

UNIVERSITY OF TECHNOLOGY SYDNEY

DOCTORAL THESIS

---

**Action for Perception: Active Object  
Recognition and Pose Estimation in  
Cluttered Environments**

---

Kanzhi WU

*A thesis submitted in fulfilment of the requirements*

*for the degree of Doctor of Philosophy*

*at the*

Centre for Autonomous Systems

School of Electrical, Mechanical and Mechatronic Systems

August 8, 2017



## Certificate of Original Authorship

I certify that the work in this thesis has not previously been submitted for a degree nor has it been submitted as part of requirements for a degree except as part of the collaborative doctoral degree and/or fully acknowledged within the text.

I also certify that the thesis has been written by me. Any help that I have received in my research work and the preparation of the thesis itself has been acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

Signature of Student:

---

Date:

---



## *Abstract*

Object recognition and localisation are indispensable competency for service robots in everyday environments like offices and kitchens. Presence of similar objects that can only be differentiated from a small part of the surface together with clutter that leads to occlusions make it impossible to detect target objects accurately and reliably from a single observation. When the sensor observing the environment is mounted on a mobile platform, object detection and pose estimation can be facilitated by observing the environment from a series of different viewpoints. Computing *Active perception* strategies, with the aim of finding optimal actions to enhance object recognition and pose estimation performance is the focus of this thesis.

This thesis consists of two main parts:

In the **first** part, it focuses on object detection and pose estimation from a single frame of observation. Using an RGB-D sensor, we propose a modular 3D textured object detection and pose estimation framework which can recognise object under cluttered environment by taking advantage of the geometric information provided from the sensor. To handle less-textured objects and objects under severe illumination conditions, we propose a novel RGB-D feature which is robust to illumination, scale, rotation and viewpoint variations, and provides reliable feature matching results under challenging conditions. The proposed feature is validated for multiple applications including object detection and point cloud alignment. Parts of the above approaches are integrated with existing work to produce a practical and effective perception module for a warehouse automation task. The designed perception system can detect objects of different types and estimate their poses robustly thus guaranteeing a reliable object grasping and manipulation performances.

In the **second** part of the thesis, we investigate the problem of *active* object detection and pose estimation from two perspectives: with and without considering the uncertainties in the motion model and the observation model. First, we propose a model-driven active object recognition and pose estimation system via exploiting the feature association probability under scale and viewpoint variations. By explicitly modelling the feature association, the proposed system can predict future information more accurately thus laying the foundation of a successful active Next-Best-View planning system even with a naive greedy search technique. We also present a probabilistic framework which handles motion and observation uncertainties in the active object detection and pose estimation problem. We present an optimisation framework which computes

the optimal control at each step, using an objective function which incorporates uncertainties in state estimation, feature coverage for better recognition confidence and control consumption. The proposed framework can handle various issues such as object initialisation, collision avoidance, occlusion and changing the object hypothesis. Validations based on a simulation environment are also presented.

## *Acknowledgements*

Firstly, I would like to express my sincere gratitude to my supervisor Prof. Gamini Disanayake for his continuous support during my PhD study, for his patience, motivation, and immense knowledge. His guidance helped me in all the time of research and writing of this thesis. I could not have imagined having a better supervisor and mentor for my PhD study. I would also like to thank my co-supervisor Dr. Ravindra Ranasinghe, the door of his office was always open whenever I ran into a trouble spot or had a question about my research or writing.

My sincere thanks also go to Assoc. Prof. Shoudong Huang and Mr Teng Zhang for the insightful discussions about SLAM and planning under uncertainty. During the past four years, I have experienced a welcoming and passionate research atmosphere in CAS and I would like to thank Prof. Dikai Liu for his management of CAS. I would like to thank Dr. Liang Zhao for patiently answering my questions about multiview geometry.

I would also like to thank my great friends and colleagues in CAS: Daobilige Su, Lakshitha Dantanarayana, Kasra Khosoussi, Raphael Falque and so many other colleagues who have helped me, too many to list here. The discussions with you not only have helped me in my academic research but have also enriched my life in the past four years.

Last but not the least, I would like to thank my parents and especially my wife Liye Sun for her understanding, support and companionship during my research career.





# Contents

<b>Certificate of Original Authorship</b>	<b>iii</b>
<b>Abstract</b>	<b>v</b>
<b>Acknowledgements</b>	<b>vii</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Background and Motivation . . . . .	1
1.2 Contributions . . . . .	3
1.3 Outline . . . . .	5
1.4 Publication List . . . . .	6
<b>2 Literature Review</b>	<b>9</b>
2.1 Single View Object Recognition and Pose Estimation . . . . .	10
2.1.1 Instance-Level Object Recognition . . . . .	10
2.1.2 Category-Level Object Recognition . . . . .	19
2.1.3 6D Pose Estimation . . . . .	19
2.2 Active Object Detection and Next-Best-View Selection . . . . .	23
2.2.1 Next-Best-View for Active Object Recognition . . . . .	23
2.2.2 Planning under Uncertainty . . . . .	25
2.3 Summary . . . . .	29
<b>I Single-View Object Detection and Pose Estimation</b>	<b>31</b>
<b>3 Textured Object Recognition and Pose Estimation</b>	<b>33</b>
3.1 Problem Formulation . . . . .	34
3.2 Algorithm Pipeline . . . . .	34
3.2.1 Support Plane Subtraction . . . . .	34

3.2.2	Feature Extraction and Matching . . . . .	37
3.2.3	Feature Clustering . . . . .	37
3.2.4	Outlier Rejection . . . . .	38
3.2.5	Pose Estimation . . . . .	39
3.2.6	Pose Combination and Refinement . . . . .	40
3.3	Experiments and Discussions . . . . .	40
3.3.1	Consistent Correspondences Search using Graph based Approach . . . . .	40
3.3.2	Multiple Objects Recognition and Pose Estimation . . . . .	42
3.3.3	Discussion . . . . .	44
3.4	Summary . . . . .	45
<b>4</b>	<b>A Novel RGB-D Feature for Challenging Object Detection</b>	<b>47</b>
4.1	A Novel Rotation, Illumination and Scale invariant RGB-D Feature . . . . .	48
4.1.1	Keypoint Detector . . . . .	48
4.1.2	Feature Descriptor . . . . .	51
	Scale Estimation and Neighbourhood Region Selection . . . . .	51
	Orientation Estimation . . . . .	53
	Descriptor Construction . . . . .	53
4.2	Experimental Results . . . . .	55
4.2.1	Evaluation Method . . . . .	56
4.2.2	Experimental Results and Analysis . . . . .	56
4.2.3	Object Recognition Dataset . . . . .	57
4.2.4	RGB-D Feature Evaluation Dataset . . . . .	58
	Viewpoint Invariance . . . . .	58
	Illumination Invariance . . . . .	58
	Scale Invariance . . . . .	60
	Rotation Invariance . . . . .	62
4.3	Object Recognition using RISAS . . . . .	64
4.3.1	Pipeline in Brief . . . . .	64
4.3.2	Object Recognition in extreme Illumination Conditions . . . . .	65
4.3.3	Less-Textured Object Recognition . . . . .	65
4.4	Conclusion . . . . .	66

<b>5</b>	<b>Object Detection and Pose Estimation for Warehouse Automation</b>	<b>69</b>
5.1	Environment Setup for APC 2015 . . . . .	70
5.2	Analysis of the Particularities of the Warehouse Pick-and-Place Problem . . . . .	72
5.3	Hardware Design of the Robotic Platform . . . . .	73
5.4	Perception Module . . . . .	76
5.4.1	Pre-Processed Prior Knowledge . . . . .	76
5.4.2	Perception System Pipeline . . . . .	78
5.5	Experimental Results . . . . .	80
5.6	Summary . . . . .	82
<b>II</b>	<b>Active Object Detection and Pose Estimation</b>	<b>85</b>
<b>6</b>	<b>Model-Driven Active Object Detection and Pose Estimation</b>	<b>87</b>
6.1	Information Rich Object Modeling . . . . .	88
6.1.1	Models for Active Object Recognition . . . . .	88
6.1.2	Information Rich Attributes for Better Prediction of Viewpoint Quality . . . . .	89
6.1.3	Feature Weighting using KD-Tree Structure . . . . .	91
	Motivation . . . . .	91
	Method . . . . .	92
	Experiments and Results . . . . .	93
	Discussion . . . . .	94
6.2	Active Object Recognition and Pose Estimation System . . . . .	95
6.2.1	Object Recognition and Pose Estimation Using a Single Viewpoint . . . . .	95
	RGB-D Segmentation . . . . .	95
6.2.2	Next-Best-View Selection using Information Rich Model . . . . .	99
6.3	Experiments and Discussion . . . . .	100
6.3.1	Case Study 1: NBV Selection in 2D environments . . . . .	101
6.3.2	Case Study 2: NBV Selection in 3D environments . . . . .	103
	Computational Cost . . . . .	108
6.4	Conclusion . . . . .	108
<b>7</b>	<b>Active Object Detection and Pose Estimation in Belief Space</b>	<b>111</b>
7.1	Preliminaries and Problem Formulation . . . . .	112

7.1.1	Notation and Preliminaries . . . . .	112
7.1.2	Problem statement . . . . .	113
7.1.3	Formulation . . . . .	113
7.2	Viewpoints Planning in General Belief Space . . . . .	114
7.2.1	MAP Estimation in General Belief Space . . . . .	115
7.2.2	Empirical Modelling of Feature Association . . . . .	117
7.2.3	Optimal Control Inference . . . . .	119
7.2.4	Objective Function Formulation . . . . .	121
7.2.5	Online Re-planning . . . . .	125
7.3	Practical Issues . . . . .	127
7.3.1	Object Pose Initialisation . . . . .	127
7.3.2	Initial Guess for Control Optimisation . . . . .	130
7.3.3	Occlusion Modelling . . . . .	131
7.3.4	Collision Avoidance . . . . .	132
7.3.5	Hypothesis Changing during Planning . . . . .	134
7.4	Simulation Experiments and Results Analysis . . . . .	135
7.4.1	Experimental Set-up . . . . .	136
7.4.2	Convergence Analysis in Optimisation . . . . .	137
7.4.3	Parameterisation . . . . .	139
7.4.4	Case Study 1: Trajectory Planning in Another Scenario . . . . .	145
7.4.5	Case Study 2: Obstacle Avoidance and Occlusion Handling . . . . .	147
7.4.6	Case Study 3: Hypothesis Changing During Planning . . . . .	149
7.5	Conclusion . . . . .	151
<b>8</b>	<b>Conclusion</b>	<b>155</b>
8.1	Single-View Object Detection and Pose Estimation . . . . .	155
8.2	Active Object Detection and Pose Estimation . . . . .	157
<b>A</b>	<b>Proofs in Chapter. 7</b>	<b>161</b>
A.1	Linearisation of (7.14) . . . . .	161
A.2	Quadratic Form Representation . . . . .	163
	<b>Bibliography</b>	<b>165</b>

# List of Figures

1.1	Outline of the thesis and relationship between chapters. . . . .	6
2.1	Outline of the reviewed literature in Chapter. 2. . . . .	9
2.2	Differences between instance-level object recognition dataset and category-level object recognition dataset. . . . .	10
2.3	General pipeline for instance-level object recognition. . . . .	11
2.4	Typical flowchart of obtaining visual vocabulary representation. . . . .	16
2.5	Flowchart of RANSAC to remove outliers in raw features matches. . . . .	18
2.6	Traditional image classification pipeline. . . . .	19
2.7	Deep learning based image classification pipeline. . . . .	19
3.1	Pipeline for single view textured object detection and pose estimation . . . . .	35
3.2	Sample image from Willow dataset . . . . .	36
3.3	Example images from reconstructed scenarios similar as Fig. 3.2. . . . .	36
3.4	Comparison results for clustering correspondences for 4 same images under different ranges using two different methods. . . . .	38
3.5	Sample images from the collected dataset which contains multiple occluded objects. . . . .	43
3.6	Objects detection and pose estimation results in cluttered environment. . . . .	44
3.7	Object detection results in a similar scenario as Willow dataset. . . . .	45
4.1	Algorithm flowchart of the keypoint detector in the proposed RISAS RGB-D feature. . . . .	48
4.2	Calculating the main normal vector of the depth image to extract RGB-D keypoints. . . . .	49
4.3	Precision-Recall curves for difference $\tau$ value in feature point extraction. . . . .	50
4.4	Comparison results for different neighbourhood selection methods for descriptor construction. . . . .	52
4.5	Algorithm flowchart of constructing descriptor of the RGB-D feature. . . . .	54
4.6	Precision recall curves under different parameterisation of descriptor construction. . . . .	55

4.7	Evaluation results on RGB-D scene dataset. . . . .	57
4.8	Example images under different viewpoint variations. . . . .	58
4.9	Precision-Recall curves under different viewpoint variations. . . . .	59
4.10	Example images and results under square root illumination variation . . . . .	60
4.11	Example images and results under square illumination variation . . . . .	60
4.12	Example images and results under cube root illumination variation . . . . .	61
4.13	Example images and results under cube illumination variation . . . . .	61
4.14	Example images and results under illumination change using ND mirror . . . . .	61
4.15	Reference image for illumination and rotation variation. . . . .	62
4.16	Example images under different scale variations. . . . .	62
4.17	Comparative matching results under different scale variations. . . . .	63
4.18	Example images of different 3D rotations. . . . .	63
4.19	Precision-Recall curves corresponding to Fig. 4.18. . . . .	63
4.20	$I_{\text{object}}$ for object recognition under illumination variations. . . . .	65
4.21	Object detection results corresponding to the object in Fig. 4.20. . . . .	66
4.22	$I_{\text{object}}$ for less-textured object detection. . . . .	66
4.23	Object detection results corresponding to the object in Fig. 4.22. . . . .	67
5.1	Targeting objects for recognition and pose estimation in APC 2015 . . . . .	70
5.2	The shelf which contains the objects used in APC 2015. . . . .	71
5.3	Difficulties in object detection and grasping. . . . .	72
5.4	Robotic platform of APC 2015. . . . .	74
5.5	Sensor configuration for Amazon Picking Challenge. . . . .	76
5.6	Pre-processed image for object detection. . . . .	77
5.7	Perception module for APC 2015. . . . .	78
5.8	Selected recognition methods for each objects in APC 2015 . . . . .	79
5.9	Single object detection results. . . . .	81
5.10	Single object detection results. . . . .	83
6.1	Dense and sparse model of the object "Fruity Bites" . . . . .	88
6.2	Object modelling using an RGB-D camera mounted on Turtlebot. . . . .	89
6.3	The number of correct matches under different scale and viewpoint variations. . .	90
6.4	Belvita biscuit boxes of different favours . . . . .	91

6.5	Feature weighting especially for similar object . . . . .	93
6.6	The system framework of object recognition and pose estimation using single RGB-D observation . . . . .	96
6.7	RGB-D segmentation results from cluttered environment. . . . .	97
6.8	Framework of the proposed active object recognition and pose estimation system . . . . .	98
6.9	Active object recognition and pose estimation using Turtlebot . . . . .	101
6.10	Optimised localisation using Parallax BA . . . . .	102
6.11	Prediction of potential matches in next frame . . . . .	102
6.12	Planned trajectory for active object recognition and pose estimation . . . . .	103
6.13	Object recognition and pose estimation results in different steps of the path . . . . .	104
6.14	Example images from the pre-collected RGB-D data. . . . .	104
6.15	Optimised camera poses using RGBD-SLAM . . . . .	105
6.16	Object recognition and pose estimation results in different steps of the path . . . . .	106
6.17	Planned trajectory for active object recognition and pose estimation . . . . .	106
6.18	Object recognition and pose estimation results in different steps of the path. . . . .	107
7.1	Probabilistic distribution of feature correspondence confidence . . . . .	119
7.2	Example: path planning for next 3 steps and the updated pose after 1 step of control execution. . . . .	125
7.3	Replanning strategy if the belief is only changed within the threshold. . . . .	126
7.4	Replanning strategy if the belief shows significant differences compared with the prediction. . . . .	127
7.5	Occlusion modelling for observation prediction. . . . .	131
7.6	Probabilistic occlusion modeling. . . . .	132
7.7	Obstacle avoidance in path planning. . . . .	133
7.8	An example of the simulation environments. . . . .	136
7.9	Optimisation results for different random sampled initial guess. . . . .	138
7.10	Comparative experiments of different $L$ values. . . . .	140
7.11	Experimental results of objective function parameterisation 1 in Table. 7.2. . . . .	142
7.12	Experimental results of objective function parameterisation 2 in Table. 7.2. . . . .	143
7.13	Experimental results of objective function parameterisation 3 in Table. 7.2. . . . .	144
7.14	Simulation environment of scenario 2. . . . .	145
7.15	Planned trajectory in the scenario 2. . . . .	146

7.16 Estimation errors in the scenario 2. . . . .	146
7.17 Environment with known obstacles. . . . .	147
7.18 Planned trajectory in environment with known obstacles. . . . .	148
7.19 Estimation errors in environment with a known obstacle. . . . .	149
7.20 Environment with both known obstacles and unknown obstacles. . . . .	149
7.21 Environment understanding at the beginning step. . . . .	150
7.22 Planned trajectory in the environment with a forbidden region and an obstacle. . .	150
7.23 Estimation errors in the environment with a forbidden region and an obstacle. . .	151
7.24 Similar object with different distributed feature. . . . .	152
7.25 Planned trajectory in the environment where object hypothesis is changed after observing new features. . . . .	152
7.26 Estimate error corresponding to the trajectory in Fig. 7.25. . . . .	153



# List of Tables

3.1	Feature matching accuracy comparison results . . . . .	37
3.2	Outlier rejection comparative experimental results with RANSAC . . . . .	42
3.3	Object detection results: accuracy, estimate error and time consumption . . . . .	43
6.1	Time consumption analysis for individual steps . . . . .	108
7.1	Optimisation results for randomly sampled initial guess . . . . .	139
7.2	Different weight parameterisation methods in the objective function. . . . .	141



# List of Abbreviations

<b>ANN</b>	Approximate Nearest Neighbour
<b>APC</b>	Amazon Picking Challenge
<b>BoW</b>	Bag of Words
<b>BRAND</b>	Binary Robust Appearance and Normals Descriptor
<b>BRIEF</b>	Binary Robust Independent Elementary Features
<b>BRISK</b>	Binary Robust Invariant Scalable Keypoints
<b>BRM</b>	Belief RoadMap
<b>CNN</b>	Convolutional Neural Networks
<b>DLT</b>	Direct Linear Transformation
<b>DoF</b>	Degree of Freedom
<b>FAST</b>	Features from Accelerated Segment Test
<b>FIRM</b>	Feedback-based Information RoadMap
<b>GLOH</b>	Gradient Location and Orientation Histogram
<b>GBS</b>	General Belief Space
<b>ISS</b>	Intrinsic Shape Signatures
<b>KPQ</b>	KeyPoint Quality
<b>LBSS</b>	Laplace-Beltrami Scale-Space
<b>LDA</b>	Linear Discriminant Analysis
<b>LLC</b>	Locality-constrained Linear Coding
<b>LM</b>	Levenburg Marquart
<b>LQG</b>	Linear Quadratic Gaussian
<b>LQR</b>	Linear Quadratic Regulator
<b>LOIND</b>	Local Ordinal Intensity and Normal Descriptor
<b>LSH</b>	Local Sensitive Hashing
<b>MDP</b>	Markov Decision Process
<b>MOPED</b>	Multiple Object Pose Estimation and Detection

<b>MPC</b>	<b>Model Predictive Control</b>
<b>NBV</b>	<b>Next-Best-View</b>
<b>ORB</b>	<b>Oriented FAST and Rotated BRIEF</b>
<b>PCA</b>	<b>Principle Component Analysis</b>
<b>PCL</b>	<b>Point Cloud Library</b>
<b>PFH</b>	<b>Point Feature Histogram</b>
<b>PnP</b>	<b>Perspective-n-Point</b>
<b>POMDP</b>	<b>Partially Observable Markov Decision Process</b>
<b>PRM</b>	<b>Probabilistic RoadMap</b>
<b>RANSAC</b>	<b>RANdom SAmples Consensus</b>
<b>RISAS</b>	<b>A Rotation, Illumination and Scale invariant Appearance and Shape Feature</b>
<b>R-CNN</b>	<b>Region-based Convolutional Neural Networks</b>
<b>RNN</b>	<b>Residual Neural Network</b>
<b>ROI</b>	<b>Region Of Interest</b>
<b>RRG</b>	<b>Rapid-exploring Randomised Graph</b>
<b>RRT</b>	<b>Rapid-exploring Randomised Trees</b>
<b>RRBT</b>	<b>Rapid-exploring Randomised Belief Trees</b>
<b>SfM</b>	<b>Structure from Motion</b>
<b>SHOT</b>	<b>Signature of Histogram Orientation</b>
<b>SIFT</b>	<b>Scale Invariant Feature Transform</b>
<b>SLAM</b>	<b>Simultaneous Localisation And Mapping</b>
<b>SORSAC</b>	<b>Spatially ORdered SAmples Consensus</b>
<b>SQP</b>	<b>Sequential Quadratic Programming</b>
<b>SURF</b>	<b>Speed-Up Robust Feature</b>
<b>SVD</b>	<b>Singular Value Decomposition</b>
<b>RANSAC</b>	<b>RANdom SAmples Consensus</b>
<b>USC</b>	<b>Unique Shape Context</b>