Robotics Research*, 4:95–108, 1985.

- [114] D. Weinshall. Qualitative depth from stereo, with applications. *Computer Vision, Graphics, and Image Processing*, 49:222–241, 1990.
- [115] J. Weng, N. Ahuja, and T. S. Huang. Optimal motion and structure estimation. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 15:864–884, 1993.
- [116] 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. IEEE Int’l Conf. Robotics and Automation (ICRA ’02)*, pages 359–365, May 2002.
- [117] R. D. Wright and L. M. Ward. *Orienting of Attention*. Oxford University Press, New York, 2008.
- [118] T. Xiang and L-F. Cheong. Distortion of shape from motion. In *British Machine Vision Conference*, pages 153–162, 2002.
- [119] T. Xiang and L-F. Cheong. Understanding the behavior of sfm algorithms: a geometric approach. *International Journal of Computer Vision*, 51:111–137, 2003.
- [120] A. L. Yarbus. *Eye Movements and Vision*. Plenum, New York, 1967.
- [121] G. S. Young and R. Chellapa. Statistical analysis of inherent ambiguities in recovering 3-D motion from a noisy flow field.
- [122] J. Zeil, A. Kelber, and R. Voss. Structure and function of learning flights. *J. of Experimental Biology*, 199:245–252, 1996.
- [123] Z. Zhang. Understanding the relationship between the optimization criteria in two-view motion analysis. In *Proc. IEEE Int’l Conf. Computer Vision and Pattern Recognition*, pages 772–777, 1998.