RECOGNISING, REPRESENTING AND MAPPING NATURAL FEATURES IN UNSTRUCTURED ENVIRONMENTS

Fabio Tozeto Ramos

A thesis submitted in fulfillment
of the requirements for the degree of
Doctor of Philosophy

Australian Centre for Field Robotics
Department of Aerospace, Mechanical and Mechatronic Engineering
The University of Sydney

February, 2007

Declaration

This thesis is submitted to the University of Sydney in fulfillment of the requirements for the degree of Doctor of Philosophy. This thesis is entirely my own work, and except where otherwise stated, describes my own research.

Fabio Tozeto Ramos

February, 2007

Reprinted with corrections and emendations March, 2008

Abstract

Fabio Tozeto Ramos
The University of Sydney

Doctor of Philosophy
February 2007

Recognising, Representing and Mapping Natural Features in Unstructured Environments

This thesis addresses the problem of building statistical models for multi-sensor perception in unstructured outdoor environments. The perception problem is divided into three distinct tasks: recognition, representation and association. Recognition is cast as a statistical classification problem where inputs are images or a combination of images and ranging information. Given the complexity and variability of natural environments, this thesis investigates the use of Bayesian statistics and supervised dimensionality reduction to incorporate prior information and fuse sensory data. A compact probabilistic representation of natural objects is essential for many problems in field robotics. This thesis presents techniques for combining non-linear dimensionality reduction with parametric learning through Expectation Maximisation to build general representations of natural features. Once created these models need to be rapidly processed to account for incoming information. To this end, techniques for efficient probabilistic inference are proposed. The robustness of localisation and mapping algorithms is directly related to reliable data association. Conventional algorithms employ only geometric information which can become inconsistent for large trajectories. A new data association algorithm incorporating visual and geometric information is proposed to improve the reliability of this task. The method uses a compact probabilistic representation of objects to fuse visual and geometric information for the association decision.

The main contributions of this thesis are: 1) a stochastic representation of objects through non-linear dimensionality reduction; 2) a landmark recognition system using a visual and ranging sensors; 3) a data association algorithm combining appearance and position properties; 4) a real-time algorithm for detection and segmentation of natural objects from few training images and 5) a real-time place recognition system combining dimensionality reduction and Bayesian learning.

The theoretical contributions of this thesis are demonstrated with a series of experiments in unstructured environments. In particular, the combination of recognition, representation and association algorithms is applied to the Simultaneous Localisation and Mapping problem (SLAM) to close large loops in outdoor trajectories, proving the benefits of the proposed methodology.

Acknowledgements

The work presented in this thesis would not be possible without the input and support of several people directly or indirectly involved. I am deeply grateful to the people listed below whose contribution to this work is more important than what I can describe in these few lines.

It was a great privilege to be advised by Hugh Durrant-Whyte. I have learnt a lot from the many discussions we had either at the university or while having some pints in a local pub. I thank him for believing in my potential as a researcher, for bringing me to Australia and for supporting me over these 4 years. His insightful vision and thoughtful guidance were essential not only for this thesis but for my career as a researcher.

I am indebted to my examiners whose thoughtful questions and suggestions helped me in improving this thesis in the final version. Paul Newman provided me with an excellent and very detailed review. I am looking forward to spending some time in his group at Oxford. Eduardo Nebot has taught me not only how to barbecue in the traditional Argentinian way but how to keep myself motivated, always looking for new research directions. His thesis report is also very much appreciated. In the last months of my Ph.D I had the pleasure of sharing an office with Dieter Fox. I am very grateful to him for the discussions and suggestions that have originated new ideas and research topics that go beyond the material presented in this thesis.

As a Ph.D student at ACFR I had the chance to expose and discuss my ideas with very knowledgeable researchers who have provided significant inputs into this thesis. Specifically, I would like to mention my three main collaborators. It has been a pleasure working with Suresh Kumar. I am very grateful to him for being my coauthor in several publications and for the innumerous discussions on techniques for dimensionality reduction that constitute part of the contributions in this thesis. Ben Upcroft has been my friend since the begining of my Ph.D. Besides the innumerous schooners of beer we had, he has taught me writing styles, robotic programming and techniques for analysis of experiments. I am very thankful to him for being my coauthor over these years and for organising the statistical learning reading group; a source of ideas for many Ph.D students. Even before coming to Australia, my faithful mate Juan Nieto was already helping me in organising my accommodation so as to make my Down Under adaptation easier. I really enjoy his friendship and the fun we had in trips and parties. I am very thankful to him for the discussions on techniques for estimation and SLAM used in this thesis.

I have immensely enjoyed being part of ACFR. Among soccer matches, beer times and barbecues, I had the chance to meet very smart people from whom I have learnt a lot. Particularly, I would like to express my gratitude to Oliver Frank, Roman Katz, Jose Guivant and Christel-loic Tisse for helping to set up sensors in the vehicle used in the last part of the thesis. I also thank Tobias Kaupp, Rosalind Wang, Alex Brooks and Ian Mahon for discussions and joint projects we have engaged in.

The path to my Ph.D would have been a lot tougher without the endless love of my wife Karin Ramos. I thank her for accompanying me throughout this journey to pursue my career goals and for all the encouragement and patience in the moments I needed her the most.

Last, but by no means least I would like to thank Mom and Dad for all the dedication and support involved in my education. This Ph.D is not only my achievement but theirs as well.

To Mom and Dad for the support and
encouragement throughout my life

Contents

Declaration i

Abstract ii

Acknowledgements iii

Contents vi

List of Figures xi

List of Tables xviii

Symbols and Notation xix

1 Introduction 1

1.1 Motivation . . . 1

1.2 Navigation in Complex Environments . . . 2

1.2.1 Compact Probabilistic Representations . . . 3

1.2.2 Robust Data Association . . . 6

1.3 Contributions . . . 7

1.4 Thesis Outline . . . 9

2 Autonomous Recognition and Mapping 11

2.1 Introduction . . . 11

2.2 Statistical Learning . . . 12

2.2.1 Discriminative and Generative Models . . . 12

2.2.2 Maximum likelihood . . . 13

2.2.3 Maximum *a posteriori* . . . 16

2.2.4 Bayesian learning 16
2.3 Dimensionality Reduction 17
2.3.1 Principal Component Analysis 18
2.3.2 Multi-Dimensional Scaling 21
2.3.3 ISOMAP 24
2.3.4 LLE 26
2.3.5 Laplacian Eigenmaps 29
2.3.6 Semi-Definite Embedding 30
2.4 Recognition 34
2.4.1 Patch-Based Object Recognition 35
2.4.2 Feature-Based Object Recognition 37
2.5 Localisation and Mapping 39
2.5.1 Topological Localisation 39
2.5.2 Stochastic SLAM 41
2.6 Summary 46

3 Recognition of Natural Features **47**

3.1 Introduction 47
3.2 Feature Extraction 48
3.2.1 Colour 49
3.2.2 Texture 50
3.2.3 Colour-Texture Cue Fusion 50
3.2.4 Experiments 52
3.3 Recognition and Segmentation from few Images 53
3.3.1 VBEM for Mixtures of Gaussians 57
3.3.2 Bayesian formulation for GMM 58
3.3.3 Model Selection 60
3.3.4 Segmentation and Recognition with VBEM 62
3.3.5 Experiments 63
3.3.6 Comparison of VBEM with EM 67
3.4 Laser-Camera Landmark Recognition 67
3.4.1 Laser Clustering and Shape Descriptors 69
3.4.2 Defining a ROI with Laser 71
3.4.3 Supervised Dimensionality Reduction 73

3.4.4 Logistic Regression . . . 78
3.4.5 Experiments . . . 79
3.5 Summary . . . 80

4 Stochastic Representation **86**
4.1 Introduction . . . 86
4.2 Model Generation . . . 88
4.2.1 The Generative Model . . . 88
4.2.2 Parameter Estimation . . . 90
4.2.3 Choice of a Suitable Number of Mixture Components . . . 91
4.3 Model Evaluation . . . 92
4.3.1 Probabilistic Inference . . . 92
4.3.2 Integration within a Bayesian Filtering Framework . . . 93
4.4 Experiments . . . 94
4.4.1 Synthetic Examples . . . 94
4.4.2 Autonomous Ground Vehicle . . . 95
4.4.3 Unmanned Underwater Vehicle . . . 96
4.4.4 Natural Feature Tracking . . . 104
4.5 Summary . . . 106

5 Place Recognition **110**
5.1 Introduction . . . 110
5.1.1 Previous Work . . . 111
5.1.2 Problem Definition . . . 115
5.2 Place Representation . . . 116
5.2.1 Dimensionality Reduction . . . 116
5.2.2 Generative model for Places . . . 117
5.2.3 Multi-Class Classification . . . 118
5.3 Implementation . . . 118
5.4 Experiments . . . 120
5.4.1 Indoor dataset . . . 120
5.4.2 Outdoor dataset . . . 122
5.5 Summary . . . 127

6 Integrating Perception with Mapping **129**

6.1 Introduction . . . 129

6.2 Related Work . . . 130

6.2.1 Mapping with Vision . . . 130

6.2.2 Relational Models and Mapping . . . 132

6.2.3 Loop Closure with Imaging and Range Sensors . . . 132

6.3 Landmark Recognition . . . 134

6.4 Landmark Representation . . . 135

6.4.1 Model Updating . . . 136

6.5 Combined Appearance-Position Data Association . . . 137

6.6 Experiments in Outdoor SLAM . . . 139

6.7 Discussion . . . 143

6.8 Summary . . . 145

7 Conclusions and Future Work **148**

7.1 Summary of Contributions . . . 148

7.1.1 Feature Representation Through Non-Linear Stochastic Model . . . 149

7.1.2 Data Association Combining Position and Appearance Estimates . . . 149

7.1.3 Combined Laser-Camera Landmark Recognition . . . 150

7.1.4 Real-time Natural Feature Recognition and Segmentation . . . 151

7.1.5 Large-Scale SLAM Demonstration in Unstructured Environment . . . 152

7.1.6 Place Recognition from Few Training Images . . . 152

7.2 Future Research . . . 153

7.2.1 Incremental Non-Linear Dimensionality Reduction . . . 153

7.2.2 Scan-SLAM Combining Laser and Appearance . . . 154

7.2.3 3D Image Reconstruction . . . 154

7.2.4 Demonstration in Urban Environments . . . 155

A Laser-Camera Calibration **156**

A.1 Camera Calibration . . . 156

A.2 Laser Calibration . . . 157

A.2.1 Optimisation Procedure . . . 157

B VBEM for Mixtures of Gaussians **161**

B.0.2 A short note about priors 161

B.0.3 Bayesian Gaussian Mixture Model 162

C Mixture of Linear Models **168**

C.1 Learning 168

C.2 Inference 171

C.3 Implementation 172

List of Figures

1.1 Typical image and laser scan from an outdoor environment. As most objects are further than the maximum range of the sensor, only 10 distances were obtained. They are insufficient for correct characterisation of the objects as shown in the image. 4

1.2 Metric map augmented with landmark pictures. The additional visual information allows better perception and decision making. 5

1.3 Gating procedure for data association. The robot with position $\hat{\mathbf{x}}_v$ and uncertainty represented by the ellipse has to associate an observation $\mathbf{z}$ to landmarks previously observed denoted by $\hat{\mathbf{x}}_{l,1}, \hat{\mathbf{x}}_{l,2}, \hat{\mathbf{x}}_{l,3}, \hat{\mathbf{x}}_{l,4}, \hat{\mathbf{x}}_{l,5}$. The gate defined by the dashed ellipse eliminates landmarks $\hat{\mathbf{x}}_{l,3}, \hat{\mathbf{x}}_{l,4}$ and $\hat{\mathbf{x}}_{l,5}$. 6

1.4 SLAM over an extensive trajectory. The uncertainty of the vehicle and feature positions increases making data association difficult before closing the loop. . 8

2.1 Maximum likelihood estimators approximates the likelihood function by a set of Delta functions centred on training points. It can therefore disregard important information. 15

2.2 Swill roll dataset. (a) Manifold surface. (b) 2000 samples used to test dimensionality reduction algorithms. The colour coding indicates neighbourhood. . 20

2.3 Swiss Roll dataset projected into a 2-dimensional space using the eigenvectors computed by PCA. 21

2.4 Results of MDS on the Swiss roll dataset. Note the similarity of this solution with the solution found by PCA. 24

2.5 (a) K-nearest neighbour ($K = 15$) graph in high-dimensional space obtained as the first step of Isomap for the Swiss roll dataset. (b) The same graph defined over the Isomap solution in low-dimensional coordinates. 27

2.6 Final Isomap solution for the Swiss roll dataset with $K = 15$. Neighbourhoods and distances are preserved indicating that Isomap has correctly learnt the nonlinear structure of the manifold. 28

2.7 Swiss roll embedding computed by LLE. The solution is neighbourhood preserving but not isometric - the manifold is deformed towards its right end. . 29

2.8 Low-dimensional representation of the Swiss roll dataset obtained using Laplacian Eigenmaps. The values of the free parameters are $K = 15$, and $t = \infty$. Neighbourhood is preserved but the shape of the manifold is significantly deformed. 30

2.9 Low-dimensional representation of the Swiss roll dataset computed by semidefinite embedding. The solution is isometric and neighbourhood preserving. . 33

2.10 Salience-based feature detector applied to a set of chair images. The variety of chair shapes makes association of features a difficult task for object recognition. 37

2.11 Occupancy grid of ACFR upstairs area overlaid with a corresponding topological map. 40

2.12 Dynamic Bayesian network representing the SLAM problem. Landmark 1 is observed at times 0 and 2, landmark 2 is observed at times 1 and k, while landmark 3 is observed at time k. The control inputs are denoted by $\mathbf{u}_k$ and the hidden state consisting of robot and landmarks positions is denoted by $\mathbf{x}_k$. Sensor and motion models are also indicated. 42

3.1 Autonomous vehicles in unstructured environments. Conventional feature extractors such as SIFT extract too many features in natural imagery as shown in the pictures on the right. The tails of the white lines depict the SIFT extracted features. These features are not useful in autonomous mapping applications where the main objective is to extract unique physical features such as trees, bush and lakes that exist in the environment. The proposed methodology uses a frequentist approach to feature extraction to detect unique natural features. 49

3.2 Gabor wavelets at different orientations and scales. 51

3.3 Sample image from underwater data sequence (a). Information content in colour (b) and texture (c) spaces. 52

3.4 Original image (a) and mutually informative feature set (b). The features that are maximally informative with respect to both the colour and texture cues are highlighted in red. They exhibit significant contrast with respect to their local neighbourhoods and are expected to persist in the environment. . . 53

3.5 Stability of the feature selection scheme over 100 images from the reef data. A sample 8 frames are shown to demonstrate the consistency of the promoted features. Features with the highest contrast displayed in red physically correspond to regions with distinctive hue and texture. 54

3.6 Block diagram of the proposed algorithm. Given an image with an object, VBEM is used to discover the number of components and their parameters. Different models are evaluated and the selected one is returned with the estimated parameters of the density $p(\mathbf{z}' \mid \mathbf{z}^I, object)$. 56
3.7 Example of images from (i) lake dataset, (ii) fence dataset, (iii) tree dataset, and (iv) coral dataset. Some of the images contain the object of interest while others have only the background. 65
3.8 Segmentation results for tree/bush in a farm environment (i); roads from an aerial images (ii) and a piece of coral in an underwater environment (iii). . . . 66
3.9 Diagram of the recognition algorithm from laser and image. In the feature extraction block, laser is used to select regions of interest in the image. The selected patch contains appearance information which is augmented with shape descriptors from laser. The dimensionality is reduced with a combination of PCA, FDA and orthonormalisation expansion. The final classification is obtained with logistic regression. 68
3.10 Projection of laser points in the image and resulting clusters found by the algorithm. Points of the same cluster are assigned the same colour. Note that laser points on the post and on the tree are in different clusters. Points on the wall, bicycles and bicycle racks are assigned to the same cluster. 70
3.11 Steps involved in the extraction of shape descriptors from a laser scan. (1) Laser points projected into the image. (2) Detected clusters. (3) Translation to the origin. (4) Rotation with respect to robot heading. (5) Polynomial fitting. 72
3.12 Selected regions of interested after clustering and projecting laser points into the images. 73
3.13 Images of trees used to train the tree recognition system. 81
3.14 Images labelled as "no trees" used to train the tree recognition system. Some of the images in this class contain parts of trees, however, the tree trunk is not evident or was not hit by the laser, making them difficult for localisation. 82
3.15 Training points projected in the direction of the first and the second eigenvalues computed with PCA+FDA+Expansion approach. Note that the first eigenvalue is able to separate the classes very well. The decision boundary computed with logistic regression is indicated by the blue line. 83
3.16 Receiver-operator curve for the tree classification problem. 85

4.1 Block diagram depicting the off-line model generation and online model inference phases in the proposed feature extraction and representation framework. 87

4.2 Graphical model for computation of parametric models from NLDR algorithms. An arrow directed into a node depicts a dependency on the originating node. The discrete hidden variable s represents a specific neighbourhood on the manifold. The feature extraction and Isomap algorithms supply the input data for learning. 89

4.3 Mixture of 32 factor analysers learnt from data sampled randomly on the S-manifold. The covariances corresponding to $p(\mathbf{z}|\mathbf{x}, s)$ are overlaid on the plot (a). The mixture model captures the local covariance structure of the data over different regions of the manifold. The covariances of $p(\mathbf{x}|s)$ (b). The uncertainty in the low dimensional state of a noisy high dimensional observation $\mathbf{z}$ is represented through the statistical model. 94

4.4 Residual variance as a function of the dimensionality computed by Isomap. Note that the first three dimensions contain most of the data set information. 95

4.5 Sample image acquired by the AGV (**top**) and low dimensional embedding of randomly sampled high dimensional image patches used as the training set (**bottom**). Ellipses represent the covariance matrices of the mixture model learnt through EM. Typical colours and textures in the environment are captured in the low dimensional representation. 97

4.6 Contour of the inferred means of the top eigenvector (**right**) on each 11×11 image patch. This state enables a clear discrimination of the sky (dark blue, range $\approx$0-50) from all other visual groups in the image. 98

4.7 Contour of inferred means of the second eigenvector. This state allows separation of the bush (range $\approx$ 0-50) from the grass and the tracks in the scene. . 98

4.8 Contour of inferred means of the third eigenvector. This state allows discrimination between the bush ($\approx$ 125) and the tracks ($\approx$ 150). Contours of the bush as well as the tracks are also apparent. 99

4.9 Residual variance as a function of the dimensionality computed by Isomap for the underwater dataset. The estimated dimensionality is 5 by verifying the dimensions that most contribute for the residual variance. 99

4.10 Sample image acquired by the UUV (**top**) and low dimensional embedding of randomly sampled high dimensional image patches used as the training set (**bottom**). Ellipses represent the covariance matrices of the mixture model learnt through EM. Typical colours and textures in the environment are captured in the low dimensional representation. 100

4.11 Inference results of a testing image. Every point corresponds to the mean of the dominant component inferred using the model learnt, given the observed features. Points are plotted in the Eigenvector 1 (brightness) vs. Eigenvector 4 (Texture dominant). The vertical axis captures a gradual change in the brightness of the image patches, while the horizontal axis models a decrease in the wavelengths of the patch texture. 101

4.12 Contour of the inferred means of the top eigenvector on each 11×11 image patch. This state physically corresponds to the brightness of the patches. . . . 101

4.13 Contour of the inferred means of the second eigenvector. This state seems to be correlated to the hue of the image patches. 102

4.14 Contour of the inferred means of the third eigenvector. This state is related to characteristic textures. 102

4.15 Inferred low dimensional states (**top**) and weighted average state estimate (**bottom**). The weighted average estimator assumes that the low dimensional state of a test sample is the weighted average of the k nearest high dimensional neighbours of the sample, with the weights being inversely proportional to the high dimensional distances. The trained manifold is represented in red and corresponding test samples are similarly colour coded in each figure. 103

4.16 Inferred low dimensional states (**top**) and weighted average state estimate (**bottom**). The trained manifold is represented in red and corresponding test samples are similarly colour coded in each figure. Good qualitative agreement is observed between the two estimates. 107

4.17 Tracking of bush and tree in 10 different frames of the sequence. Green patches correspond to tree while magenta patches correspond to bush. 108

4.18 KL divergence between “tree” patches in two frames (blue line). KL divergence between “bush” patches in two frames (red dot-dash line) and KL divergence between “bush” and “tree” in the same frame. Every tenth frame from the entire image sequence is used in these computations. 109

5.1 Images used for place recognition. As can be seen, blurring, occlusions, changes in illumination and moving objects such as people and cars are some of the issues a place recognition algorithm has to cope with. 111

5.2 Pioneer AT used in our experiments. 115

5.3 Learning and classification procedures and complexities for place recognition. In this diagram, N is the number of samples; M is the number of components; D and d are the size of the high and the low dimensional spaces; k is the size of the width of patches; n_s and n_a are the number of scales and angles for the Gabor convolution. 117
5.4 Residual variance of Isomap as a function of the number of dimensions. 119
5.5 Sample images of the indoor dataset 121
5.6 The generative model learnt for the kitchen. Points are plotted on the direction of the two largest eigenvalues of the essential features. Ellipses correspond to the covariance matrices of the components learnt with VBEM. The association between the patches and their location in the real scene is also indicated. 122
5.7 Graphical representation of the confusion matrix for the indoor experiment, using the human labels as the ground-truth. Darker cells represent larger values. 123
5.8 Training images used to learn the kitchen model. The resulting model was able to recognise wider views of this place such as in Figure 5.6. 124
5.9 Sample images of the outdoor dataset. 125
5.10 The same as Figure 5.6 but for the "ACFR-park" model. 126
5.11 Graphical representation of the confusion matrix for the outdoor experiment, using the human labels as the ground-truth. Darker cells represent larger values. 127

6.1 Recognised landmarks mapped during SLAM. 135
6.2 Appearance space for the landmarks computed through nonlinear dimensionality reduction. Each point in the graph represents the mean of a landmark from several observations. 138
6.3 Vehicle model. 140
6.4 Utilitary car used in the experiments. For the results reported, the front laser and the low-resolution camera are not used. 140
6.5 Estimated trajectory using odometry only. 142

6.6 Estimated trajectory and landmarks position with conventional EKF-SLAM. For comparison, the estimated trajectory with GPS is represented with the green line. The EKF-SLAM becomes inconsistent after the first large loop which is not closed properly. The error in landmarks pose and vehicle trajectory is thus propagated to the rest of the map. The numbers in the graph indicate successive positions of the trajectory, starting with position 1 at (0,0). Note that the GPS signal is not available in some areas of the park. 143
6.7 Map and vehicle trajectory obtained using the described methodology. 144
6.8 The same as Figure 6.7 with results overlaying an aerial picture of the area. The four images below the map show typical objects of the environment such as people, buses and buildings, and detected landmarks. 146
6.9 Multiple views of one of the landmarks acquired during SLAM. As with this particular landmark, during outdoor exploration, the vehicle observes landmarks at completely different viewpoints. The yellow crosses represent laser points projected on the object. As laser scans and images are not perfectly synchronised, misalignment exists especially when the vehicle is turning (last image on the right). 147

A.1 Checkboard pictures used to calibrate laser and camera. 158
A.2 Calibration procedure for the computation of the extrinsic parameters between laser and camera. The coordinate system used for each sensor and for the calibration board are indicated in green. 159
A.3 Projection of laser points in the checkboard used for calibration. 160

List of Tables

3.1 Recognition performance of the proposed algorithm for different datasets. . . 64
3.2 Recognition performance of VBEM and EM (AROC) in fitting mixtures with different number of components. The trees dataset was used in this experiment. 67
3.3 Comparison between EM-BIC and VBEM for object recognition (AROC) for the tree and the coral datasets. 67

5.1 Precision and recall results for the indoor dataset. 123
5.2 Precision and recall results for the outdoor dataset. 124

Symbols and Notation

Unless otherwise stated, the following notation is used: Matrices are capitalised and vectors are in bold type. A set of vectors is both capitalised and bold type. There is no distinction in notation between probabilities and probability densities.

Symbol	Meaning
$\cdot$	dot product
$\otimes$	outer product
$\succeq$	used to indicate the constraint that a matrix is positive semi-definite, e.g. $K \succeq 0$
$\lceil x \rceil$	smallest integer $\geq x$
$\Vert\,.\,\Vert$	a particular norm, usually L_2
$\vert K \vert$	determinant of matrix K
$\nabla \mathbf{f}_{\mathbf{u}}$	partial derivative of $\mathbf{f}$ w.r.t $\mathbf{u}$
$L^{\#}$	pseudo-inverse transpose of L
$\mathbf{z}^T$	the transpose of vector $\mathbf{z}$
$\mathbf{x}_{k\vert k-1}$	the subscript means estimated at time k given observations up to time $k-1$
d	dimensionality of the output space
$d^2_{r,s}$	square of the distance between points r and s
$\mathrm{diag}\,(W)$	a vector containing the diagonal elements of matrix W
$\mathcal{D}\,(\pi;\lambda)$	Dirichlet distribution for π parameterized by λ
D	dimensionality of the training data
δ_{ij}	Kronecker delta function
$\mathrm{eig}\,(W)$	eigenvector of matrix W
E	residual
η	binary matrix $n \times n$ indicating neighbourhood

<u>Symbol</u>	<u>Meaning</u>
$\mathcal{F}_M$	negative free energy
$\mathcal{G}$	graph
$h\left(x\right)$	differential entropy of random variable x
$\mathbf{h}(\mathbf{x})$, $\mathbf{h}_{\mathbf{x}}$	observation model
$\mathbf{H}$	centring matrix
I	identity matrix or an image
$I\left(x, y\right)$	input image
I_C	colour space
I_T	texture space
$\mathcal{I}\left(x, y\right)$	mutual information between random variables x and y J
$\log\left(z\right)$	natural logarithm (base e)
$\mathcal{L}$	log-likelihood or Lagrangian; clear from the context
m	number of components in a mixture model
M	probabilistic model
n	number of training cases
$\mathcal{N}\left(\mathbf{x} \mid \boldsymbol{\mu}, \Sigma\right)$	the variable $\mathbf{x}$ has a Gaussian distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix Σ
$O\left(\,.\,\right)$	big Oh
$p\left(\,.\,\right)$	probabilistic function or probability
$\mathbf{P}_{k\mid k-1}$	estimated covariance for the state vector at time k
$\Phi(\mathbf{X})$	embedding cost function for LLE
$q_{\boldsymbol{\theta}}\left(x\right)$	free distribution on x
$\mathbf{Q}_k$	estimated covariance of the control inputs noise at time k
$\mathbf{R}_k$	covariance of the observation noise
$\mathbb{R}$	the real numbers
s	discrete hidden variable with multinomial distribution
S	scatter matrix
$\mathbf{S}_k$	innovation covariance matrix
$\mathcal{S}\left(\mathbf{x}; \rho, \boldsymbol{\Lambda}, \omega\right)$	Student-t distribution for $\mathbf{x}$ parameterized by ρ, Λ and ω
$\mathcal{T}(\mathbf{X})$	sum of pairwise squared distances of points in $\mathbf{X}$
$\mathrm{Tr}(K)$	trace of matrix K
$\boldsymbol{\theta}$	set of model parameters
$\tau\left(d^G_{r,s}\right)$	inner product matrix of $d^G_{r,s}$

Symbol	Meaning
Υ	transformed points in semi definite embedding
$\mathbf{U}^k$	control inputs up to time k
$\mathbf{W}_k$	Kalman gain matrix
$\mathcal{W}(\Gamma; \alpha, \mathbf{B})$	Wishart distribution for matrix Γ parameterized by α and $\mathbf{B}$
$\mathbf{x}$	hidden random variable (vector)
$\mathbf{x}_k$	state vector at time k
$\mathbf{x}_{v,k}$	vehicle pose at time k
$\mathbf{x}_{m,l}$	map with landmark positions
χ^2	Chi-squared distribution
$\mathbf{Z}$	$D \times n$ matrix of training inputs $\{\mathbf{z}_i\}_{i=1}^n$ or set of observations
$\mathbf{Z}^k$	observations up to time k
$\mathbf{z}_i$	the ith training input or observation

Chapter 1

Introduction

1.1 Motivation

This thesis is concerned with the problem of building stochastic models of unstructured environments from sensory information. These models are used to address the main perceptual tasks such as recognition, representation and mapping. The methods and algorithms described are applied to specific robotics problems in aerial, terrestrial and underwater domains. The combination of detection, representation and association results in more reliable robots that can operate robustly in complex natural environments. This chapter motivates the thesis and presents the main problems robots face when operating in unstructured domains.

Over the past ten years mobile robotics research has primarily focused on problems related to navigation such as localisation and map building. The first step in building an autonomous robot is to provide the ability to navigate safely in an unknown environment while keeping an internal estimate of its position with respect to the surrounding world. This must be achieved despite noisy measurements, irregular terrain, dynamic environments, different weather conditions and many other complexities.

When a robot estimates its position with respect to incrementally mapped environmental features, the problem is known as simultaneous localisation and mapping (SLAM). Conventional stochastic solutions to SLAM involve the computation of covariance matrices which are in general of complexity $O(n^2)$ in the number of features (or landmarks) in the map. Much effort has been made to reduce the complexity of SLAM algorithms. The best SLAM algorithms can now deal with many thousands of landmarks (Guivant and Nebot 2001b; Thrun et al. 2002; Paskin 2003; Bosse et al. 2004). However, in many cases, position information alone is not enough to navigate robustly. Association of map features becomes difficult as errors in position estimates increase. Furthermore, extraneous objects might exist in the environment and their inclusion in the map can have disastrous consequences. For these reasons, reliable detection and association of landmarks plays a major role in autonomous localisation and mapping. When multiple sensors are combined to detect, associate and represent objects and landmarks, this becomes part of the wider *perception* problem.

Humans can detect and create internal representations of thousands of objects that enable them to classify observed objects as belonging to a particular class, despite variations in shape, colour or size. Such knowledge is acquired over many years of learning involving interpretation of sensory information and creation of models that are robust to complexity. Endowing robots with such capability would considerably enhance autonomy and reliability. In mapping and localisation, for example, a robot would be able to map only specific objects in the environment that are known to be reliable landmarks, and identify dynamic objects commonly found in real applications such as humans and cars.

The benefits of reliable perception in unstructured environments are analysed in two problems. In navigation, the robot has to detected and represent landmarks to create consistent maps. With a large number of detected landmarks the problem is how to associate them correctly despite position uncertainty. Furthermore, many tasks require descriptive maps that possess more information than just feature positions. For these tasks compact probabilistic representations can provide higher level models for decision making. These problems are detailed with illustrative examples below.

1.2 Navigation in Complex Environments

The robust identification of landmarks for localisation and mapping requires the classification of an object as static or dynamic and the selection of landmarks that are easier to identify and associate. Appearance properties such as shape, colour and texture can extract interesting features that are not apparent in purely geometric models. A major issue is the creation of accurate models that can account for the variability of appearance expected in unstructured environments. Appearance models can be divided into two classes: models for recognition encode a general description of the landmark class and must account for all variability within the class in both shape, colour and texture; models for association encode the information necessary to distinguish one particular landmark from others. Models can be generated as the robot navigates by creating unsupervised representations or learnt from training data. As it is difficult to provide extensive datasets that capture the variability of natural environments, this thesis concentrates on the unsupervised approach to building representations which is addressed in Chapter 4.

Most current outdoor robotics is based on sensors that provide geometric (range and bearing) information. Although these sensors are accurate, they only provide geometric profiles of objects which are, in general, insufficient for recognition. Conversely, imaging sensors provide richer information such as shape, texture and colour. They are passive, do not consume much power and, in most cases, are less expensive than ranging sensors. The main problem stopping their applicability to outdoor robotics is the difficulty in interpreting the complex information provided.

As opposed to indoor robotics where large patterns such as walls and doors are easily identified with range sensors, outdoor applications are characterised by the lack of geometric structure. Moreover, the existence of far-field objects introduces another problem as landmarks may lie beyond the maximum range of the sensor. This can be seen in Figure 1.1 where a laser scan is plotted in a typical outdoor image. Because most objects are outside the maximum range of the sensor, only 10 (3%) out of 361 readings obtained from the laser scan are valied and useful measurements. These points are separated into three clusters with very similar spatial configuration. While two of these clusters are caused by reflections from trees, the third cluster results from a person. It can be seen that the identification of the person is difficult from only the range graph, but is possible from the associated image. This gives an idea of the importance of imagery in interpreting the world in outdoor robotics.

The main issue in the use of appearance information in unstructured environments is the difficulty in computing models able to encode the complexity of the data. The creation of appearance models for recognition involves learning generative or discriminative models from training data. The lack of structure in the environment imposes many difficulties, of which the need for extensive datasets is the most challenging. To address this issue, Bayesian inference is used in this thesis to show how few training examples can be used to create generative models for recognition and segmentation of natural features.

Although visual information can provide most of the necessary features for recognition, the task can be computed more efficiently when the search for the object is constrained to specific areas in the image. Range readings can be used to reduce the image area where objects are more likely to be found. An algorithm using this idea is presented in Chapter 3. This algorithm extracts shape information from laser readings and fuses it with appearance features to recognise landmarks in a discriminative fashion.

The ability to interpret complex environments is a key challenge for robots. Throughout this thesis algorithms are presented to address this problem combining multiple sensors and creating stochastic representations from training data.

1.2.1 Compact Probabilistic Representations

In many applications, a map with only the position of landmarks does not contain enough information for higher-level decision making. Sometimes it is desirable to have additional visual information that includes properties such as colour, shape and texture that can be

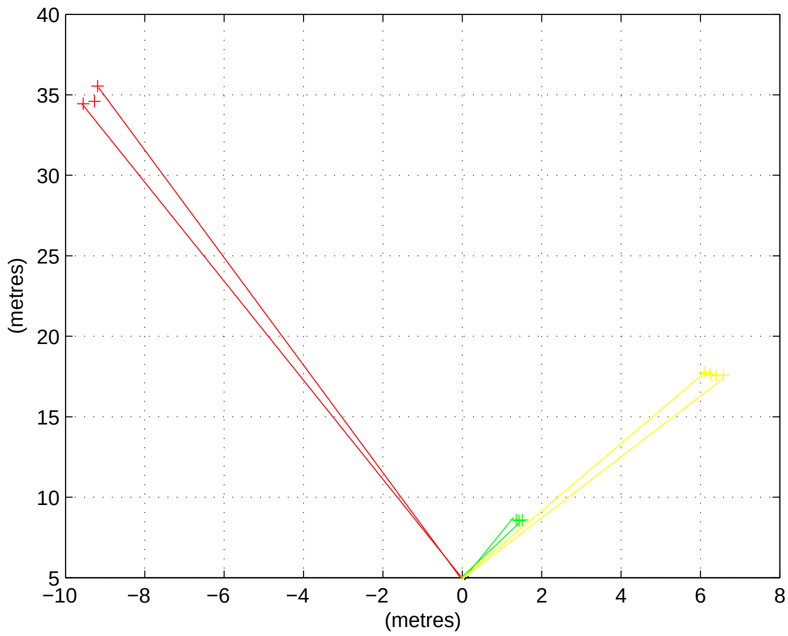

Figure 1.1: Typical image and laser scan from an outdoor environment. As most objects are further than the maximum range of the sensor, only 10 distances were obtained. They are insufficient for correct characterisation of the objects as shown in the image.

used to identify objects of interest for particular applications. Examples include rescue, inspection missions and underwater exploration.

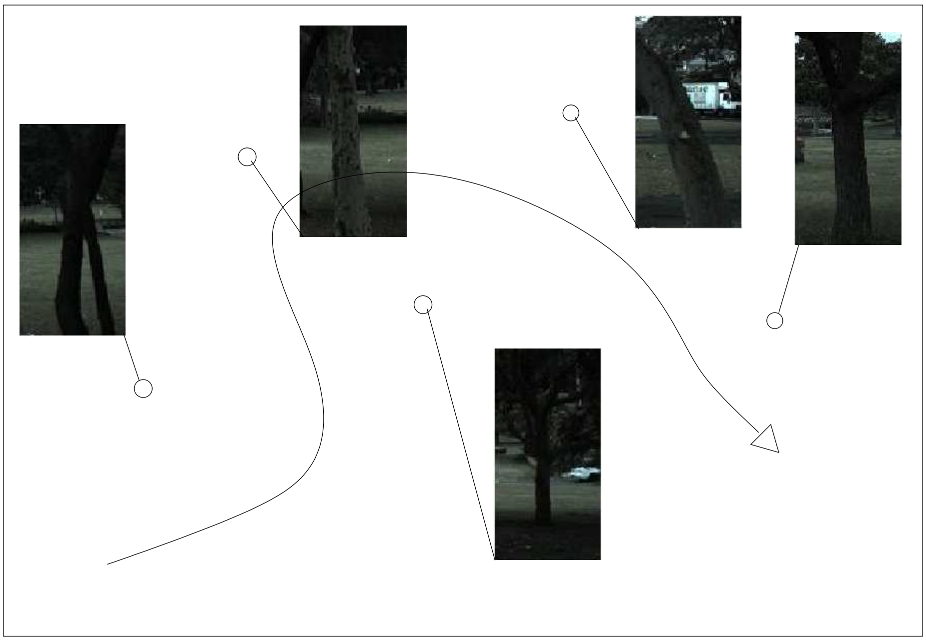

Figure 1.2: Metric map augmented with landmark pictures. The additional visual information allows better perception and decision making.

Conventional SLAM algorithms create maps of point features representing the centre of landmarks. Those features are in general detected with range sensors and the centroid is extracted from the profile of the range measurements obtained. Whilst this type of map helps localisation, it does not contain the necessary information to, for example, classify trees according to their specie or to distinguish objects.

To illustrate this problem, Figure 1.2 depicts a metric map with pictures of identified landmarks. Map and pictures were generated by solving the SLAM problem in an urban environment. The inclusion of visual information allows better recognition and association of landmarks. Additionally, it provides information for higher level decision making, for tasks that go beyond navigation. From this example, it can be concluded that the combination of range sensors and imaging sensors for building maps improves the value of SLAM. Accurate range measurements and richer appearance information from cameras are thus complementary.

This thesis proposes a methodology to build compact probabilistic models to encode visual features. The framework is developed with non-linear dimensionality reduction techniques associated with statistical learning. The information from cameras can thus be incorporated into a SLAM framework to yield maps that have at the same time accurate position estimates and higher level feature information, such as landmark images obtained at different viewpoints. The probabilistic visual representation for natural features is described in

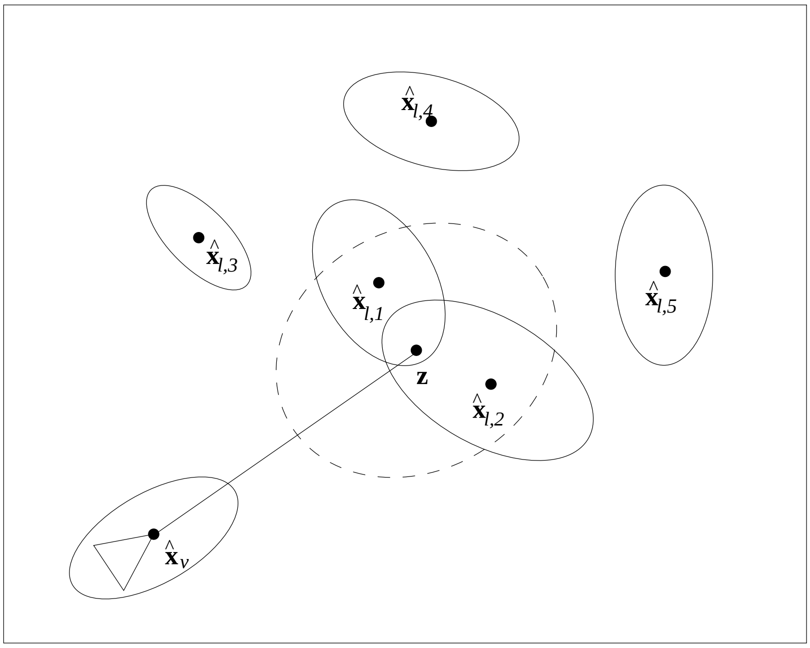

Figure 1.3: Gating procedure for data association. The robot with position $\hat{\mathbf{x}}_v$ and uncertainty represented by the ellipse has to associate an observation $\mathbf{z}$ to landmarks previously observed denoted by $\hat{\mathbf{x}}_{l,1}, \hat{\mathbf{x}}_{l,2}, \hat{\mathbf{x}}_{l,3}, \hat{\mathbf{x}}_{l,4}, \hat{\mathbf{x}}_{l,5}$. The gate defined by the dashed ellipse eliminates landmarks $\hat{\mathbf{x}}_{l,3}, \hat{\mathbf{x}}_{l,4}$ and $\hat{\mathbf{x}}_{l,5}$.

Chapter 4 of this thesis.

1.2.2 Robust Data Association

A major problem when performing localisation and mapping in large outdoor environments is reliable data association. As the uncertainty over the position of landmarks and vehicle grows, correct association can be very difficult. If for some circumstance an incorrect data association hypothesis is accepted and introduced in the estimation process, the filter may become inconsistent compromising the whole map. The traditional approach for data association is to use a technique known as *gating* (Blackman and Popoli 1999), explained in detail in Chapter 2.

Gating computes a hypothesis test to eliminate associations that are unlikely to be true considering the uncertainty in robot and landmark positions. This is illustrated in Figure 1.3. The uncertainty on robot pose is represented by the ellipse around its expected position denoted by $\hat{\mathbf{x}}_v$. At a particular instant, the robot makes a new observation $\mathbf{z}$. The problem is then to associate $\mathbf{z}$ to some of the landmarks previously detected $(\hat{\mathbf{x}}_{l,1}, \hat{\mathbf{x}}_{l,2}, \hat{\mathbf{x}}_{l,3}, \hat{\mathbf{x}}_{l,4}, \hat{\mathbf{x}}_{l,5})$ or classifying it as a new landmark. The dashed ellipse represent landmarks that are within the gate, and are possible associations for observation $\mathbf{z}$. This eliminates landmarks $\hat{\mathbf{x}}_{l,3}, \hat{\mathbf{x}}_{l,4}$

and $\hat{\mathbf{x}}_{l,5}$ from consideration. Among the remaining landmarks, the decision is to a associate $\mathbf{z}$ to either $\hat{\mathbf{x}}_{l,1}$ or $\hat{\mathbf{x}}_{l,2}$, or to a new landmark. Strategies to address this problem consider the nearest neighbour (NN) as the most likely. When the distance to the nearest neighbour is further than a defined distance, the measurement is considered as coming from a new landmark. Although this methodology is computationally efficient and provides accurate results for small trajectories it clearly fails when position uncertainties are large. Gating and NN become unreliable and not suitable for large scale SLAM problems.

The detection of loop closure in SLAM is a significant issue when a robot follows an extensive path. Observed landmarks are initialised in the map and have to be recognised and correctly associated when the robot re-observes them in a trajectory with loops. When navigating without closing the loop, the uncertainty of the robot location grows making further associations more difficult. Furthermore, the number of landmarks used in large-scale SLAM can be significant, increasing the complexity of the problem. This is illustrated in Figure 1.4 where SLAM is performed with hundreds of features. The uncertainty of the vehicle and feature positions grow making data association difficult.

This thesis address the problem of loop closure and data association in large-scale environments by integrating high-level appearance models into the data association process. A demonstration of these ideas in an extensive unstructured environment is provided in Chapter 6.

1.3 Contributions

The main contributions of this thesis are:

1. **Stochastic representation of objects through non-linear dimensionality reduction.** A non-linear statistical model encoding a neighbourhood-preserving dimensionality reduction is proposed as a representation for features in natural environments. This representation is able to distinguish similar objects such as trees and bushes. Efficient inference in this model is formulated and tested with inputs of thousands of dimensions. Inference operations result in mixture of Gaussians that can be integrated in a non-linear filtering scheme.

2. **Landmark recognition system using a laser and camera.** A new algorithm that combines laser and camera information for object detection in outdoor environments is presented. The algorithm is based on a combination of unsupervised and supervised dimensionality reduction methods that fuses information from these two sensors to discover the most discriminative dimensions. Having points mapped to a lower-dimensional space, logistic regression is applied for the final classification. The

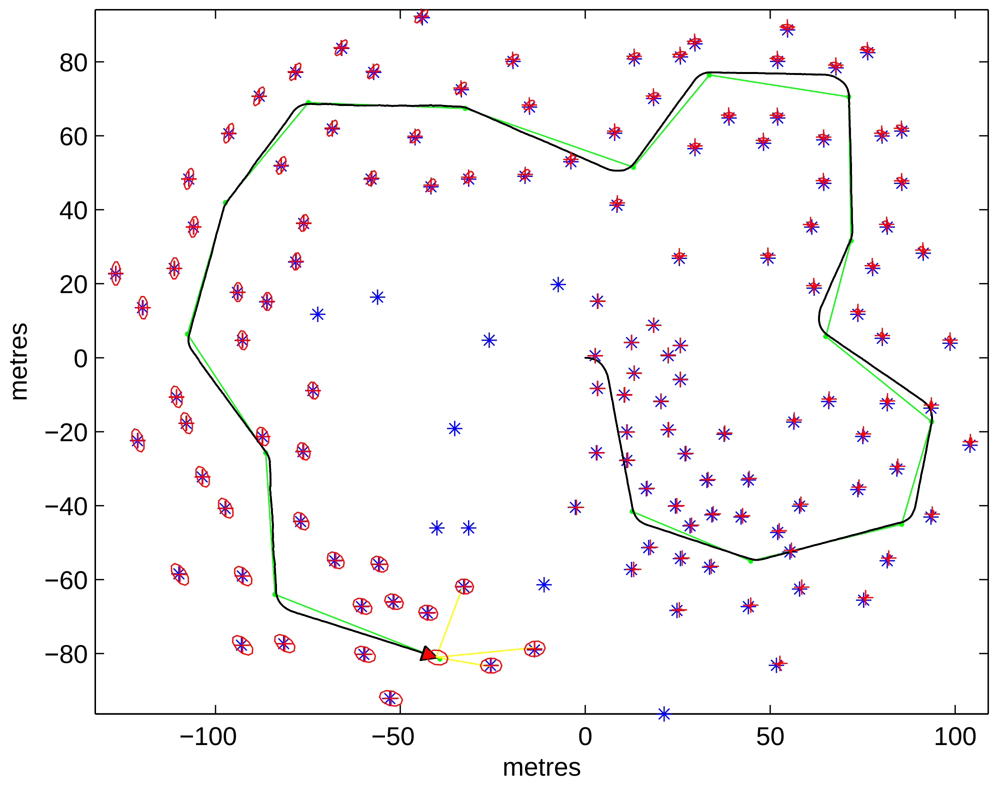

Figure 1.4: SLAM over an extensive trajectory. The uncertainty of the vehicle and feature positions increases making data association difficult before closing the loop.

algorithm can process about 5 frames per second and is able to detect objects up to 30 metres from the robot. As opposed to most computer vision algorithms for object recognition that have to search for objects across the whole image, this new approach uses laser to identified regions of interest, and only information in these regions are processed. The resulting real-time algorithm works with high-resolution images to incorporate texture information.

3. **Data association algorithm combining appearance and position properties.** With a probabilistic representation both position and appearance can be used to associate measurements with landmarks using the normal gating technique. The result is a more robust data association algorithm that combines complementary clues. When data association using only position is statistically sufficient, appearance information is not required. When position information is not sufficient, the augmentation of landmark models with appearance information significantly helps in selecting the best hypothesis.

4. **Real-time detection and segmentation of natural objects from few training images.** A fully Bayesian learning methodology using variational calculus is derived for multivariate mixtures. Generative models for object and non-object image patches are learnt from less than 10 images. The algorithm can be used both to segment and classify natural objects and is tested in aerial, underwater and terrestrial domains. Results demonstrate that the algorithm is more stable and accurate than conventional maximum likelihood techniques.

5. **Large-scale deployment of the algorithms for a challenging outdoor simultaneous localisation and mapping problem.** The combination of detection, representation and data association significantly improve results over conventional SLAM algorithms while additionally providing statistical models of landmarks and images acquired at different distances and view points. The framework is tested in an unstructured environment where conventional approaches fail.

6. **Real-time place recognition system tested in indoor and outdoor environments.** The combination of dimensionality reduction and statistical modelling can be applied to the problem of place recognition from images. Generative models from places are learnt from a reduced dataset of images and labels. Images are divided into smaller patches and a classification scheme is proposed where instead of classifying patches individually, uses the whole set of patches in the image to draw its decision. Results are demonstrated in indoor and outdoor experiments.

1.4 Thesis Outline

This thesis is organised as follows:

Autonomous Recognition and Mapping

In Chapter 2, the basic concepts of object recognition, dimensionality reduction and simultaneous localisation and mapping are presented. A review of previous work in these areas is provided and current linear and nonlinear dimensionality reduction techniques detailed. Algorithms for localisation and mapping are explained and recent research on combining vision and range sensors for simultaneous localisation and mapping emphasised.

Recognition of Natural Features

Algorithms for detection and segmentation of natural features are presented in Chapter 3. The chapter starts by describing a Bayesian framework for learning generative models for object recognition from images. These models can be trained from few images and additionally used for segmentation. When more training samples or other sensors are available the combination of dimensionality reduction and discriminative learning can provide a faster solution for object detection. Using these ideas, a method for classification from laser and visual information is proposed and tested for recognition of trees in an urban park.

Stochastic Representation

The problem of representing natural features with probabilistic models is presented in Chapter 4. The probabilistic model described represents a regression function from raw sensor data to a low-dimensional space where essential properties are preserved. The representation has a form of a mixture of linear models with uncertainty encoded by a set of Gaussians. Experiments demonstrate the potential of the approach for intelligent information compression and abstraction.

Place Recognition

The problem of place recognition from images is explored in Chapter 5. Given a set of images from particular places, the robot has to recognise its location. The problem can be seen as a multi-class classification task. However, rather than learning a single classifier, the approach creates generative models for each place. This has the advantage of being incremental, i.e. if classification is required for additional places, models already learnt can still be used and new data is incorporated to other models.

Integrating Perception with Mapping

Perceptual models described in previous chapters are combined and applied to the problem of simultaneous localisation and mapping in Chapter 6. Reliable landmark recognition eliminates the problem of navigating in an environment with dynamic objects while landmark appearance representation can significantly improve data association. The combination of position and appearance information for data association is discussed in the chapter, with further outdoor experiments reporting the benefits of the new approach.

Conclusions and Future Work

Chapter 7 concludes the thesis by analysing the experimental results obtained with the methodology proposed in Chapters 3, 4, 5 and 6. Directions for future work and open issues regarding representation from multi-sensory information and their application to robotics are then discussed.