Long Term Appearance-based Mapping with Vision and Laser

D. Phil Thesis

Mobile Robotics Group
Department of Engineering Science
University of Oxford



Supervisor
Professor Paul Newman

Rohan Paul
St. Catherine's College

Trinity Term, 2012

Rohan Paul
St. Catherine's College

Doctor of Philosophy
Trinity Term, 2012

Abstract

This thesis is about appearance-based topological mapping for mobile robots using vision and laser. Our goal is life-long continual operation in outdoor unstructured workspaces.

We present a new probabilistic framework for appearance-based mapping and navigation incorporating spatial and visual appearance. Locations are encoded probabilistically as random graphs possessing latent distributions over visual features and pair-wise euclidean distances generating observations modeled as 3D constellations of features observed via noisy range and visual detectors. Multi-modal distributions over inter-feature distances are learnt using non-parametric kernel density estimation. Inference is accelerated by executing a Delaunay tessellation of the observed graph with minimal loss in performance, scaling log-linearly with scene complexity.

Next, we demonstrate how a robot can, through introspection and then targeted data retrieval, improve its own place recognition performance. We introduce the idea of a dynamic sampling set, the onboard workspace representation, that adapts with increasing visual experience of continually operating robot. Based on a topic based probabilistic model of images, we use a measure of perplexity to evaluate how well a working set of background images explains the robot's online view of the world. Offline, the robot then searches an external resource to seek additional background images that bolster its ability to localize in its environment when used next.

Finally, we present an online and incremental approach allowing an exploring robot to generate apt and compact summaries of its life experience using canonical images that capture the essence of the robot's visual experience-illustrating both what was ordinary and what was extraordinary. Leveraging probabilistic topic models and an incremental graph clustering technique we present an algorithm that scales well with time and variation of experience, generating a summary that evolves incrementally with the novelty of data.

Funding

This research was supported through the Rhodes Scholarship by the Rhodes Trust (2008-2011), Graduate Scholarships from St. Catherine's College (2008-09 and 2011-12) and by BAE Systems under Mapping and Navigation project (2011-2012).

Statement of Authorship

This thesis is submitted to the Department of Engineering Science, University of Oxford, in fulfilment 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.

Rohan Paul, St. Catherine's College

Copyright ©2012

Rohan Paul

All rights reserved.

To my family and teachers for their unwavering love, support and guidance.

Acknowledgements

As my DPhil journey draws to a close, I remain immensely grateful to those who supported me in this endeavour. First and foremost, I wish to thank my advisor Paul Newman for his guidance, energy and support throughout the last four years. Paul is an inspiring individual, mentor and researcher and I have learned a great many things from him. To be enthused by Paul's vision and being a part of something incredible called MRG has been a privilege. My heartfelt gratitude to the entire MRG crew for not only teaching me interesting stuff but for also sharing the ups and downs of research.

Special thanks to Mark Cummins for helping me find my feet in the first few months, and for his many words of wisdom. Thanks to Ingmar Posner for his mentorship and some fantastic discussions on machine learning and autonomy. I thank Alastair Harrison, Alex Stewart and Benjamin Davis for all their hardwork in mantaining the lab platforms, systems and tools.

I wish to particularly thank Ian Baldwin and Winston Churchill for sharing this journey with me, and for their friendship and peer support. Thanks to Mike Smith for collecting the New College data set with us. I would like to thank Janet Sadler for her administrative help, and Justine Glazebrook for proof reading thesis chapters. I would like to mention Byron Byrne my Graduate Advisor for his useful advice.

I am grateful to Daniela Rus for advice on star clustering that eventually led to the visual summarization work reported in this thesis. I wish to thank Rudolf Triebel for collaborative work on unsupervised learning and Peter Corke for very enjoyable work on visual localisation. Special thanks to Andrew Zisserman and VGG members for useful feedback and pointers on vision related topics. I would like to thank Dave Blei, Zoubin Ghahramani, Carl Rasmussen and Yee Whye Teh for their outstanding talks at the Machine Learning Summer School, that inspired my interest in machine learning and played a key role in later research.

I am immensely grateful to the Rhodes Trust for giving me the privilege of studying at Oxford. The scholar and staff community at the Rhodes House has been like a family for me. Special thanks to the Warden, Don Markwell and Mary Eaton for always being supportive and for encouraging me to think about the broad impact of my research. To my friends, who made the Oxford experience so special: Varun Gulshan, Arghya Sengupta, Dhvani Mehta, Nina Hall, Emma Preston, Sid Easwaran, Mark Sheehan, Melis Anahtar, Alia Whitney-Johnson and Todd Gingrich; thanks for everything.

I am greatly indebted to my high school teachers Vidhu ma'am, Bhargav ma'am and "Bade" bhaiya for sparking my interest in science and engineering. I thank my professors at IIT Delhi, M. Balakrishnan, P. V. M. Rao and Subhashish Banerjee, who laid the foundation for this endeavour.

Lastly, and most importantly, I thank my parents, Drs. Shashi and Vinod Paul, and my brother, Ketan who have given me unconditional love, support and encouragement throughout my life. This thesis is dedicated to them.

Contents

1 Introduction **1**

1.1 Overview 1

1.2 Outline of the Problem 2

1.3 Contributions 3

1.4 Thesis Structure 5

1.5 Publications 5

2 FAB-MAP 3D: Topological Mapping with Vision and Laser **8**

2.1 Appearance-based Topological Mapping 10

2.1.1 Place Recognition or "Loop closing" 10

2.1.2 Notable Place Recognition Systems in Robotics 12

2.1.3 Overview of FAB-MAP System 16

2.2 Related work 23

2.2.1 Use of Spatial Information in Computer Vision Tasks 23

2.2.2 Use of Geometric Information in Workspace Modeling 25

2.2.3 Combining Visual and Range Data for Robot Mapping 27

2.3 Topological Mapping with Spatial and Visual Appearance 28

2.3.1 Illustrative Example 30

2.3.2 Probabilistic Model for Locations 34
2.3.3 Learning Distributions over Feature Distances 41
2.3.4 Probabilistic Navigation and Mapping 46
2.3.5 Accelerating Graph Likelihood Computation 53
2.4 Evaluation 57
2.4.1 Platform and data set 57
2.4.2 Processing pipeline 59
2.4.3 Precision-Recall Performance 61
2.4.4 Loop Closures 64
2.4.5 Learning Spatial Appearance 67
2.4.6 Efficiency and Timing 71
2.4.7 Limitations 72
2.5 Conclusions 73

3 Self-Help: Seeking Perplexing Images for Ever Improving Navigation 76

3.1 Related Work 78
3.1.1 Latent Topic Models and Applications 78
3.1.2 Novelty Detection in Images and Videos 82
3.1.3 Related Work in Active Learning 83
3.2 Probabilistic Topic Models 85
3.2.1 Illustration 86
3.2.2 Latent Dirichlet Allocation (LDA) 91
3.3 Retrieving Thematically Similar Images 96
3.4 Is an Image Perplexing? 101

3.5 Application to Topological Mapping . . . 104
3.5.1 Recap of FAB-MAP Formulation . . . 104
3.5.2 The Sampling Set . . . 105
3.5.3 Incremental Sampling Set Construction . . . 106
3.6 Results . . . 110
3.6.1 Data set and Platform . . . 110
3.6.2 Learning Visual Topics . . . 111
3.6.3 Retrieval of Thematically Similar Images . . . 112
3.6.4 Detection of Novel Images . . . 116
3.6.5 Incremental Improvement in Loop Closure Detection . . . 116
3.7 Chapter Summary . . . 123

4 Online Visual Workspace Summaries 126

4.1 Related work . . . 128
4.1.1 Canonical Views Selection . . . 128
4.1.2 Organizing Image Collections using Meta-data . . . 129
4.1.3 Summarization and Layout . . . 130
4.1.4 Video summarization . . . 131
4.1.5 Applications in Mobile Robotics . . . 132
4.2 Image Representation in Topic Space . . . 134
4.3 Star Clustering and Online Organization . . . 136
4.3.1 Clustering with Star Sub-graphs . . . 138
4.3.2 Online Operation . . . 144
4.4 Incremental Visual Summary Generation . . . 145
4.5 Results . . . 148

4.5.1 Vocabulary and Topic Learning 148
4.5.2 Visual Summaries 148
4.5.3 Thematic Annotation 153
4.5.4 Topical Clusters 153
4.5.5 Clustering Quality 157
4.5.6 Efficiency and Timing 160
4.5.7 Additional Results 164
4.5.8 Other Insights 167
4.6 Chapter Summary 171

5 Conclusions and Future Work **173**

5.1 Summary of Contributions 173
5.2 Future work 176
5.3 Concluding Remark 181

"The search for truth is in one way hard and in another way easy, for it is evident that no one can master it fully or miss it wholly. Each adds a little to our knowledge of nature, and from all the facts assembled there arises a certain grandeur."

Aristotle, 350 B.C.

Chapter 1

Introduction

1.1 Overview

A robot perceives the world through sensors, which are noisy and ambiguous. An intelligent robot must answer a fundamental question: given a number of noisy sensor observations what does the world look like and where am I located? Essentially, it needs a map representation of the environment. Constructing and localizing within a map is a central problem in robotics and a pre-requisite for subsequent planning and action tasks.

One category of robot mapping deals with *metric* maps that provide a fine-grained metrically accurate picture of the world. However, these maps are usually only locally consistent as cumulative measurement errors often lead to global inconsistencies, particularly in large environments. In this thesis we explore *topological* maps that capture the higher-order spatial connectivity of the environment expressed as a graph containing significant places in the workspace with semantic links connecting them. A key element of topological mapping is the *loop closing* problem:

the robot's ability to recognize if it is back to the same place. A central theme of this thesis is topological mapping and loop closure detection using appearance information alone. We investigate algorithms to model the visual and spatial characteristics of places, learn *a-priori* which appearances are common and summarize topical themes in the robot's workspace.

The second theme that permeates this thesis is *life-long* learning: a robot's ability to improve performance with widening experience. We visualize a future where robots operate for long periods of time in dynamic unstructured environments. To be effective, the robot must adapt to its environment and refine its understanding with new observations of the world. It must maintain and continually update its representation of what is common and rare in its environment. The key competency we seek is *introspection*: to determine what was perplexing or unexplained, that further drives active information gathering and assimilation, allowing continual performance improvement with widening experience of the mobile robot.

1.2 Outline of the Problem

We bring the two themes of topological mapping and life-long learning together and tackle the following three questions in this thesis:

- How does a robot model the spatial and visual appearance of its workspace and reason intelligently about the topological structure of its environment?
- How does a robot adapt to its operating environment and improve its place recognition performance over time by identifying perplexing visual themes and searching for relevant models from past experience?

- How does a continuously exploring robot, accruing visual information, summarize its experience discovering thematic structure in the environment?

1.3 Contributions

The contributions made by this thesis are structured within the following themes:

FAB-MAP 3D: Topological Mapping with Vision and Laser

We present a novel probabilistic framework for appearance-based mapping and navigation using spatial and visual appearance data. Like much recent work on appearance based navigation [34, 35] we adopt a bag-of-words approach in which positive or negative observations of visual words in a scene are used to discriminate between already visited and new places. We add an important extra dimension to the approach: we explicitly model the spatial distribution of visual words as a random graph in which nodes are visual words and edges are distributions over distances. Care is taken to ensure that the spatial model is able to capture the multi-modal distributions of inter-word spacing and account for sensor errors both in word detection and distances. Crucially, these inter-word distances are viewpoint invariant and collectively constitute strong place signatures and hence the impact of using both spatial and visual appearance is marked. We provide results illustrating an increase in precision-recall area compared to a state-of-the-art visual appearance only systems.

Self Help: Seeking Out Perplexing Images for Ever Improving Navigation

We demonstrate how a robot can, through introspection and then targeted data retrieval, improve its own performance. It is a step in the direction of life-long learning and adaptation and is motivated by the desire to build robots that have plastic competencies which are not baked in. We consider a particular instantiation of this problem in the context of place recognition. Based on a topic based probabilistic model of images, we use a measure of perplexity to evaluate how well a working set of background images explains the robot's online view of the world. Offline, the robot then searches an external resource to seek out additional background images that bolster its ability to localize in its environment when used next. In this way the robot adapts and improves performance through use.

Online Visual Workspace Summaries

Someday mobile robots will operate continually. Day after day, they will be in receipt of a never ending stream of images. In anticipation of this, we present an online and incremental approach allowing an exploring robot to generate apt and compact summaries of its life experience using canonical images that capture the essence of the robot's visual experience-illustrating both what was ordinary and what was extraordinary. Leveraging probabilistic topic models and an incremental graph clustering technique we present an algorithm that scales well with time and variation of experience, generating a summary that evolves incrementally with the novelty of data.

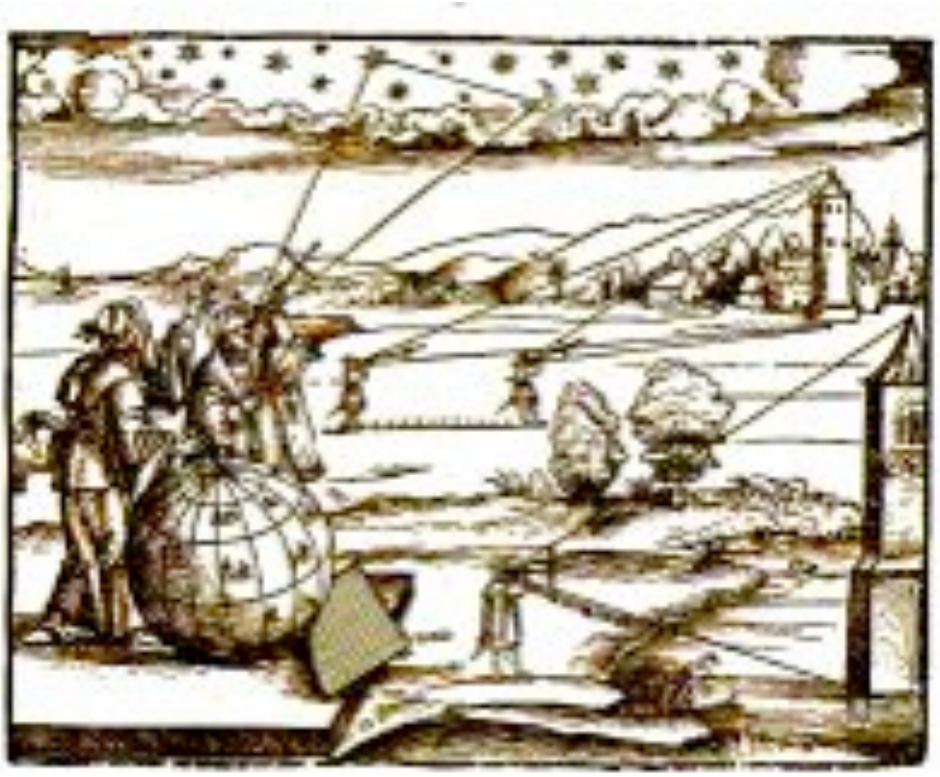
1.4 Thesis Structure

This thesis is structured as follows. Chapter 2 presents an overview of existing appearance-based mapping techniques in robotics and describes FAB-MAP [34], an existing state-of-the-art place recognition system. This is followed by "FAB-MAP 3D", a novel probabilistic formulation that incorporates 3D spatial information and demonstrate performance improvement by deploying a vision-cum-laser system. In Chapter 3 we present a technique called "Self-Help" that learns a relevant model of the visual world through introspection and targeted data retrieval demonstrating asymptotic improvement in place recognition performance over time. This chapter concludes the work on topological mapping. In Chapter 4 we view vision-based topological mapping as compression and turn our attention to the problem of generating visual summaries, presenting an online algorithm that can summarize the life experience of a mobile robot exploring a workspace. Chapter 5 concludes this thesis by summarizing contributions and outlines future avenues for research.

1.5 Publications

FAB-MAP 3D presented in Chapter 2 was published in the proceedings of the IEEE International Conference on Robotics and Automation (ICRA) in 2010 [134], and was recognized as Best Vision Paper Finalist. The system was subsequently presented at the RGB-D Workshop at Robotics Science and System (RSS) Conference 2010 and European Robotics Forum 2011. The systems work related to New College Vision and Laser Data set used in Chapter 2 contributed to a data paper in the International Journal of Robotics Research (IJRR) in May 2009 [174]. The framework described in Chapter 3 appeared in the proceedings of ICRA 2011 [135] and

was presented at the Life-long Autonomy and Exploration Workshop, ICRA 2011. The work described in Chapter 4 appeared in the proceedings of ICRA 2012 [136]. Related applications of presented techniques appeared in the proceedings of the American Association of Artificial Intelligence (AAAI) Conference 2012 [186] and the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) 2012 [137].



Introductio Geographica, Petrus Apianus, 1533 A.D.

Chapter 2

FAB-MAP 3D: Topological Mapping with Vision and Laser

This chapter concerns non-metric navigation and mapping in appearance space - a by-product of which is loop closure detection. In essence, we want to have a robot create a topological representation of its trajectory represented by a graph in which each node is a distinct place and edges represent transitions between places. There has been substantial work on this topic in recent years (Section 2.1 provides an overview) most of which has used a single sensing modality - usually vision. In this work, we provide and test a formulation which uses not only the visual appearance of scenes but also aspects of its geometry. The approach presented here, titled FAB-MAP 3D, has its roots in the FAB-MAP algorithm [34] which has recently been shown to operate in realtime over trajectories of 1000km with high precision [35, 33, 34, 32]. The essence of FAB-MAP is that it learns a probabilistic model of scene appearance online using a generative model of visual word observations and a sensor model which explains missed observations of visual words. We take the same

approach in this work but have the added complication that the observation of spatial ranges between words is coupled to the observation of pairs of visual words incorporating range information from Lidar, stereo or structure from motion. We capture this interaction via a random graph which models a distribution over word occurrences as well as their pairwise distances. We describe how through non-parametric Kernel Density Estimation we can learn interesting and suitable distributions over inter-word distances and also accelerate inference by executing a Delaunay tessellation of the observed graph. Finally, there is also an important prima facie advantage of using distances because they are invariant under rigid transformation and that is precisely what we require of a place descriptor in topological navigation. We shall demonstrate our system and show improved performance over vision-only sensing in an outdoor setting. We must stress though that throughout this work we only need intra-scene distances which can be derived in a local frame, nowhere do we require a single metric picture of the world.

The remainder of this chapter is structured as follows. We begin by discussing the topological mapping and loop closure detection problem reviewing notable place recognition systems, including an overview of FAB-MAP probabilistic mapping system, Section 2.1. This is followed by a review of related work in computer vision and robotic mapping that incorporates spatial and geometric information, Section 2.2. Next, we present FAB-MAP 3D probabilistic navigation and mapping formulation in Section 2.3. Implementation and performance evaluation is presented in Section 2.4. Finally, Section 2.5 concludes this chapter.

2.1 Appearance-based Topological Mapping

2.1.1 Place Recognition or "Loop closing"

Place recognition is a key problem in robot mapping and navigation. Metric mapping using traditional SLAM approaches has an inherent drawback that positional uncertainty grows over time due to accumulating errors from dead-reckoning or scan-matching methods resulting in globally inconsistent maps. Figure 2.1 shows an example. Recognizing previously visited places using appearance information can provide additional positional constraints significantly improving the accuracy of the resulting map. Additionally, the ability to recognize previously-visited places also allows a mobile robot to discover the topology of operating workspace, in essence creating *topological* maps that capture the higher-order workspace connectivity, see Figure 2.2. Here, the environment as a graph where nodes represent places and arcs capture connectivity via transitions. An exploring robot incrementally initializes new places as nodes until the robot re-visits the place and the loop is closed. Further, place recognition has other applications like assisting the robot in service tasks like recognizing charging stations, key landmarks or goal positions etc.

Detecting loop closures is challenging due to two main reasons as illustrated in Figure 2.3. Firstly, different parts of the robot's workspace may look the same to the sensors, a problem termed as perceptual aliasing. This occurs due to objects like foliage, walls, road markings or repeating architectural elements like windows, railings etc that are pervasive in the workspace. A successful system must learn which features are common in the environment and adequately base the loop closure decision on more distinctive features. Secondly, real world is dynamic and sensory appearance of a place can vary across time. As a mobile robot revisits a place

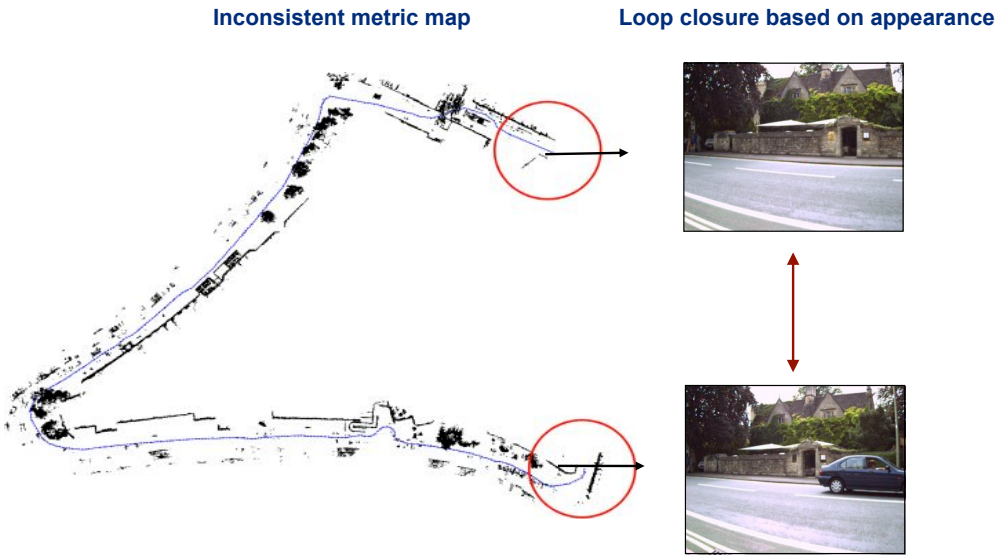


Figure 2.1: Importance of loop closure detection. Accumulation of incremental errors can lead to metrically inconsistent maps. The figure shows a laser scan-matching based map resulting at the end of a traversal along a triangular loop where measurement errors (particularly around corners) grow over time causing the two ends of the map to diverge. Loop closure detection using another modality like camera imagery can provide additional metric constraints, yielding a consistent map. Figure courtesy Alastair Harrison and Mark Cummins.

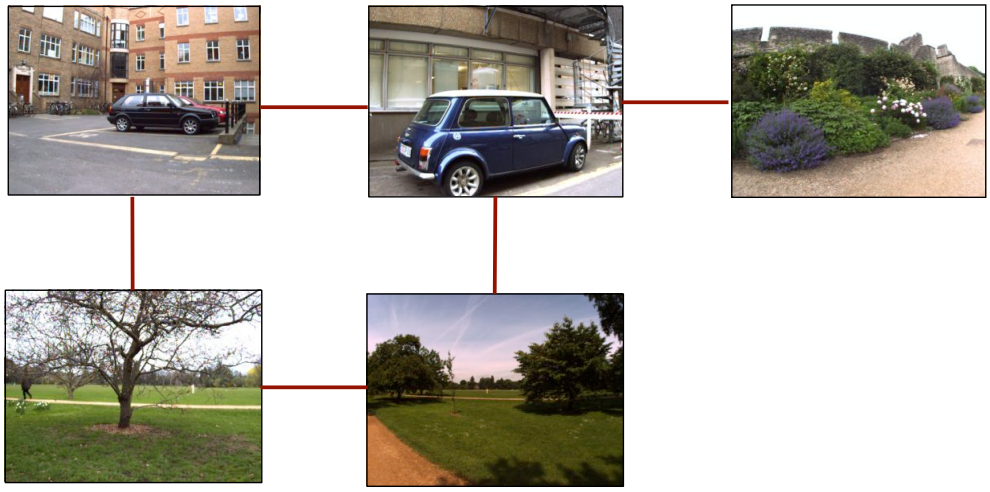


Figure 2.2: Topological map of a workspace using camera imagery. Nodes represent places and arcs correspond to transitions. New places are initialized during exploration until the robot re-visits a place and a loop is closed. Topological maps capture the high-level connectivity in the workspace giving a higher-order representation of the environment.

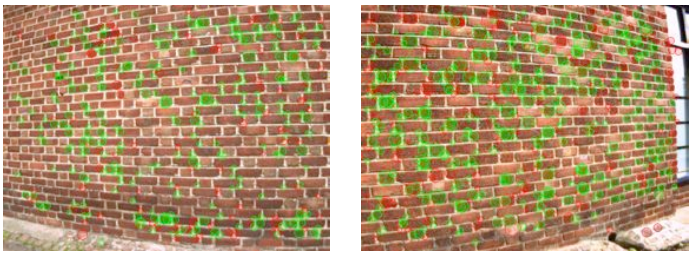
the scene might look different due to dynamic elements like cars, people etc. The same location can look different due to illumination, weather or view point changes causing different features to be observed. Hence, there is a need to account for the minor changes in scene appearance when a place is revisited.

Place recognition or loop closure detection has been an active area of research in recent years within robotics and computer vision communities. In the next section we survey some prominent works relevant to this thesis.

2.1.2 Notable Place Recognition Systems in Robotics

Global image similarity-based methods

Early efforts focused on using global image similarity to recognize loop closures. Ho and Newman [77] present a method where pair-wise similarity between all scenes is encoded as a similarity matrix and loop closures are detected by finding statistically



(a)



(b)

Figure 2.3: Challenges in appearance-based loop closure detection: (a) Perceptual aliasing: Repetitive structures like brick walls in the environment cause different places to look remarkably similar. (b) Scene variability: Appearance of the same place can vary due to dynamic objects, variable lighting or viewpoint changes. Images from [34].

significant sequences of similar scenes. Perceptual aliasing is ameliorated by a rank reduction step by an eigen value decomposition of the similarity matrix, in effect removing the common themes (principle eigen vectors) permeating the environment. However, the solution scales cubically and is not posed in a probabilistic framework. The recent localization-only work by Milford [121] takes the same approach of finding similar trajectories where image comparison is done using the normalized cross-correlation measure. In [170], Singh et al. first build a manhattan-world model for indoor environments and then match panoramas using extracted gist features. In a related work, Levin et al. [105] use global colour histograms and colour moment information to determine scene matches.

Feature-based methods

A second category of approaches rely on matching salient features extracted from appearance data. Silpa-Anan and Hartley [166, 167] present a correspondence matching scheme based on SIFT features [109] to detect loop closures and further present optimizations using KD-trees, voting and an improved RANSAC method. Angeli et al. [3] present an incremental loop-closure detection scheme using a bag-of-words approach coupled with epipolar geometric checks. In another effort, Milford and Wyeth [120] present a biologically inspired approach for modeling place appearance based on neural networks where "place nodes" encode local visual appearance. Konolige et al. [97] present a topological mapping technique where nodes consist of stereo keyframes and edges represent constraints from visual odometry and feature similarity. A vocabulary tree [129] is maintained that provides loop closure candidates later matched by the place recognition system involving a strong geometric check. This approach builds on a similar earlier work by Fraundorfer et al. [52].

Probabilistic approaches

As a departure from direct image comparison methods, the work by Cummins and Newman [32, 34] presents a probabilistic framework that learns a generative model for place appearance and then infers the likelihood that two observations arise from the same location model. The technique involves learning visual word correlations addressing scene variability and ameliorates perceptual aliasing by accounting for observations very common in the environment. The FAB-MAP formulation motivates this work and is summarized in the next section. A few works have proposed improvements over the original FAB-MAP system by use of better approximations and folding-in additional information. As an example, Maddern et. al. [113] incorporate vehicle odometry information and use a particle filter to track the vehicle over the continuous trajectory with an interpolated place appearance model. Mei et al. [116] eliminate arbitrary trajectory discretization to form places and instead reason directly on co-visible landmarks as observed by different views of a moving camera. They introduce the notion of dynamic bag-of-words based on finding cliques in the co-visibility graph.

Combining metric and appearance information

A few authors have explored combining both appearance and metric information for loop closure detection. Olson [131] presents an algorithm that uses both camera-derived and laser-derived features. Given a number of potential matches, the algorithm extracts the subset that is most self-consistent based on the spectral properties of the pair-wise compatibility matrix. Blanco et al. [13] introduce a method to reconstruct the robot's path in a hybrid discrete-continuous state space which naturally combines metric and topological maps. In [66], Gutmann and Konolige present

an incremental mapping scheme called Local Registration and Global Correlation (LRGC). The loop closure detection strategy involves integrating a set of local scans into a map patch and using this as a template to find matches in the old map. A set of heuristics are applied to verify potential loop closures. Goedeme et al. [61] present a different approach based on Dempster-Shafer theory where loop closure hypothesis are evaluated by combining evidence from neighbouring hypotheses.

Data association

Most existing approaches make data association decisions based on the maximum likelihood estimates, assigning the current observations to the most likely place already in the map or a new place as determined by the place recognition system. Such a hard assignment can cause potential failure when a sequence of observations later in time gives evidence for an alternate topology, especially when perceptual aliasing is high. Ranganathan et al. [150] address this issue by maintaining a distribution over the set of all possible topological maps resulting from associating observations with map locations. Approximate inference is carried out using a Markov Chain Monte Carlo (MCMC) sampling yielding topology estimates from the posterior distribution. In [146], the authors extend this work by presenting an online particle filter based algorithm. In a recent effort, [132] Olson et al. present an approach for maintaining multiple map hypothesis extensible to multi-robot scenarios.

2.1.3 Overview of FAB-MAP System

In this section we present an overview of FAB-MAP [34] a state-of-the-art loop closure detection system that forms the basis for FAB-MAP 3D. The central idea is to pose the place recognition problem probabilistically and compute how likely

it is that two observations came from the same location in the map or from an unexplored place instead of computing the observation match score directly. This involves learning a generative model of place appearance that accounts for imperfect sensors. Further, the formulation addresses perceptual aliasing in the environment and accounts for scene change across visits. For details please refer to [34, 33, 35]. Readers familiar with this work can proceed directly to subsequent sections.

Formulation

The world is modeled as a set of discrete locations where each location is described as a distribution over appearance words. Incoming sensory data is converted into a bag-of-words representation [173] and the likelihood that the current observation originated from a location in the robot's map or from a new place is estimated This yields a probability distribution over locations and is used to make a data association decision: initializing a new place in the topological map or updating the appearance model for an existing place.

Formally, scene images are represented as a collection of words from a vocabulary of size $|v|$. The observation at time k is denoted as $Z_k = \{z_1, \ldots, z_{|v|}\}$, where z_i is a binary variable indicating the presence (or absence) of the i^{th} word of the vocabulary and $\mathcal{Z}^k$ is used to denote the set of all observations up to time k. The features are based on SURF descriptors [7] and quantized to visual words [173] through an offline clustering phase form training set of images. At time k, the robot's map consists of n_k discrete and disjoint locations $\mathcal{L}^k = \{L_1, \ldots, L_{n_k}\}$. Each of these locations has an associated appearance model. Variables z_i represent the observed features. Hidden variables, e_i represent the event that an object which generates observations of type z_i exists at the location forming the location model as the

set $\{p(e_1 = 1|L_i), \ldots, p(e_{|v|} = 1|L_i)\}$. Locations independently generate words e_j. Words generate observations z_k which are interdependent. A detector model relates feature existence e_i to feature detection z_i and is specified as a false positive rate $p(z_i = 1|e_i = 0)$ and false negative rate $p(z_i = 0|e_i = 1)$.

Estimating Location

Calculating $p(L_i|\mathcal{Z}^k)$ is formulated as a recursive Bayes estimation problem. Let $p(L_i|\mathcal{Z}^{k-1})$ be the prior belief about the location, $p(Z_k|L_i, \mathcal{Z}^{k-1})$ represents the observation likelihood, and $p(Z_k|\mathcal{Z}^{k-1})$ forms the normalization term. Considering the current and past observations as independent, given the location, the observation likelihood can be expressed as:

$$p(L_i|\mathcal{Z}^k) = \frac{p(Z_k|L_i, \mathcal{Z}^{k-1})p(L_i|\mathcal{Z}^{k-1})}{p(Z_k|\mathcal{Z}^{k-1})} \tag{2.1}$$

Learning the higher-order conditional dependencies between appearance words is intractable and hence the joint distribution $p(Z_k|L_i)$ is approximated via the Chow Liu tree expansion, Equation 2.3. Here z_r is the root of the tree and z_{p_q} is the parent of z_q in the Chow Liu tree [26]. The observation factors $p(z_q|z_{p_q}, L_i)$ in Equation 2.3 can be further expanded as Equation 2.4.

$$p(Z_k|L_i) = p(z_n|z_1, z_2, ..., z_{n-1}, L_i)p(z_{n-1}|z_1, z_2, ..., z_{n-2}, L_i)...p(z_2|z_1, L_i)p(z_1|L_i) \tag{2.2}$$

$$p(Z_k|L_i) \approx p(z_r|L_i) \prod_{q=2}^{|v|} p(z_q|z_{p_q}, L_i) \tag{2.3}$$

$$p(z_q|z_{p_q}, L_i) = \sum_{s_{e_q} \in \{0,1\}} p(z_q|e_q = s_{e_q}, z_{p_q}) p(e_q = s_{e_q}|L_i) \tag{2.4}$$

The terms in the above equation can be estimated using the prior probabilities, conditionals and the detector model parameters [34]. The Chow Liu tree expansion approximates the joint distribution, $p(z_{1,}z_2 \dots z_{|v|})$ by the nearest tree-structured Bayesian network in the context of the Kullback-Leiber divergence. It allows each variable in the approximate distribution to be conditioned on at most one other variable hence forming a tree thereby capturing first-order conditional dependencies. Further Chow and Liu [26] presented a polynomial time algorithm to select the best such distribution. The algorithm proceeds by constructing the mutual information graph with n nodes (each representing the variable z_i) and $\frac{n(n-1)}{2}$ edges, where each edge (z_i, z_j) has weight equal to the mutual information $I(z_i, z_j)$ between variables i and j.

$$I(z_i, z_j) = \sum_{z_i \in \Omega, z_j \in \Omega} p(z_i, z_j) \log \frac{p(z_i, z_j)}{p(z_i)p(z_j)} \tag{2.5}$$

Mutual information between two random variables expresses the extent to which the knowledge of one conveys about the state of the other variable, Equation 2.5. It is zero if the variables are independent and positive otherwise. Computing the maximum weight spanning tree of this graph gives the optimal tree structured network. Additionally, the conditional probabilities required for inference can be obtained from word co-occurrences in training data. Figure 2.4 shows a partial visualization of the Chow Liu Tree. The system effectively learns the presence of objects and feature correlations in the data in an unsupervised manner.

The second key component is the denominator term $p(Z_k|\mathcal{Z}^{k-1})$ in Equation 2.1. To account for the possibility of the observation originating from an unmapped

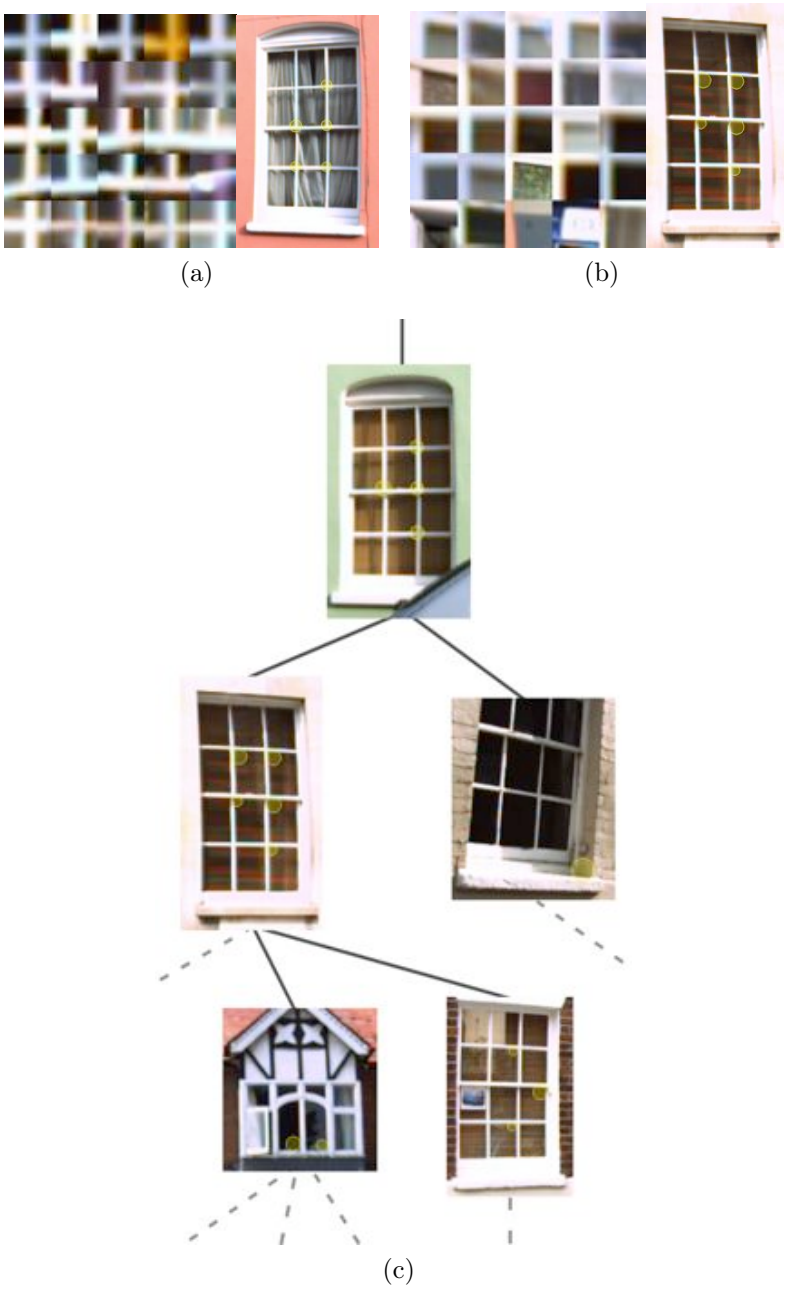


Figure 2.4: Learning with Chow Liu trees. (a) A sample word in the vocabulary. Typical image patches (left) and a representative interest point in an image (right). Features quantized to this word typically appear on top-left window corners. Its most correlated word appears in (b). Features quantized to the most correlated word typically appear on cross pieces of windows. (c) Visualization of a section of the Chow Liu tree. Each word in the tree is represented by a typical example. Images from [34].

place, the likelihood term is partitioned into the set of mapped and unmapped places where the set of unmapped places is approximated through sampling, Equation 2.6.

$$p(Z_k|\mathcal{Z}^{k-1}) \approx \sum_{m \in M} p(Z_k|L_m)p(L_m|\mathcal{Z}^{k-1}) + p(L_{new}|\mathcal{Z}^{k-1}) \sum_{u=1}^{n_s} \frac{p(Z_k|L_u)}{n_s} \tag{2.6}$$

Here, n_s is the number of samples used and $p(L_{new}|\mathcal{Z}^{k-1})$ is the prior probability of being at a new place. The samples approximating the distribution over natural images in the environment are obtained via random selection from a large collection of observations of the environment. This approximation to the partition function allows the system to cope with perceptual aliasing. The sampling set captures visually repetitive features like foliage, railings etc. Hence, the probability mass is split appropriately in perceptual aliasing cases thereby allowing loop closure detection only when the scene appearance is highly distinctive.

Effectiveness and Limitations

The FAB-MAP system has been demonstrated on several public image data sets. In [34], authors report results on the New College data set (testing robustness to perceptual aliasing) and the City Center data set (testing matching ability in presence of scene change). At 100% precision the system achieves 48% recall for the New College data set and 37% recall for City Center data set. In [33], the authors present an approximate inference technique based on the use of concentration inequalities that yields 25x-50x speed improvements for the core likelihood computation with minimal degradation in accuracy. A probabilistic bail out condition is derived for multi-hypothesis testing employing Bennett's inequality where the probability of an



(a) False negative example.



(b) False positive example.

Figure 2.5: FAB-MAP limitations. The recall rate is low, typically picking up 40% of the loop closures. An example of a false negative pair shown in (a). Additionally, there can be rare cases when an incorrect loop closure is declared. A false positive loop closure appears in (b). Images from [35].

erroneous result is a user defined parameter. In [35] authors presented a highly scalable formulation suitable for very large scale navigation based on inverted index and defining a sparse approximation to the FAB-MAP model and demonstrated reliable operation over a 1000km trajectory.

Although the FAB-MAP system has been remarkably successful on vast scales there is evident scope for improvement. One of the main limitations is low recall rate. The system typically picks 40% of the loop closures. Figure 2.5a shows an example of a loop closure pair not detected by FAB-MAP. Note that the system is designed to trade precision over recall since declaring a wrong loop closure can

cause catastrophic mapping failure. However, there are still some cases when an incorrect data association data occurs due to perceptual aliasing and major scene change. Figure 2.5b illustrates an example. One of the causes is the simplistic image representation used for inference. The perceived image is expressed as a set of visual words where the spatial arrangement information is lost. Further, only occurrence or absence of visual words is considered. Information in multiple occurrences of visual features, for example in repetitive structures like railings or windows is discarded. The insight that incorporating spatial information could improve place recognition performance forms the basis for FAB-MAP 3D which we introduce in Section 2.3.

Before we delve into the formulation, we first discuss relevant efforts from vision and robotics community in the next section that incorporate geometric or spatial information.

2.2 Related work

2.2.1 Use of Spatial Information in Computer Vision Tasks

Image retrieval systems attempt to locate all occurrences of an object in a large database given a representative query image in presence of variable illumination, partial occlusion, transformation and perspective changes. Present day large scale image retrieval systems rely on bag-of-words representation and vector space model of information-retrieval. The spatial layout of the visual words in the query image encodes the 2D spatial structure of the underlying objects populating the scene. Hence, state-of-the art systems attempt to incorporate the local spatial information (ordering or layout) and global relationships (epipolar geometry) to verify, disambiguate and re-rank retrieval results.

Sivic et al. [173] incorporate spatial information by requiring nearest-neighbour matches that satisfy (i) loose similarity requiring that neighbouring matches in the query region lie in a surrounding area in the retrieved images and (ii) strict similarity requiring same spatial layout (using affine transformation available from matched regions). Philbin et al. [139] use a spatial verification step that estimates a transformation between the query region and each target image based on how well feature locations are predicted by the estimated transformation. They experimented with a set of transformations including combinations of translation, scale, shear etc. and used shape information in affine-invariant image regions to generate hypotheses from corresponding features. Chum et al. [28] introduced query expansion that infers a generative model of the object from the query region. This is coupled with a fast and robust procedure to estimate an affine homography between the query and target image. In a recent work [138], Perdoch et al., proposed a memory-efficient representation of geometry associated with visual words. The local geometry is represented as ellipses and local affine frames. A compact representation is attained by discretization of similar ellipses to prototypes. The best prototypes are learnt from k-means clustering in the space of ellipses leading to the concept of geometric vocabularies. Another approach called geometric min-Hash [27], incorporates local geometry of interest points in a hashing based retrieval system.

Researchers studying object recognition have explored the use of spatial information in recognition. A part-and-structure model was introduced by [50] for describing visual objects and was used as a metric for image matching decisions. Burl et al. [22] introduced the constellation model that models an object class, like faces, as a set of parts arranged in a deformable spatial configuration. Here "parts" can be thought of as small image patches placed down at the appropriate positions. They present

a probabilistic framework and introduce a soft detection approximation to the joint distribution, combining both local photometry (part match) and the global geometry (shape likelihood). Fergus et al. [48] extended this model to provide a probabilistic representation for object shape, appearance, occlusion and relative scale where parameter learning is based on expectation-maximization in a maximum-likelihood setting. A similar approach was used in face recognition by Burl et al. in [23].

In [1], Agarwal and Roth introduced an object detection approach using a sparse part-based model for objects. The central idea is to learn a vocabulary of object parts from a set of image samples of the object class. Images are represented using parts from this vocabulary with their spatial configuration also encoded following which a classifier is learnt using this representation. This method is related to the work presented in this chapter. However, there are notable differences, particularly in the problem setup. The Agarwal and Roth approach is aimed at object detection for which a training image corpus is available. However, place recognition for outdoor robotic applications involves recognising the entire location (scene) perceived in the image collected online. This would require learning a 1-vs-all discriminative classifier for each new place encountered against all places currently in the map, with training time scaling linearly for each map insertion. Note that learning all possible object models populating a scene is computationally infeasible. Futher, the spatial relations are constrained to 2D image space which may not be invariant under large view point changes. A more accurate approach would model distances in a 3D Euclidean space.

2.2.2 Use of Geometric Information in Workspace Modeling

Serratosa et al. [159], surveyed several graph techniques for identifying objects in indoor scenes for mobile robots. Their approach builds on the notion of attribute

graphs and random graphs introduced by the structural pattern recognition community. An attribute graph is used to represent an observed pattern with vertices representing a pattern primitives (like structural components) and arcs, the relationships between them (like distances), for example planar graphs representing 2D views of a rigid 3D object [187].

To probabilistically model variability in the observed attribute graphs the notion of a random graphs were introduced by Wong et al. [203]. Nodes in a random graph represent probability distributions over the state of pattern primitives and arcs represent distributions over attribute relationships. Hence, attribute graphs can be understood as probabilistic outcomes of an underlying random graph (different 2D views of a 3D object). For practical applications, the joint probability distributions become intractable. Wong et al. used independence assumptions on vertices and arcs for computational feasibility [203]. Serratosa et al. extended this framework to function-described graphs incorporating higher-order relations between node pairs and present an approximate inference scheme [164]. In [159], this approach was demonstrated on a mobile robot equipped with a colour camera where the attribute graph was obtained via colour segmentation and creating adjacency links. Random graphs were learnt for simple objects like cups, boxes etc. from multiple 2D views of the object rotated about different axes.

In a recent work, Ranganathan et al. develop a place recognition technique for indoor robotic applications [147]. They present a 3D generative model for places via objects modeling their shape and appearance with position estimates from a stereo depth map. Individual object models are learned in a supervised manner from partially segmented and labeled training images and Markov Chain Monte Carlo (MCMC) approximate inference is used for data association between image

features and objects while inferring the place label.

2.2.3 Combining Visual and Range Data for Robot Mapping

A few efforts have explored fusing appearance information from cameras and depth information from laser scanners for robot mapping applications. Biber et al. [9] describe an approach to acquire and render realistic 3D models of indoor environments using a 2D laser scanner and a panoramic camera. A metric map is built using laser scan-matching based SLAM engine followed by extraction of walls, lines and ground plane from the laser point cloud. Texture is rendered on the walls by warping images from a cross-calibrated camera onto the extracted plane. Stamos et al. [176] present a similar approach that involves extracting laser-based features and registering image and range data for photorealistic 3D model rendering. Tapus et al. [180] represent places using multi-modal features termed "fingerprints" including colour bins from cameras and corners from laser scanner. They use Partially Observable Markov Decision Process (POMDP) for mapping and global localization. Miura et al. [122] generate probabilistic grid maps using an omnidirectional stereo camera coupled with a laser range finder, where the less-accurate stereo depth map is refined using range information from planar laser scans. Local grid maps are first built individually for each modality using temporal integration and later merged using a logical integration rule.

Recently, Kinect[1]-style RGB-D sensors have gained popularity predominantly for indoor operation. These cameras produce both a colour image and per-pixel depth estimates over a short range of a few meters. RGB-D cameras have been applied for dense indoor mapping, [76] and [86] are notable examples. The key idea is to

[1]http://en.wikipedia.org/wiki/Kinect

use visual features and associated depth values to obtain an initial alignment, and then jointly optimize over the sparse feature matches and dense point cloud alignment. For ensuring global map consistency a keyframe-based loop closure detection is used based on inlier detection using RANSAC coupled with vocabulary tree based matching as presented in [129]. However, the loop closure step relies only on visual feature matching without using the depth information.

We now conclude our survey of related works and in the next section present a topological mapping approach that models the spatial and visual appearance characteristics of scene appearance.

2.3 Topological Mapping with Spatial and Visual Appearance

The motivation for incorporating range information is two fold. Prior to this the work, the FAB-MAP framework only modeled the presence or absence of a word at a location and did not incorporate their spatial arrangement. As a result, the system assigned equal likelihoods to two places if the same visual words were seen in two places even if their spatial arrangement was different, Figure 2.6. Secondly, FAB-MAP discards the number of times a word appears in a scene which is important for scenes containing repetitive visual features as illustrated in Figure 2.7. This is addressed in FAB-MAP 3D because by using the range between occurrences of visual words we are implicitly counting word occurrence. Crucially, these inter-word distances are viewpoint invariant and collectively constitute strong place signatures, Figure 2.8.

The impact of using both spatial and visual appearance is marked. We provide

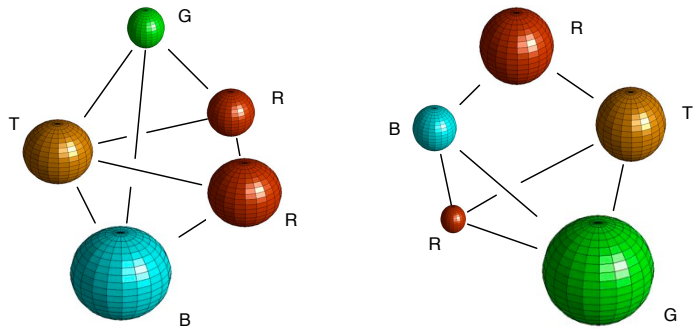
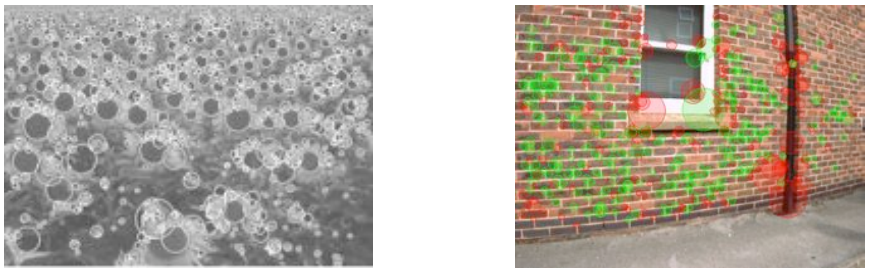


Figure 2.6: An illustrative example showing the significance of spatial information. The two scenes have the same visual words {**R**ed, **G**reen, **B**lue, **T**an} but different configurations (pairwise distances). The FAB-MAP framework considers both places to be the same. However, FAB-MAP 3D captures the spatial information through the random graph model and infers the places to be different. Note that sphere sizes differ due to perspective view.



(a) (b)

Figure 2.7: Multiple occurrences of visual words are common. (a) SURF features detected in a sunflower field and (b) brick walls in urban workspaces. FAB-MAP 3D attempts to capture the spatial distribution of visual words including the number of times they occur in a repetitive environment.



Figure 2.8: Why distances in 3D? Three images taken by a camera moving horizontally in front of a building with four corner features on the window in view marked with connecting edges. The imaged 2D distances between features vary across the image sequence. However, the underlying 3D graph structure (and hence inter-feature 3D distances) remain invariant under rigid transformations.

results illustrating an increase in precision-recall performance compared to a state-of-the-art visual appearance only systems in Section 2.4. In the next sub-section we discuss an illustrative example giving an intuition for the approach followed by the mathematical formulation of the algorithm.

2.3.1 Illustrative Example

In order to build an intuition for the probabilistic mapping framework and the random graph formulation, we discuss an illustrative toy example in this section. The precise mathematical formulation is presented in the next section.

In this simple example, the world is assumed to be composed of four types of visual words: *square*, *circle*, *triangle* and *diamond* as shown in Figure 2.9. A mobile robot observes the scene as a graph via noisy visual and range sensors. Note that the graph is assumed planar in this example. Visual features can be observed as present, absent or occur multiple times at varying distances as shown in Figure 2.9(left). The visual and spatial characteristics of the scene are encoded probabilistically as likelihoods over word occurrences and pair-wise distances, see Figure 2.9(right).

The likelihood over feature existence is certain for features detected in the scene. However, for the features not observed in the scene like the *diamond-type* features, a lower probability is assumed, accounting for a possible detection error by the visual sensor. The spatial appearance is characterized by maintaining a belief over inter-feature distances stored as a pair-wise matrix accounting for all feature pairs including multiple feature occurrences of the same type. As an example, Figure 2.10a highlights the bimodal distribution encoding the belief that *square-type* features are observed at short as well as long distances (with varying frequency) in the scene. Similarly, Figure 2.10b emphasizes the unimodal distribution between *circle* and *triangle* features representing the fact that they are observed to be the same distance in two occurrences. Further, for pair-wise features that are not observed to co-occur an uninformed prior distribution is assumed.

During the first traversal, the robot observes each location as a scene graph and encodes it probabilistically as a random graph using the procedure outlined above. At a loop closure event when the robot re-visits a place, it observes a similar scene graph with minor changes due to possible sensor errors, scene change, illumination or viewpoint differences causing false positives, missing features as well as features appearing at different distances as shown in Figure 2.11.

The loop closure probability is estimated by calculating the likelihood of the random graph model generating the perceived features including the observed relative distances. Since a probabilistic model explains the visual and spatial appearance for a location, a loop closure event can be detected even with scene change and noisy sensor measurements, providing more robustness compared to a simplistic approach that directly computes the matching similarity between the observed graphs. Note that we also need to handle the case when highly repetitive scenes like brick walls

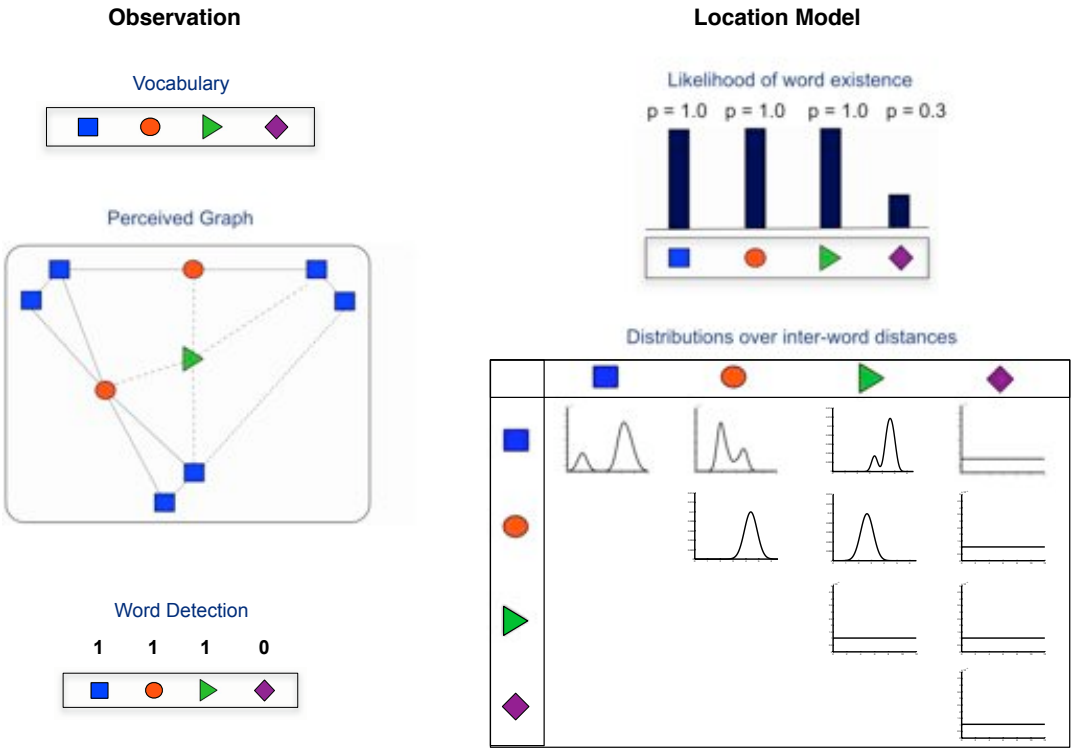
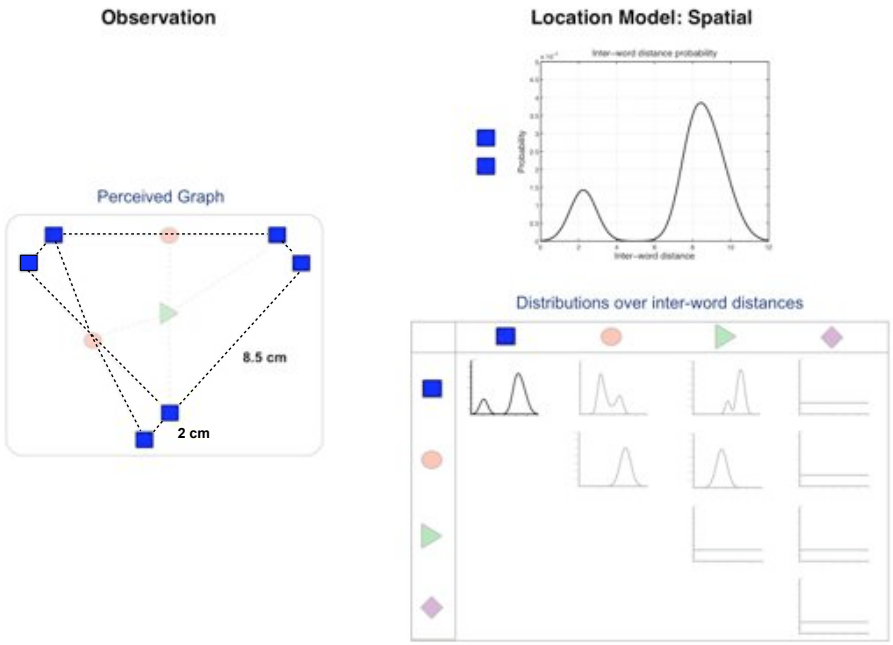
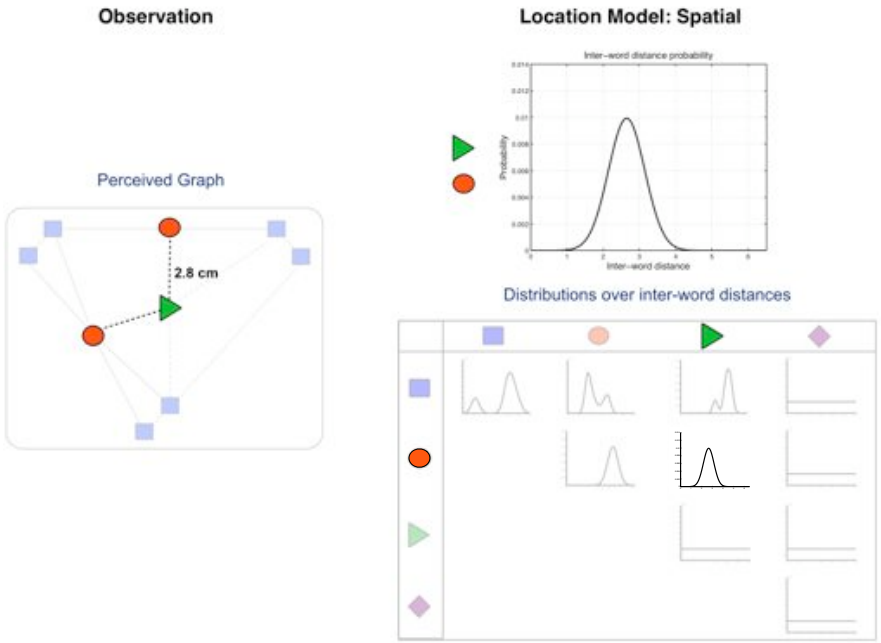


Figure 2.9: Illustrative toy example. The visual world is assumed to consist of four feature types: *square*, *circle*, *triangle* and *diamond*. The robot perceives the scene as a graph of visual features through noisy sensors (left) and encodes it probabilistically as a random graph (right). The visual appearance is captured as likelihoods over word existence and spatial appearance is characterized by belief over pair-wise distances stored as a matrix. The graph is assumed planar in this example.



(a) For the example scene, the *square-type* features are observed at short and long distances (with varying frequency). Hence, a bimodal distribution represents the belief over inter-word distances.



(b) For the example scene, a unimodal distribution represents the belief that the *circle* and *triangle* features were observed to be the same distance in two observed occurrences.

Figure 2.10: Illustrative example highlighting the type of distributions over feature distances encoded by the model.

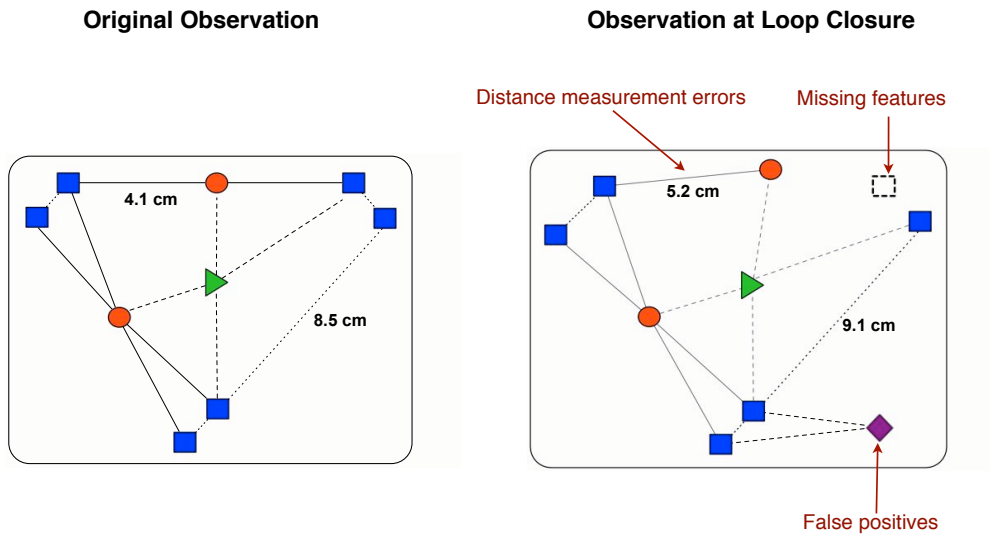


Figure 2.11: Illustrative example. Graph observed during the first visit (left) and at loop closure event (right). Note that the graph observed during the re-visit is similar to the first observation but can possess differences like false positives, missing features or different range measurements due to sensor errors or scene change.

can cause a false loop closure to be declared. Hence, in addition to encoding scenes observed during the mapping process, we also model the *average* place which allows us to determine if the observation came from a place in the map or from a new place in the environment. Next, we formalize the intuition developed with this example and present the algorithm in the following section.

2.3.2 Probabilistic Model for Locations

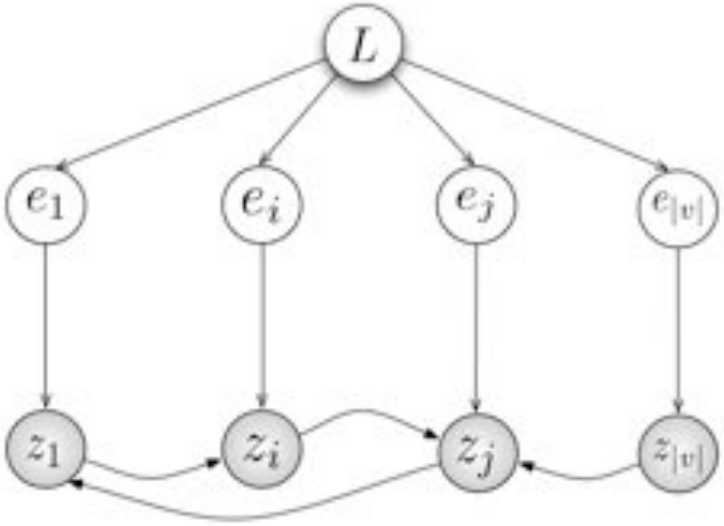
The world is modeled as a set of independent and disjoint locations. A mobile robot collects image and range observations of the environment and computes the probability that the observation comes from a known location in the topological map or from a new place. The observed scene is described by salient 2D visual features detected in the captured image quantized to visual words learnt offline [173]. We

assume that the vehicle is equipped with a range measuring sensor, e.g., a laser range finder or a stereo camera. The range sensor assigns 3D positions (vehicle-relative) to the visual features detected in the image forming a 3D graph of visual features characterizing the location. Note that we do not require large scale 3D reconstruction but simply a way to calculate the range between visual words.

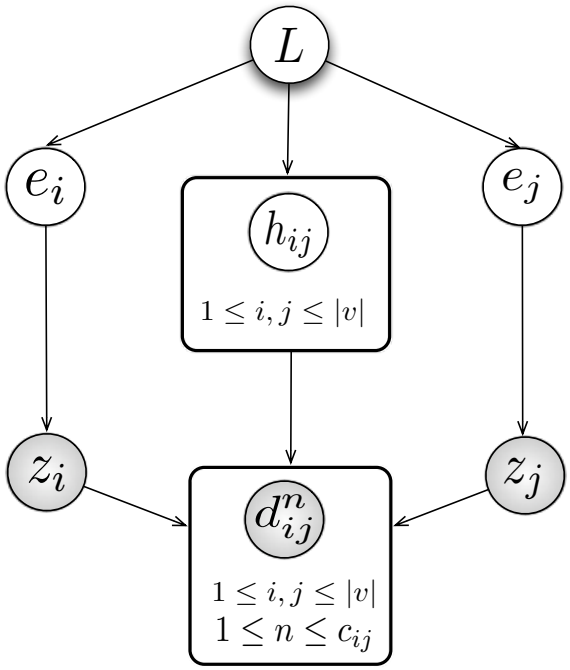
Random Graph Representation

Each location is modeled as a random graph with vertices and edges capturing the visual and spatial appearance respectively. Vertices represent likelihoods over latent scene elements that generate observed features and edges capture the belief over euclidean distances between scene elements that give rise to distance measurements in the observed scene. Formally, let $L_k = \{E_k, H_k\}$ represent the location model where E_k represents a random vertex set $\{e_i | 1 \leq i \leq |v|\}$, binary variable e_i is the event that the i^{th} word exists at the location[2]. The visual appearance of a place is characterized by the set $\big\{p(e_1 = 1|L_k), \ldots, p(e_{|v|} = 1|L_k)\big\}$, an estimate of the probability that each word exists at the location, Equation 2.7. The random arc set, H_k captures the spatial appearance and consists of variables, $\{h_{ij} | 1 \leq i, j \leq |v|\}$ where each h_{ij} maintains a discrete probability distribution (histogram) over euclidean distances in 3D space between the i^{th} and the j^{th} visual word in the vocabulary, Equation 2.8. Collectively, these distributions represent the belief over distances between all pairwise feature generating latent scene elements (including multiple occurrences) at each location.

[2]We ask for the reader's forbearance for the counter-intuitive naming of the vertex set as E. This notation is used for consistency with previous FAB-MAP papers.



(a) Visual component.



(b) Spatial component.

Figure 2.12: Generative Model: Locations independently generate object features, e_i which produce observations, z_i detected by the visual sensor (top). First order correlations exist for word observations. Additionally, locations possess distributions over word pair distances, h_{ij} which give rise to observed distances conditioned on the observations z_i and z_j of each word pair (bottom). The model includes distance observations from multiple occurrences of a word.

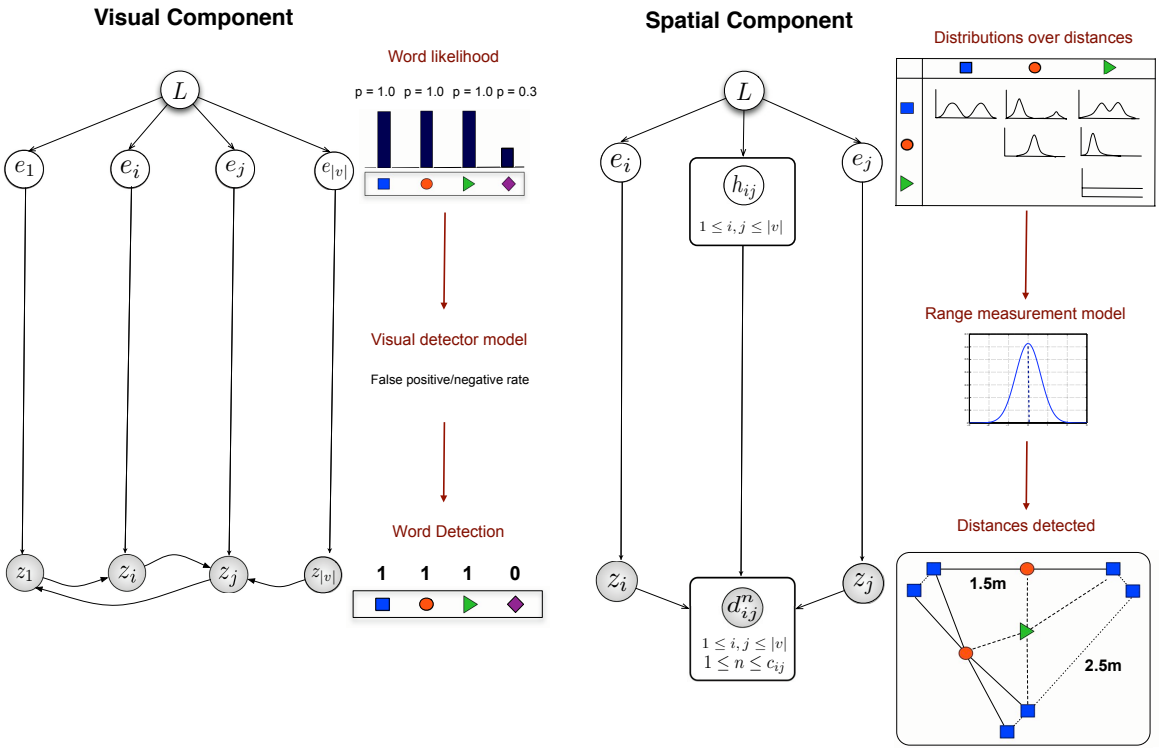


Figure 2.13: Graphical model juxtaposed with the toy example. The figure illustrates the latent distributions over feature existences and word distances generating the observed graph (observed features and distances) via noisy range and visual sensor models.

$$L_k^{visual}: \quad E_k = \{p(e_1 = 1|L_k), p(e_2 = 1|L_k) \ldots, p(e_{|v|} = 1|L_k)\} \tag{2.7}$$

$$L_k^{spatial}: \quad H_k = \{p(h_{ij} = r|L_k)|1 \leq i, j \leq |v|, 0 \leq r \leq R_{max}\} \tag{2.8}$$

$$H_k = \begin{bmatrix} h_{1,1} & h_{1,2} & \ldots & \ldots & h_{1,|v|} \\ h_{2,1} & h_{2,2} & & & \vdots \\ \vdots & & h_{i,j} & & \vdots \\ \vdots & & & \ddots & \vdots \\ h_{|v|,1} & \ldots & \ldots & \ldots & h_{|v|,|v|} \end{bmatrix}$$

Observations

An observation of a local scene is represented by a graph, $G_k = \{Z_k, D_k\}$ consisting of observed features including their relative distances. Here, Z_k represents the vector $\{z_1, \ldots, z_{|v|}\}$ where each z_i is a binary variable indicating the presence (or absence) of the i^{th} visual word in the scene. The set of observed spatial distances between visual features is denoted by D_k and also includes the distances perceived between words of the same type. Let c_{ij} be the count of all pairwise distances observed between the i^{th} and the j^{th} word. Note that c_{ij} exceeds one when either i^{th} and/or the j^{th} word occurs multiple times and c_{ij} is zero when either word is not observed. Formally, the set of observed spatial distances, D_k can be represented as $\{d_{ij}^n|1 \leq i, j \leq |v|, 1 \leq n \leq c_{ij}\}$.

$$G_k^{visual}: \quad Z_k = \{z_1, z_2 \ldots, z_{|v|}\} \tag{2.9}$$

$$G_k^{spatial}: \quad D_k = \{d_{ij}^n | 1 \leq i, j \leq |v|, 1 \leq n \leq c_{ij}\} \tag{2.10}$$

$$D_k = \begin{bmatrix} d_{1,1}^{c11} & d_{1,2}^{c12} & \ldots & \ldots & d_{1,|v|}^{c_{1,|v|}} \\ d_{1,2}^{c12} & d_{1,2}^{c12} & & & \vdots \\ \vdots & & d_{i,j}^{c_{ij}} & & \vdots \\ \vdots & & & \ddots & \vdots \\ d_{|v|,1}^{c_{|v|,1}} & \ldots & \ldots & \ldots & d_{|v|,|v|}^{c_{|v|,|v|}} \end{bmatrix}$$

Range and Visual Detector Models

Figure 2.12 presents the generative model for the spatial and visual appearance of locations. A location, L independently generates objects features, e_i which produce observations, z_i detected by the visual sensor. A detector model, Det_{visual} for the appearance sensor connects hidden variables for feature existence to the observed variables representing feature detection, Equation 2.11.

$$Det_{visual} \begin{cases} p(z_i = 1 | e_i = 0), & \text{false positive rate.} \\ p(z_i = 0 | e_i = 1), & \text{false negative rate.} \end{cases} \tag{2.11}$$

In addition, each location possesses distributions over word pairs, h_{ij} which give rise to measured distances, d_{ij}. The uncertainty in the range measuring process is modeled as a Gaussian conditional density, Det_{range} centered on the discrete ranges of h_{ij} and parameterized by variance, σ_{range}, Equation 2.12.

Note that more sophisticated range detector models can be employed based on the likelihood fields of the underlying range sensor. For example, [162] provides a detailed exposition on accurate models for the laser range finder incorporating local measurement noise, unexpected objects, random measurements and failures. This can be used to build more accurate error models for word distances using range sensors. Similarly, accurate models based on physical characteristics are possible in case distances are obtained using depth from stereo cameras.

$$Det_{range} = \quad p(d_{ij}|h_{ij} = x) \sim \quad N(x, \sigma_{range}) \tag{2.12}$$

Figure 2.13 displays the graphical model juxtaposed with the toy example discussed in section 2.3.1. The figure illustrates the observed scene graph (visual features and 3D distances) as explained by latent distributions over feature existence and word distances connected via visual and range sensor models.

Discussion: Generative Model

The random graph location model provides a natural way to capture *both* the visual and spatial appearance of a place. The visual component consists of probabilities that feature generating words exist at the location. The spatial relationships are modeled by the likelihoods over possible distances between any two words. The visual and range detector models explain the noisy way in sensors perceive the world connecting likelihoods with observations in the generative model.

A key aspect is that inter-word 3D distances between visual features (instead of absolute positions in Euclidean space) constitute spatial appearance. These pro-

vide viewpoint invariance since inter-word distances for objects in a scene remain approximately same when a robot revisits a place. Finally, the model incorporates spatial information from multiple occurrences of a word by maintaining distributions over distances between words of the same type. At first glance, it may appear that maintaining likelihoods over pair-wise distances at each location may be infeasible due to space complexity. However, as we will demonstrate in the results section, the matrix is very sparse, only a small number of features are observed to co-occur, making the approach tractable. Note that currently multiple occurrences are not utilized in visual appearance model where inference is based only on the presence or absence of a visual word.

2.3.3 Learning Distributions over Feature Distances

The spatial appearance of a place is characterized by probability distributions over word distances. We attempt to learn informative priors for these distributions from observed training data. Two classes of techniques can be employed: *parametric* and *non-parametric*. In *parametric* methods, a specific functional form (governed by a finite number of parameters) is assumed for the underlying distribution and the parameter values are determined from the data set. A limitation of this approach is that the chosen density might be a poor model of the distribution that generates the observed data resulting in inferior predictive performance [12]. Visual words observed by outdoor mobile robots can appear at highly varied distances, e.g., features detected on foliage are commonly seen at short distances and features on repeated structures like brick walls or railings can display large variations exhibiting multi-modal behaviour. To represent these complex multi-modal distributions, without assuming a particular functional form, we adopt a *non-parametric* density

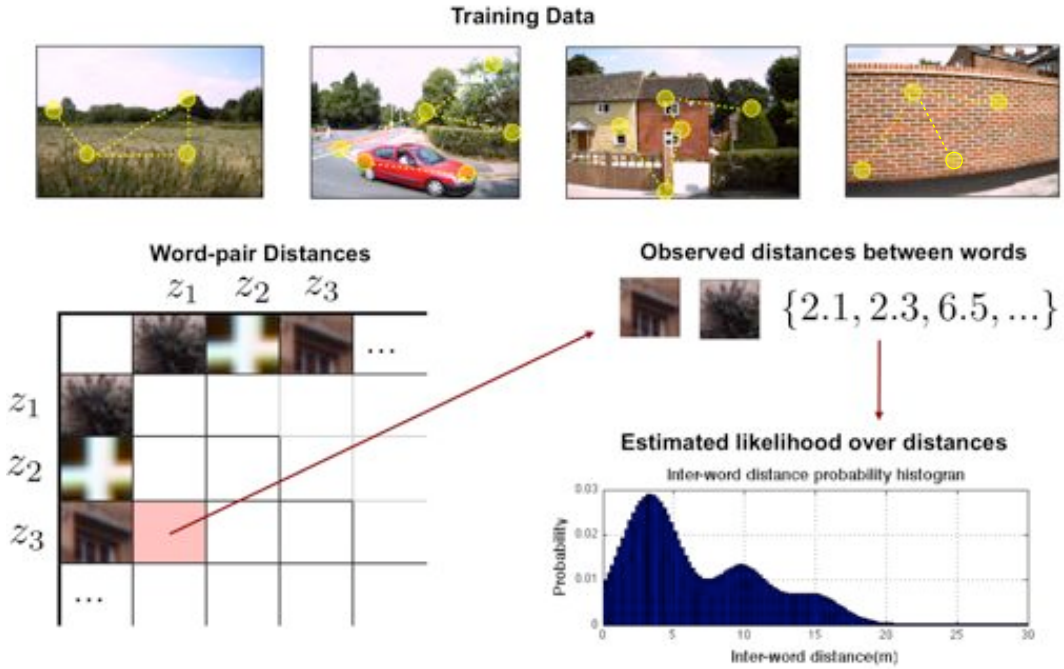


Figure 2.14: Learning distributions over inter-feature distances offline from training data. All observed distances between word-pairs (shown as a matrix) including multiple occurrences of the same word are recorded. A distribution over distances is learnt through kernel density estimation. In case a word pair is not observed to co-occur, an uninformed prior is assumed. Although, the size of the word-pair matrix is large (square of the vocabulary size), the matrix is highly sparse since only a small number of words co-occur or appear multiple times, making the approach tractable in practice.

representation.

Histogram Approximation to Continuous Distributions

Formally, let L_{max} be the maximum distance between any two words in the scene. This is typically the maximum range of a distance measuring sensor. The continuous range is sub-divided into bins b_k each of length Δ. Let R be the total number of bins in each histogram ($R = \frac{L_{max}}{\Delta}$) . Let $p(h_{ij} = b_k)$ represent the cumulative density in

bin, b_k, where $1 \leq k \leq R$, Equation 2.13.

$$p(h_{ij} = b_k) \quad = \quad p((k-1)\Delta \leq h_{ij} \leq k\Delta) \tag{2.13}$$

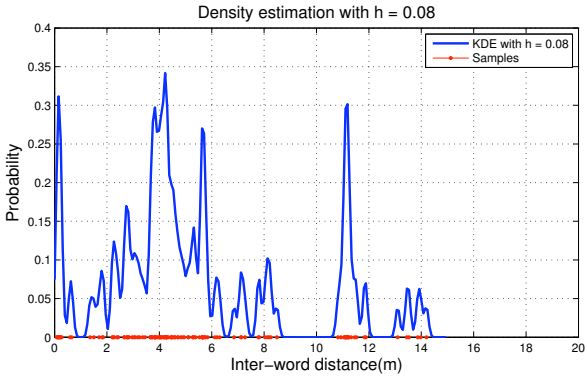
Note that continuous distance values are discretized as bins and the probability histogram represents a piecewise constant approximations to the continuous distribution [162].

Learning via Kernel Density Estimation

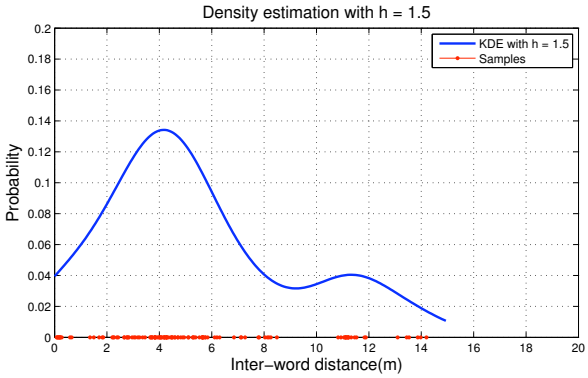
Given finitely many inter-word distance samples from training data, we estimate probability distributions using Kernel Density Estimation (KDE) [90] further discretized as a histogram. Probability density at a point x is estimated through a linear combination of kernel functions centered on training data $\{x_i\}_{i=1...N}$, where samples $\{x_i\}$ are assumed independent and identically distributed (i.i.d.) according to an underlying distribution using Equation 2.14. Here, the kernel $K(u)$ satisfies the conditions $K(u) \geq 0$ and $\int K(u)du = 1$. The most widely used kernel is the Gaussian of zero mean and unit variance for which the KDE can be written as Equation 2.15.

$$\hat{p}(x) \quad = \quad \frac{1}{Nh} \sum_{i=1}^{N} K(\frac{x - x_i}{h}) \tag{2.14}$$

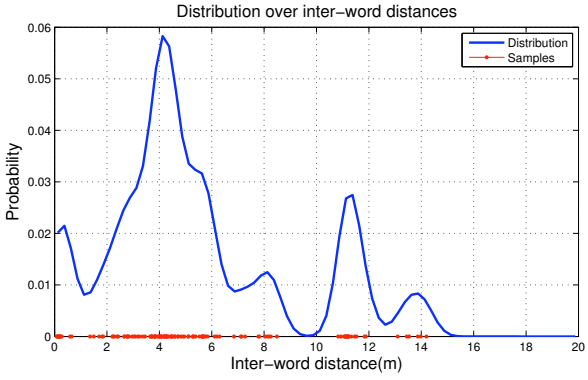
$$\hat{p}(x) \quad = \quad \frac{1}{N\sqrt{2\pi}h} \sum_{i=1}^{N} \exp -(\frac{(x - x_i)^2}{2h^2}) \tag{2.15}$$



(a) $h = 0.08$



(b) $h = 1.5$



(c) $h = 0.4$

Figure 2.15: Effect of kernel bandwidth on density estimates. The figure shows kernel density estimates (curve along y-axis) for observed data (red points along x-axis) with varying kernel bandwidth, h indicated with each graph. (a) Kernel width too small leads to over fitting leading to a very noisy estimate. (b) Bandwidth too large leads to over-smoothening leading to a coarse bimodal estimate. (c) An intermediate bandwidth leads to the optimal density estimate.

The kernel function is characterized by a bandwidth (h) that determines the accuracy of the model. Kernels too narrow lead to over-fitting and very wide bandwidths lead to under-fitting [90]. A number of techniques have been proposed for data-driven bandwidth selection. These methods minimize the *asymptotic mean integrated square error* (AMISE) between the estimate, $\hat{p}(x)$ and the actual density, $p(x)$. The most successful methods rely on estimation of *density derivative functionals* through the *solve-the-equation plug-in method* [152]. We used an implementation of an efficient $\epsilon - exact$ approximation algorithm for optimal bandwidth estimation based on the *Improved Fast Gaussian Transform* (IFGT) [204]. The algorithm has computational complexity linear in the number of training points. Once the optimal bandwidth is estimated Equation 2.15 is used for estimating the histogram probabilities.

Smoothing Density Estimates

Occasionally, due to limited training data there are few range samples for rare word pairs. Such sampling error can cause probability estimates for some histogram bins to take degenerate values of 0 or 1. To mitigate this effect, the maximum likelihood probability estimates must be smoothed and then re-normalized. A number of smoothing techniques exist [36], typically of the form as Equation 2.16. We assume uniform prior, i.e., $p_{prior} = \frac{1}{L_{max}}$ and set $K = \sqrt{N}$, where N is the number of training samples. In case no training samples are seen, each bin is assigned a flat prior.

$$p_{smooth} = (\frac{N}{N+K})p_{mle} + (\frac{K}{N+K})p_{prior} \tag{2.16}$$

There is scope for further improvements in inter-word density modeling. As distances are non-negative scalar quantities, kernels with positive support like gamma kernels can be used for density estimation, where mixture parameters can learnt through expectation maximization [25]. Alternatively, a logistic transformation could be used to constrain the density estimates to positive values bounded by maximum range, L_{max}. For simplicity, in this work, the probability mass assigned to infeasible values was ignored during histogram approximation to the density estimate.

Further, density estimation can be done in an incremental fashion where new data is incorporated in the model as it becomes available. In [68], authors present an efficient recursive density approximation that relies on propagation of density modes. A probability density is represented as a weighted sum of Gaussians whose number, weights, means and covariances are updated incrementally with new data. Starting from the previous modes of the density, a variable-bandwidth mean shift is employed to detect the new modes. This incremental approach has been applied to image processing tasks like background modeling [68] and object tracking [69].

2.3.4 Probabilistic Navigation and Mapping

Estimating Location

At time k, the workspace is modeled as a collection of n_k discrete and disjoint locations $\mathcal{L}^k = \{L_1, \ldots, L_{n_k}\}$. Given the set of observations (perceived graphs) till time k denoted as $\mathcal{G}^k$ graphs, we estimate the likelihood of the robot being at each location L_n (represented probabilistically as a random graph). Calculating $p(L_n|\mathcal{G}^k)$, the likelihood over locations can be posed as a recursive Bayes estimation problem:

$$p(L_n|\mathcal{G}^k) = \frac{p(G_k|L_n, \mathcal{G}^{k-1})p(L_n|\mathcal{G}^{k-1})}{p(G_k|\mathcal{G}^{k-1})} \tag{2.17}$$

Here, $p(L_n|\mathcal{G}^{k-1})$ is the prior estimate of the robot's location, $p(G_k|L_n, \mathcal{G}^{k-1})$ represents the observation likelihood and the term $p(G_k|\mathcal{G}^{k-1})$ is the total likelihood normalizing the distribution. Observations are assumed to be conditionally independent given location. Thus, $p(G_k|L_i, \mathcal{G}^{k-1})$ is approximated as $p(G_k|L_i)$. The likelihood that the observed graph was generated by location L_n is factored as two terms: (i) $p(Z_{k,}|L_n)$, likelihood of the visual appearance given location and (ii) $p(D_k|Z_{k,}L_n)$, likelihood of the observed spatial distances conditioned on visual observations and location.

$$\begin{aligned} p(G_k|L_n, \mathcal{G}^{k-1}) &\approx p(G_k|L_n) \\ &= p(\{Z_{k,}D_k\}|L_n) \\ &= p(D_k|Z_{k,}L_n)p(Z_{k,}|L_n) \end{aligned} \tag{2.18}$$

The visual appearance likelihood term is expanded using the Chow-Liu approximation [26], Equation 2.3. As reviewed earlier in Section 2.1.3, this expansion approximates the discrete joint distribution $p(z_1, z_2, \ldots, z_{|v|})$ by the closest tree-structured Bayesian network according to the Kullback-Leiber (KL) divergence criteria.

Conditioned on the visual observation and location, the spatial likelihood term is estimated as Equation 2.19. Pairwise distance edges in the graph are considered independent of other edges given observations of their end points. The likelihood of observing a pairwise distance $p(d^n_{ij}|z_i, z_j, L_n)$ is factored in terms of the prior belief

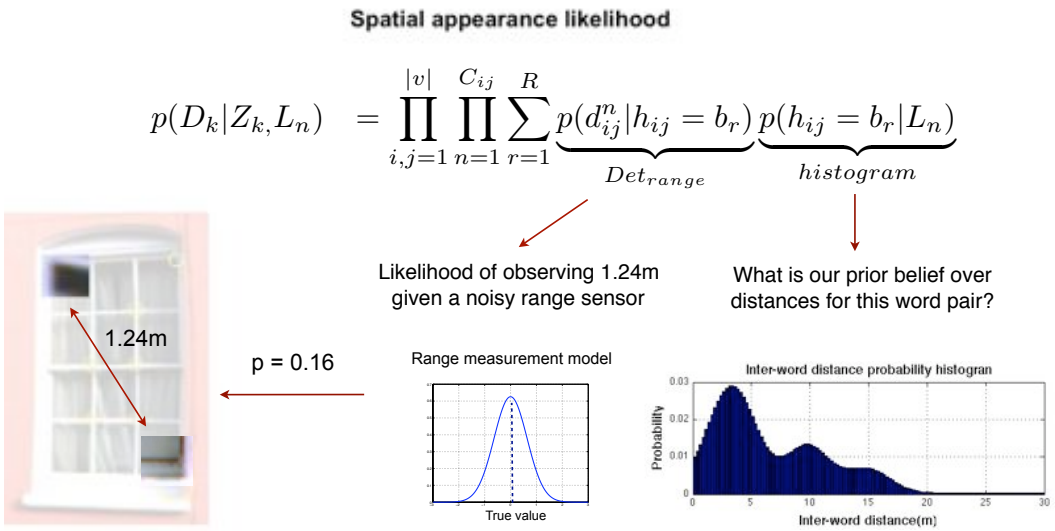


Figure 2.16: A diagrammatic representation for calculating spatial appearance likelihood term for a location conditioned on the observed features. As an example, the likelihood of the observed distance between two corner features appearing on windows is determined by combining the prior distribution over inter-word distances learnt via KDE during training with the Gaussian range sensor model and marginalizing over bins for the histogram. Note that all word-pairs including multiple occurrences are accounted independently.

over the distance $p(h_{ij} = b_r|L_n)$ from histogram, h_{ij} and the probability of observing the distance given belief, $p(d_{ij}^n|h_{ij} = b_r)$ via the range detector model. The likelihood is obtained by marginalizing over the discrete range estimates, b_k from histogram, h_{ij}. The range detector model is assumed independent of location.

$$p(D_k|Z_k,L_n) = \prod_{i,j=1}^{|v|} \prod_{n=1}^{|C_{ij}|} p(d_{ij}^n|z_i,z_j,L_n) \tag{2.19}$$

$$p(D_k|Z_k,L_n) = \prod_{i,j=1}^{|v|} \prod_{n=1}^{|C_{ij}|} \sum_{r=1}^{R} \underbrace{p(d_{ij}^n|h_{ij}=b_r)}_{Det_{range}} \underbrace{p(h_{ij}=b_r|L_n)}_{histogram} \tag{2.20}$$

Evaluating the Normalization Term

The normalization term $p(G_k|\mathcal{G}^{k-1})$ represents the total likelihood of the observation G_k. An observation can come from the set of locations currently in the robot's map (M) as well as the set of all previously unknown locations ($\overline{M}$). Hence, the term $p(G_k|\mathcal{G}^{k-1})$ can be expressed as:

$$p(G_k|\mathcal{G}^{k-1}) = \sum_{m\in M} p(G_k|L_m)p(L_m|\mathcal{G}^{k-1}) + \sum_{u\in\overline{M}} p(G_k|L_u)p(L_u|\mathcal{G}^{k-1}) \quad (2.21)$$

The second term involves summation over all unmapped places and cannot be directly computed. The summation is approximated through mean field approximation[3] [91] by constructing an *average place* model, $L_{avg} = (E_{avg,}H_{avg})$.

$$p(G_k|\mathcal{G}^{k-1}) \approx \sum_{m\in M} p(G_k|L_m)p(L_m|\mathcal{G}^{k-1}) + p(G_k|L_{avg}) \sum_{u\in\overline{M}} p(L_u|\mathcal{G}^{k-1}) \quad (2.22)$$

The visual appearance component of the average place, E_{avg} is constructed by assigning e_i variables their marginal probabilities from training data. In similar vein, the spatial appearance of the average place, H_{avg} is the set of marginal histograms for each word pair, where each histogram is a density estimate learnt from all pairwise distance samples observed in training data.

This formulation also addresses the perceptual aliasing problem: the fact that different parts of the environment appear the same to robot's sensors. e.g., similar

[3]As described in [34], a superior alternative to the *mean-field* approximation is the *sampling* based approximation. The current implementation uses the *mean-field* approach due to limited training data. Efforts are underway to collect a larger data set using the recently acquired platforms in the group.

looking foliage and brick walls appear commonly while navigating outdoors. The visual appearance model for the average place learns which features are common in the environment, like words appearing on foliage have high marginal probabilities. Additionally, the spatial appearance model for the average place learns what distances words commonly appear at. Hence, it can learn that features detected on brick walls commonly appear at repeated distances. Overall, the system matches an observation to a location only when *both* visual and spatial appearance is distinctive.

Location Prior

The location estimation Equation 2.17 requires the prior probability of being at a particular location, $p(L_n|\mathcal{G}^{k-1})$. This can be obtained by applying a motion model to the current estimate of the robot's location in the topological map [34]. If the robot is at place i at time t, it is likely to be at one of the places $\{i-1, i, i+1\}$ at time $t+1$, with equal probability. With an explicit topological map this can be generalized to a multi-way probability. As an example, if the present location has three neighbours in the topological map then the probability mass is spread equally for the four locations. This holds when the current location in the map has known neighbours. For a node with unknown neighbours, like the case of a newly added place in the map, part of the probability mass is assigned to the new place and the rest is distributed uniformly over the remaining places in the map. The partition is decided by the probability that a topological link with unknown end point leads to a new place which is a user defined parameter. In case, such assumptions on the motion model are not valid, the prior can be left uniform. As noted in [34], the effect of the location prior is weak and does not significantly affect performance. However, a uniform prior can lead to an increase in the false positive rate. There

is scope for improving the location prior using some metric information e.g., from a visual odometry engine [125]. A better location prior can also reduce the likelihood of a wrong data association decision that can adversely effect performance.

Updating Location Model

A new location in the topological map is initialized with the average place model learnt from training data where word generators exist with marginal probability and word-pair histograms are initialized to prior distributions using Equations 2.23 and 2.24.

$$p(e_i = 1|L_{new}) = p(e_i = 1) \tag{2.23}$$

$$p(h_{ij} = b_r|L_{new}) = p(h_{ij} = b_r) \tag{2.24}$$

When an observed graph relates to a location in the map, the random graph model for the location is updated using current observations, according to prior belief and the sensor models. For the visual component, the probability of feature existence, $p(e_i = 1|L_n)$ is updated as Equation 2.25. Additionally, the observed word-pair distances are used to obtain the posterior distributions over word distances. The probability estimates for each histogram bin, $p(h_{ij} = b_r|L_n)$ is updated as Equation 2.26. We assume that observations of a word or an observed distance between a word-pair does not convey information about either existence or pairwise distances of other words.

$$p(e_i = 1|L_n, \mathcal{G}^k) = \frac{p(z_i|e_i = 1)p(e_i = 1|L_n, \mathcal{G}^{k-1})}{p(z_i|L_n)} \tag{2.25}$$

$$p(h_{ij} = b_r|L_n, \mathcal{G}^k) = \frac{p(d^n_{ij}|b_r)p(h_{ij} = b_r|L_n, \mathcal{G}^{k-1})}{p(d^n_{ij}|L_n)} \tag{2.26}$$

The data association decision for observations and locations is based on maximum likelihood criterion. Loop closures are accepted only when the loop closure probability exceeds a user defined threshold, $p_{accept} = 0.999$ as suggested in [34]. There is scope for improving the data association using multiple hypothesis approaches as proposed in [132], [150] and [149].

Smoothing Location Likelihoods

A smoothing step is applied to the likelihood estimates according to Equation 2.27 where σ is the smoothening parameter and n_k is the number of places currently in the map. Likelihood smoothening prevents the system from asserting a loop closure based on a single observation. As noted in [34], the inference engine is highly dependent on the estimation of the normalization term $p(G_k|\mathcal{G}^{k-1})$ which encompasses the observation likelihood from a new place. Ideally, this term should fully represent the distribution over the set of *all* observations of the environment and the *mean-field* approach or the *sampling* approach should capture the visual and spatial richness of the workspace. However, due to computational limitations and finite amount of data available we can only approximate this term and there can be repetitive environmental features that the system does not capture. Hence, there is a possibility that the system declares a loop closure when similarity is actually due to perceptual aliasing arising in the visual or spatial component.

$$p(G_k|L_n) \leftarrow \sigma p(G_k|L_n) + \frac{(1-\sigma)}{n_k} \tag{2.27}$$

In [34], authors present an example where FAB-MAP declares an incorrect loop closures with two images of a railing. This occurs due to the fact that there were no examples of railings in the sampling set and the system does not know that the large number of features generated on railings are actually correlated and common. A similar behaviour can take place for the spatial component in FAB-MAP 3D. For example, there can be some common foliage features that do not co-occur in the training set and hence the system assumes a flat prior for its inter-word distribution for the *average* place model. During inference, the system can incorrectly assert high spatial similarity for these features in different parts of the workspace with a very little probability mass being assigned to the new place.

2.3.5 Accelerating Graph Likelihood Computation

While estimating location via the recursive Bayes update as Equation 2.17, the per-scene computation complexity is largely governed by (i) the graph likelihood or spatial likelihood term $p(D_k|Z_{k,}L_n)$ from Equation 2.19 and (ii) corresponding posterior histogram density updates $p(h_{ij} = b_r|L_n, \mathcal{G}^k)$ from Equation 2.26. Both computations depend on the number of pair-wise distance histograms from the underlying random graph model that are included for estimating these terms. Here, we investigate a graph tessellation approach that selects a linear sub-set from a quadratic number of pair-wise distances and hence reduces the number of histogram updates involved. In Section 2.4.3, we experimentally verify that the approximation

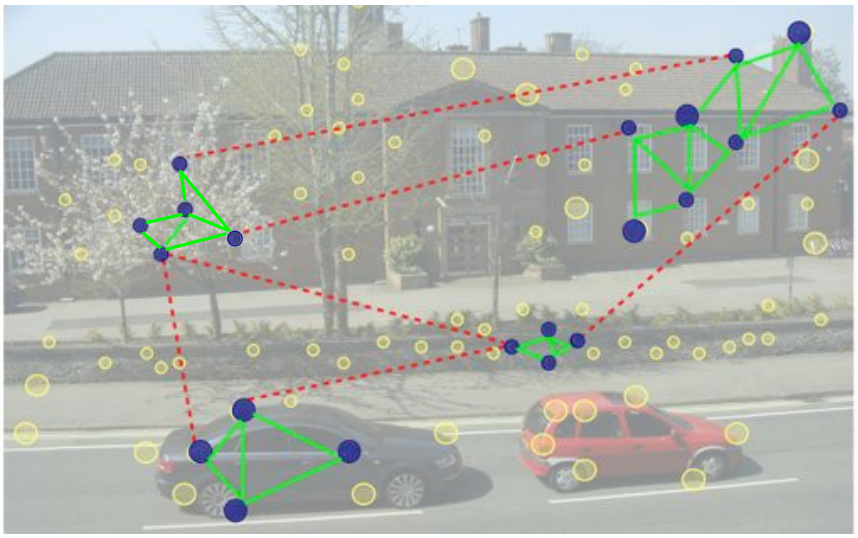


Figure 2.17: Intuition for using Delaunay tessellation for the observed graph. Visual features (shown as circles) emanate from underlying objects in the scene and hence possess local spatial correlations. Features separated by large distances arising from different objects are less likely to be correlated. Hence, we consider distances along the triangulation edges connecting neighbouring points (green edges) and do not consider the long range edges (red dashed edges). The triangulation is obtained via Delaunay tessellation that prevents skewed triangles, in effect connecting local neighbours.

results in minimal loss in performance.

Consider an observed scene graph where N_f represents the number of visual words detected. Considering all pair-wise distance edges in the observed 3D graph results in $\frac{N_f(N_f-1)}{2} \approx O(N_f^2)$ individual histogram updates while calculating the spatial likelihood or posterior histogram probabilities. However, note that visual features perceived in a workspace originate from objects that populate the scene. Hence, features generally possess high local spatial correlations for example features detected on a window. Similarly, features separated by large distances are spatially less correlated. Using this intuition, we would like to consider distances only to *neighbouring* points, where by *neighbouring* we imply a pair of points whose cells in the Voronoi tessellation share an edge.

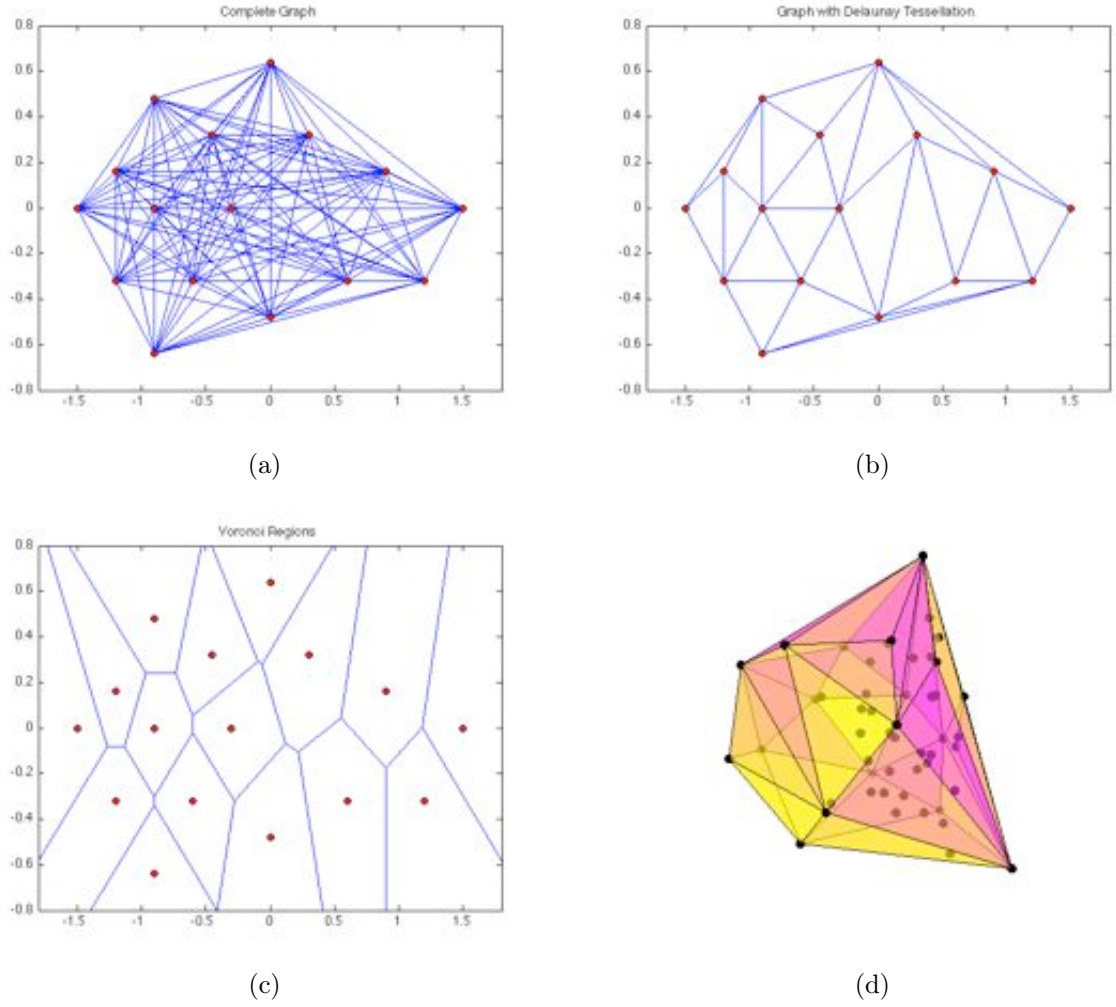
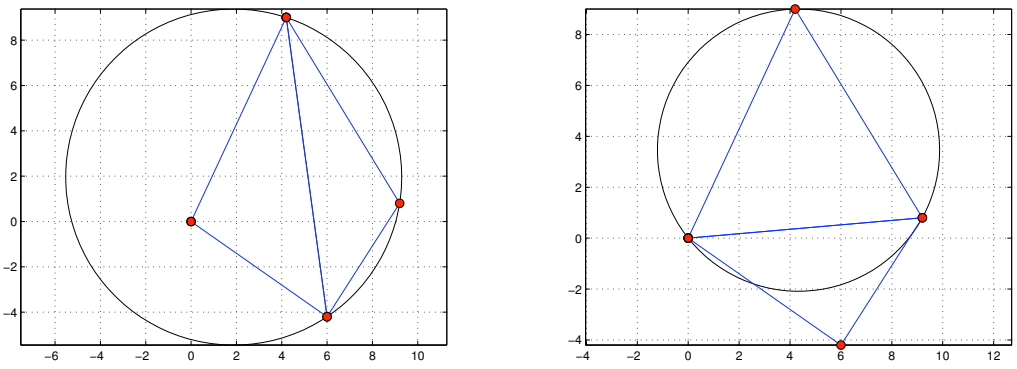


Figure 2.18: (a) A complete 2D graph of 16 nodes with 120 edges. In the complete graph likelihood computation each pair-wise distance (120 node-pair distances) leads to a histogram update. (b) Delaunay tessellation of the same 2D graph into triangles such that the corresponding circumcircles do not contain points other than the triangle vertices. In the approximate likelihood computation, only distances arising from the tessellated triangle edges lead to histogram updates (36 node-pair distances). (c) Voronoi regions of the same graph. Delaunay triangulation represents the the dual graph of Voronoi tessellation. (d) Delaunay tessellation for a 3D graph leads to division of the graph into pyramids with the similar property valid on circumspheres of the tessellated pyramids.



(a) (b)

Figure 2.19: An example with four 2D points illustrating the Delaunay conditions. In (a) since one of the vertices lies within the circumcircle of a triangle the desired condition is not satisfied. However, switching the common vertical edge with a horizontal edge results in a valid Delaunay triangulation (b) since none of the vertices lie within the circumcircles for each triangular simplex. Note that the resulting triangulation leads to neighbourhood connectivity and prevents a skewed tessellation.

Formally, we compute the Delaunay tessellation [38] of the 3D graph that results in a division of the graph into tetrahedrons (simplices) such that no data point is contained in any circumsphere of the simplices. The Delaunay condition on circumspheres prevents the tessellation to return skewed tetrahedrons, in effect connecting points to local neighbours, Figures 2.18 and 2.19. Alternatively, the tessellation can be viewed as maximizing the minimum angle of the tetrahedrons, in effect connecting neighbouring points. Note that the Delaunay tessellation is the dual of Voronoi tessellation and applies not only to R^d but to general metric spaces.

We restrict the graph likelihood computation only to edges of the tessellated graph which scales $O(N_f)$ compared to $O(N_f^2)$ for the complete graph. We use an implementation based on *Qhull*[4], a standard computational geometry package, to compute the tessellation [6]. The 3D Delaunay tessellation algorithm possesses

[4]Available at http://www.qhull.org/

$O(N_f \log N_f)$ complexity [6]. Under certain cases, a valid tessellation does not exist due to numerical issues or coincident points. In such cases we can use nearest-neighbour criterion to pick relevant edges. An experimental evaluation of several implementations of 3D Delaunay tessellation appears in [108].

2.4 Evaluation

2.4.1 Platform and data set

The topological mapping algorithm was tested on image and laser data obtained from a mobile robot. Data collection was performed using a vehicle shown in Figure 2.21. Imagery was captured at 3Hz from a Point Grey Ladybug 2 camera and laser range data was obtained from a SICK LMS 291 laser, scanning 90 degrees at 75Hz with 0.5 degree resolution. The laser was mounted so as to scan in a vertical plane normal to the vehicle's forward motion. The camera and the laser scanner were cross-calibrated experimentally. The data set was gathered within New College, Oxford in an environment of medieval buildings enclosing an oval lawn and cambered tarmac space[5]. The New College site possesses repetitive architectural features causing perceptual aliasing and is also traversed by people, thereby testing the system's robustness to scene change. The full data set is divided into three epochs (i) Campus, (ii) Parkland and (iii) Campus-Parkland based on visual characteristics. A detailed description of the platform and captured data appears in [174].

[5]The data set is available at: http://www.robots.ox.ac.uk/NewCollegeData/

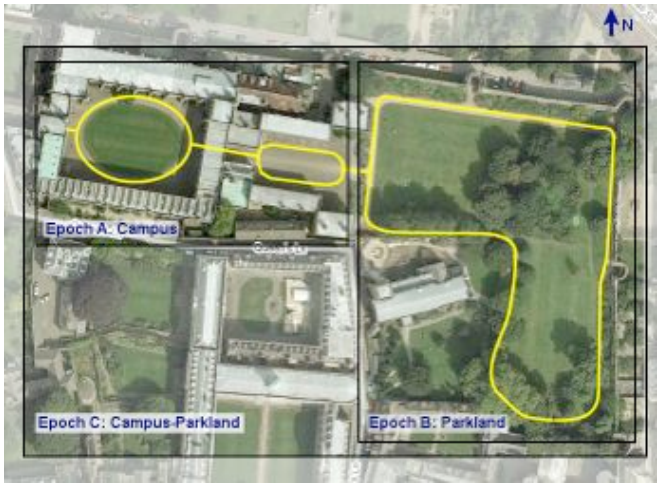
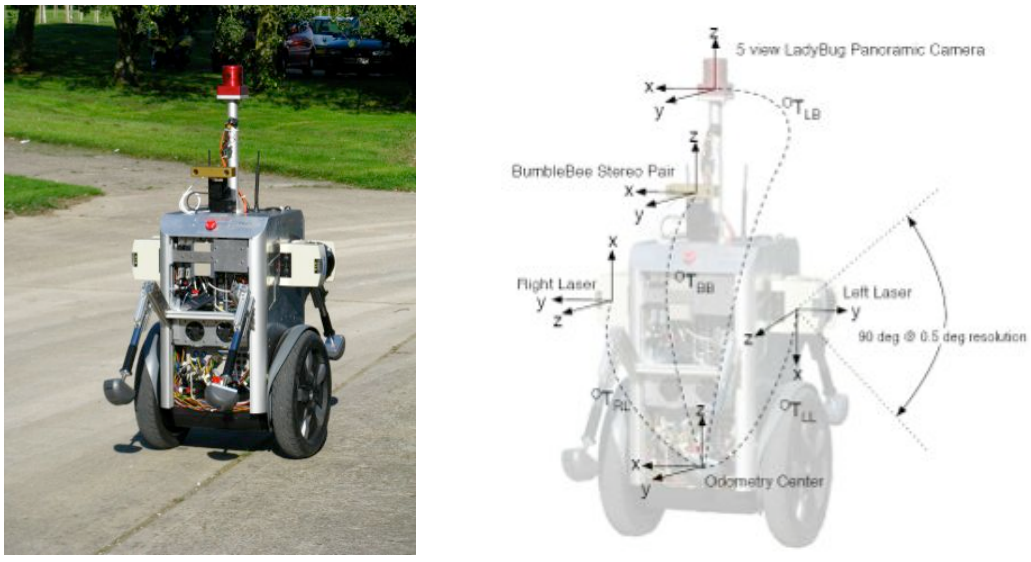


Figure 2.20: Aerial view showing the experimentation site, New College, Oxford with the robot's approximate trajectory shown in yellow. This entire data set is divided into three epochs based on visual characteristics.



(a) (b)

Figure 2.21: (a) Robotic platform used for experimentation with sensors and (b) coordinate frame centers shown.



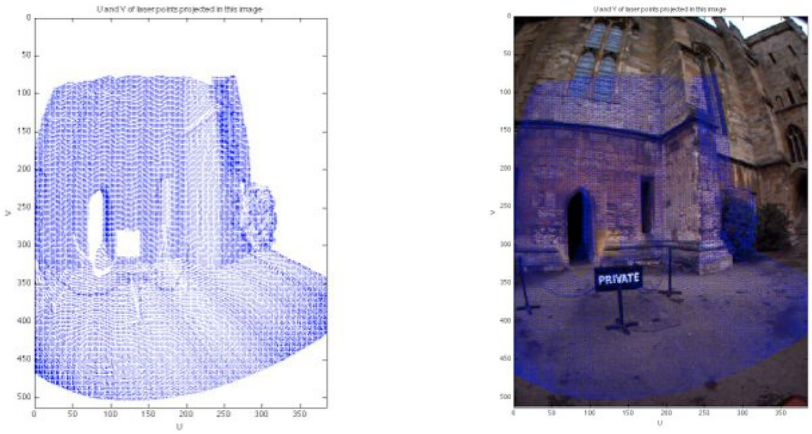
(a) (b)

Figure 2.22: Close up views of (a) Point Grey's Ladybug 2 omnidirectional camera and (b) SICK LMS 291 range finder used on the robotic platform.

2.4.2 Processing pipeline

Each image collected by the robot was converted to a *bag-of-words* representation by first extracting SURF features [7] and then quantizing them against a fixed vocabulary to obtain visual words for the image, yielding Z_k. The vocabulary was generated by clustering SURF features obtained from training images where cluster centers correspond to vocabulary visual words. A vocabulary size of $11K$ words was used. The SURF descriptor also determines the scale at which the feature was detected and approximates the feature size in image space.

The next task is to determine the inter-word distances for the perceived scene. Laser scans obtained in a 16sec window around the image capture time are back-projected into camera view. For each visual word detected in the image, proximal laser points are determined lying within a radius equal to the feature size (indicated by the SURF feature scale) from the detection centre. The visual word is assigned a 3D coordinate by taking an inverse radially-weighted average of the 3D coordinates of all nearby laser points. Hence, the position of each visual word is known in 3D space and all pairwise distances can be computed forming the set, D_k. The resulting observation graph, $G_k = \{Z_k, D_k\}$ is then passed on to the inference engine.



(a) (b)

Figure 2.23: Obtaining 3D coordinates for visual words by projecting laser points into camera frame after cross-calibration. (a) Laser point cloud. (b) Laser point cloud projected in the camera image.

The visual appearance likelihood term requires learning a Chow-Liu tree for the visual vocabulary. The Chow-Liu tree is constructed via a procedure outlined in [34] that proceeds by constructing the mutual information graph from word co-occurrence data available from the training set and subsequently computing the maximum weight spanning tree.

The marginal pairwise distance histograms were determined from a training set of 400 observations with visual and laser range data. Although the set of all pairwise distance histograms is very large ($11K \times 11K$) a relatively smaller number, 557491 word pairs were observed to co-occur in the data set (0.46% occupancy). For the other word pairs a uniform prior over ranges was assumed. Since the word co-occurrence matrix is very sparse, the number of histograms required for inference is tractable in practice. For space efficiency, only a single global copy of the marginal word-pair histograms is maintained. While initializing a new place model, only the

modified distance histograms are maintained locally. For word-pairs not seen in the observed graph, marginal distributions are assumed and can be obtained from the global copy while computing the observation likelihood term.

The final ingredient is the detector model. As suggested in [34] the visual detector model parameters were set to $p(z_i = 1|e_i = 0) = 0$ and $p(z_i = 0|e_i = 1) = 0.39$. The variance for the range detector, σ_{range} was set to 1.5m. Although, as noted by [104] the range uncertainty for LMS laser scanners (for close range) is $\approx$ 3cm, the range detector model also incorporates (i) uncertainty arising from vehicle odometry errors that affect projection of laser scans into image space that are taken a few seconds before or after image time and (ii) slight errors in cross-calibrating the laser ranger and the camera. Note that there may be scope for learning the detector model parameters empirically via grid search on a larger data set. However, in our experiments the results were not found to be very sensitive to this parameter.

2.4.3 Precision-Recall Performance

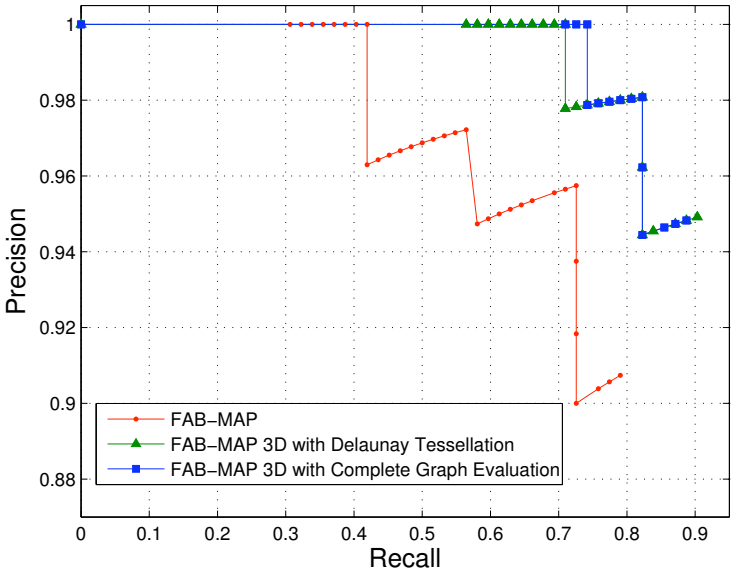
Precision-recall performance was evaluated on a test set consisting of 117 images. Ground truth was obtained from GPS data and determined by hand in sections where reception was intermittent. The precision-recall curves were obtained by varying the probability threshold at which loop closures are accepted. For comparing the core inference aspect of the system, the prior probability of being at a location was kept uniform (no motion model) for both implementations. In our setting, precision refers to the fraction of true image matches among those declared by system, Equation 2.28. Similarly, recall refers to the proportion of the total possible loop closure matches that exceed the probability threshold, Equation 2.29

$$precision = \frac{true\,positives}{true\,positives + false\,positives} \tag{2.28}$$

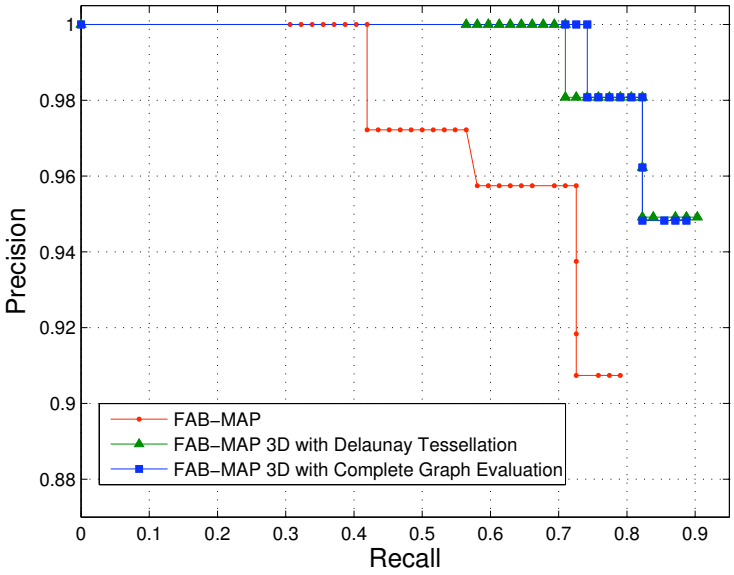
$$recall = \frac{true\,positives}{true\,positives + false\,negatives} \tag{2.29}$$

Figure 2.24 plots the precision-recall curves. We specifically compare the recall performance at 100% precision since an incorrect loop closure declaration can cause catastrophic mapping failure if the output of the place recognition system is coupled with a metric-mapping system [34]. FAB-MAP 3D with complete graph evaluation achieved a maximum recall of 74% at 100% precision. The FAB-MAP algorithm had a lower recall of 42% at 100% precision. FAB-MAP 3D with accelerated graph inference based on Delaunay tessellation achieved 100% precision at 71% recall. Hence, the graph approximation has a marginally lower recall than the full graph computation but still significantly outperforms FAB-MAP.

Precision-recall curves possess a distinctive sawtooth shape as shown in Figure 2.24a. This behaviour can be understood as follows: upon lowering the threshold, if the $(k+1)^{th}$ loop closure result is a false positive (nonrelevant) then the recall is same as for the previous k results by the precision value is lower. However, if the $(k+1)^{th}$ result is a true positive then both precision and recall increase and the curve moves up and right. As suggested in information retrieval literature [114], it is useful to plot the *interpolated* precision at a certain recall level r as the highest precision found at any recall level $r' \geq r$, Equation 2.30.



(a)



(b)

Figure 2.24: (a) Precision-recall curves comparing FAB-MAP, FAB-MAP 3D with complete graph evaluation and FAB-MAP 3D with accelerated graph inference on the New College data set. Note the scale. FAB-MAP 3D has a higher recall of 74% at 100% than FAB-MAP that has 42% recall at 100% precision. The graph for the accelerated approach partially overlaps with FAB-MAP 3D with complete evaluation. The accelerated approach has marginally lower recall of 71% but still performs better than FAB-MAP. Plots with interpolated precision appear in (b)

$$precision_{interp} = \max_{r' \geq r} precision(r') \tag{2.30}$$

The justification for using interpolated precision is that it is reasonable to decrease the decision threshold (thereby increasing k) if it increases the percentage of true positive results among the declared results. The precision-recall curves with interpolated precision for the experiment are plotted in Figure 2.24b.

2.4.4 Loop Closures

Figure 2.26 illustrates examples of true loop closures declared by FAB-MAP 3D but not detected by FAB-MAP. The perceived 3D graphs for one such loop closure pair appear in Figure 2.25. The spheres represent visual words and colours indicate the word type. The graphs are similar but possess minor differences caused by dynamic objects, slight viewpoint changes as well as sensor errors causing missed detections and false positives.

The FAB-MAP framework only considers words presence and does not find the two places to be distinctively similar. However, the probabilistic random graph framework in FAB-MAP 3D models both word presence and spatial characteristics. Utilizing the extra spatial information, the system infers high likelihood of the perceived graphs originating from the same location (or random graph), thereby declaring loop closure with high confidence.

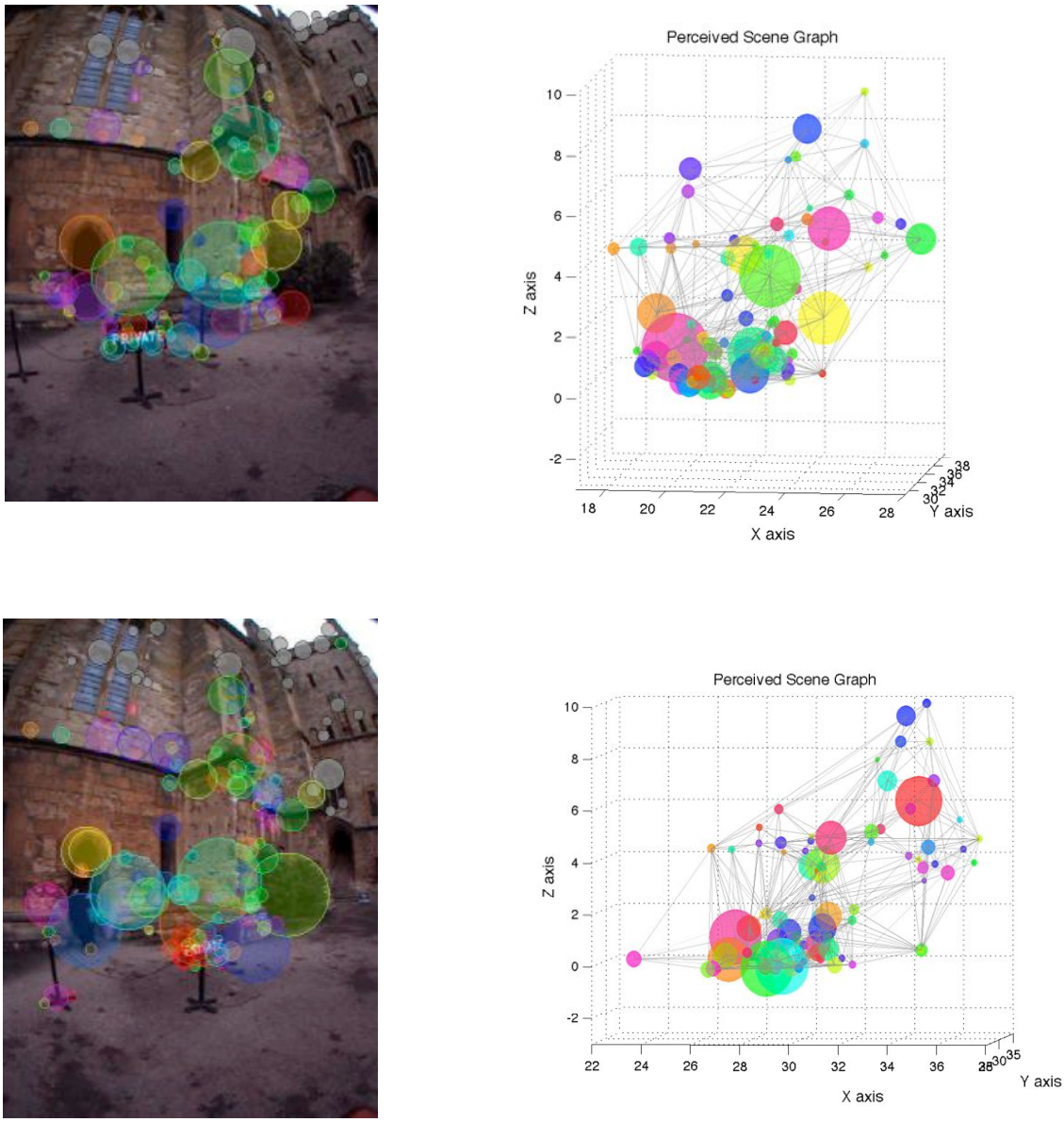
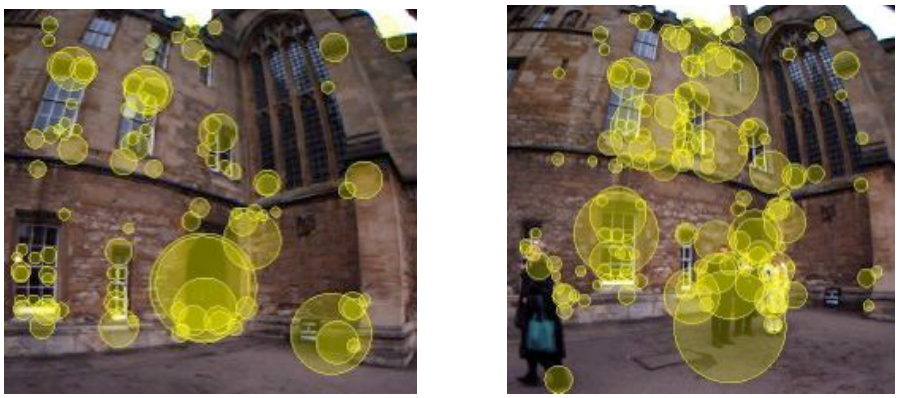


Figure 2.25: An example of a true loop closure detected by FAB-MAP 3D with high confidence whereas FAB-MAP assigned close to zero loop closure probability. The observed 3D graphs for both scenes appear on the right. Spheres represent visual words and colours indicate word type. The broad similarity in the graphs enables FAB-MAP 3D to infer loop closure. Image features shown in gray did not have associated range information due to limited coverage of the cross-calibrated laser scanner, hence excluded from the graph likelihood computation.



(a)



(b)

Figure 2.26: Examples of true loop closures detected by FAB-MAP 3D using spatial similarity and not detected by FAB-MAP.

2.4.5 Learning Spatial Appearance

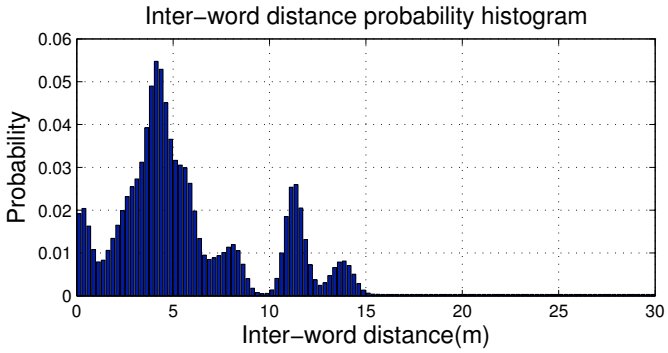
Figure 2.27 presents an example of learning word-pair distance probability histograms. Figure 2.27a shows a visual word that typically appears on the upper half of windows, occurring repetitively in the environment (Figures 2.27c and 2.27d). Figure 2.27b plots the probability distribution histogram modeling distances between multiple occurrences of this word. The variable inter-word distances due to repetitive structure are captured by the multi-modal distribution learned through kernel density estimation (KDE) with optimal bandwidth selection.

Note that smoothing prevents probability estimates for (unlikely) large distances from becoming zero. This provides support for a possible later observation in this range to be incorporated via recursive Bayes updates for each bin during the data association step. Figure 2.3.3 illustrates the effect of smoothing the KDE estimate with a uniform prior distribution using Equation 2.16.

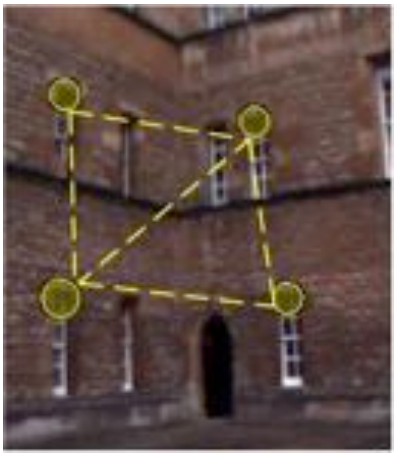
Figure 2.29 displays the posterior belief update during data association for two distributions over inter-word distances. The prior distributions (left) learnt offline via KDE are transformed to posterior distributions (right) via a Gaussian range sensor model ($\sigma_{range} = 1.5$m) given an observed 3D distance measurement between visual features using Equation 2.26. It is noteworthy that the distance measurements of 3.2m and 6.7m respectively (shown with a red line), corresponded to the less-likely mode in the prior belief but captured the dominant mode after the posterior update.



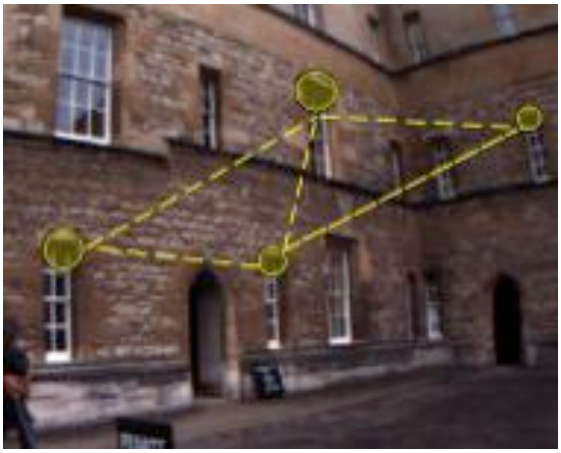
(a)



(b)

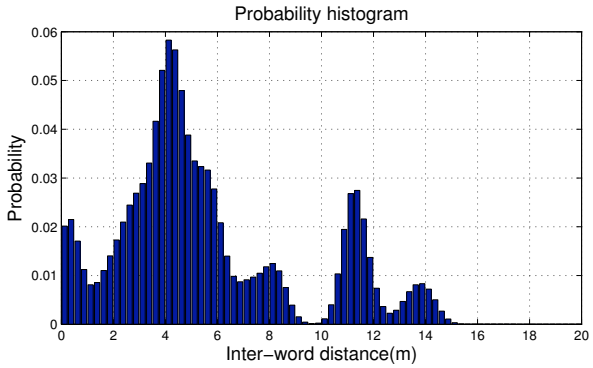


(c)

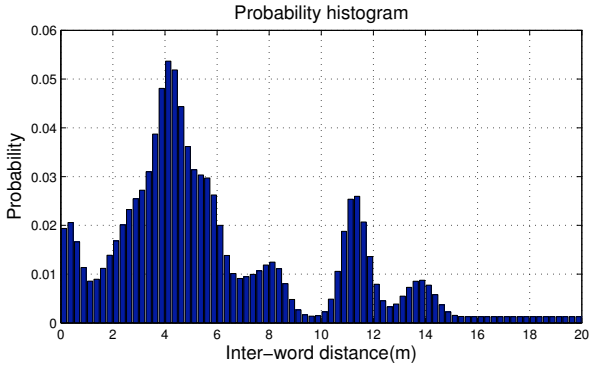


(d)

Figure 2.27: Learning word pair histograms. (a) A visual word that typically appears on the upper half of windows. (b) Probability distribution for distances between multiple occurrences of this word. The word appears repetitively, generally at regular distances (c) and (d). The variable inter-word distances due to repetitive structure are captured by the multi-modal distribution learned through kernel density estimation with optimal bandwidth estimation. Also note that smoothing prevents likelihood estimates for (unlikely) large distances from becoming zero. This provides support for a possible later observation in this range which would be incorporated via recursive Bayesian updates for each bin during data association.



(a)



(b)

Figure 2.28: Effect of smoothing histograms over distances. (a) Kernel density estimate (KDE) from training samples. Note that probability estimates for distances beyond 15m are zero. (b) Probability estimate after smoothing with a uniform prior using Equation 2.16. This prevents probability estimates for large distances (although unlikely) from becoming zero. Hence provides support for a possible later observation in this range to be incorporated via recursive Bayes updates during the data association step.

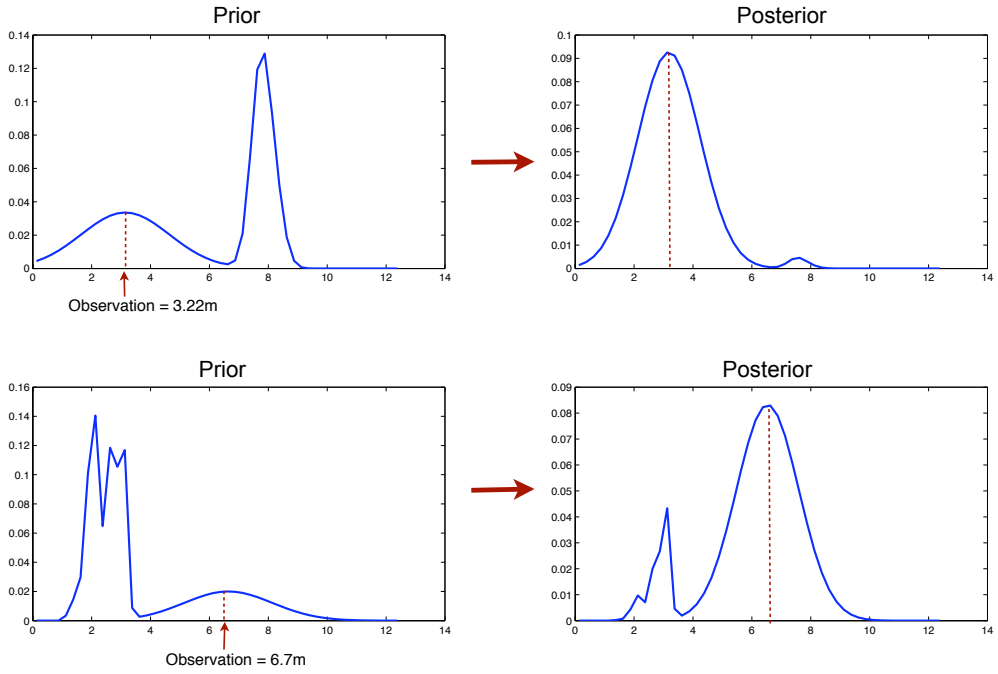
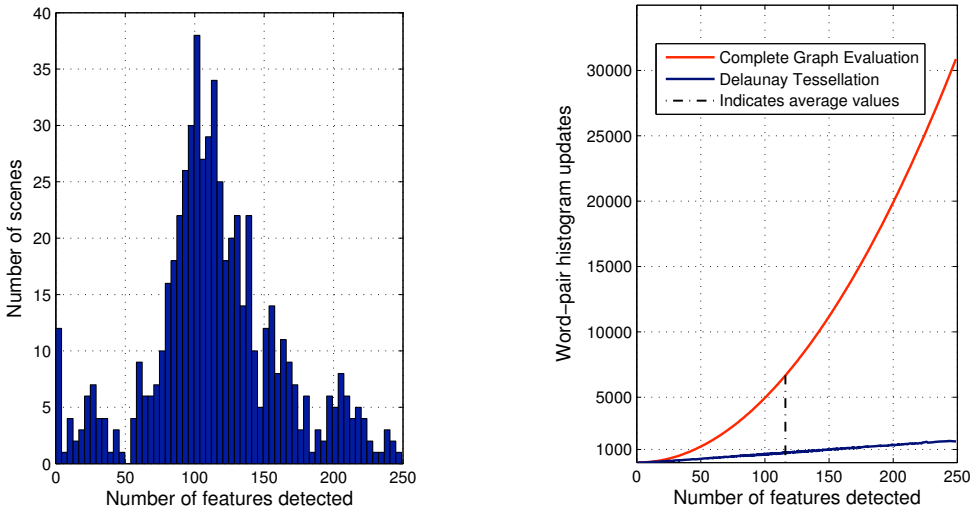


Figure 2.29: Prior distributions over inter-feature distances (left) learnt offline via KDE are transformed to posterior distributions (right) via a gaussian range sensor model ($\sigma_{range} = 1.5$m) given an observed 3D distance measurement between visual features using recursive Bayes update (Equation 2.26). Note, the distance measurements of 3.2m and 6.7m respectively (shown with a vertical red line), corresponded to the less-likely mode in the prior belief but captured the dominant mode after the posterior update.



(a) (b)

Figure 2.30: (a) Number of scenes vs. number of features detected, N_f for the combined set of training and test sets (517 locations). Average $N_f = 116$ (std. dev. 48 and median 112). (b) Number of word-pair distance probability histogram updates vs. N_f. The histogram updates required, scales quadratically in N_f for the complete graph evaluation and linearly for computation with Delaunay tessellation, illustrating the speed-up of the approximate scheme.

2.4.6 Efficiency and Timing

The computational overhead for FAB-MAP 3D over FAB-MAP was measured as: (i) *online*, average 314ms inference time per place in the map including tessellation and (ii) *offline*, 4.5hours for one-off density estimation procedure on a 2.66 GHz Intel Core 2 Duo machine running a MATLAB implementation.

Figure 2.30a plots the number of scenes vs. number of features detected in each scene (N_f), representing scene complexity, for the combined set of training and test sets (517 locations). The average N_f was found to be 116 (standard deviation 48, median 112). The main additional computational cost of FAB-MAP 3D over FAB-MAP stems from the number of inter-word distance probability histograms updated

per scene.

Figure 2.30b plots the number of histogram updates vs. number of detected features per scene for the data set. The number of histogram updates required scaled quadratically in scene complexity for the complete graph evaluation and linearly for computation with Delaunay tessellation, illustrating the advantage of the approximate method with a marginal decrease in performance, see Figure 2.24. Computing the tessellation scales log-linearly with N_f, however the *Qhull* implementation is very fast adding little overhead cost (average 12ms per scene).

2.4.7 Limitations

Figure 2.31 illustrates practical limitations of the laser ranging based approach for estimating 3D positions for visual words. Different positioning of the camera and the lasers can cause a disparity in the viewing or sensing regions of the two sensors. Laser scans projected into the camera frame may not cover the entire image causing absence of range estimates for some visual features. Figure 2.31a shows a rectangular patch on the background wall with no projected laser points, since for the scanning laser this region appears in the shadow of the foreground sign board. Dynamic objects like people, vehicles etc. can cause scene change between the image obtained by the camera and the laser scans taken later in time, causing incorrect point clouds to appear in the scene as shown in Figure 2.31c. Computing the inverse-radially weighted-average of the 3D coordinates of the projected laser points within feature radius provides some robustness to such cases.

Note that surfaces axis-parallel to the laser scanner yield very few reflections and hence sparse range estimates for a surface which could possess many visual features. Additionally, the projection accuracy for the laser scans within a time window of



(a) (b) (c)

Figure 2.31: Practical limitations. Laser scan points projected into camera frame are shown in blue. (a) Laser scanning range is insufficient to cover the entire image causing absence of range estimates for visual words detected in the upper portion of the image. (b) The rectangular patch seen on the background wall forms the shadow for the signpost for the laser scanner. Hence, no laser points project in this region. (c) Dynamic objects like moving people can cause incorrect point clouds to appear in the scene. The occurs due to the time interval between the when the camera images the scene and the time when the laser scans the same region.

image capture time hinges on integrating odometry information to obtain robot pose prior to projection. The accumulation of odometry error makes the projections of laser scan taken further in time less accurate. The current system used odometry from the segway base. This can potentially be improved through visual odometry available from stereo cameras mounted on the robot.

2.5 Conclusions

In this chapter we introduced a probabilistic framework for appearance based topological mapping called FAB-MAP 3D. In this formulation, locations are represented as random graphs and a generative model is learnt over word occurrences as well as their spatial distributions. This approach provides substantial and compelling

improvement in precision-recall performance over the existing FAB-MAP system. By capturing spatial information, the algorithm reduces the number of false positives and shows a decrease in false negative rate, particularly in scenes possessing a large number of common words where a loop closure decision hinges on spatial information. The framework shows robustness to perceptual aliasing as well as scene change. The system scales linearly with the number of places in the map. We also presented a method for accelerating graph inference based on Delaunay tessellation of the observed graph, that scales log-linearly with scene complexity.

In the next chapter, we will explore a second improvement for the FAB-MAP system that allows the system to obtain relevant image-based representation of its workspace leading to asymptotic improvement in performance.

"Perplexity is the beginning of knowledge."

Khalil Gibran, *A Tear and A Smile* (1950)

Chapter 3

Self-Help: Seeking Perplexing Images for Ever Improving Navigation

This chapter concerns having a robot actively seek data to improve its understanding of the world. The big picture motivation of the work is to enable robot longevity and here we consider a specific instantiation of this problem - that of asymptotically improving scene recognition with a camera.

We make use of the FAB-MAP algorithm [34, 35] which probabilistically associates a current view of the world (image) taken by a robot with a previously visited or a new place. FAB-MAP requires priming with a set of images (referred to as a sample set) which in concert, statistically represents the appearance of the robot's workspace. For operation in urban settings, for example, one equips it with a sample set containing random images of cities and towns. There is an obvious shortcoming here - the robot is constrained to work in settings in which its sample set has sufficient explanatory power. If moved into surroundings quite different from those represented by its sampling set the performance drops - nothing is as expected and

everything is astounding. Further, FAB-MAP has been successfully demonstrated to create maps exceeding 1000km [35]. Consequently, the topological map generated representing the cumulative visual experience of the robot can far exceed the size of the onboard sampling based workspace representation. Hence, the ability to enrich the onboard workspace representation by incorporating new visual themes discovered during exploration is a key competency required for life-long topological mapping.

In this chapter, we demonstrate how by producing a generative model for the observed images based on latent visual topics we can actively grow a customized sample set by incorporating well chosen examples from an external corpus which is more representative of the workspace the vehicle is experiencing. In this way, we replace the inflexibility of a static *a-priori* sample set with a plastic, dynamic one and show that this affords a performance improvement over time. One could think of this as a robot actively seeking to widen its experience, better understand its surroundings and becoming less perplexed with time.

Our problem setup is as follows. A mobile robot must maintain a compact onboard sample set summarizing the visual appearance of its environment. The robot explores the environment collecting image data and identifies the most perplexing images based on its current sample set. It then searches the least explained or novel images in a large repository of image data (or past data sets collected by the robot) finding examples with similar thematic content. Next, the robot retrieves relevant samples and assimilates them into the sample set, thereby improving its representation and performance.

The rest of this chapter details this framework and presents the following components:

- use of Latent Dirichlet Allocation (LDA) based topic models to extract a low-dimensional thematic representation for images incorporating word co-occurrence statistics in an unsupervised manner, Section 3.2
- finding images similar in thematic content from an external repository applying the language-model based information retrieval approach, Section 3.3
- identifying the most novel images seen by the robot given its current representation using a perplexity-based measure, Section 3.4
- application to FAB-MAP, presenting an algorithm for constructing a representative sampling set, Section 3.5
- experimentation on real data sets collected by a mobile robot, Section 3.6.

In the next section we present an overview of literature relevant to this chapter before presenting the incremental learning algorithm in subsequent sections.

3.1 Related Work

This chapter builds on research in the areas of probabilistic topic models, saliency or novelty detection and active learning applied to vision, machine learning, text mining and language modeling domains. In this section we survey some relevant works.

3.1.1 Latent Topic Models and Applications

Topic models based on Latent Dirichlet Allocation (LDA) were introduced in the seminal work of Blei et al. [17] in the context of statistical text analysis. The

authors presented a Bayesian generative model for multinomial data and a variational Bayes approximate inference technique demonstrating superior performance for language modeling, classification and collaborative filtering tasks compared to probabilistic latent semantic analysis (pLSA) [79] and mixture of unigram models [127]. Griffith and Steyvers [64] extended this work and introduced an efficient collapsed Gibbs sampler for approximate inference and demonstrated an application involving unsupervised discovery and tracking of scientific topics from a corpus of journal abstracts.

Several researchers have explored modifications and improvement over the original LDA graphical model. The basic LDA model assumes independence in the word generation process and does not model correlation between topics present in a document. In [192], Wallach introduced the bigram topic model that extends LDA by incorporating a notion of word order, building on prior work by MacKay and Peto in [112]. Graber and Blei [20] introduced syntactic topic models (STM) where latent topic distributions can be learnt using both semantic (document-level features) and syntactic cues (local word context). In [15] Blei and Lafferty introduced correlated topic model (CTM) allowing topics in a document to exhibit correlations incorporated via a logistic normal distribution instead of Dirichlet priors in the model. In [14], the authors developed a dynamic topic model (DTM) that captures temporal evolution of topics in a sequentially organized document corpus. The approach involves using state space models on the natural parameters of the multinomial distributions that represent the topics. Aimed at document network analysis, Chang and Blei [24] developed relational topic model (RTM) that jointly models documents and links between them (conditioned on their topical content) and can be used for network summarization, document-word and network-link prediction. The work of

Blei and McAuliffe [16] presented supervised latent Dirichlet allocation (sLDA) for modeling documents and their labeled attributes. The model is motivated by applications involving attribute prediction for unseen documents, e.g., movie ratings predicted from reviews where sLDA learns latent topics that are highly predictive of the response.

The above mentioned topic models require the number of topics to be specified *a-priori* or estimated via model selection on a hold-out set. In [181], Teh et al. introduced hierarchical Dirichlet process (HDP), a non-parametric hierarchical Bayesian model that additionally infers the number of topics from data. A key aspect of Bayesian non-parametric models is their ability to automatically adjust their complexity with observations and generalize with relatively little training data.

Within the computer vision community, Sivic et al. [172] employed LDA for discovering object categories in image corpora and present a hierarchical model for unsupervised discovery of object class hierarchies. In [44], Li and Perona present a novel approach for unsupervised learning and recognition of natural scene categories. The algorithm is based on a modification of the LDA model where an explicit category variable is introduced for classification. In [193], Wang et al. address image classification and annotation in a unified framework and present a probabilistic model extending sLDA to jointly model image class label and annotations. Philbin et al. [141] introduced geometric LDA (gLDA) for unsupervised particular object discovery in image corpora. Here, images are considered as mixtures of object and latent topic models incorporate both the type and spatial layout of visual words in a geometrically consistent manner. In a related effort, Wang and Grimson [197] formulate spatial LDA where a spatial prior is used grouping visual words in close proximity into the same topic. As a key non-parametric example, Sudderth et al.

[178] introduced transformed Dirichlet processes to infer the spatial structure in visual scenes, explicitly handling the uncertainty in the number of object instances present in a given image.

Within robotics, Endres et al. [42] applied LDA for unsupervised discovery of object classes in 3D laser range data. The authors demonstrated experiments in indoor workspaces where inferred topics represented object categories like tables, chairs etc. Recently, Bayesian non-parametric approaches have gained prominence since their Bayesian aspect prevents over-fitting while their non-parametric nature allows data-driven model selection preventing under-fitting. In [41], Joho et al. introduced a hierarchical generative model for describing indoor scenes like a kitchen table with household items where objects occur in recurrent configurations not independent of each other. The proposed model combines Dirichlet processes and Beta processes to infer the latent object constellations in a scene allowing both scene segmentation and completion of a partially specified scene. Joseph et al. [92] proposed modeling mobility patterns of ground vehicle (or targets) using a Gaussian Process (GP) mixture with Dirichlet process (DP) prior allowing representation of an *a-priori* unknown number of patterns. In a related work [93], the authors presented a data-driven infinite mixture model of battery health characteristics, aimed at predicting time to failure events from current and voltage behaviours.

In other applications, Wang and Mori [199] introduced semi-latent topic models for discovering human action sequences in videos. Rodriguez et al. [154, 155] applied correlated topic model for tracking in unstructured crowded scenes, where words correspond to low-level quantized motion features and topics correspond to motion behaviours. Within information retrieval Wei and Croft [200] used LDA-based topic models for ad-hoc document retrieval extending efforts by Liu et al. [107].

3.1.2 Novelty Detection in Images and Videos

The detection of novel and salient elements in information streams like images and videos has been explored by several researchers. In [87], Itti and Baldi present a surprise-detection model for video streams, simulating early vision receptors. The model computes low-level spatial and temporal surprise at each location in the image and results were found significantly correlated with eye movements of humans watching complex video streams. In [75], Hendel et al. address the problem of identifying atypical events in surveillance videos using the notion of Bayesian surprise: where an event is considered "surprising" if its occurrence leads to a large change in the probability of the world model. The corpus of video events is modeled using LDA and Bayesian surprise is measured by computing the change in the Dirichlet prior for the LDA model as a result of each video event's occurrence. In [148], Ranganathan and Dellaert use Bayesian surprise for detecting landmarks in the context of topological mapping for robots using appearance and laser measurements.

Several authors have explored unsupervised learning of activity behaviours in video streams possessing crowded and dynamic scenes aimed towards categorization, segmentation, mining and anomaly detection applications. Vardarajan et al. [191] applied probabilistic Latent Semantic Analysis (pLSA) for unsupervised activity analysis and presented measures based on normalized likelihood and adaptive Bhattacharya measure to detect abnormal behaviour. The authors also highlighted the importance of normalizing abnormality measures with respect to document length. In [198], Wang et al. present a hierarchical Bayesian model describing atomic activities as distributions over low-level visual features and interactions as distributions over atomic activities. They improve the standard LDA [17] and HDP [181] models to allow simultaneous grouping of words into topics and documents into

clusters (sharing the same prior over topics) in an unsupervised manner. Additionally, abnormal activities can be identified as those possessing low-likelihood under the learnt model. Hospedales et al. [84] present a different formulation termed Markov chain topic models (MCTM) for learning multi-object behaviours incorporating spatio-temporal cues. The MCTM model attempts to combine elements of dynamic Bayesian networks (e.g. HMMs) and Bayesian topic models (e.g. LDA) allowing clustering of visual events into activities and behaviours, additionally correlating global behaviours over time. In [85], the authors present an extension called joint topic model for identifying rare and subtle behaviours. The approach incorporates weak supervision (few positively labeled examples) and a shared latent structure between the common and rare behaviour classes.

Boiman and Irani [18] recognize unusual activities and salient regions in videos and images by considering the complexity required to represent the visual data as a composition of basis dictionary video patches. They attempt to probabilistically compose a new observed image region or a new video segment using patches extracted from previous visual examples. Regions in the observed data which can be composed using large contiguous chunks of data from the database are considered very likely, whereas regions in the observed data which cannot be composed from the database or can be composed only using small fragmented pieces are regarded as unlikely. In another related work, Zhang et al. [208] discuss novelty detection for adaptive filtering applications.

3.1.3 Related Work in Active Learning

The problem of actively seeking image data to improve place recognition performance addressed in this chapter closely related to the active learning paradigm in

machine learning. The central idea behind active learning is that a learning algorithm can achieve greater accuracy with fewer training labels if it is allowed to choose the data from which it learns [165]. An active learner acquires new training data by posing queries, in the form of unlabeled data instances to be labeled by an oracle (e.g., a human annotator) and then updates the model by incorporating the new data point. Vast literature exists in this domain. Here we restrict ourselves to only relevant works.

Joshi et al. [94] address multi-class image classification using SVM classifiers and propose margin-based uncertainty (measured as entropy) and best-versus-second-best (BvSB) criteria for actively seeking training data. Holub et al. [81] propose an information-theoretic criteria for selecting images that maximizes the expected information to be gained about the entire pool of unlabeled data available and also propose an entropy-based stopping criteria. Hospedales et al. [83] analyse active learning criteria that optimise both rare class discovery and classification simultaneously using a combination of generative and discriminative classifiers. Kapoor et al. [95] present an active learning approach for object categorization using a GP classifier. Here, data points with maximum uncertainty obtained by combining the posterior mean and variance estimates are queried for labels and used to improve the classifier. The theoretical work of Krause and Guestrin [99] investigates active learning of GPs with applications to monitoring spatial phenomena. The authors analyze the performance difference between sequential and a-priori sampling strategies and present an efficient non-myopic sequential algorithm for observation selection.

The problem of actively selecting the most informative points from a large collection has also been explored in the context of sparse approximations to GPs. The work of Lawrence et al. [103] and Seeger et al. [163] are notable examples. In

the text classification domain, Tong and Koller [185] present an active learning algorithm using SVMs with a notion of a version space. Further, McCullum and Nigam [115] introduce a combination of query-by-committee (QBC) active learning [53] with expectation maximization (EM) for improved learning performance. The recent work by Tellex et al. [177] explores active information gathering for human-robot dialog. The authors formulate an information-theoretic strategy for asking targeted clarifying questions to resolve the robot's uncertainty about the mapping between phrases and aspects of the workspace.

We now conclude our survey of relevant literature. In the next section we describe topic models based on LDA, a key mathematical tool that we will apply in the current and next chapter.

3.2 Probabilistic Topic Models

Probabilistic topic models represent documents as a mixture of intermediate latent topics. Given a collection of documents such as scientific abstracts, each represented as a bag-of-words vector, the model is able to learn common topics such as *ecology*, *astronomy* etc. in an unsupervised manner [64]. Using the approach by Sivic et al. [173], images can be represented as a vector of visual words. Hence, we use the terms "documents" and "images" interchangeably.

Latent Dirichlet Allocation (LDA) is a widely used probabilistic topic model [17]. LDA is a hierarchical Bayesian generative model for a collection of discrete data possessing tractable inference to estimate topics and topic proportions. In this work, we apply LDA based topic models to identify perplexing observations and for retrieving images similar in thematic content. In the following sub-sections we

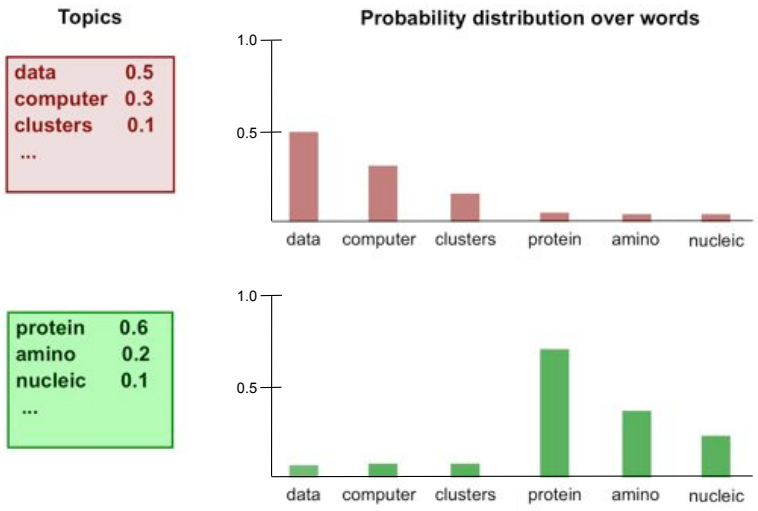


Figure 3.1: Probabilistic topics as distributions over words. The figure illustrates two topics: *computer science* and *biology*. Characteristic words for each topic e.g., *protein* and *amino* for topic *biology* are assigned higher weights. Each word occurs with varying proportions in each topic.

review the LDA generative model and inference. In the next section, we present an illustration of the model with a toy example followed by a mathematical review. For detailed exposition on LDA please refer to [17], [64] and [74]. Readers familiar with the model can proceed directly to Section 3.3.

3.2.1 Illustration

Probabilistic topic models describe the generative process of forming document corpora via latent themes or topics. Topics are distributions over words and each document is a distribution over topics. Different documents can have varied mixing proportions of each topic. The topics themselves are latent and represented as hidden variables that combine probabilistically to generate observed document.

Figure 3.1 presents an illustration. Two topic distributions corresponding to *computer science* and *biology* are shown. Each topic assigns higher probabilities to

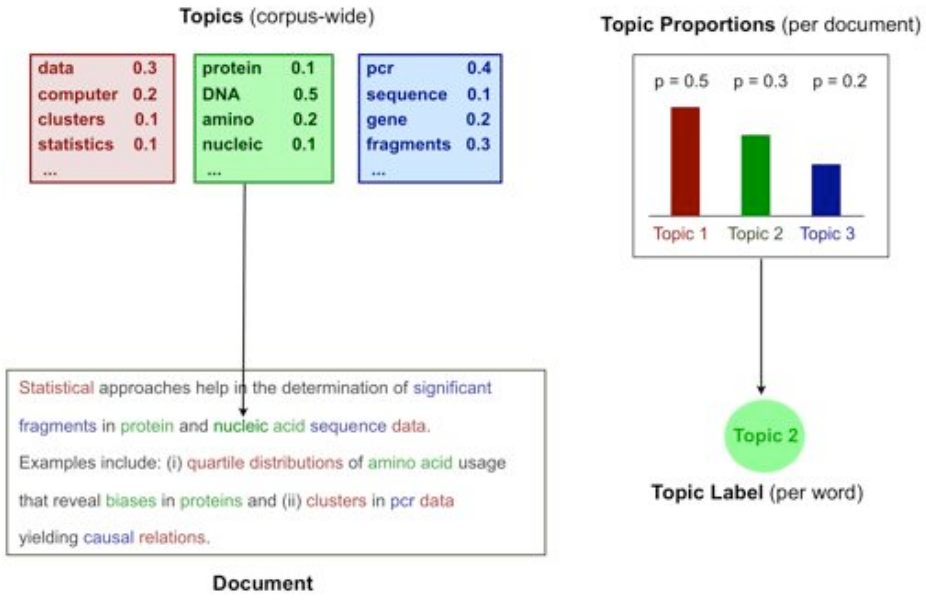


Figure 3.2: The LDA model describes document generation as a two stage process. First, the topic mixing proportions are determined for the document (top right). Next, each word is generated by first sampling a topic label (e.g. *biology*) from the distribution over topics followed by choosing a word (e.g., *nucleic*) from the corresponding topic distribution over words.

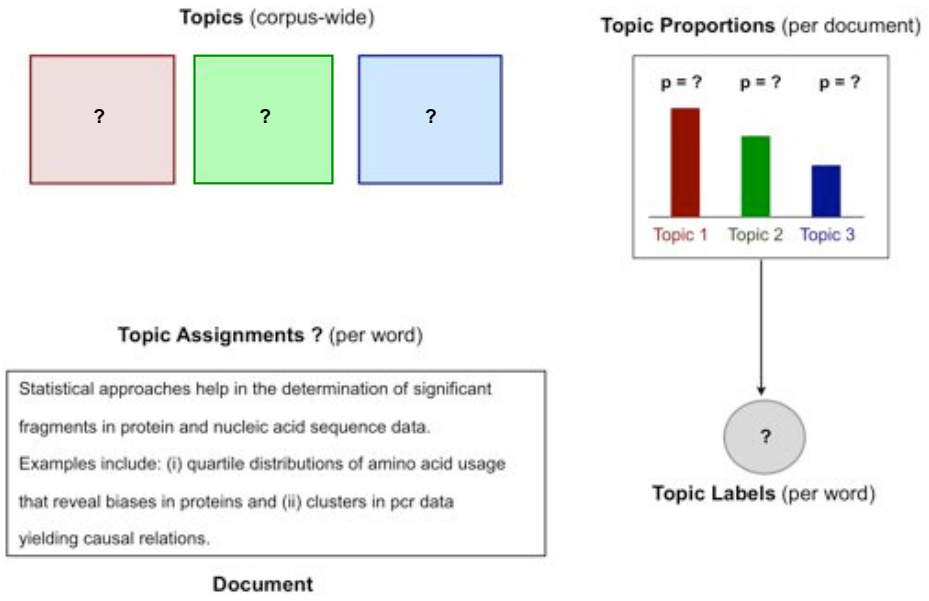


Figure 3.3: The collection of words forming a document are the only observed variables. The latent structure including corpus-wide topic distributions, per-doc topic proportions and per-word topic label must be inferred via posterior inference.

characteristic words and every term occurs in all topics with varying probabilities. Given topics, the generation of a document begins by first picking the mixing proportion for topics specific to the document, represented as a multinomial distribution. A word is generated by first sampling a topic (say *biology*) according to document-wide topic distribution followed by sampling the topic distribution to pick a word (say *nucleic*). Each word is generated independently using the sample procedure resulting in a document, see Figure 3.2. Different documents can be characterized by their topic proportions.

Note that the observed data only consists of the words organized as documents in a large corpus as shown in Figure 3.3. The latent corpus-wide topics, document-specific topic proportions and topic assignments for observed word tokens are inferred using posterior inference. The latent topic structure is typically estimated

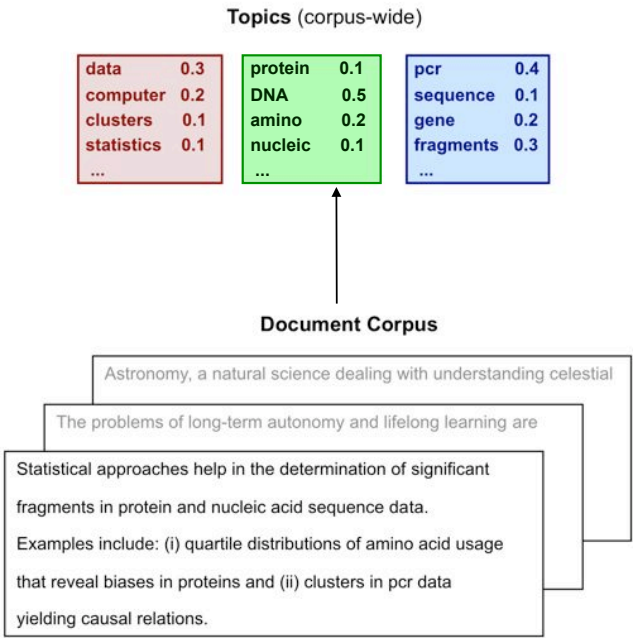


Figure 3.4: Topic distributions are typically estimated once from a large document corpus.

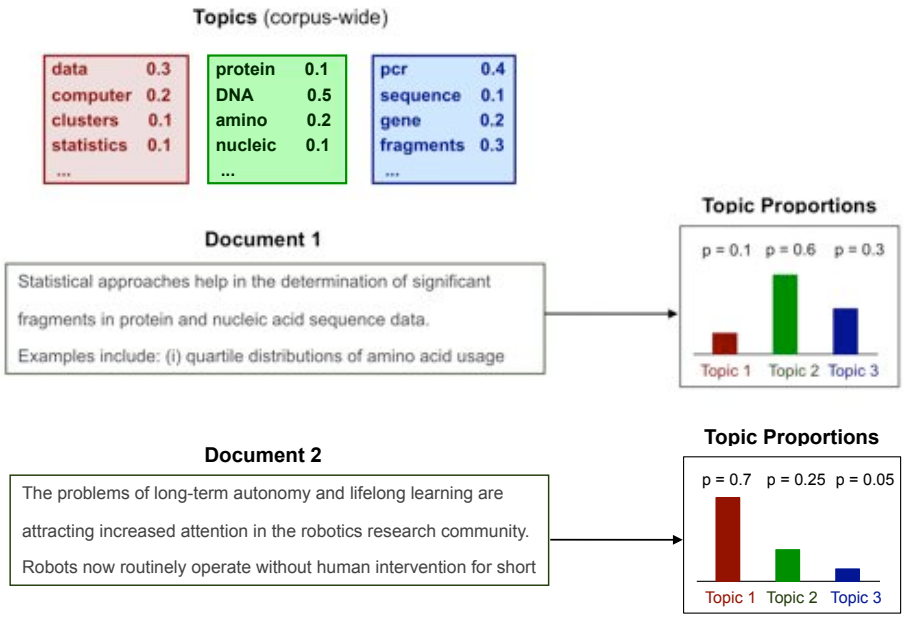


Figure 3.5: Once corpus-wide topics have been learnt, topic proportions can be estimated for each newly acquired document yielding a lower-dimensional representation for the observed document.

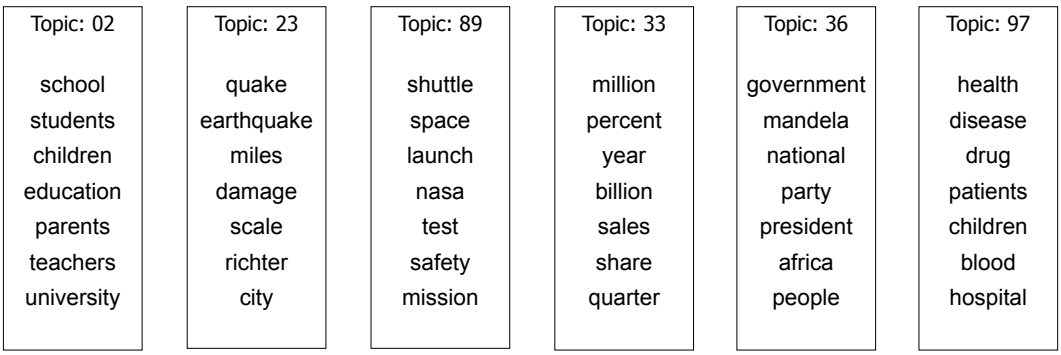


Figure 3.6: Representative topics learnt on a news article corpus (2246 articles, 100 topics). Most probable words are shown for each topic. Note that the discovered topics possess a coherent theme.

once for a large corpus, Figure 3.4. A new document or query can be situated in the thematic model by estimating its topic proportions conditioned on the learnt topic distributions, Figure 3.5.

As a practical example, Figure 3.6 illustrates topic learning on the *Associate Press*[1] data set commonly used by text retrieval community. A 100 topic model was learnt on the 2246 articles in the data set. The figure illustrates a few sample topics with most probable words are displayed (a stop list of common words was applied). Note that topics discovered possess interpretable themes like *space*, *education*, *health* etc. Applied to visual tasks, topics model the co-occurrence of visual features and the mixture of hidden topics refers to the degree to which a certain object or scene element is contained in the image, giving rise to a low dimensional description of the coarse visual thematic content.

Next, we review the LDA mathematical formulation formally describing the generative model and posterior inference for topics and topic proportions.

[1] *Associate Press* news article corpus. Available at: http://www.cs.princeton.edu/~blei/lda-c/index.html

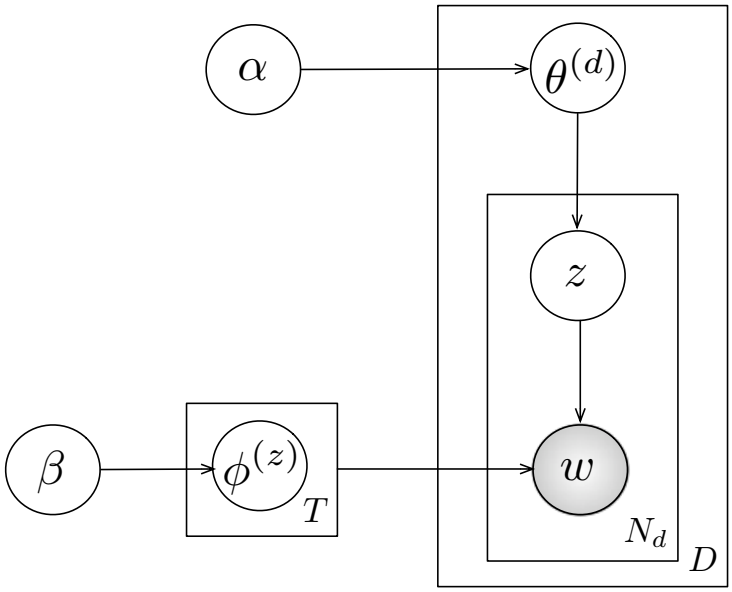


Figure 3.7: LDA Generative Model. Topics, $\phi^{(z)}$ are multinomial distributions over vocabulary words (with Dirichlet prior β). The generative process for a document, d begins by sampling a distribution over topics, $\theta^{(d)}$ (with Dirichlet prior α). Document words are generated by first drawing a topic label z from topic-proportion $\theta^{(d)}$ and then sampling a word w from $\phi^{(z)}$.

3.2.2 Latent Dirichlet Allocation (LDA)

Generative Model

A document d from a corpus of D documents consists of a set of words $(w_1, w_2, \ldots, w_{N_d})$, where w_i is a single word occurrence from a vocabulary of size W. The model postulates T topics, each characterized by a multinomial distribution over words $P(w|z)$. The generative process (Figure 3.7) for each word begins by first sampling a topic label z_i from the multinomial distribution over T topics for the given document $P(z|d)$ followed by sampling a word w_i from the distribution over words $P(w|z)$ for the sampled topic label. Hence, the likelihood of a word in a document can be obtained via marginalization over intermediate topics using Equation 3.1.

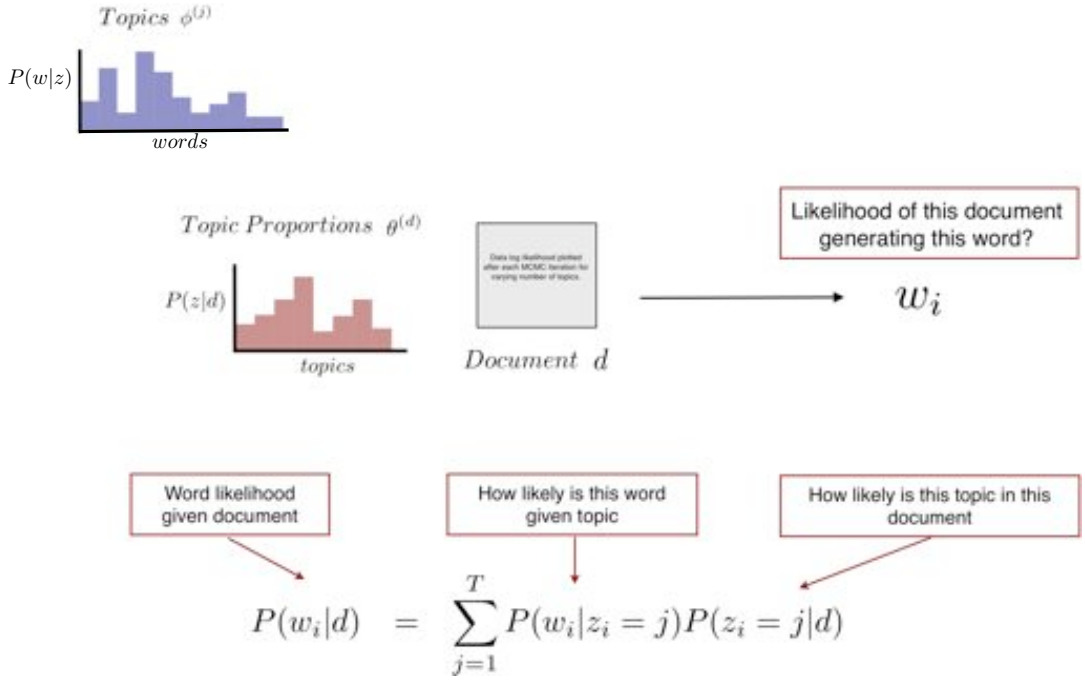


Figure 3.8: Given learnt topics ϕ, the estimated topic proportions form a topic-based probabilistic generative model for a document d. The likelihood of observing a word w_i given the document model can be obtained via marginalizing over the latent topics aggregating word likelihood under each topic.

$$P(w_i|d) = \sum_{j=1}^{T} P(w_i|z_i = j)P(z_i = j|d) \tag{3.1}$$

The topic-proportion $P(z|d)$ for each document d and word-likelihoods $P(w|z = j)$ for each topic j are abbreviated as $\theta^{(d)}$ and $\phi^{(j)}$ respectively [64]. Dirichlet priors are placed on topics and topic proportions. Hence, $\theta^{(d)} \sim Dirichlet(\alpha)$ and $\phi^{(j)} \sim Dirichlet(\beta)$. The Dirichlet is a probability distribution over a simplex i.e. positive vectors that sum to one with the form given in Equation 3.2. Note that topics are a W dimensional Dirichlet and topic proportions are a T dimensional Dirichlet.

$$P(\theta|\alpha) = \frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)} \prod_{i=1}^{T} \theta_i^{\alpha_i - 1} \tag{3.2}$$

The Dirichlet distribution belongs to the exponential family and is *conjugate* to the multinomial distribution. The hyper-parameters α and β control the mean and sparsity of the sampled multinomial distributions. The hyper-parameters are assumed to be symmetric and sparse by the model.

Estimating Topics and Topic Proportions

Estimating the set of topics and topic-proportions from observed word tokens requires reversing the generative process. For each observed word w let z be the topic indicator variable. The goal is to estimate the topic distributions that best describe the data by evaluating the posterior distribution $P(\mathbf{z}|\mathbf{w}, \alpha, \beta) \propto P(\mathbf{w}|\mathbf{z}, \beta)P(\mathbf{z}|\alpha)$. Here, $\mathbf{w} = \{w_1, w_2, \ldots w_n\}$ represents word tokens in the corpus and $\mathbf{z} = \{z_1, z_2, \ldots z_n\}$

constitutes the topic assignment vector. Following Griffiths et al. [64], the posterior distribution $P(\mathbf{z}|\mathbf{w})$ can be written as Equation 3.3. Further, the joint distribution over words and tokens $P(\mathbf{w}, \mathbf{z})$ is factored as $P(\mathbf{w}|\mathbf{z})P(\mathbf{z})$ with the likelihood and the prior terms expressed as Equations 3.4 and 3.5.

$$P(\mathbf{z}|\mathbf{w}) = \frac{P(\mathbf{w}, \mathbf{z})}{\sum_{\mathbf{z}} P(\mathbf{w}, \mathbf{z})} \tag{3.3}$$

$$P(\mathbf{w}|\mathbf{z}, \beta) = \left(\frac{\Gamma(W\beta)}{\Gamma(\beta)^W}\right)^T \prod_{j=1}^{T} \frac{\prod_w \Gamma(n_j^{(w)} + \beta)}{\Gamma(n_j^{(\cdot)} + W\beta)} \tag{3.4}$$

$$P(\mathbf{z}, \alpha) = \left(\frac{\Gamma(T\alpha)}{\Gamma(\alpha)^T}\right)^D \prod_{d=1}^{D} \frac{\prod_j \Gamma(n_j^{(d)} + \alpha)}{\Gamma(n_\cdot^{(d)} + T\alpha)} \tag{3.5}$$

Exact inference is intractable and approximated via Markov Chain Monte Carlo (MCMC) using collapsed Gibbs sampling in the state space of topic labels[64]. The Markov chain is initialized by a random assignment of topic labels $\mathbf{z}$. Subsequent states are reached by sequentially sampling each variable z_i from a distribution conditioned on observed words and current assignment of all other topic labels. The desired conditional distribution is expressed as in Equation 3.6. Here $\mathbf{z}_{-i}$ refers to the current topic assignments of all other word tokens and $n_{-i}^{(\cdot)}$ is the count excluding the current assignment z_i.

$$P(z_i = j|\mathbf{z}_{-i}, \mathbf{w}, \alpha, \beta) \propto \left[\frac{n_{-i,j}^{(w_i)} + \beta}{n_{-i,j}^{(.)} + W\beta}\right]\left[\frac{n_j^{(d_i)} + \alpha}{n_{-i,\cdot}^{(d_i)} + T\alpha}\right] \tag{3.6}$$

$$\hat{\phi}_j^{(w)} = \frac{n_j^{(w)} + \beta}{n_j^{(.)} + W\beta} \tag{3.7}$$

$$\hat{\theta}_j^{(d)} = \frac{n_j^{(d)} + \alpha}{n_{\cdot}^{(d)} + T\alpha} \tag{3.8}$$

Equation 3.6 expresses the conditional distribution for topic label z_i assigned to word w_i as a product of the likelihood of word w_i under topic j and the probability of topic j in document d_i. Upon convergence after sufficient iterations topic labels are recorded and the maximum likelihood multinomial estimates for topics and topic proportions are obtained using Equations 3.7 and 3.8 respectively.

Inference for New Documents

The estimation of topic distributions $\phi^{(j)}$ is typically done once for a large corpus. Subsequently, for any new document, the topic proportions $\theta^{(d)}$ can be inferred via Gibbs sampling using Equation 3.9. Here, the topic assignments to words appearing in the document are sampled conditioned on the learned topic distributions $\hat{\phi}^{(w)}$ that remain fixed during Gibbs sampling.

$$P(z_i = j|\mathbf{z}_{-i}, \mathbf{w}) \quad \propto \quad \left(\hat{\phi}_j^{(w)}\right) \left(\frac{n_j^{(d_i)} + \alpha}{n_{-i,\cdot}^{(d_i)} + T\alpha}\right) \tag{3.9}$$

This concludes the mathematical review of LDA based topic model and inference. To summarize, topic models provide a low-dimensional representation of bag-of-words data capturing their thematic content via word-occurrences. LDA, being a probabilistic generative model allows us to compute the predictive likelihood of any document or image. We use this for detecting novel images and to retrieve similar images from a data base. This is presented in the next section.

3.3 Retrieving Thematically Similar Images

Given a large repository of images, our task is to determine images similar to a query image. Note that our application does not require precise geometric matches. Instead, we seek images similar in thematic content. Training the LDA model on the data base images yields topic distributions ϕ and topic proportions θ^d for images in the corpus, measuring the degree to which each topic occurs in the image. This forms a probabilistic topic-based representation $P(w|d, \theta^d, \phi)$ for each data base image.

Retrieving thematically similar images from the data base requires a ranking $P(d_b, \theta^b|d_a, \phi)$ for each data base image d_b given a query image d_a, where the probability is interpreted as the likelihood of being thematically similar to the query image, Equation 3.10. Under a uniform prior for all data base images this probability is proportional to $P(d_a|d_b, \theta^b, \phi)$, the likelihood of the query image d_a under the model θ^b for each corpus image d_b.

$$P(d_b, \theta^b | d_a, \phi) \quad = \quad P(d_a | d_b, \theta^b, \phi) P(d_b, \theta^b) / P(d_a) \tag{3.10}$$

Here we have adopted the language-modeling perspective to query-document similarity [114]. A document is similar to a query if its document model has a high predictive-likelihood of generating the query. The retrieval framework involves building a language model for each document in the corpus and ranking the documents using the likelihood of the query according to the document model.

Formally, for a given query image d_a presented as a sequence of N_a words, the likelihood of originating from document d_b is expressed as Equation 3.11, where the term w_j^a represents the j^{th} word in d_a. The likelihood of each word under the model $P(w_j^a | d_b, \theta^b, \phi)$ is evaluated by marginalizing over top+654ics using Equation 3.12. Please see Figures 3.8, 3.9 and 3.10 for an illustration.

$$P(d_a | d_b, \theta^b, \phi) \quad = \quad \prod_{j=1}^{N_a} P(w_j^a | d_b, \theta^b, \phi) \tag{3.11}$$

$$P(w_i^a | d_b, \theta^b, \phi) \quad = \quad \sum_{j=1}^{T} P(w_i^a | z_i = j, \phi) P(z_i = j | \theta^{(d)}) \tag{3.12}$$

Intuitively, the LDA-based model is not limited only to literal words in a document, but instead implicitly describes the document with many other topically-linked words. By mapping documents to a low-dimensional topic space representation, the model can semantically associate ones with similar topics, even though the documents themselves might have few words in common. Additionally, in [82],

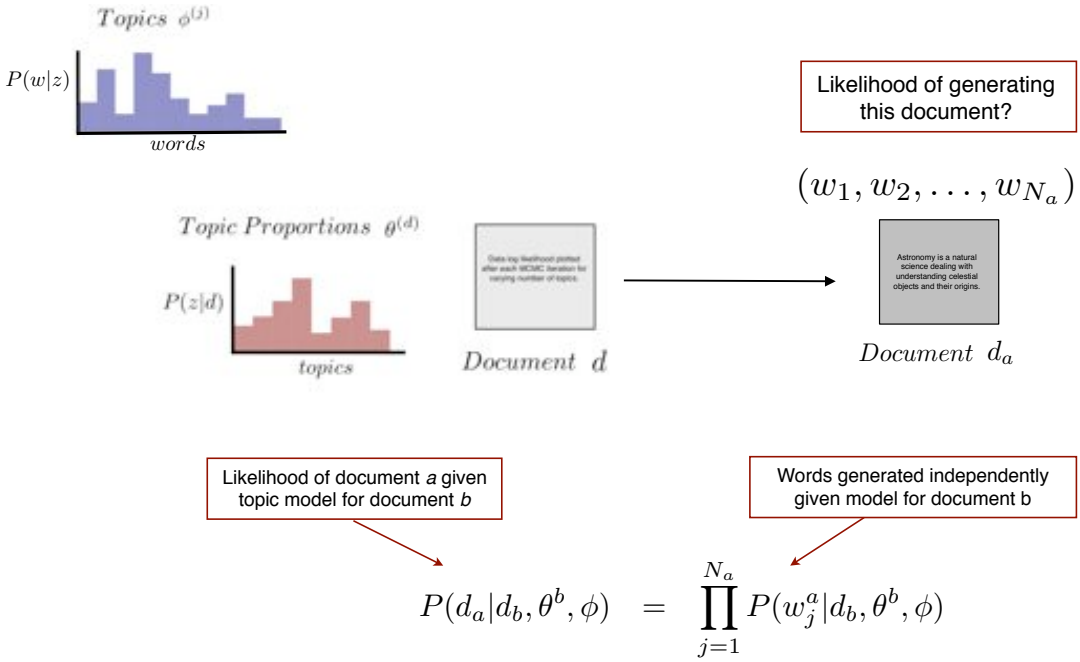


Figure 3.9: The likelihood of generating a collection of words given a topic-based model for a document can be obtained by multiplicative aggregation of the individual word-likelihoods under the independence (or bag-of-words) assumption.

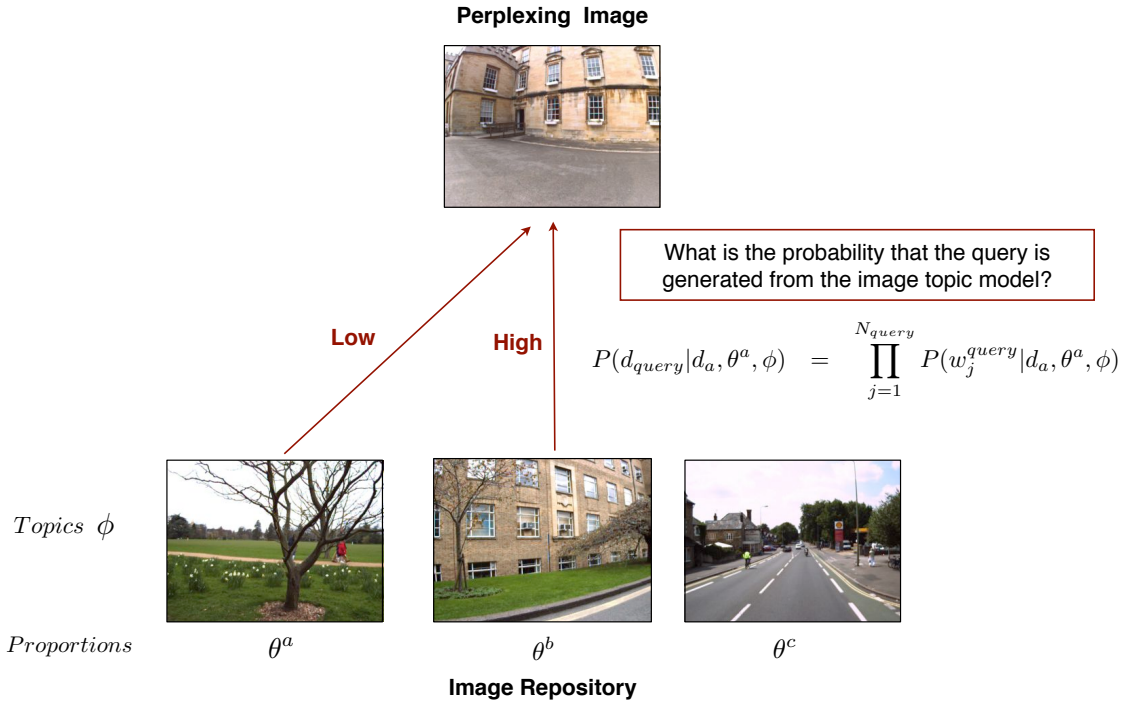


Figure 3.10: Language model based retrieval. A database image is similar to a query if its topic based model has a high predictive-likelihood of generating the query. In the example, due to topical similarity, the repository image containing the side of a building has a higher likelihood of generating the query image compared to an image taken in park area.

authors provide quantitative results to show that an LDA based approach performs better than simpler measures like cosine distance or Jenson-Shannon divergence on image retrieval tasks.

The query-likelihood probabilities provide a distribution over the corpus images according to thematic similarity. The most probable images from the corpus can be sampled to form the retrieved set. As discussed in [206], topic model based estimates must be smoothed for retrieval with a unigram model and Dirichlet smoothing (see [107] and [82]). Smoothing involves adjusting the maximum likelihood estimator to compensate for data sparseness. Further, it allows a more accurate estimation of

the document language model to explain the non-informative words in the query.

Let d_a denote the query document containing words w_a and M_b represent the probabilistic model for document d_b in the corpus. Assuming independence in word generation the likelihood of d_a under model M_b is given as:

$$P(d_a|M_b) = \prod_{j=1}^{N_a} P(w_j^a|M_b) \tag{3.13}$$

As presented earlier in Section 3.2, probabilistic topics ϕ and topic proportions θ_b are used to build a topic based model M_b^t for the document. Hence, word likelihood is given as:

$$P_T(w_j^a|M_b^t) = \sum_{j=1}^{T} P(w_j^a|z_j = k)P(z_j = k|\theta^b) \tag{3.14}$$

The accuracy of the document model for retrieval task can be improved by smoothing with a unigram model that assumes a multinomial distribution over word occurrences. As suggested by [200] and [82], we apply Jelinek-Mercer smoothing, linearly interpolating the LDA-model $P_T(w_j^a|M_b^t)$ with the unigram model $P_U(w_j^a|M_b^u)$ using weighting factor λ.

$$P(w_j^a|M_b) = \lambda P_T(w_j^a|M_b^t) + (1-\lambda)P_U(w_j^a|M_b^u) \tag{3.15}$$

Under the unigram language model for d_b, the maximum likelihood (ML) word estimates $P_U(w_j^a|M_b^u)$ are determined simply using relative counts.

$$P_{ML}(w_j^a|M_b^u) = \frac{c(w_j;d_b)}{\sum_{j=1}^{N_b} c(w_j;d_b)} \tag{3.16}$$

Here, $c(w_j;d_b)$ denotes the number of occurrences of word w_j in the document. The maximum likelihood estimator generally under estimates the probability of any word unseen in the document. Following [206], this estimate is smoothed using a fallback model that discounts probabilities for words seen in the text and assigns extra probability mass to unseen words. A unigram model learnt on the document corpus, D provides the fallback model given as:

$$P_{ML}(w_j^a|D) = \frac{\sum_{i=1}^{M} c(w_j;d_i)}{\sum_{i=1}^{M}\sum_{j=1}^{N_i} c(w_j;d_i)} \tag{3.17}$$

Bayesian smoothing with a Dirichlet parameter μ yields the following estimate for $P_U(w_j^a|M_b^u)$:

$$P_U(w_j^a|M_b^u) = \left(\frac{N_b}{N_b+\mu}\right) P_{ML}(w_j^a|M_b^u) + \left(1-\frac{N_b}{N_b+\mu}\right) P_{ML}(w_j^a|D) \tag{3.18}$$

For experiments in this paper, the smoothing parameter values were set to $\lambda = 0.8$ and $\mu = 50$ consistent by values suggested in literature [82] and [206].

3.4 Is an Image Perplexing?

We now address the task of determining the novelty (or redundancy) of an observed image given an image corpus. As discussed in the previous section, the LDA gen-

erative model allows us to calculate the document likelihood $P(d_a|d_b, \theta^b, \phi)$ via a topic model representation, Equation 3.11. Note that the computed likelihood is dependent on the query length. We seek a length normalized measure such that the measure computed for a series of images is not biased by the number of visual words that would appear in the image.

We compute the per-word predictive-perplexity [74] of the observed image given the topic-based model of an image from the corpus using Equation 3.19. Perplexity indicates the uncertainty in predicting a single word. Intuitively, a perplexity per-word of k implies that to generate each observed word, the model is as confused as if it were picking uniformly from k equally likely options [102]. In our setting, the maximum possible value of perplexity can equal the vocabulary size W. Note that minimum perplexity value can be 1 indicating no model uncertainty in word prediction. Importantly, a model that better captures word co-occurrences requires fewer possibilities to select words, yielding a *lower* perplexity per-word for new data.

$$\text{perplexity}_{\text{pw}}(d_a|d_b, \theta^b, \phi) \quad = \quad \exp\left\{\frac{-\log P(d_a|d_b, \theta^b, \phi)}{N_a}\right\} \tag{3.19}$$

The next step is determining the redundancy of a new unseen document in relation to the corpus. A new document is redundant if its information content is covered by documents present in the corpus. Similarly, documents highly dissimilar to the ones seen previously contain new information and hence considered novel. Zhang et al. [208] explored redundancy detection in the context of adaptive information filtering. We extend and adapt the framework and define the redundancy of document d_i vis-à-vis a document d_j as the negative of the predictive perplexity as given in Equation 3.20 where θ^j is the topic proportion estimated for d_j. The

redundancy of an observation given a document corpus, D is determined by setting it equal to the maximally similar value in all $\mathrm{R}(d_i|d_j)$ as given in Equation 3.21.

$$\mathrm{R}(d_i|d_j) = -\mathrm{perplexity}_{\mathrm{pw}}(d_i|d_j, \theta^j, \phi) \tag{3.20}$$

$$\mathrm{R}(d_i|D) = \max_{d_j \epsilon D} \mathrm{R}(d_i|d_j) \tag{3.21}$$

$$d_{\mathrm{Novel(Q)}} = \arg\min_{d_j \epsilon Q} \mathrm{R}(d_i|D) \tag{3.22}$$

Finally, given an observed image set Q, the most novel image $d_{\mathrm{Novel(Q)}}$ pertaining to an existing corpus D is selected as the minimally redundant sample, Equation 3.22. Intuitively, the novel images thus identified are the ones least explained by the image corpus. Hence, we seek images similar to the perplexing ones and augment those to the corpus forming an improved representation.

In the following section, we bring the topic model based retrieval and novelty detection systems presented above and present an algorithm to actively evolve an appropriate image based representation of robot's workspace for improving appearance based topological mapping.

3.5 Application to Topological Mapping

We now apply our framework to an appearance based mapping algorithm FAB-MAP [34, 35]. FAB-MAP is a loop closure detection system that allows a navigating robot to determine if the current observation comes from a known location in its map or from a new one. Our goal will be to show that over time we can improve its performance with experience. A detailed overview of FAB-MAP appears in Section 2.1.3. Here we summarize the key FAB-MAP Bayes equations and discuss the significance of the sampling set. This is followed by presenting the algorithm to construct a sampling set over time with to improve mapping performance.

3.5.1 Recap of FAB-MAP Formulation

At time t, the robot's workspace has n_t locations $\mathcal{L}^t = \{L_1, \ldots, L_{n_t}\}$ where each location L_i has an associated appearance model represented by a distribution over appearance words. When the robot collects a new observation Z_t, we compute the distribution over locations given the observation $p(L_i|\mathcal{Z}^t)$ formulated as a recursive Bayes estimation problem:

$$p(L_i|\mathcal{Z}^t) = \frac{p(Z_t|L_i, \mathcal{Z}^{t-1})p(L_i|\mathcal{Z}^{t-1})}{p(Z_t|\mathcal{Z}^{t-1})} \tag{3.23}$$

where $\mathcal{Z}^t$ is the set of all observations up to time t, $p(Z_t|L_i, \mathcal{Z}^{t-1})$ is the observation likelihood of the observation given the location L_i. The normalization term $p(Z_t|\mathcal{Z}^{t-1})$ is the total likelihood of the observation, Z_t. An observation can come from the set of locations currently in the robot's map (M) as well as the set of all previously unknown locations ($\overline{M}$). Hence, the denominator is expressed as:

$$p(Z_t|\mathcal{Z}^{t-1}) = \sum_{m \in M} p(Z_t|L_m)p(L_m|\mathcal{Z}^{t-1}) + \sum_{u \in \overline{M}} p(Z_t|L_u)p(L_u|\mathcal{Z}^{t-1}) \tag{3.24}$$

The second term involves summation over all unmapped places and is approximated via sampling location models L_u [34]. This is instead approximated by sampling observations Z and using them to form location models. Observation likelihood $p(Z_t|L_u)$ is evaluated for each sample and Equation 3.24 is expressed as:

$$p(Z_k|\mathcal{Z}^{k-1}) \approx \sum_{m \in M} p(Z_k|L_m)p(L_m|\mathcal{Z}^{k-1}) + p(L_{new}|\mathcal{Z}^{k-1}) \sum_{u=1}^{n_s} \frac{p(Z_k|L_u)}{n_s} \tag{3.25}$$

Here, n_s is the number of samples used and $p(L_{new}|\mathcal{Z}^{t-1})$ is the prior probability of being at a new place, uniformly distributed among samples. This yields the total probability of the observation originating from a place not in the map. The resulting distribution over locations is used to decide if a new location is to be added to the map or not.

3.5.2 The Sampling Set

The sampling set forms the robot's present compact representation of the workspace visual appearance. It represents a Monte Carlo sampling-based approximation to the true distribution over visual appearance of the operating environment.

The sampling set is critical for performance since it ameliorates the perceptual aliasing problem: the fact that different parts of the environment appear the same to robot's sensors. e.g., similar looking foliage and brick walls appear commonly

while navigating outdoors. A representative sampling set would possess several examples of such common and repetitive visual features. Hence, the normalization step distributes the probability mass preventing a false loop closure declaration.

In [34], the authors describe the sampling set construction process as randomly sampling images from traversal data sets obtained from previous runs of the robot. Specific examples of common features (potential perceptual aliasing cases) were explicitly added through visual inspection to make the sampling set representative of the application environment. In the next section we forego the idea of a static and hand-crafted sampling set and present dynamic representation that evolves incrementally with novel experience.

3.5.3 Incremental Sampling Set Construction

We now present an algorithm to actively seek data and evolve a representative sampling set through introspection, by harnessing the topic model based saliency detection and thematic retrieval presented earlier in Sections 3.3 and 3.4.

We assume that the robot has access to a large repository containing imagery or visual topological maps collected during past exploration of the environment. The repository can be considered as the entire visual experience of the robot from which an appropriate and compact sampling set is to be extracted forming an onboard representation to be used for topological mapping. Loop closure images and near-duplicate images captured from the same location (while the robot is stationary) are excluded from the database and hence the image samples can be considered independent and identically distributed. Visual topic distributions ϕ are estimated using the database images using the unsupervised MCMC procedure outlined in Section 3.2. Topic proportions θ are inferred using learnt topics for all images in the

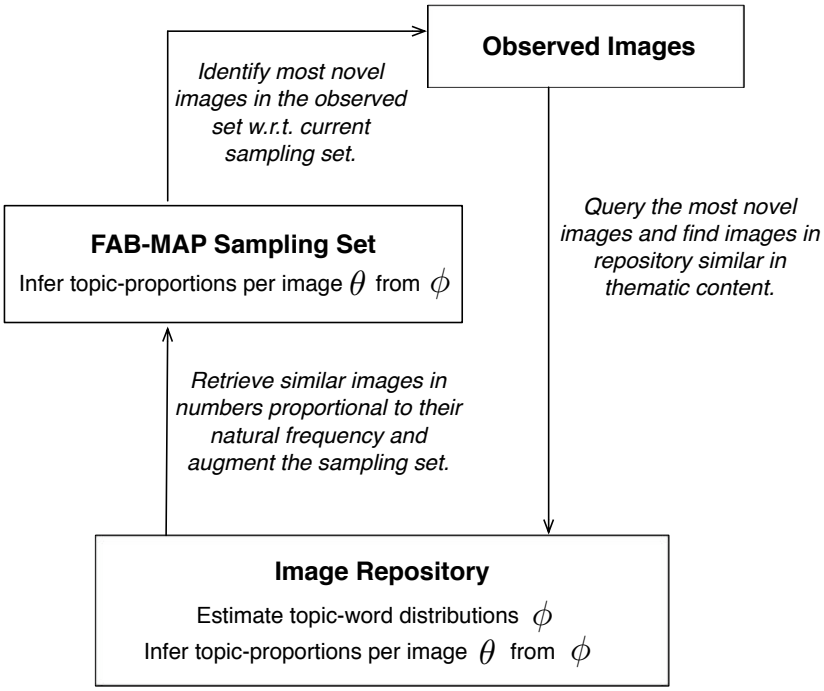


Figure 3.11: Sampling set evolution. Topics ϕ are estimated from an image repository and used to infer topic proportions θ for sampling set and repository images. Online, the robot collects observations from which most novel images are identified. These perplexing images are searched in the repository for thematically similar images which are retrieved (respecting the underlying distribution) and augmented to the evolving sampling set.

database and the sampling set.

During online exploration, the robot collects imagery and uses FAB-MAP for topological mapping with the environment appearance characterized collectively by the onboard current sampling set. Offline, the robot executes an introspection-improvement loop. It first identifies the most perplexing (or minimally redundant) images from the observed image set acquired during exploration conditioned on the current sampling set. The perplexing images from the environment are the ones least-represented (or explained) by the topic model based sampling set representation.

The next task is to improve the current sampling set by augmenting images thematically similar to the perplexing ones. This is accomplished by querying the most perplexing images in the image repository and augmenting the sampling set. The cycle can be iterated after each data collection run by the robot. Figure 3.11 illustrates the improvement cycle.

Algorithm 3.1 outlines the key steps in the procedure. The offline stage consists of feature extraction and vocabulary learning followed by topic learning and inference for the database set of images. The online or incremental phase consists of interleaved operation and introspection stages. The sampling set is initialized using a single image picked randomly from the database. Within each improvement cycle, the perplexing images are determined using Equations 3.21 and 3.22. For each selected novel image, a topic model based similarity to each database image is estimated using Equations 3.11 and 3.19. This yields a multinomial distribution over the set of database images. Thematically similar images in the repository are retrieved respecting their natural occurrence frequency in the environment (as represented in the database). For example, images containing foliage features are very common

Algorithm 3.1 Incremental Sampling Set Evolution

```
// OFFLINE
// Database Image Corpus, DB
// Vocabulary, V
// Vocabulary Size, VS
// Number of Topics, T
// Topic Distributions, φ
// Topic Proportions, θ
// Features and vocabulary generation

    V ← Vocab_Generation(VS, DB)

// Topic and Topic proportions estimation

    φ ← Topic_Distribution_Est(T, DB, VS)
    θ ← Topic_Proportions_Est(DB, φ)

// ONLINE
// Sampling Set, SS
// Observed Image Set, ObS
// Initialize sampling set.

    SS[0] ← Random image from DB.

for each improvement cycle, t

    // Operate
        ObS[t] ← Collect data from robot.
        Use FAB-MAP on ObS[t] with SS[t].
    // Introspection
    for each image d_i ∈ ObS[t]
        // Estimate topics proportions
        θ^{d_i} ← Topic_Proportions_Est(d_i, φ)
        // Estimate redundancy
        R(d_i|SS[t]) using eq:3.21
    end
    // Determine the most novel image
        d_Novel ∈ ObS[t] using eq:3.22
    // Retrieval
    for each image d_j ∈ DB
        // Estimate likelihood using eq:3.11,3.19
        Mult_dist[j] ← p(d_Novel|d_j, θ^j, φ)
    end
        Ret_samples ← Sample from Mult_dist
    // Improve sample set
        SS[t+1]←SS[t] ∪ Unique(Ret_samples)

end
```

compared to rare images of sign boards. Hence, we retrieve a greater number of images when the likelihood distribution is more uniform. This is accomplished by sampling the likelihood multinomial a fixed number of times and selecting only the unique samples, thereby accepting fewer samples when the distribution is peaked.

The retrieved samples are then augmented forming a new sampling set. Hence, we create a compact sampling set targeted to the workspace environment of the exploring robot. In essence, each iteration of the introspection loop improves the sampling-based approximation to the distribution over the visual appearance of the environment. This impacts the estimation of the *new* place probability guarding against perceptual aliasing. This is achieved by picking the most uncertain or perplexing images that promise to be most informative given the current observations and the sampling set[2]. We assume that the robot is continually exploring and encounters varied workspaces in each traversal. Hence, we do not seek to discriminatively optimise for loop closure detection for a particular traversal, rather we improve the sampling set distribution which is expected to generalise to subsequently encountered environments. In the next section, we present experimental results using data collected from a mobile platform.

3.6 Results

3.6.1 Data set and Platform

We tested the system on data collected from a mobile robot. The characteristics of data sets used in evaluation are summarized in Table 3.1. A collection of 2800 images from 28km of urban streets and parks using the robot's camera formed

[2]Note that there is scope other criteria like information gain etc.

Table 3.1: Summary of data sets used in experiments.

Data set	**No. of Images**	**Length**	**Features per scene (median)**	**Loop closures**	**Environment**
Urban	2849	28 km	160	—	Streets, parks and motorways
New College	2146	1.9 km	89	564	Medieval buildings and gardens
City Center	2474	2 km	116	1121	Roads with vehicles/pedestrians

the database. Images were captured perpendicular to the robot's motion and did not overlap. Images were converted to a bag-of-words representation [173] by first extracting SURF features [7] and later quantizing them against a fixed vocabulary of size $11K$.

3.6.2 Learning Visual Topics

Topic models were estimated on this data set using Gibbs sampling as discussed in Section 3.2. In order to determine the number of iterations required to ensure MCMC convergence to the target distribution we used data log-likelihood as a measure [64]. Figure 3.12a plots log-likelihood with each iteration for varying number of topics T. Dirichlet priors were set to $\alpha = 50/T$ and $\beta = 0.1$. For the runs with topics 3, 50 and 200 the log-likelihood stabilizes within 200 iterations. However, convergence for the run with 500 topics is much slower and stabilizes after 600 iterations. We also experimented with observing multiple runs of the Markov chain and convergence rates were similar. Hence, the number of sampling iterations was set to 600.

The LDA model requires the number of topics to be specified. This can be cast as a model selection problem. The appropriate number of topics modeling the data set was determined by maximizing the data likelihood given topics $P(\mathbf{w}|T)$ assuming

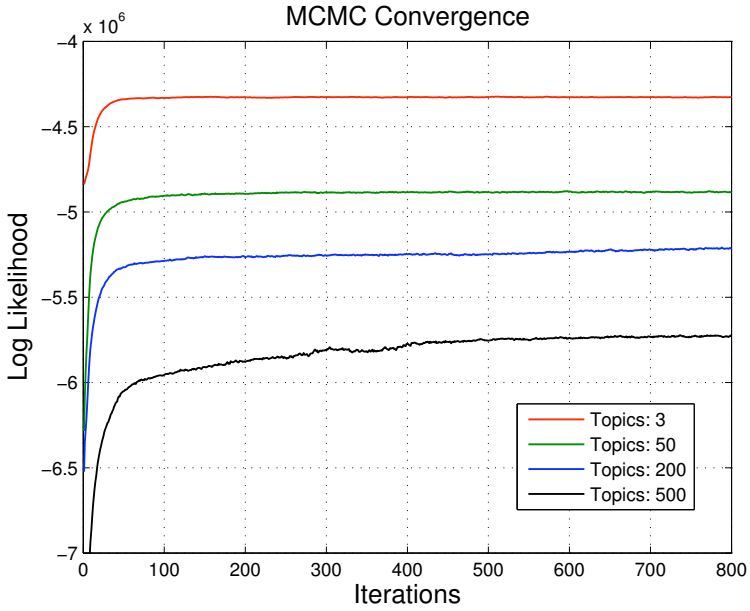
a uniform prior on the number of topics. As suggested in [64], $P(\mathbf{w}|T)$ can be approximated using $P(\mathbf{w}|\mathbf{z}, T)$ where topic-labels $\mathbf{z}$ are obtained from the Gibbs sampler upon convergence. Figure 3.12b plots the result. The data log-likelihood upon convergence for the data set peaks with 50 topics.

Figure 3.13 shows examples of visual topics discovered on the urban data set. The most probable visual words characterizing a topic occur along rows and typical occurrences are shown along columns. Co-occurring words frequently get assigned to a particular topic. Figure 3.13a illustrates a topic that represents visual words commonly occurring on foliage and trees. Topics in Figures 3.13b and 3.13c capture edge or corner type features appearing on walls. Figure 3.13d illustrates a topic modeling words seen on the horizon line constituting the top of walls and houses.

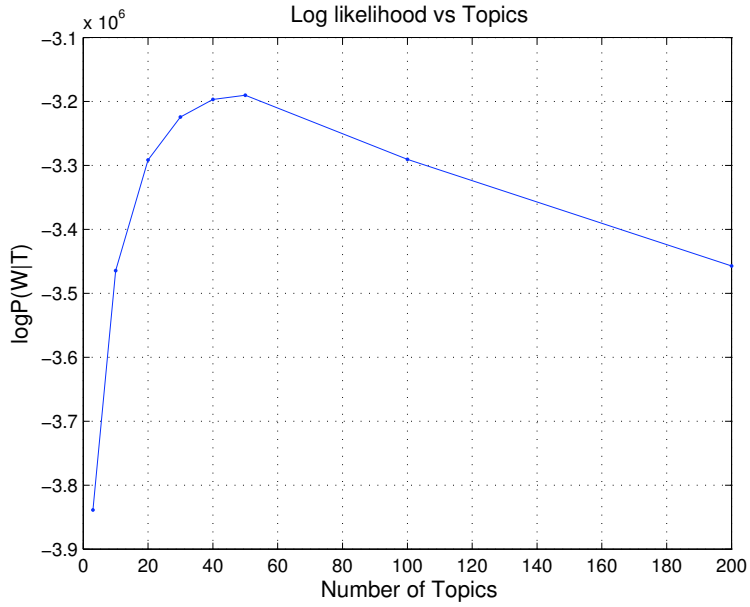
Figures 3.13e and 3.13f illustrate the effect of using very few (under-fitting) and too many topics (over-fitting) for image representation. Using very few topics, $T = 3$ leads to very general topics. Figure 3.13e shows an example where the most probable visual words constitute features foliage, windows and walls. On the other hand, using a very large number of topics, $T = 200$ gives rise to very specific topics, Figure 3.13f, where the topic particularly captures regular features commonly observed on walls.

3.6.3 Retrieval of Thematically Similar Images

Figure 3.14 presents instances of topic-model based image retrieval. The query image is highlighted and the most similar images retrieved from the database are shown subsequently. Note that the system returns images that are thematically similar as opposed to strict geometric matches. This is the key to our approach. While querying to find images similar to a perplexing image (say an image of a building) we do not seek exact instances of the same building. Rather, we aim to retrieve examples



(a) Log likelihood convergence during MCMC.



(b) Log likelihood Log likelihood variation with topics.

Figure 3.12: (a) Iterations for MCMC convergence. Data log-likelihood plotted after each MCMC iteration for varying number of topics. For 3, 50 and 200 topics the chain converges within 200 iterations. Convergence for 500 topics is much slower and stabilizes after 600 iterations. Hence, the number of Gibbs sampling iterations was set to 600. (b) Selecting the number of topics, T for the urban data set. The plot shows the data log-likelihood upon convergence for varying number of topics. The data log-likelihood peaks for 50 topics. Hence, we set $T = 50$. A uniform prior is assumed on topics.



(a) $T = 50$



(b) $T = 50$



(c) $T = 50$



(d) $T = 50$



(e) $T = 3$



(f) $T = 200$

Figure 3.13: Illustrative visual topics discovered on the urban data set (a-d). Five most probable visual words for each topic appear along rows and common instances occur along columns. The number of topics, T is 50 for figures (a-d). A topic capturing visual words co-occurring on foliage and trees is shown in (a). Topics shown in (b) and (c) represent edge or corner features appearing on walls. Words observed on the top of foliage and houses constitute topic (d). Using too few topics $T = 3$ leads to very general topics as shown in (e). Note that the most probable words consist of features on walls, foliage, windows etc. As a contrast, using too many topics, $T = 200$ yields highly specific topics as illustrated in (f). Here, the topic specifically captures regular features commonly found on walls.



(a) Retrieval of thematically linked images of buildings.



(b) Retrieval of similar parkland and foliage images.

Figure 3.14: Illustrative results for topic-model based image retrieval. Query image is shown top left and five most similar images from the database are shown subsequently. Note that the retrieved images represent a common visual theme.

of the class of similar looking buildings which better represents a common mode of visual appearance and is accomplished via a low-dimensional topic representation. By mapping visual features on buildings to a topic that probabilistically models co-occurring words on buildings e.g. Figure 3.13f, other relevant images containing similar visual features are obtained.

3.6.4 Detection of Novel Images

We tested the novelty detection component of the system by first learning a topic model on a data set where visual appearance was restricted and later estimating the perplexity of images from another data set possessing a wider variation in appearance including imagery similar to the topic data set.

The New College data set [174] was used for topic learning. This data set consists of 2146 images from a 2.1 km traversal within a typical college in Oxford. The visual appearance is restricted to medieval buildings and parkland areas. Typical images are shown in Figure 3.15a. A restricted set of 564 images were used from the data set after removing loop closure pairs.

Conditioned on the learnt topic model, images from the urban data set were presented to the system and redundancy values were computed using Equation 3.21. Images of brick walls and foliage were found to be most redundant due to high similarity to visual themes in the New College Data set as shown in Figure 3.15b. Images of sign boards, road vehicles and modern, regular shaped buildings, were found to be most novel or least redundant, Figure 3.15c. Since no such examples are present in the New College Data set, the learnt topic model is highly perplexed to encounter these feature sets typically found outdoors.

3.6.5 Incremental Improvement in Loop Closure Detection

In this section, we evaluate the improvement in FAB-MAP loop closure detection performance using the incremental learning algorithm. We used the urban data set as the image database used for constructing the sampling set. The City Center and New College data sets (introduced previously) were used as the observed image set,



(a) Typical images from the New College Data set used for topic learning.



$R = -1.17 \times 10^3$ $R = -1.32 \times 10^3$ $R = -1.18 \times 10^3$

(b) Least novel (most redundant) images in urban data set.



$R = -7.86 \times 10^3$ $R = -6.54 \times 10^3$ $R = -1.05 \times 10^4$

(c) Most novel (least redundant) images in the urban data set.

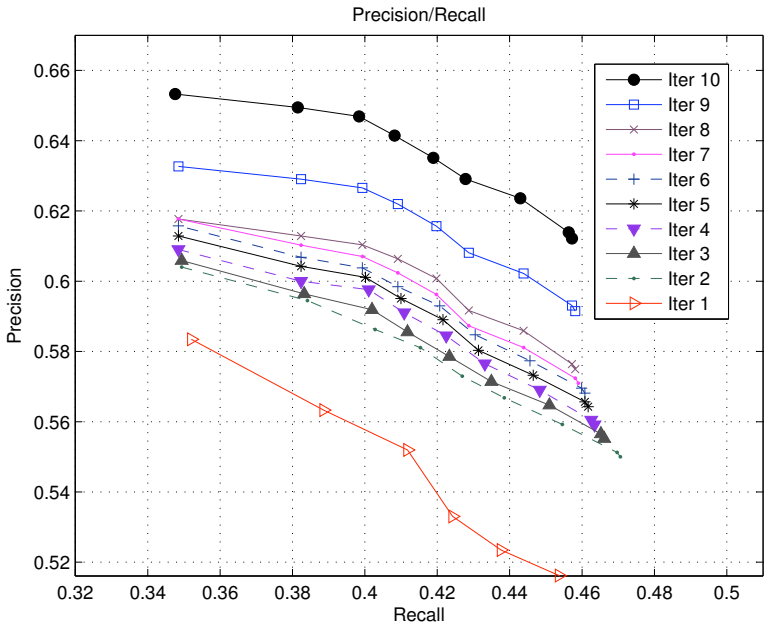
Figure 3.15: Detecting novel images. Topics were learnt on the New College Data set with typical images of medieval buildings and parklands, shown in (a). For images in the urban data set, redundancy values were computed and listed below each image. (b) Brick wall and foliage images were least novel due to high similarity to visual themes in the New College Data set. (c) Images of road signs, cars and modern buildings were found most novel.

representing the imagery collected online by the mobile robot. The City Centre data set is 2km in length, possessing 2474 images and provides a challenging setup for image matching due to considerable scene change in images. This data set does not overlap with the urban data set used for learning topics.

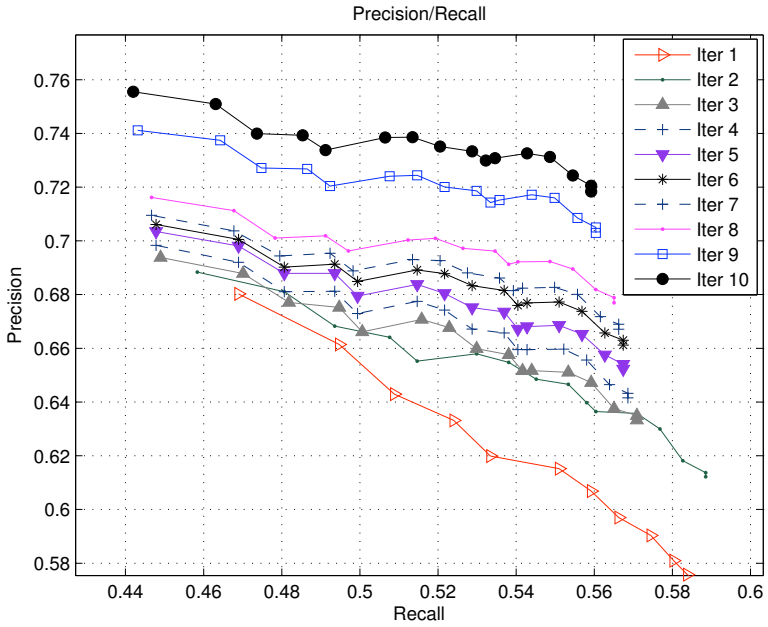
The observed data set was split into epochs or improvement cycles and presented sequentially to the incremental sampling set construction algorithm. The sampling set was initialized with a single randomly-selected image from the database. In the experiments, 10 improvement cycles were executed. Each iteration consisted of an operation phase: using the current sampling set for loop-closure detection using FAB-MAP on the observed data set followed by an introspection phase: augmentation to the sampling set using targeted data retrieval triggered by perplexing images. Within each epoch, the top 15% most perplexing images were identified from the acquired imagery for thematic retrieval. A maximum of 50 similar samples per query were retrieved. Note that this also denotes the number of times images are sampled from the similarity multinomial distribution over data base images where the unique samples are augmented to the sampling set.

FAB-MAP requires the visual detector model as an input. Following [34], the false positive rate and the false negative rates were set to $p(z_i = 1|e_i = 0) = 0$ and $p(z_i = 0|e_i = 1) = 0.39$ respectively. Ground truth was generated using GPS and was determined via visual inspection in places where the signal was intermittent. In order to test the core inference component, the prior probability over map locations was left uniform.

Loop closure performance was determined using precision-recall curves obtained after each improvement cycle. Precision and recall values were calculated by varying the probability threshold at which loop closures are accepted. In this setting,



(a) City Center data set.



(b) New College data set.

Figure 3.16: Asymptotic increase in FAB-MAP loop closure detection performance. The precision-recall curves are plotted after each improvement cycle. The curves move upwards with each iteration (indicating improved performance) as an appropriate sampling set is evolved. Maximum precision increases from (a) 58.3% to 65.3% for City Center data set and from (b) 68.0% to 75.5% New College data set respectively.

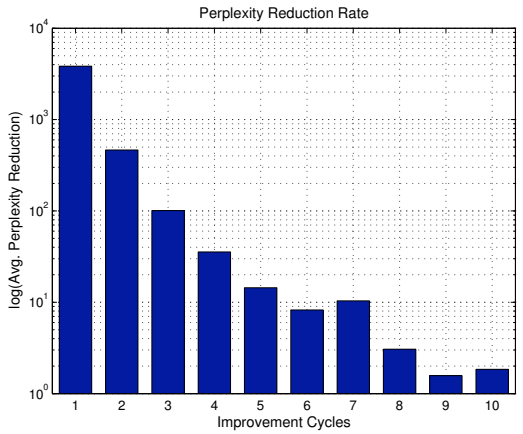
precision refers to the fraction of correct loop closures from those exceeding the acceptance probability threshold and similarly, recall is the proportion of true image-to-image matches that exceed the probability threshold. Figure 3.16 plots the resulting precision-recall curves after each iteration. For both data sets, the curve moves upwards in each improvement cycle indicating an improved performance as a better sampling set is evolved by the algorithm. Maximum precision increases from 58.3% to 65.3% for the City Center data set and from 68.0% to 75.5% for the New College data set.

Next, we evaluate the reduction in perplexity (or equivalent redundancy increase) in each epoch using Equation 3.26. This is determined as the difference in the redundancy estimates of the imagery collected within an epoch; obtained using the current and the subsequently improved sampling set after each learning iteration. The cumulative difference is normalized by the size of the epoch.

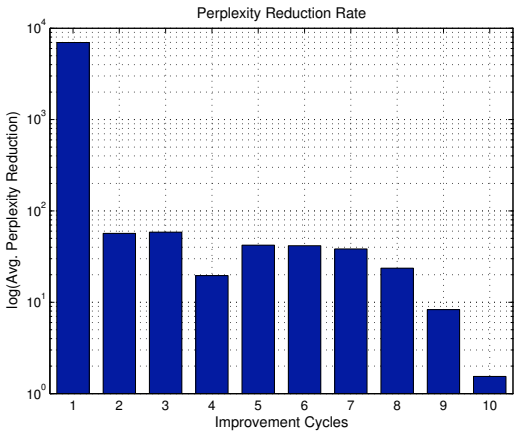
$$\text{Perplexity Reduction}(t) = \sum_{d_i \in \text{ObS}_t} \frac{\text{R}(d_i|\text{SS}_{t+1}) - \text{R}(d_i|\text{SS}_t)}{|\text{ObS}_t|} \tag{3.26}$$

Figures 3.17a and 3.17b illustrate the results. Note the logarithmic scale on y-axis. For the City Center data set, perplexity reduction is higher during the initial iterations and decreases in later epochs. For the New College data set, the decrease in perplexity reduction is more gradual before tapering in later epochs. Overall, the system is less perplexed over time.

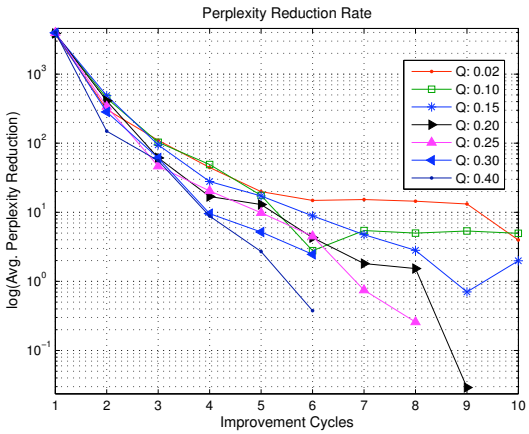
In Figures 3.17c and 3.17d perplexity reduction curves are plotted by varying the fraction of most perplexing images queried in each epoch from 0.02 till 0.40. Using a larger query fraction results in a more significant decrease in perplexity reduction in each epoch, consequently saturating faster. The perplexity reduction curves appear



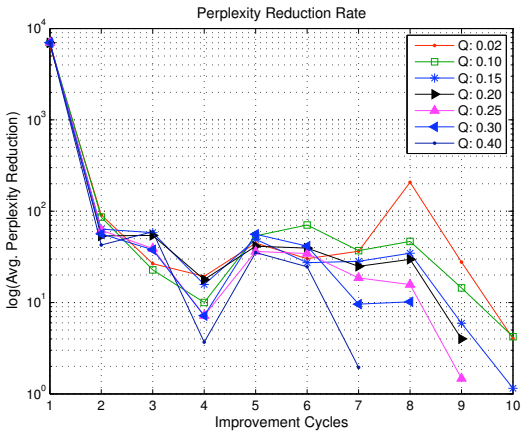
(a) City Center data set.



(b) New College data set.



(c) City Center data set.



(d) New College data set.

Figure 3.17: (a-b) Perplexity reduction in each epoch for the City Center and New College data set using Equation 3.26. Top 15% most perplexing are queried in the data base. Perplexity reduction is higher during initial iterations shows an overall decline in later epochs for City data set. The decrease is more gradual for New College data set. (c-d) Perplexity reduction in each improvement cycle by varying the fraction of most perplexing images in each epoch, Q as 0.02, 0.10, 0.15, 0.20, 0.25, 0.30 and 0.40. Overall, using a higher query fraction results in a faster decline in perplexity reduction per epoch. The perplexity reduction curves appear less sensitive to the query fraction for the New College data set which can be attributed to stronger correlations between query samples. Note the log scale on the y-axis. Points corresponding to epochs where no reduction in perplexity is observed have not been plotted on the log scale.

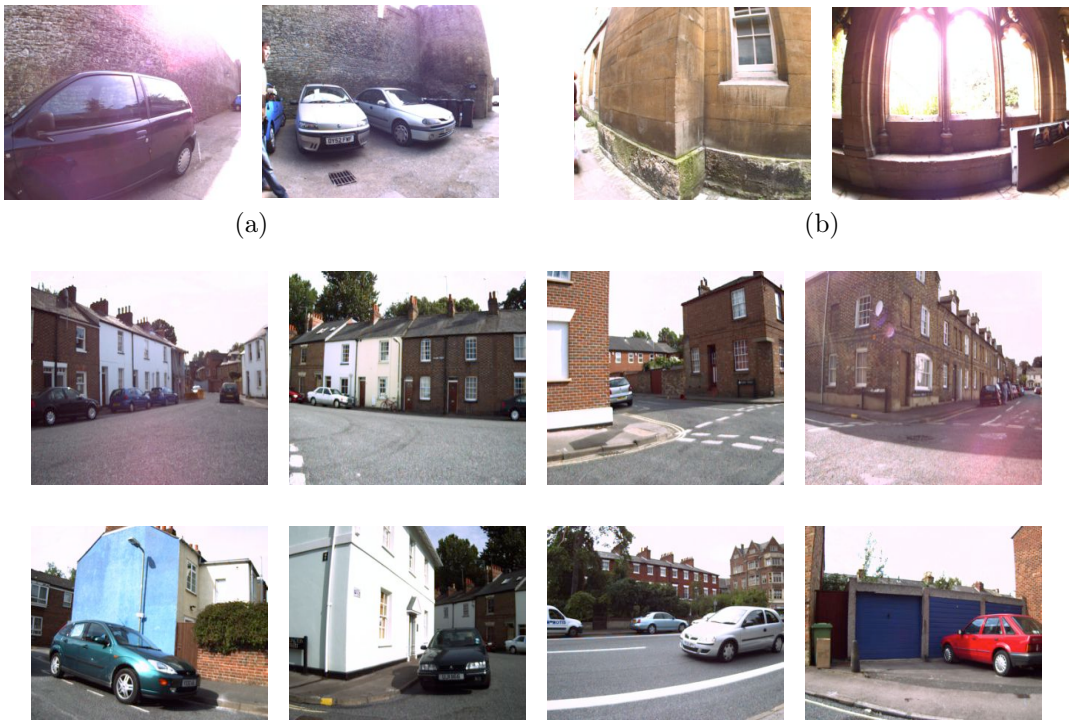


Figure 3.18: (a-b) Loop closure pairs declared as false positives prior to an improvement cycle but correcly labeled as true negatives using the augmented sampling set. (c) Representative images added to the sampling set in the current improvement cycle. Note the images of buildings and cars. Inclusion of these images indicates that these features are common in the operating environment and hence discards false positives arising due to perceptual aliasing.

less sensitive to the query fraction for the New College data set compared to the City Center data set. The New College data set possesses a more restrictive visual appearance compared to the City data set. Hence, the batch of highly perplexing query samples from an epoch are likely to be strongly correlated. Consequently, the union of retrieved samples can provide similar and overlapping information.

Figure 3.18 presents an example of incorrect loop closure declarations that are remedied by an augmented sampling set after executing an improvement cycle. Figures 3.18a and 3.18b show two loop closure pairs declared as false positives in the

previous cycle but correctly labeled as true negative loop after the improvement cycles. Figure 3.18c shows representative images that are augmented to the sampling set in the current cycle. Note the images containing buildings, windows and cars. Inclusion of these examples in the sampling set allows the systems to learn that these feature sets are common in the environment and hence correctly discards false positives that may arise due to perceptual aliasing.

3.7 Chapter Summary

We have shown how a robot can, through introspection and then targeted data retrieval, improve its own navigation performance in the context of topological mapping and localization. It is a step in the direction of lifelong learning and adaption and was motivated by the desire to build robots that have plastic competencies which are not baked in. They should react to and benefit from use.

We have considered a particular instantiation of this problem in the context of place recognition using the FAB-MAP algorithm. We used LDA based topic models to enable the calculation of a measure of image perplexity viz-a-viz a sample set which is supposed to be representative of all scenes experienced by the robot. This measure guides a retrieval of additional images that share topics with the confusing ones. The sample set is thus extended to better explain the images that the robot is seeing at run time - in a sense, the robot is customizing itself to its surroundings.

Although we have demonstrated adaption in the context of a particular problem this is part of a bigger picture - one in which lifelong learning is achieved by robots actively and continually seeking out training data or experience as a result of on going use.

This chapter concludes our work on topological mapping and navigation. In the next chapter we investigate the problem of generating online visual summaries that incrementally organize the cumulative visual workspace experience (e.g., a topological map) of an exploring robot.



Galerie de Vues de la Rome Moderne (Picture gallery with views of modern Rome) Giovanni Panini, 1758 A.D.

Chapter 4

Online Visual Workspace Summaries

Consider the following: a robot is sent out into the world day after day continually taking pictures of its environment; implicitly accruing an ever richer picture of its world. It is gaining experience. The question we ask in this chapter is how should that robot summarize its day, its week, or even its working "life time" when asked? Immediately, it is interesting to think of this as the flip side to the vast amount of research which exists on metric workspace mapping. That corpus of work summarizes the experience of a mobile robot metrically - it produces crisp, sometimes almost architectural drawings of the robot's workspace.

In this work however, we swap metric summaries for visual summaries. We want the robot to produce a story board of canonical images which capture the essence of the robot's visual experience - illustrating both what was ordinary and what was extraordinary. Here, we systematically address this question in a way that scales well with time and variation of experience. We seek a summary that evolves incrementally with the novelty of data - it should grow with saliency of experience and not merely duration. To be sure, if the robot stood still for a year in static

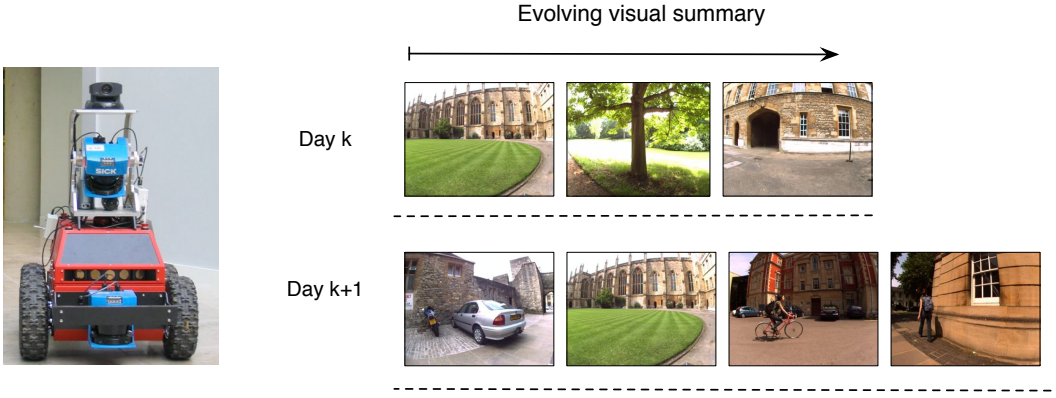


Figure 4.1: The task is to allow a continually exploring robot to incrementally generate a visual summary of its traversal using canonical images that capture the common as well as the salient aspects of its workspace.

world we would not welcome a lengthy precis!

At a high level we proceed in the following way. Each image is characterized as a mixture of visual topics mapping to a point in topic vector space. We incrementally organize these images using an online graph clustering technique. The structure of this graph is used to generate a visual summary of a robot's experience. Importantly, the graphical organization evolves over time as new imagery is collected by the robot. We show that this naturally results in an ever-improving workspace summary.

This chapter is arranged as follows. The next section overviews related literature. Section 4.2 concerns the use of topic space representation for images. This is followed by a discussion on the star clustering algorithm used for online organization in Section 4.3. The visual summary generation algorithm is laid out in Section 4.4. Detailed experimental results appear in Section 4.5 followed by concluding remarks in Section 4.6.

4.1 Related work

4.1.1 Canonical Views Selection

A few research efforts have addressed the problem of canonical-view selection or scene summarization for a given collection of images aimed towards recognition, retrieval, 3D reconstruction or browsing applications. For a collection of 2-D views of 3-D objects and a similarity metric based on object silhouettes, Denton et al. [40], select a bounded canonical set of views that are as dissimilar to each other while being as similar as possible to non-canonical views. The algorithm aims at view selection for view-based 3-D object recognition application and relies on semi-definite programming relaxation (extending [39]) with the lower and upper bounds on the cardinality of the canonical set specified. In a related work, Hall and Owen [67] define canonical views of objects as the most unique views among all views, subject to orthogonality constraints, essentially extracting the rarest datum or outliers in the data set. They compute a lower dimensional eigen model for a set of images (vector of normalized greyscale values) and extract the least likely view that is orthogonal to the previously selected views in an incremental manner.

The work of Simon et al. [169] addresses canonical views selection for summarizing large scenes (instead of single objects), using images from multi-user online photo collections (like Flickr.com). Instead of inferring views from geometry or from a set of uniformly sampled views, the approach relies on the population of photographers to provide the likelihood distribution over camera viewpoints and focusses on computing clusters and peaks in this distribution using visual features and tag-data. In [106] Li et al. present a system that combines 2D appearance and 3D geometric constraints to extract scene summaries and build models from unstructured inter-

net image collections. The approach involves extracting global *gist* descriptors [130] from images, over-clustering using k-means followed by geometric verification to retain only those clusters whose images share common 3D structure. Each remaining cluster is represented with a single iconic image with geometric relationships represented by an *iconic scene graph*.

Ladikos et al. [100] build *region graphs* from large image collections based on the extent of overlapping sub-regions (representing evidence for objects) instead of requiring the detailed 3D structural information in the scene. Using the graph a reduced image set is constituted by greedily picking images with higher number of overlapping regions with neighbouring images. In [160] Schaffalitzky and Zisserman present a technique for matching unordered image sets using a combination of image invariants and multiple-view matching techniques. Clusters or connected sub-sets of views are formed yielding multiple view tracks allowing a 3D reconstruction of cameras and points. In a related approach Snavely et al. [175] extract a *skeletal set* from an unordered image collection that yields a reconstruction with bounded loss of quality compared to the full image set. The joint covariance is represented as an image graph with edges represent the relative pose uncertainty where the *skeletal* subgraph is determined as the minimum number of interior nodes that span the entire graph while achieving the desired uncertainty bound.

4.1.2 Organizing Image Collections using Meta-data

A few authors have explored using tagged meta-data information instead of image content for summarizing image collections for offline browsing or search applications. Jaffe et al. [89] use hierarchical clustering on image geo-tags followed by a ranking according to representativeness based on textual tags and temporal infor-

mation. Clough et al. [30] organize an image collection using only textual tags using the concept of subsuming tags, i.e., a tag t_i subsumes t_j if the set of images tagged with t_i is a super set of images tagged with t_j. In [161] Schmitz employs the subsumption model to induce ontologies from Flickr images incorporating a probabilistic framework to handle noise in image tags. Motivated by camera phone applications, Pigeau and Gelgon [142] use time stamps and geolocation information for images for unsupervised clustering of images. A Gaussian mixture model is assumed and clustering is done independently for spatial and temporal components using the Expectation Maximisation (EM) algorithm, model complexity determined via integrated completed likelihood criterion [11].

4.1.3 Summarization and Layout

The task of creating a single representative summary image (or collage) from a set of images has been studied by some researchers. Wang et al. [195] create a *picture collage* by first extracting salient regions for each image by adopting the visual attention model [88] followed by determining the 2D spatial arrangement on a canvas to maximize visibility of each element. The problem is formulated probabilistically and MAP inference is carried out using an efficient MCMC approach. In a related work [196] authors construct *video collage*, a static summary of a video sequence where temporal structure of the sequence is preserved while optimally laying out regions of interest extracted from individual frames. In [157] Rother et al., present a technique for automatically synthesizing a *digital tapestry*, a large output image from a given set of input images by stitching together salient and spatially compatible image blocks from the input set. The problem is formulated as a multi-class labeling problem using Markov random field and optimized via graph-cut based algorithm.

In a subsequent work titled *AutoCollage* [156] authors improve the scalability and robustness by introducing a multi-stage optimization and explicit region of interest selection. In [2] Agarwal et al. present *digital photomontage* that combines parts of a set of photographs into a single composite picture using graph-cut minimization and gradient-domain fusion aimed towards a seamless and artifact-free visual output. In an earlier related work, Szeliski et al. [179] addressed building panoramic mosaics from image sequences from a hand held camera addressing issues of alignment, moving objects and exposure variations.

4.1.4 Video summarization

Summarization of video content for compression, browsing, surveillance and retrieval task has been an active area of research for several years. Initial efforts for video summarization focused on dividing the video sequence into shots and then picking key-frames from each shot. Ueda et al. [188] represented each shot by its first and last frames. Ferman and Tekalp [49] clustered frames in each shot and selected the frame closest to the center of the largest cluster as the keyframe. Gong and Liu [62] use singular value decomposition (SVD) on the entire video sequence to to pick key frames forming a summary.

As an alternative to keyframe-extraction, Acha et al. [151] introduce *stroboscopic* movies where where multiple dynamic instances of a moving object are played simultaneously. The algorithm extracts moving objects tracks using background subtraction and then minimizes temporal and background consistency costs for object tracks yielding the optimal summary. Pritch et al. [144] improve this method and demonstrate synopsis generation from multiple static surveillance cameras. Further, Simakov et al. [168] investigate the requirements for a similarity metric used

for spatio-temporal alignment in the above mentioned techniques and present a bi-directional similarity measure for visual data maximizing visual saliency and coherence criteria. Mei et al. [117] present *video collage*, similar to efforts cited in Section 4.1.3, where the single image summary is created from a video sequence. The technique involves picking most representative images from the video stream, extracting regions of interest from these images and arranging the patches in a canvas loosely preserving temporal coherence.

4.1.5 Applications in Mobile Robotics

Within mobile robotics, the task of summarizing visual maps has recently gained attention. Girdhar and Dudek [56] present a method for online extraction of k-most novel images from an image corpus using set-theoretic surprise measuring the fitness of an image as a summary image. The approach requires the summary size to be specified and is aimed towards identifying salient aspects of data. Note, in the context of summary generation both the common and salient aspects must be represented and learnt online. In [148], Ranganathan et al. use Bayesian surprise for identifying salient landmarks for topological mapping with vision and laser features. In another related work [97], Konolige et al. present *view based maps*, an online large-scale mapping technique for constructing topological maps with stereo data. The map is pruned by extracting relevant keyframes using a distance based heuristic causing the number of selected keyframes to scale with map length. The approach is closely related to intelligent sub-sampling of video sequences for structure-from-motion (SfM) for improving efficiency and robustness as discussed in [51] and [153].

The technique of Booij et al. [19] extracts a representative set of images from appearance-based topological maps for indoor robot localization. An appearance

graph is build connecting image pairs that can allow 3D reconstruction of the local space (using the 8 point algorithm [70]) and the sparse subset is obtained by finding the Connected Dominating Set (CDS) of the graph [65]. In a related effort Zivkovic et al. [209] use graph partitioning techniques (normalized graph cuts) to find groups of similar images in the image graph. However, the technique is offline and requires the number of clusters to be specified. The localization system presented in [98] clusters images taken by a navigating robot according to time stamps. However, it is assumed that each place is visited only once otherwise loop-closure images might be assigned to different clusters due to time difference between visits. Authors in [158] retain only the most salient image features from images instead of discarding whole images. A graph theoretic formulation is used to automatically extract an optimal set of landmarks per region in the environment so that each image in the region can still be localized.

Finding informative image subsets has found applications in the planetary exploration domain where remote explorers typically possess limited onboard memory or processing capabilities. The system requires a subsampling algorithm to downlink only pertinent samples for efficient bandwidth utilization and higher quality science return. Haden et al. [72, 73] apply an iterative k-means clustering on diverse image features like texture, colour, time and spatial information etc. A similar clustering based approach is applied for aerial imagery in [71]. Thomson et al. [182] leverage an expert labeled data set indicating science content of expected imagery and use image texture analysis in conjunction with context sensitive hidden markov model (HMM) for adaptive downlink. In a subsequent paper [183], authors use Gaussian Process (GP) regression to model the science content and pose the subset selection as an active learning problem where the agent picks the optimal set that maximizes

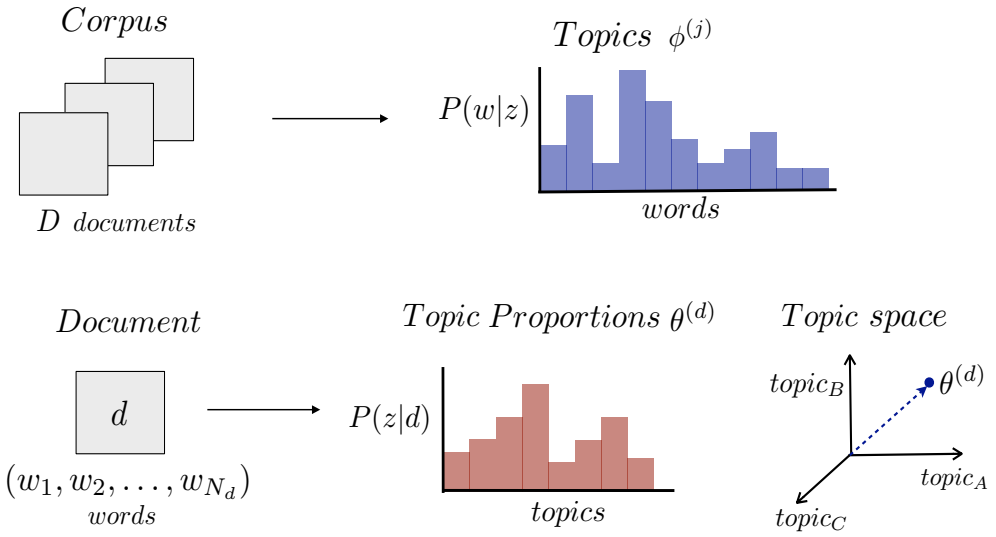


Figure 4.2: Topic estimation and inference. Topics are distributions over words and estimated once from an image corpus. Learnt topics are used to estimate the vector of topic proportions for an observed image mapping to a point in topic vector space. Using a cosine similarity metric online star clustering is used to organize images into topical clusters.

information gain over the unknown image contents.

4.2 Image Representation in Topic Space

In this section, we discuss how an image can be encoded thematically using probabilistic topic model representation. The use of probabilistic topic models based on Latent Dirichlet Allocation (LDA) was discussed earlier in Section 3.2. A brief overview is presented here for completeness and continuity. Readers with familiarity may proceed directly to Section 4.3.

Probabilistic topic models have their genesis in information retrieval, hence we leverage an analogy between documents and images. At the lowest level an image is described as a list of visual words [173]. We can now think of an image as a document of visual words represented as a point in vector space where each dictionary

word represents an orthogonal axis. Visual words in an image are not independent and arise from objects characterizing the scene. Features emanating from a common object, frequently co-occur across multiple images. Topic models [64] represent documents as a mixture of intermediate latent topics. Topics are distributions over words and probabilistically capture co-occurring features. Each document or an image is a distribution over topics and different documents can possess varied topic proportions. Topic distributions are estimated once offline from a large corpus. Online, topic proportions are estimated for each image, see Figure 4.2. The vector of topic proportions maps an image to a point in topic-space. Typically, the number of topics is much smaller than the vocabulary size leading to considerable dimensionality reduction. Image similarity can be measured via cosine distance in topic space. Since topics provide a lower-dimensional thematic representation, images with common topics can get associated even if they have few words in common.

Latent Dirichlet Allocation (LDA) is a widely used probabilistic topic model [17] for which topic estimation is tractable. LDA is a hierarchical Bayesian generative model and describes document formation as: (i) picking a multinomial distribution over topics specifying the likelihood of each topic in the document and (ii) generating constituent words by sampling topic proportions to obtain a topic label followed by sampling the word from the selected topic distribution over words. Inference involves reversing the generative process to recover the topics and the topic proportions per document. This is approximated using an MCMC Gibbs sampling procedure in the state space of topic labels for observed words using the update rule as given in Equation 4.1. Here, z variable is a topic indicator variable one for each observed word w and α, β parameterize Dirichlet priors placed on topic and topic proportion distributions. The number of topics and vocabulary sizes are referred to

as T and W. After sufficient sampling iterations, topic labels are recorded and used to form maximum likelihood multinomial estimates for topic and topic proportion distribution [64].

$$P(z_i = j|\mathbf{z}_{-i}, \mathbf{w}, \alpha, \beta) \quad \propto \quad \left[\frac{n_{-i,j}^{(w_i)} + \beta}{n_{-i,j}^{(\cdot)} + W\beta}\right] \left[\frac{n_{j}^{(d_i)} + \alpha}{n_{-i,\cdot}^{(d_i)} + T\alpha}\right] \tag{4.1}$$

After obtaining a suitable representation of images in topic space, our next task is to incrementally organize the imagery collected by a robot and generate a visual summary of the traversal.

4.3 Star Clustering and Online Organization

We use the star clustering algorithm [4] to compute a topic-driven organization of the robot's image collection. The star clustering algorithm is an efficient clustering algorithm that identifies the underlying thematic structure of a document collection and organizes it using topical clusters, as long as the documents can be compared using a similarity metric.

The algorithm guarantees a minimum similarity between data points in the cluster and number of clusters are naturally discovered by the procedure. This is in contrast to algorithms like k-means where the number of clusters are specified *a-priori* but no guarantees hold on the intra-cluster similarity. Hence, for star clustering the desired similarity threshold is a free parameter that affects the resolution at which clustering is obtained. This behaviour makes the algorithm suitable for our application where the natural partition of the images is not known to the traversing robot. Based on saliency and the underlying similarity distribution of the images

Algorithm 4.1 Star Clustering [3]: INSERT (α, L, G_σ)

```
α.type ← satellite
α.degree ← 0
α.adj ← empty
α.centers ← empty
for all β in L
    α.degree ← α.degree + 1
    β.degree ← β.degree + 1
    InsertAdjList(β,α.adj)
    InsertAdjList(α,β.adj)
    if (β.type = center)
        InsertCenterList(β,α.adj)
    else
        β.inQ ← true
        Enqueue(β,Q)
    endif
end for
α.inQ ← true
Enqueue(α,Q)
UPDATE(G_σ)
```

encountered the summary evolves over time accurately capturing both the common and uncommon aspects of the explored visual workspace.

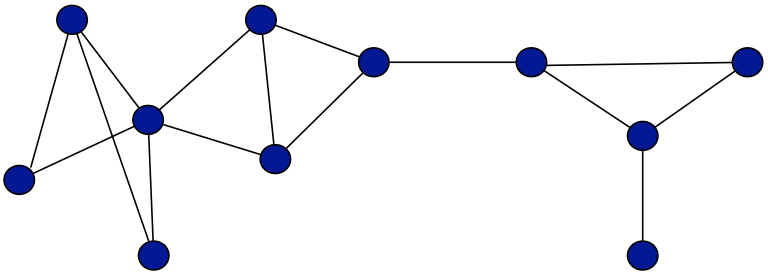
Further, the star clustering algorithm can be run online and is computationally very efficient. The ability to incrementally determine the topic organization of an image collection makes it especially suitable for our problem setting, where the data collection from a mobile robot is incremental in nature. Next, we present a mathematical overview and summarize the algorithm. Readers with greater familiarity can skip directly to Section 4.4 that presents the summary generation procedure.

4.3.1 Clustering with Star Sub-graphs

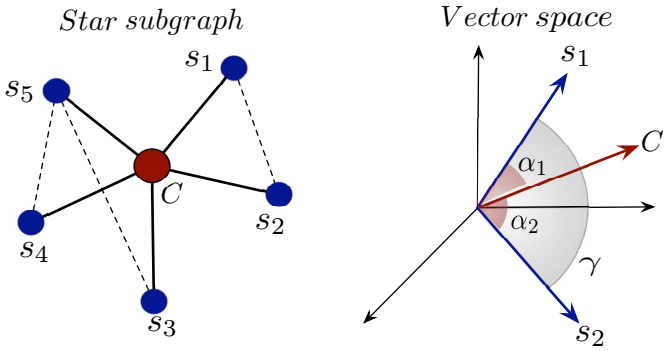
The image corpus is represented via an undirected and weighted similarity graph, $G = (V, E, w)$ where vertices correspond to images and weighted edges represent cosine similarity in topic space. The similarity graph can be studied at various thresholds of pair-wise document similarity, σ. The associated thresholded similarity graph, G_σ is obtained from G by removing edges with pairwise similarity less than σ, see Figure 4.3a. The clustering algorithm covers G_σ with a computationally efficient cover of dense star-shaped subgraphs [4]. A star-shaped subgraph on $m+1$ vertices consists of a star center and m satellite vertices, where there exist edges between the star center and each of the satellite vertices as shown in Figure 4.3b. Note that edges between satellite vertices may or may not exist.

The optimal clustering as shown in Figure 4.3c is obtained by forming a *minimal* vertex cover for the graph with *maximal* star-subgraphs resulting in the following properties for each vertex: (i) a star center is not adjacent to another star center and (ii) every satellite vertex is adjacent to at least one center vertex of equal or higher degree. The number of clusters is naturally induced as the size of the dense cover. For each cluster in the graph the cluster center acts as its exemplar. Satellite vertices displaying multiple themes can be associated with multiple clusters, Figure 4.3c. Note that the minimal cover may not be unique, in some cases more than one covers can satisfy the above conditions.

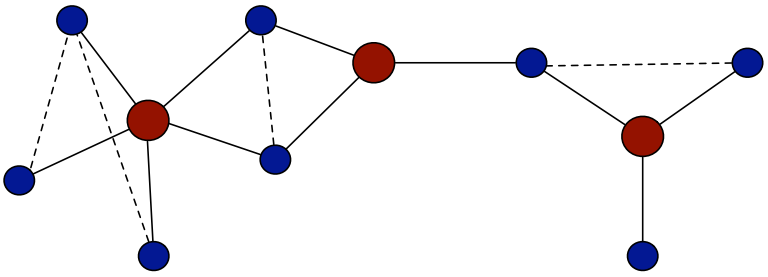
A key aspect of the algorithm is the bound on the intra-cluster similarity for a graph covering with star sub-graphs. Figure 4.3b(right) illustrates the center vertex C and any two satellite vertices S_1 and S_2 for the star-subgraph in the implied topic space. By definition, edges exist between the center and the satellite vertices for each star-subgraph. Let α_1 be the angle between the position vectors for C and S_1.



(a) The clustering procedure begins by computing a similarity graph, G_σ with each image as a node with links indicating similarities exceeding a specified threshold σ.

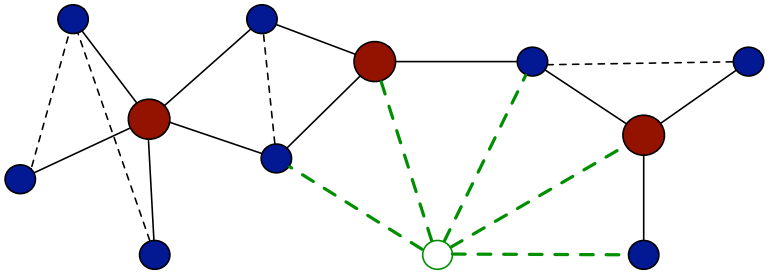


(b) An example of a star-shaped subgraph with center C and five satellite vertices (left). Each node in the graph maps to a point in a vector space where pairwise similarity is endowed using cosine distance metric. By construction, center-satellite similarities are at least σ. The vector space geometry with cosine distance ensures that expected satellite-satellite similarities are high leading to dense clusters.

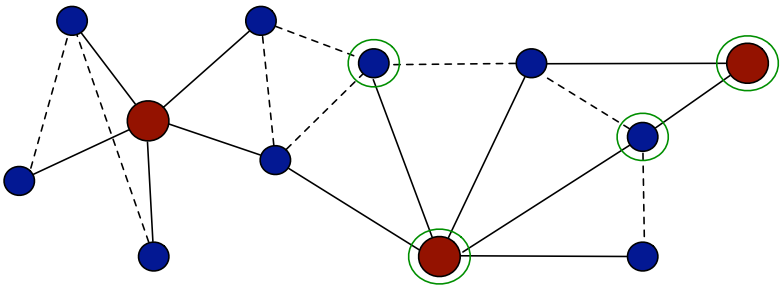


(c) The graph organized into clusters using a minimal cover with star-shaped subgraphs. The cluster centers compactly summarize the visual experience of the robot. Note that an image (possessing varied themes) can belong to multiple clusters.

Figure 4.3: Clustering with star-shaped subgraphs.



(a) A new data point may introduce additional links in the similarity graph (green) affecting adjacency and hence the validity of current minimal star cover.



(b) Inconsistent stars are broken and re-arranged to incorporate the new point. The green circles indicate positions where graph modifications took place. The number of stars broken determine the insertion running time. On real graphs, the average number of stars broken is small, thereby yielding an efficient and incremental clustering approach.

Figure 4.4: Star clustering re-organization upon insertion of a new data point.

Algorithm 4.2 Star Clustering [3]: DELETE (α, G_σ)

```
for all β in α.adj
    β.degree ← β.degree + 1
    DeleteCenterList(α,β.adj)
end for
if (α.type = satellite)
    for all β in α.centers
        for all μ in β.adj
            if (μ.inQ=false)
              μ.inQ ← true
              ENQUEUE(μ,Q)
            end if
        end for
    end for
else
    for all β in α.adj
        DeleteCenterList(α,β.centers)
        β.inQ ← true
        ENQUEUE(β,Q)
    end for
end if
UPDATE(G_σ)
```

Algorithm 4.3 Star Clustering: UPDATE (G_σ)

```
while (Q ≠ empty)
φ ← ExtractMax(Q)
if (φ.centers = empty)
    φ.type ← center
    forall β in φ.adj
        InsertCenterList(φ,β.centers)
    endfor
else
    if (∀δ ∈ φ.centers, δ.degree < φ.degree)
        φ.type ← center
        forall β in φ.adj
            Insert(φ,β.centers)
        endfor
        forall δ in φ.centers
            δ.type ← satellite
            forall μ in δ.adj
              Delete(δ,μ.centers)
            if (μ.degree ≤ δ.degree ∧ μ.inQ = false)
              μ.inQ ← true
              Enqueue(μ,Q)
            endif
            endfor
        endfor
        φ.centers ← empty
    endif
endif
φ.inQ ← false
endwhile
```

Similarly, let α_2 denote the angle between the position vectors for C and S_2. Hence, $\cos\alpha_1 \geqslant \sigma$ and $\cos\alpha_2 \geqslant \sigma$. By examining the geometry of the star-subgraphs, the maximum angle between satellites can be $(\alpha_1 + \alpha_2)$ and consequently the *minimum* satellite-satellite similarity is given by:

$$\cos\gamma \quad \geqslant \quad \cos(\alpha_1+\alpha_2) \tag{4.2}$$

$$\cos\gamma \quad \geqslant \quad \cos\alpha_1\cos\alpha_2 - \sin\alpha_1\sin\alpha_2 \tag{4.3}$$

However, as proved in [4], the *expected* satellite-satellite similarity is given by Equation 4.4. The expected pairwise similarities are high and imply dense clustering of data. As an example, for $\sigma = 0.6$, $\cos\alpha_1 = 0.65$ and $\cos\alpha_1 = 0.90$ the *minimum* satellite-satellite similarity is 0.25 but the *expected* similarity is a much higher value of 0.71.

$$\cos\gamma \quad \geqslant \quad \cos\alpha_1\cos\alpha_2 + \frac{\sigma}{\sigma+1}\sin\alpha_1\sin\alpha_2 \tag{4.4}$$

Note that if the graph was covered by cliques instead of star-subgraphs, the resulting clustering would guarantee clusters of similarity σ due to resulting edges between satellite vertices as cliques are fully connected sub-graphs by definition. However, obtaining a clique cover is an NP-complete problem, hence intractable for large data sets [111, 4]. Star clustering approximates the clique cover with star-subgraphs, guaranteeing only high *expected* quality but gaining in computational efficiency.

4.3.2 Online Operation

Star clustering can be undertaken incrementally with each arriving data point with a potential re-arrangement of existing stars. Figure 4.4 presents an example. For each inserted vertex, its degree and adjacency list is computed and the following cases are examined: if the new vertex is not adjacent to a star center, then the inserted vertex is added as a star center forming a new cluster; if the inserted vertex is adjacent to a center vertex with higher degree, then the inserted vertex becomes the satellite for the center vertex. The graph is re-arranged in two cases: (i) when all centers adjacent to the inserted vertex have degree lower than the new vertex or (ii) vertex insertion increases the degree of an adjacent satellite beyond the degree of its associated star center. Under these conditions existing stars are broken and satellites are re-examined. However, the number of graph re-arrangement operations during insertion are usually small which we verified experimentally in Section 4.5.

Table 4.1: Data for vertices maintained.

$v.type$	satellite or center
$v.degree$	degree of v
$v.adj$	list of adjacent vertices
$v.centers$	list of adjacent centers
$v.inQ$	flag specifying if v is being processed

The star clustering algorithm can be implemented using an undirected graph data structure, vertex structure containing fields listed in Table 4.1 and a priority queue sorted by vertex degree for processing the list of satellite vertices that have the possibility of being promoted to center status. From [4], the pseudo code for *Insert*, *Update* and *Delete* operations appears in Algorithm 4.1, 4.2 and 4.3. Further, we employed an optimized version of the algorithm that saves operations by predicting the future status of a satellite vertex or other star-satellite status changes induced

by the inserted vertex [4]. Note that the graph structure can also be similarly re-arranged once data point is removed from the graph structure.

The online algorithm is computationally efficient with expected complexity for inserting n vertices as $O(n^2 \log n^2)$, asymptotically linear in the size of the graph $\Theta(n^2)$ within logarithmic factors. The expected size of the star cover is $O(\log n)$. Importantly, the star cover obtained by the incremental algorithm is the *same* as one of the minimal covers (since the minimal cover might not be unique) obtained by the batch algorithm where a greedy star cover is obtained for the thresholded graph for the entire corpus.

Note that the star clustering algorithm bears some resemblance to the popular spectral clustering technique [126] that has been applied to topological mapping as in [21]. Spectral clustering relies on computing the eigenvectors for the normalized affinity matrix, but is an offline algorithm requiring the number of clusters to be specified *a-priori*. Some efforts have explored incremental updates for spectral clustering, for example [190] and [128], though incremental updates to the eigenvectors can introduce errors that can accumulate substantially over time.

4.4 Incremental Visual Summary Generation

As discussed in the previous section, the star clustering procedure organizes images collected by a mobile robot incrementally into star clusters in topic space. At any time instant, the cluster centers represent a visual summary of the robot's traversal, Figure 4.5. The clustering threshold specifies the resolution at which the summary is generated. For each image in the corpus the associated cluster centers provide a thematic annotation in terms of the current summary images. Hence, the robot's

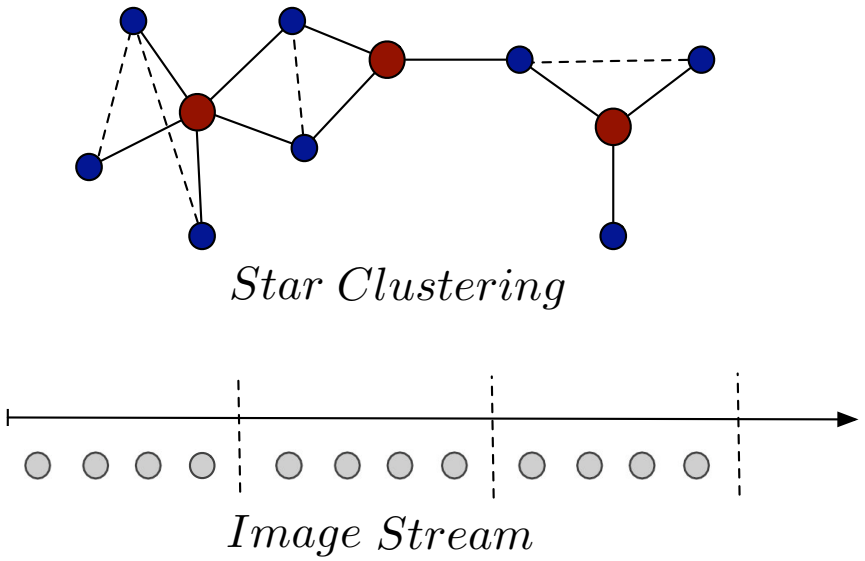


Figure 4.5: Visual summary generation. Images collected by the mobile robot are organized incrementally into star clusters. The star centers form a visual summary of the traversal. As the star clustering adapts with each new collected image, the summary and thematic annotation improves over time with increasing experience.

trajectory can be understood as a combination of segments, each annotated by summary images to which each trajectory image are presently assigned. Since clusters adapt with each new collected image, the summary and the thematic annotation improves over time with increasing experience.

Algorithm 4.3 presents the pseudo-code for the procedure. The visual vocabulary [173] and topic distributions over words [64] are learnt offline using a representative set of images. Online, each perceived image is first converted to a bag-of-words representation followed by estimating the topic proportions conditioned on the topic distributions learnt offline. Next, the image represented as a distribution over topics is inserted into the evolving graph structure and organized into star clusters. The current set of centers form the desired visual summary of the traversal. In the next section, we bring the described components together and present experiments on data collected from a mobile platform.

Algorithm 4.4 Incremental Visual Summary Generation

```
// OFFLINE
// Image Corpus, C
// Vocabulary Size, VS
// Vocabulary, V
// Number of Topics, T
// Topic Distributions, φ

    // Feature extraction and vocabulary generation
    V ← Vocab_Generation(VS, C)

    // Topic distributions estimation using Gibbs sampling
    φ ← Topic_Distribution_Estimation(T, C, V)

// ONLINE
// Star Clustering, G
// Threshold, σ
// Visual Summary, S
for each image collected at time t, I_t

        // Generate bag-of-words for image
        W_t ← Bag_Of_Words(I_t, V)

        // Infer topic proportions for image
        θ_t ← Topic_Proportion_Inference(W_t, φ)

        // Incremental star clustering
        G_t ← Insert(G_{t-1}, θ_t, σ)

        // Current Visual Summary
        S_t ← Get_Star_Centers(G_t)

end
```

4.5 Results

4.5.1 Vocabulary and Topic Learning

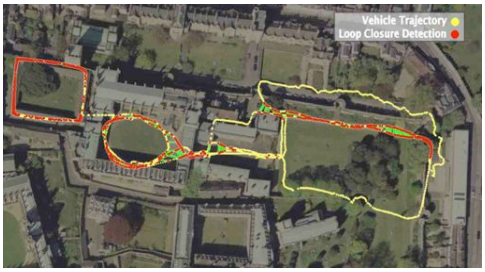
A data set traversing streets and park areas was collected consisting of 2874 images, recorded 10m apart and perpendicular to the robot's motion. Image samples were non-overlapping, excluded loop closures pairs and hence approximated independent samples from the observation distribution and used for vocabulary and topic learning. A visual vocabulary [173] of approximately 11k visual words was generated by clustering SURF features [7] extracted from this data set. Each image was represented as a multinomial of visual words by first extracting SURF features and then quantizing against the learnt vocabulary. Topic distributions were estimated using a Gibbs sampling procedure outlined in Section 4.2. The Markov chain was randomly initialized and was run till convergence for varying number of topics, T: ranging from 3 till 100. The Dirichlet priors were set as $\alpha = 50/T$ and $\beta = 0.1$ as suggested in literature. Iterations required to ensure MCMC convergence was experimentally obtained to be at least 200 and was found consistent across multiple re-starts. The number of topics was selected through the Bayesian model selection approach of maximizing the data log-likelihood given topics [64] which was found to peak for 50 topics.

4.5.2 Visual Summaries

The visual summarization algorithm was run on two data sets: (i) New College data set consisting of 1355 images from cloisters, quad area, parks and facades characteristic medieval buildings in Oxford and (ii) City Center data set comprising of 1683 images taken in dynamic urban environments including roads, buildings,

vehicles and pedestrians. There was no geographical overlap with the urban data set used for vocabulary and topic learning. For online experiments, images from the entire data set were presented sequentially and were incrementally organized into star clusters. Topic proportions for each image were estimated using topic distributions learnt from the urban data set.

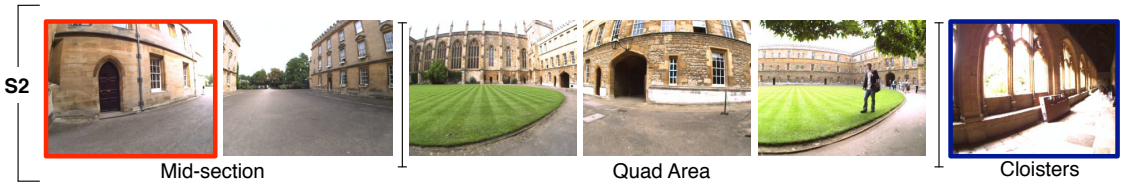
The visual summary at three time instants for the New College data set is shown in Figure 4.6. The robot began operation in the cloisters area and after covering two loops took an exit into the adjacent quads. Figure 4.6b shows the summary at that instant consisting of images of dominant windows and stone walls seen in the area. The robot then traversed the quad area and reached the mid-section. New clusters emerged as shown in Figure 4.6c containing varied views of the explored area. Note that images of walls found in the earlier summary are now absent. They were found similar to walls in the quad area and hence now appear as satellites. Figure 4.6d presents the summary at the end of the data set, containing new images from the parks and modern buildings seen later by the robot. The summary images are shown with labels indicating the geographical region where they were recorded. Note that the cluster centers summarizing the traversal appear visually distinct indicating that the star covers capture different appearance modes present in the traversed environment. Figure 4.7 illustrates the summary for the City Center data set. The initial summary consists of foliage, parks and railings, Figure 4.7b. The robot then explores roads and building areas and hence the summary is refined with new representative clusters including vehicles, roads, buildings etc.



(a) Aerial view of New College with GPS plots for the robot's trajectory.



(b) Summary after traversing the cloisters with images of large windows, medieval walls etc.



(c) Summary at the start of the mid-section after traversing the quad and cloisters area.



(d) Summary at the traversal end including examples of parks, building and entrance areas.

Figure 4.6: Incremental visual summary generation while traversing New College shown at three instants. The most recently added cluster center image and the first encountered are marked with red and blue borders respectively. Parameters: $\sigma = 0.6$ and $T = 50$. The sections where the images were taken have been hand-labeled to facilitate interpretation. Note that clusters evolve over time and capture the dominant visual themes encountered by the mobile robot.



(a) Satellite image with overlaid GPS tracks for the City Center data set.



(b) Summary while exiting parklands after collecting 180 images

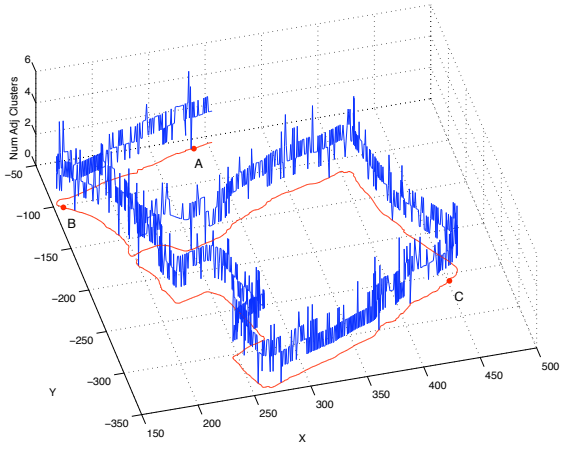


(c) Visual summary when the robot reaches Science Area buildings. New cluster centers for building features become prominent. Iteration 310.



(d) Visual summary at the end of traversal, iteration 1683.

Figure 4.7: Incremental visual summary generation for City Center data set shown at three instants. The most recently added cluster center image and the first encountered are marked with red and blue borders respectively. Parameters: $\sigma = 0.6$ and $T = 50$. The sections where the images were taken have been hand-labeled to facilitate interpretation. Note that clusters evolve over time and capture the dominant visual themes encountered by the mobile robot.



(a)



(b)

Figure 4.8: Topical annotation of the the robot's trajectory. (a) GPS trajectory for City Center data set with the number of adjacent centers plotted at each point along the path. Three time instants are indicated. (b) Observed images at three time instances during traversal shown with the associated cluster centers (using the final clustering at the end of traversal). The assigned cluster centers accurately capture the visual theme in the selected images (left column). Examine the first row. The observed parkland image (top left) is assigned to two cluster centers. The first center (middle) image was captured prior to the observed image and the second cluster center was collected later, indicating that the topical annotation for each image on the trajectory improves with time.

4.5.3 Thematic Annotation

Every image collected by the robot is assigned to clusters in the collection which can be considered as an annotation in terms of topical themes (cluster centers) learnt from exploration till now. Figure 4.8 illustrates an example. The GPS trajectory for the City Center data set is shown with the number of adjacent centers (corresponding to the final clustering at the end of traversal) plotted at each point along the path. Images collected by the robot at three instants is shown with the assigned cluster centers. The assigned cluster centers accurately capture the visual themes in the perceived images. Note that star clustering permits multi-cluster membership accounting for images that can be explained via multiple themes in the data set. Since the star clustering evolves incrementally with time, the topical annotation for each image along the trajectory also improves with increasing experience.

4.5.4 Topical Clusters

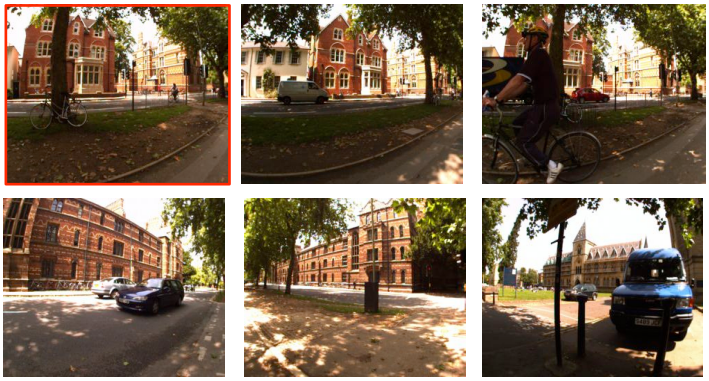
Figure 4.9 illustrates three representative clusters obtained at the end of the traversal. Cluster center image (indicated in red) and five randomly picked satellite images are shown. The cluster shown in Figure 4.9a typically consist of vehicles that generate a large number of features. The topic model learns that these feature co-occur and maps them to a common theme. The cluster shown in Figure 4.9b consists of similar images of foliage and trees in parks. Figure 4.9c presents a cluster containing images of buildings and trees as viewed from a sidewalk. Note that clusters possess a common visual theme as opposed to exact matches and hence topically organize images collected by the robot. Secondly, a relatively small number of topics (50) yielded topical clusters compared to the dictionary size of 10k, indicating significant



(a) Clustered images consist of feature sets on vehicles and the horizon. City Center data set: Cluster 46, $\sigma = 0.8$

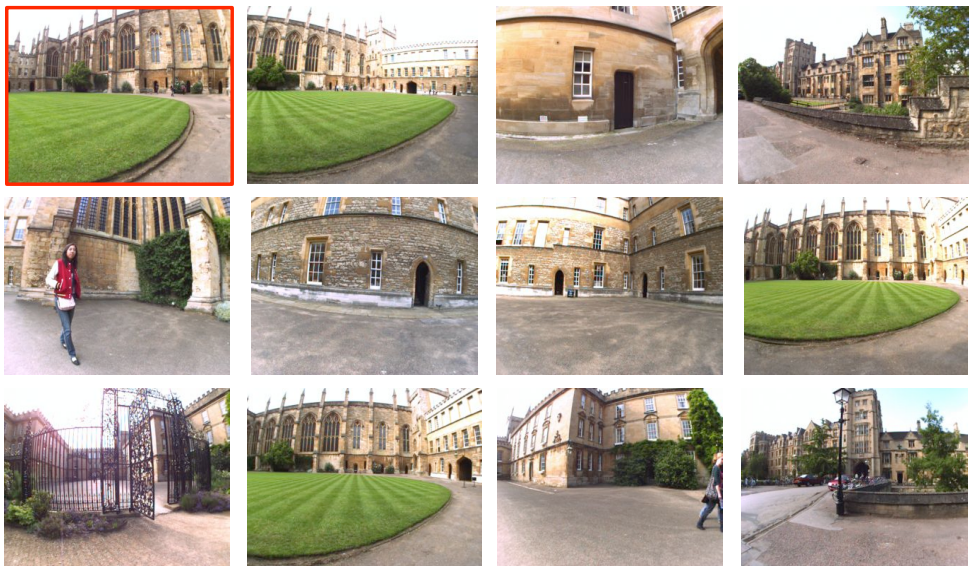


(b) Thematic cluster of parkland images. New College data set: Cluster 8, $\sigma = 0.6$



(c) Images in the cluster display common feature sets appearing on buildings and trees. City Center data set: Cluster 29, $\sigma = 0.7$

Figure 4.9: Representative clusters from the City Center and New College Data sets. Images in a cluster possess similar visual topics. Cluster center is indicated in red. Images displayed were randomly sampled from the total set of images within each cluster.



(a) Displaying cluster 29 at $\sigma = 0.6$ with 12 sampled images from 202 cluster images.

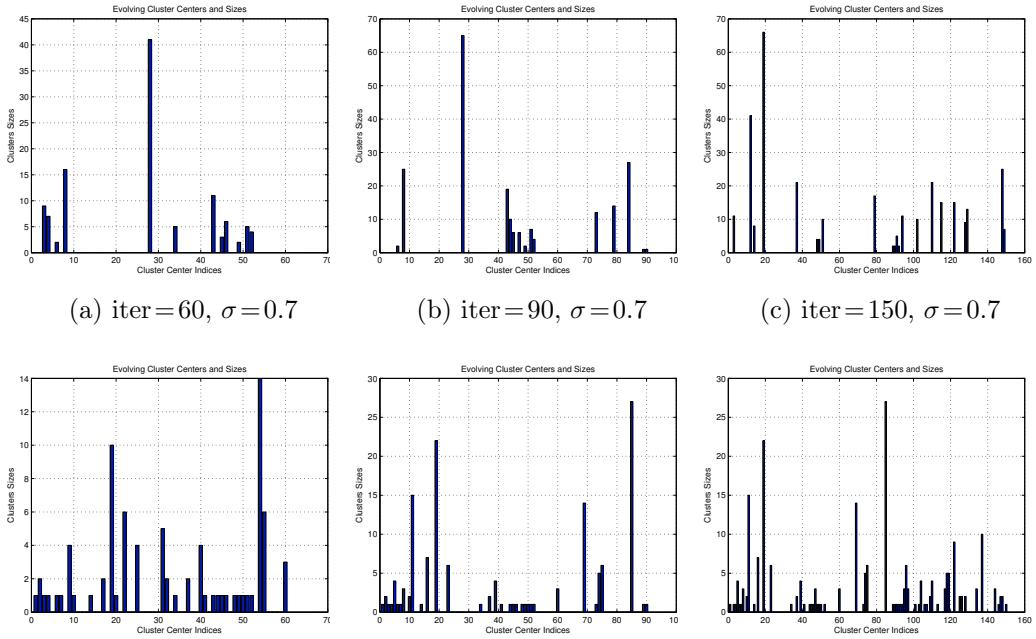


(b) Displaying cluster 29 at $\sigma = 0.7$ with 8 sampled images from 55 cluster images.

Figure 4.10: Two clusters with centers in the New College Quad obtained with varying thresholds of $\sigma = 0.6$ and $\sigma = 0.7$. Cluster centers are marked with red and correspond to approximately the same location and viewpoint within the Quad. (a) Clustering at lower thresholds results in larger clusters with less specific visual themes. (b) Increasing the threshold results in smaller clusters with higher similarity.

dimensionality reduction.

Using a higher similarity threshold, σ causes higher intra-cluster similarity, and generally results in smaller but more numerous clusters. Figure 4.10 compares two clusters obtained at $\sigma = 0.6$ and $\sigma = 0.7$, selected such that their respective clus-



(a) iter = 60, $\sigma = 0.7$ (b) iter = 90, $\sigma = 0.7$ (c) iter = 150, $\sigma = 0.7$

(d) iter = 60, $\sigma = 0.8$ (e) iter = 90, $\sigma = 0.8$ (f) iter = 150, $\sigma = 0.8$

Figure 4.11: Cluster center indices (x-axis) and corresponding cluster sizes (y-axis) during online insertion for the New College traversal. The plots are displayed after collecting: 60, 90 and 150 images and for two threshold values: 0.7 and 0.8. The vertices assigned as the cluster center evolve over time as images are collected incrementally. Using a higher threshold yields smaller but numerous clusters.

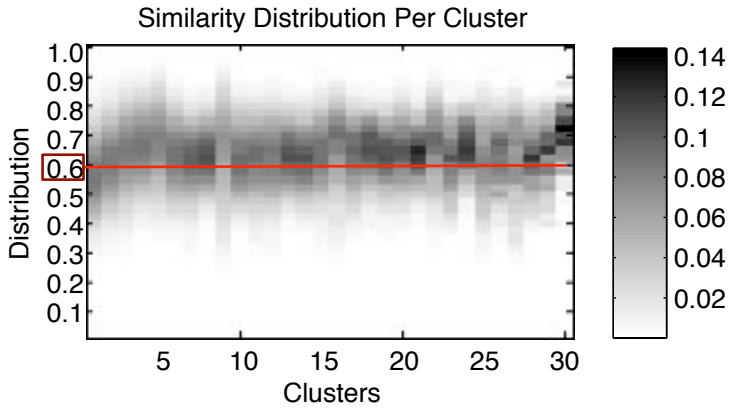
ter center images were taken at the same location in the New College quad. Both clusters possess a coherent visual theme consisting of medieval buildings with some foliage features. Cluster images for $\sigma = 0.7$ display higher similarity and are primarily from the same quad area, compared to the cluster at a lower threshold that consisted of images from the quad, mid-section and other parts of the college, hence possessing greater variability.

Figure 4.11 displays evolving star cluster centers and clusters sizes during online insertion for the New College traversal. The plots show three different time instants, i.e. after inserting 60, 90 and 150 images for two corresponding similarity thresholds $\sigma = 0.7$ and $\sigma = 0.8$. The vertices assigned as cluster centers evolve over time as images are collected incrementally. Also note that clustering at a higher threshold yields smaller but more numerous clusters.

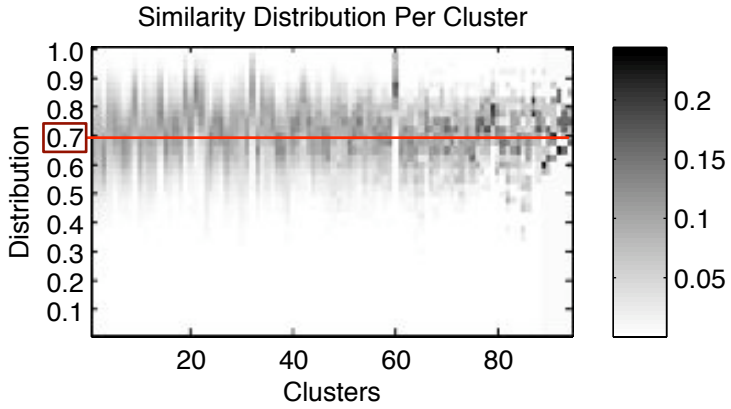
4.5.5 Clustering Quality

Next, we explore the cluster quality obtained at a specified threshold. For each cluster in the resulting star clustering, the distribution of all pair-wise similarities between satellite vertices was calculated. The empirical distribution was approximated with a probability histograms using bin size of 0.025. Further, in order to mitigate the effect of variable cluster sizes and sampling error, the probability estimates were smoothened [36].

Figures 4.12 and 4.13 plot the distributions for Science Area and New College data sets for thresholds $\sigma = 0.6$ and $\sigma = 0.7$ respectively. As discussed in Section 4.3, Equation 4.4 gives the expected similarity between satellite vertices in a star-subgraph. For a clustering at threshold, σ, the center satellite similarities are at least σ. Hence, from Equation 4.4, the expected satellite-satellite similarity is σ,

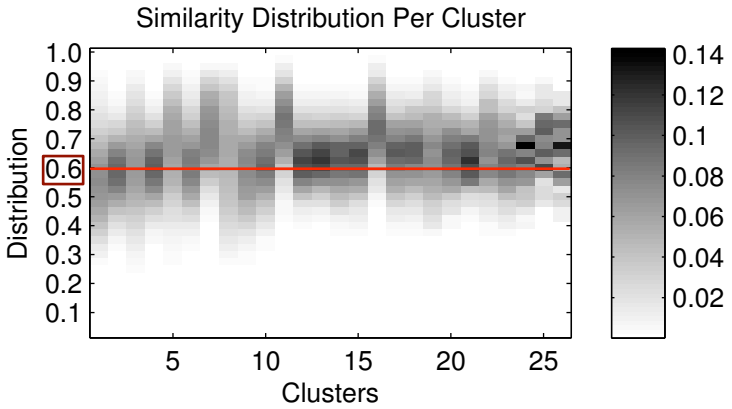


(a) Similarity distribution for clustering at $\sigma = 0.6$.

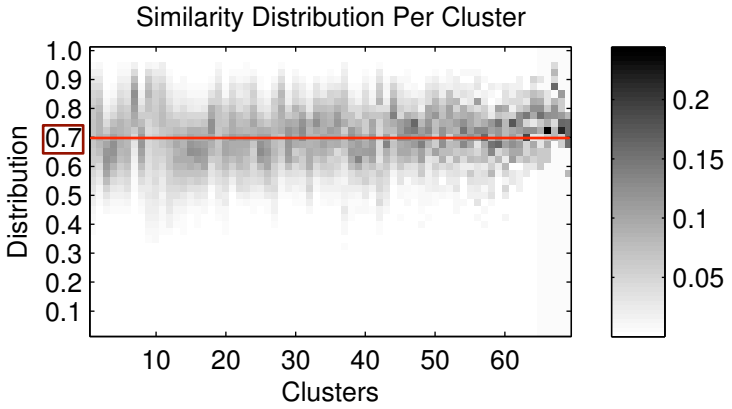


(b) Similarity distribution for clustering at $\sigma = 0.7$.

Figure 4.12: Distributions of all pair-wise similarities between satellite vertices for the Science Area data set with thresholds 0.6 and 0.7. The histogram for each cluster is plotted vertically. The threshold used is indicated with a horizontal red line. Note that the expected similarity values are close to σ indicating that star clusters are reasonably dense with high expected pair-wise similarities between satellites.



(a) Similarity distribution for clustering at $\sigma = 0.6$.



(b) Similarity distribution for clustering at $\sigma = 0.7$.

Figure 4.13: Distributions of all pair-wise similarities between satellite vertices for the New College data set for thresholds 0.6 and 0.7. The histogram for each cluster is plotted vertically. The threshold used is indicated with a horizontal red line. Note that the expected similarity values are close to σ indicating that star clusters are reasonably dense with high expected pair-wise similarities between satellites.

Table 4.2: Online insertion statistics for New College Data set with varying thresholds and topics, $T = 50$.

	New College Data set			
Threshold	$\sigma = 0.5$	$\sigma = 0.6$	$\sigma = 0.7$	$\sigma = 0.8$
Number of clusters	12	26	71	328
Number of edges (x10^5)	3.96	1.87	0.69	0.20
Insertion time/iter (msec)	4.12	9.92	0.95	0.17
Total insertion time (sec)	5.59	13.44	1.28	0.24
Avg. stars broken/iter	0.07	0.52	0.69	0.70
Total stars broken	100	707	935	954

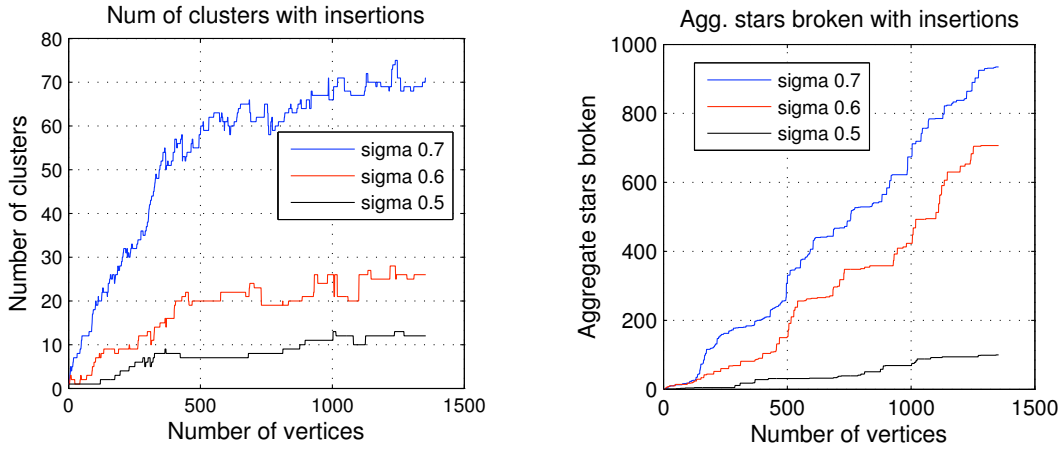
Table 4.3: Online insertion statistics for the City Center Data set with varying thresholds and topics, $T = 50$.

	City Center Data set			
Threshold	$\sigma = 0.5$	$\sigma = 0.6$	$\sigma = 0.7$	$\sigma = 0.8$
Number of clusters	14	30	98	454
Number of edges (x10^5)	7.22	3.54	1.20	0.26
Insertion time/iter (msec)	5.19	5.53	4.59	0.27
Total insertion time (sec)	8.74	9.31	7.74	0.45
Avg. stars broken/iter	0.06	0.44	1.53	0.78
Total stars broken	112	749	2579	1309

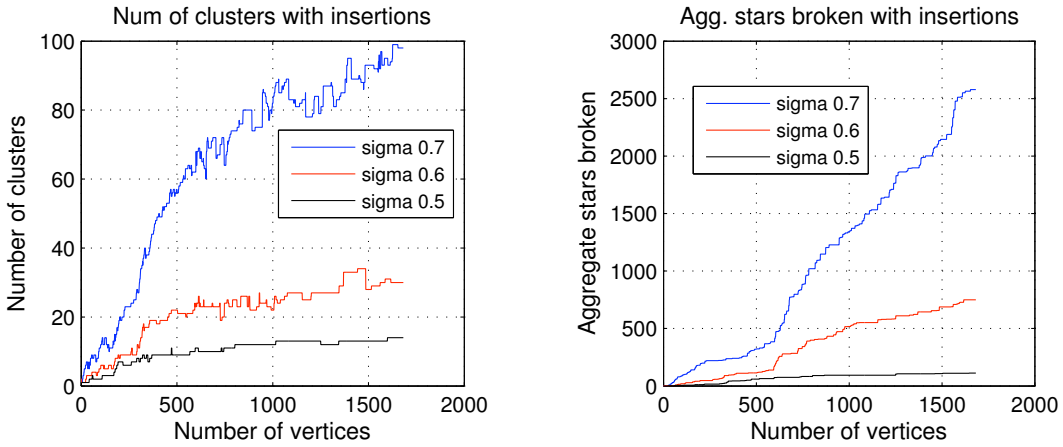
plotted as a horizontal line in Figures 4.12 and 4.13. The expected similarity values for clusters were found close to σ indicating that the star clusters obtained are reasonably dense and imply high expected pair-wise similarities between satellites.

4.5.6 Efficiency and Timing

Tables 4.2 and 4.3 present the online clustering statistics for varying thresholds: 0.5, 0.6, 0.7 and 0.8 for the two data sets. A higher similarity threshold reduces the number of edges in the graph (increasing sparsity) resulting in an increase in the number of clusters (size of the minimal cover). The number of clusters obtained varied from 12 to 328 for the New College and from 14 to 454 for the City Center



(a) Plots for New College data set.

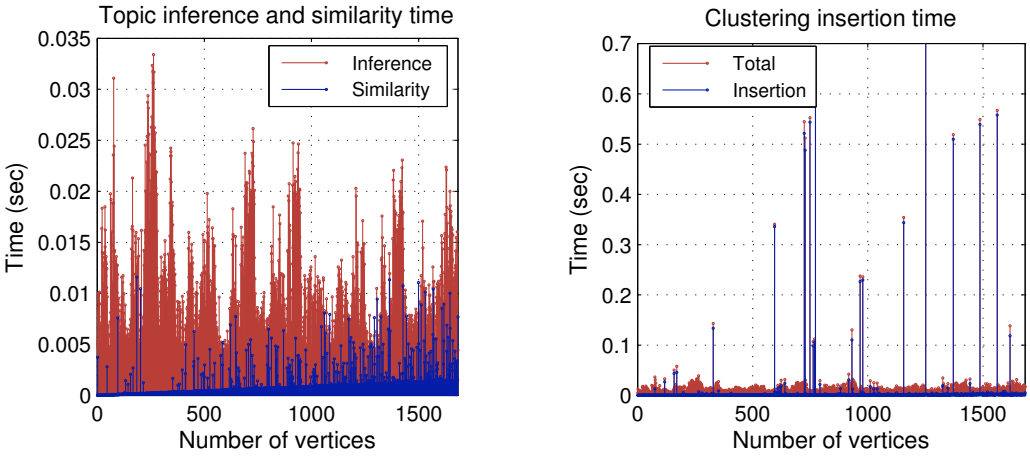


(b) Plots for City Center data set.

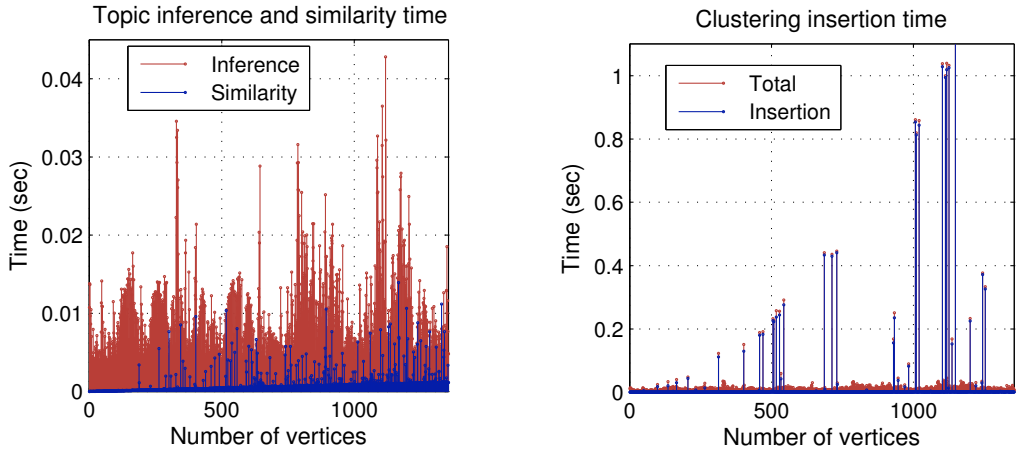
Figure 4.14: Number of clusters and aggregate number of stars broken during insertion iterations.

data set for $\sigma = 0.5$ and $\sigma = 0.8$ respectively.

The average number of stars broken during insertion indicates the work done to re-arrange the existing graph when a data point is incorporated. Notably only a small number of stars are broken per insertion on average. For example, while inserting 1355 images in the New College data set at $\sigma = 0.6$, a total of 707 stars were broken - approximately 0.52 broken stars per insertion. The average number of stars broken was less than 0.78 for all runs except for $\sigma = 0.8$ with the City



(a) Plots for City Center Data set. Clustering at $\sigma = 0.6$.



(b) Plots for New College Data set. Clustering at $\sigma = 0.6$.

Figure 4.15: Topic inference, similarity matrix computation time (sec) during insertion (left). Plots for insertion and total time are given in (right). Running time is primarily determined by the online insertion operation peaks are recorded in iterations when a large number of stars are broken. Note the scale in both graphs.

Table 4.4: Running time components for New College Data set. Total 1355 images with $\sigma = 0.6$ and $T = 50$.

	Inference	**Similarity**	**Clustering**	**Total**
Cumulative (sec)	11.93	1.26	13.45	26.64
Median (ms)	7.32	0.63	0.49	9.16
Min. (ms)	3.78	0.01	0.01	3.84
Max. (ms)	42.80	13.97	1104.70	1113.79
Avg.± std. (ms)	8.81 ± 4.92	0.93 ± 1.34	9.92 ± 80.12	19.66 ± 80.29

Center data set where a total of 2579 stars were broken (avg. 1.53 stars broken per iteration) while inserting 1689 images. The running time depends on the size of the graph, stars broken and the underlying similarity distribution for the data set. Total insertion time ranged from 0.24sec to 13.44sec yielding a small average insertion time of less than 10msec, making the approach practical for online operation.

Figure 4.14 plots the number of clusters and aggregate stars broken during each insertion iteration. Overall, the number of clusters increase over time as images are incrementally added. The cluster count grows rapidly as the robot begins exploring the environment. Over time the clusters capture the topical modes in the visual data and hence the growth rate begins to decline. Significant periods are observed when perceived images are added to existing clusters without increasing the total count, interspersed with occasions when the count marginally increases or decreases when clusters are refined due to new vertices. The graph for a run with a lower threshold shows a more prominent saturation effect and always lies below the graph with a higher threshold.

Next, we analyze the running time components (topic inference, similarity computation and star clustering updates) during online insertion for both data sets with $\sigma = 0.6$. Figure 4.15 and Tables 4.4 and 4.5 provide the details. The topic proportion inference time varies according to the number of words in the scene and

Table 4.5: Running time components for City Center Data set. Total 1683 images with $\sigma = 0.6$ and $T = 50$.

	Inference	**Similarity**	**Clustering**	**Total**
Cumulative (sec)	15.71	1.88	9.32	26.92
Median (ms)	7.86	0.81	0.75	10.24
Min. (ms)	3.67	0.01	0.01	3.79
Max. (ms)	33.40	11.60	1855.76	1861.86
Avg.± std. (ms)	9.34 ± 4.94	1.12 ± 1.38	5.54 ± 58.99	15.99 ± 59.20

was on average 8.81 ± 4.92msec and 9.34 ± 4.94msec per scene respectively for New College and City Center data sets. For each vertex to be inserted the adjacency list is determined by computing its similarity to all existing nodes in the graph. The similarity computation time grows linearly with number of vertices and did not exceed 13.97msec and 11.60msec while inserting 1355 and 1683 images respectively in sequence during the experiment. The overall running time is dominated by the clustering algorithm and is low for most insertions. A few large peaks are observed during insertions when a large number of stars are re-arranged increasing the insertion time, though such events occurred infrequently. The median total insertion times were recorded as 9.16msec and 10.24msec respectively.

4.5.7 Additional Results

We tested the incremental summary generation algorithm on the Wildcat autonomous platform shown in Figure 4.16. Data was collected using a Point Grey's Grasshopper 2 high-resolution camera mounted on top of the vehicle viewing sideways. The vehicle began operation at our field site in Begbroke, Oxfordshire traversing motor ways leading to Woodstock town before returning back to Begbroke.

Figure 4.17 displays the evolving summary at $\sigma = 0.7$. The first image (with a

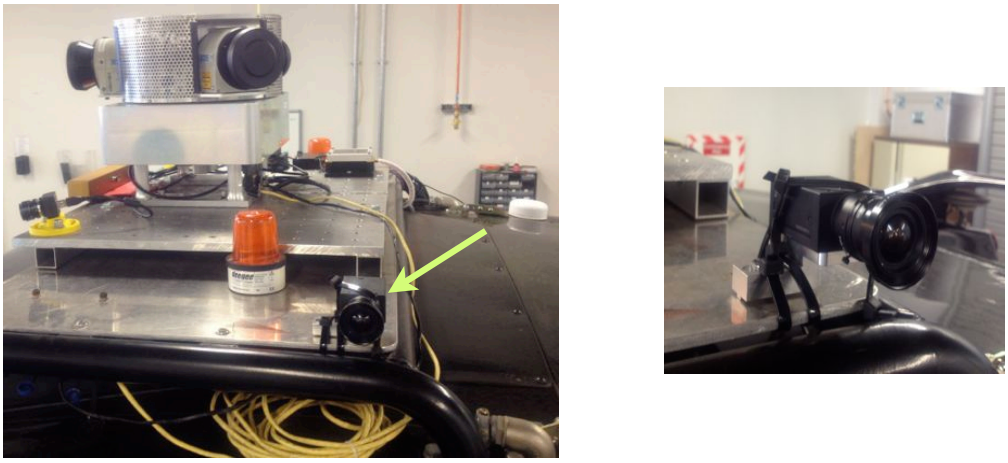


Figure 4.16: The Wildcat autonomous platform (top) with mounted sensor suite. Point Grey Grasshopper 2 camera (below) mounted on top of the vehicle facing sideways indicated with an arrow.



(a) Iteration: 14



(b) Iteration: 48



(c) Iteration: 250



(d) Iteration: 308

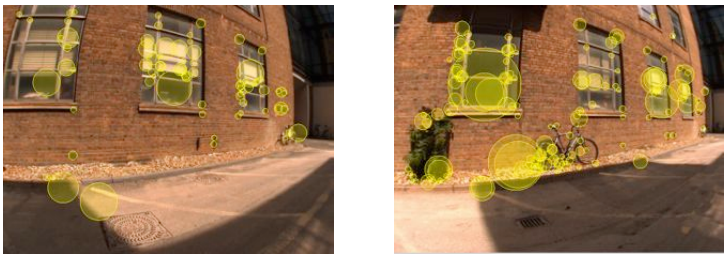
Figure 4.17: Evolving visual summary from the Wildcat autonomous platform mounted with a side facing Point Grey Grasshopper 2 camera. The current observed image is indicated in red and the summary images appear in increasing temporal sequence from left to right and top to bottom. Threshold $\sigma = 0.7$.

red border) shows the current view of the robot followed by the summary images arranged in temporally increasing sequence from left to right and from top to bottom. The initial summaries contain different views of the test site followed by views from the road seeing fields and foliage as the vehicle exits the site, Figures 4.17a and 4.17b . The later summaries contain scenes from the town that include buildings, car park, intersections and city roads as shown in Figures 4.17c and 4.17d.

4.5.8 Other Insights

Topic Model Representation

Figure 4.18 highlights the advantage of topic space representation over a basic bag-of-words representation. Figure 4.18a shows an example where the image pair was taken during two visits to the same location. The second image shows a large number of features appearing on a bicycle which was absent during first visit. These images were found to be in the same clusters using the topic model representation ($\sigma = 0.5$) and in different clusters using visual words representation (even for a threshold as low as $\sigma = 0.05$). Since the bag-of-words representation considers words independently, a large number of features observed on the cycle makes the image pair highly dissimilar. In contrast, the topic representation probabilistically captures co-occurring features and maps both images to common topics, assigning images as highly similar and consequently to the same cluster. Figure 4.18b illustrates a similar example where images were captured sequentially 1m apart. A large number of features appear on repetitive railings and are more frequent in one image compared to the other. The topic model representation probabilistically learns these as correlated, mapping them to common topics yielding high similarity in topic



(a) Loop closure pair.

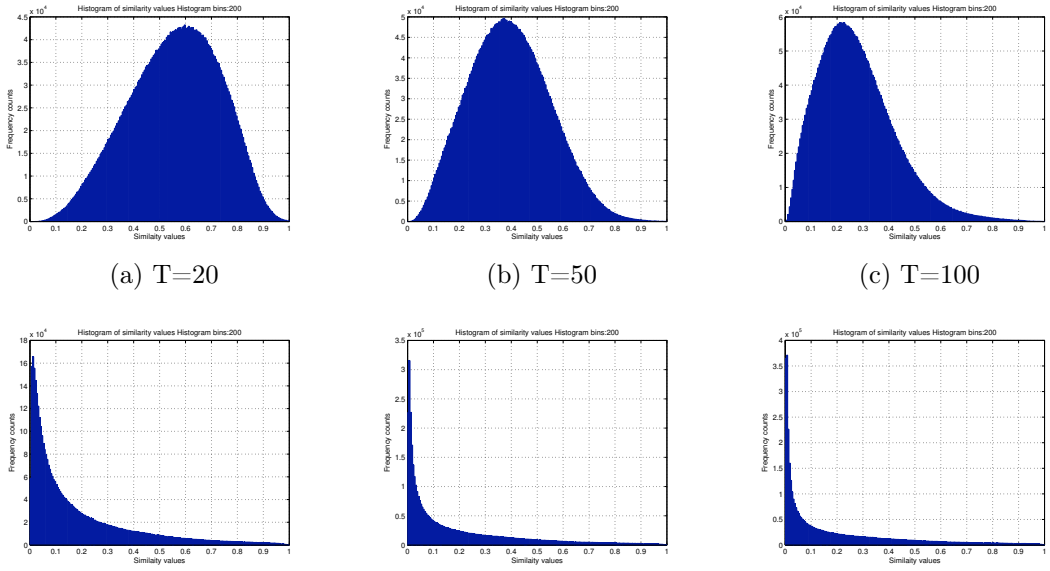


(b) Images collected 1m apart.

Figure 4.18: Image pair captured while (a) re-visiting a location and (b) 1m apart with slight viewpoint change that were assigned to the same clusters using topic model representation and different clusters with the bag-of-visual words representation.

space. The bag-of-words representation considers features independent and hence find the images dissimilar due to high number of repetitive features on railings in one of the images.

In Figure 4.19 we examine the distribution of similarity values for all image pairs in the data set, varying the number of topics for the underlying topic space representation. Simpler models with a lower number of topics ($T = 20$ or $T = 50$) yield a coarse image representation. Most image pairs have similarity around 0.6 and 0.4 indicated by the curve maxima. As the model complexity increases with greater number of topics, the curve shifts towards the origin. Effectively the similarity measure becomes more discerning with increasing topics.



(a) T=20 (b) T=50 (c) T=100

(d) T=500 (e) T=800 (f) T=1000

Figure 4.19: Histogram of pair-wise image cosine similarity for the New College data set, varying number of topics T. As model complexity increases with greater number of topics, the curve shifts towards the origin.

Limitations

Scenes possessing high dynamic contrast, shadows or featureless structures can pose two challenges for our system. First, including them during clustering can yield clusters difficult to visually interpret. Figure 4.21 shows an example image cluster where due to high contrast features were only detected in regions where lighting gradients were observed and not on the physical structure of objects like cars and walls in the scene. Hence, the clusters formed mainly due to similar lighting artifacts in images. Secondly, the total number of features observed in such scenes is much lower (< 70) than average ($\simeq 400$). A smaller number of features in a scene can cause topic inference to result in an uninformative distribution over topics. In some cases, such a distribution can seem highly similar to nearly all the scenes in the



Figure 4.20: Example of images clustered due to features arising from lighting effects.



Figure 4.21: Examples images where very few features are detected owing to over-exposure due high dynamic contrast or textureless structures like a white fence that were excluded from the clustering algorithm.

collection and hence become centers for dominant clusters involving nearly all the images.

In order to mitigate this effect, we excluded images containing features fewer than 70 from the clustering algorithm and collected them separately as a stop list of images. A few examples are shown in Figure 4.20 where very few features were found due to high dynamic contrast, glare or a dominant shadow on a majority of the scene. Note that the clustering algorithm supports an efficient *delete* operation. Hence, a human operator can browse through the image clustering and remove cases with imaging problem. The resultant clustering is guaranteed to be the same as would have been obtained from online (or batch) clustering executed on a corpus excluding the deleted examples.

4.6 Chapter Summary

In this chapter we demonstrated an online incremental approach for generating visual summaries of a robot's workspace. We employed a topic vector space representation for images and an efficient graph-based star clustering algorithm for online organization into thematic clusters forming a compact summary for the robot's visual experience. Importantly, the thematic organization improves with new data collected by the robot resulting in an ever improving workspace summary. We demonstrated the approach on data collected from a mobile platform operating in varied workspaces and presented a detailed evaluation including qualitative and quantitative results.

This chapter culminates our investigation into ever-improving appearance-based topological mapping. We now conclude this thesis by summarizing contributions and outlining avenues for future research in Chapter 5.

"Every finish marks a fresh start. From the end spring new beginnings."

Pliny the Elder, 79 A.D.

Chapter 5

Conclusions and Future Work

This chapter summarises thesis contributions and outlines directions for future research.

5.1 Summary of Contributions

This thesis contributes to the development of appearance-based topological mapping techniques aimed towards life-long robotic operation in outdoor workspaces. The specific contributions are summarized below and are structured according to chapter themes as they appear in this thesis.

Chapter 2: FAB-MAP 3D

- **Probabilistic model for locations incorporating spatial and visual appearance.** We presented a novel formulation where locations are modeled as random graphs possessing latent distributions over word occurrences and pair-wise euclidean distances. Observations are modeled as 3D constellation

of visual features as generated by the underlying probabilistic model via noisy range and visual detectors.

- **Inferring observation likelihood.** We presented inference as a Bayes' filter computing the likelihood of the observed graph given a location model. The graph likelihood term is factorized over the visual and spatial terms permitting an effective combination of word-correlation learning step introduced in prior FAB-MAP work and the spatial likelihood contribution that is conditioned on the observed features.

- **Learning spatial appearance.** Using non-parametric Kernel Density Estimation we demonstrated how distributions over feature distances can be learnt efficiently in a data-driven way forming an informative spatial prior over inter-feature distances before online navigation.

- **Accelerating graph inference.** We presented a method for accelerating graph inference with minimal loss in performance using Delaunay tessellation of the observed graph, scaling log-linearly with scene complexity. The approach relies on the heuristic that objects populating a scene display local spatial correlations. Hence long range distance computations can be ignored in favour of short range distances.

- **Improved precision-recall performance.** We demonstrated the system on a mobile platform equipped with a cross-calibrated camera and Lidar system and showed improved precision-recall performance over vision-only sensing in an outdoor environment.

Chapter 3: Self-Help

- **Introspection and active data acquisition for topological mapping.** We presented a framework that allows a robot to actively improve its own topological mapping performance through introspection and targeted data retrieval. We introduced the idea of a "dynamic" sampling set or onboard workspace representation that adapts with increasing experience as opposed to a "static or hand crafted" representation. We showed how by producing a generative model of observed images using latent visual topics we can actively grow a representative sample set by incorporating well chosen examples from an external corpus.

- **Novelty detection and thematic retrieval.** Using LDA-based probabilistic topic models, we presented a measure of perplexity to evaluate how well a working set of background images explain the robot's online view of the world. This allows the robot to identify the least explained (or most novel) images seen during traversal conditioned on its current workspace representation. Further, we showed how images thematically similar to the confusing ones can be assimilated into the sampling set using a language-model based information retrieval approach.

- **Continual improvement in place recognition performance.** We demonstrated the system on real data sets from a mobile robot operating in an urban environment. Experiments demonstrated incremental sampling set construction starting from a single image yielding a continual improvement in FAB-MAP performance.

Chapter 4: Incremental Visual Summaries

- **Visual workspace summaries.** We presented an incremental approach that allows an exploring robot to generate representative and compact summaries of its visual experience using canonical images capturing both the ordinary as well as the salient aspects of its workspace.

- **Incremental clustering in topic space.** We used a combination of probabilistic topic models that provide a lower-dimensional thematic image representation and an efficient graph clustering algorithm based on incremental star-subgraph covers to generate visual summaries of the robot's workspace. The summaries evolve incrementally with new data, growing with saliency of experience and not merely duration.

- **Experiments.** We demonstrated incremental visual summary generation using imagery collected from mobile platforms operating in varied outdoor settings.

5.2 Future work

In this section we outline possible directions for future research. The discussion is structured according to chapter themes as they appear in this thesis.

Chapter 2: FAB-MAP 3D

- **Correlations between word-pair distances.** The current place model assumes independence between each word-pair distance. However, distances

observed in a graph of visual words originate from objects possessing spatial locality due to the inherent structure of the object. e.g., distances between visual features observed on buildings are spatially correlated. Learning the full joint distribution is clearly computationally intractable. It would be interesting to explore a first-order approximation than can improve inference.

- **Using stereo data.** Stereo cameras provide an alternative to laser data for depth estimation. Although this eliminates the need for a separate range sensor FAB-MAP 3D, the disadvantage is that depth estimates are less-reliable and the field of view is restricted for conventional stereo cameras. It would be worthwhile to evaluate this system with stereo data, particularly using a wide baseline stereo camera system as explored in [55].

- **Multiple occurrences of visual words.** The current formulation utilizes the cardinality of the visual word detections only in the geometric inference engine. The core inference on visual appearance discards this information. The work by [150] attempts to incorporate multiple counts however the model assumes word independence. A generalization to non-binary word occurrences can improve inference though learning correlations may become more difficult.

- **Data driven graph tessellation.** The present Delaunay tessellation relies on the heuristic that local neighbourhood features are correlated and emerge from common objects. However, objects in the real world are of complex shape and sizes and this heuristic might not hold true. There may be gains in exploring better data-driven or adaptive tessellations [31], particularly as the size of the observed graph becomes large for example with panoramic cameras and rotating Lidar scanners.

Chapter 3: Self-Help

- **Deletion or forgetting of samples.** In the present formulation, relevant images are only added to the sampling set. Over time, it might be desirable to remove images no longer relevant and keeping the sampling set size bounded. One possible solution is to prune the most perplexing images from the sampling set given the set of images collected by the robot. Formal exploration of image removal strategies is a potential area of investigation.

- **Inferring the number of topics from data.** The number of topics is assumed known and fixed in the LDA model. In the current formulation, the number of topics was determined via a model selection step by maximizing the log-likelihood of the data given topics. There is scope for improvement using Bayesian non-parametric topic models such as HDPs [181] where the number of topics can also be inferred during posterior inference and new documents can exhibit previously unseen topics.

- **Online adaptation of topic distributions.** Presently, topic distributions are learnt *a-priori* from an image collection and remain static during operation. Topic distributions can be adapted using acquired data during silent periods when the robot is not collecting new imagery. There is scope to explore techniques like incremental Gibbs sampling [5] or recently introduced variational inference for online-LDA [78] and online-HDPs [194].

- **Reducing sensitivity to outliers.** The current perplexity-based saliency measure can be sensitive to outliers as it cannot distinguish between an outlier data point and an unmodeled image representing a new topical mode in data. Outlier images typically arise due to sensor failure or lighting effects. For

example, images with poor contrast or high dynamic range can cause very few features to be detected or images with over-exposed regions can produce feature originating purely due to lighting effects. Such rare images can be highly perplexing to the learnt topics and would lead to erroneous retrieval steps. A few approaches can be considered. First, image quality can be assessed as a pre-processing step, explicitly filtering images with very few features or high proportion of saturated pixels. Another approach could be to incorporate *density-weighting* into the query criteria and only query if the sample comes from a dense region in the topic space [165, 83]. The third approach would be to aggregate samples and learn a generative model for the query image [28] making the retrieval step less sensitive to noise.

Chapter 4: Incremental Visual Summaries

- **Limiting adjacency computation.** While summarizing large data sets, a major bottleneck is computing the adjacency of an inserted image using similarity to all previously obtained images. An approximation worth exploring would be to restrict the similarity computation to only cluster center images and use Equation 4.4 to infer the expected similarity to satellite vertices. Since the average intra-cluster similarity is high, the loss in accuracy is expected to be minor, though a detailed investigation is necessary to derive an error bound.

- **Redundancy removal and hierarchical clustering with coresets.** Often, when the vehicle is moving slowly or collecting imagery at a high frame rate, multiple overlapping and similar images get logged. Hence, for scalability reasons, instead of sending all collected images to the summarizer, a prelimi-

nary clustering can be undertaken to remove redundancy and then summarize only the representative images. Recent work in computational geometry on *coresets* [45, 47, 46] can be applied to this problem. The idea is to to extract subsets from the incremental data stream so as to incur an ϵ-bounded approximation to the resulting star clustering, yielding a data driven two-stage clustering approach. Preliminary investigation into this approach is giving encouraging results.

- **Clustering using multiple attributes.** There is scope to extend the summarization algorithm to incorporate spatial and temporal attributes in addition to visual similarity. For example, consider a geologist who may wish to specify spatial regions of interest for collecting representative imagery. One solution would be to use a spatial prior e.g., using GPS distances and combine them with the visual similarity (according to a user-selectable weight factor) to allow clusters to form in the specified spatial regions.

- **Summarization using data from other sensor domains.** Extending the algorithm to other sensor domains like laser point clouds or RGB-D depth camera data is also worth exploring. The work of [42] describes a way to learn topics based on spin images from indoor laser point clouds. Using this method as a pre-processing step, the same framework can be used to summarize a collection of laser point clouds. Similar ideas can be explored for RGB-D images using depth information in addition to visual information.

- **Topological localization using prior graph structure.** In this thesis we restricted ourselves to the scenario of an exploring robot where the summary was being generated online with incoming data. If the robot is operating in

the same workspace over multiple sessions, the star clustering graph structure generated in the previous run can be used as a topological localizer. An incoming image can be assigned to the most similar cluster center, yielding a trajectory over the graph.

5.3 Concluding Remark

In this thesis we presented: FAB-MAP 3D, a place recognition algorithm that probabilistically captures spatial distribution of visual features observed by a mobile robot equipped with vision and laser sensors; Self-Help, an approach based on probabilistic topic models allowing a robot to improve its own loop-closure performance with use through targeted image retrieval and an incremental approach for generating semantically meaningful summaries for visual maps generated by an exploring robot. This work contributes to the development of appearance-based topological mapping for mobile robots aimed towards life-long continual operation in unstructured outdoor workspaces.

Bibliography

[1] Shivani Agarwal and Dan Roth. Learning a sparse representation for object detection. In *Proceedings of the 7th European Conference on Computer Vision-Part IV*, pages 113–130. Springer-Verlag, 2002. 25

[2] A. Agarwala, M. Dontcheva, M. Agrawala, S. Drucker, A. Colburn, B. Curless, D. Salesin, and M. Cohen. Interactive digital photomontage. In *ACM Transactions on Graphics (TOG)*, volume 23, pages 294–302. ACM, 2004. 131

[3] A. Angeli, D. Filliat, S. Doncieux, and J.A. Meyer. Fast and incremental method for loop-closure detection using bags of visual words. *IEEE Transactions on Robotics, Special Issue on Visual SLAM*, 24(5):1027–1037, 2008. 14

[4] J.A. Aslam, E. Pelekhov, and D. Rus. The star clustering algorithm for static and dynamic information organization. *Journal of Graph Algorithms and Applications*, 8(1):95–129, 2004. 136, 138, 143, 144, 145

[5] A. Banerjee and S. Basu. Topic models over text streams: A study of batch and online unsupervised learning. In *SIAM Data Mining*, 2007. 178

[6] C.B. Barber, D.P. Dobkin, and H. Huhdanpaa. The quickhull algorithm for convex hulls. *ACM Transactions on Mathematical Software (TOMS)*, 22(4):469–483, 1996. 56, 57

[7] H. Bay, T. Tuytelaars, and L. Van Gool. Surf: Speeded up robust features. *Computer Vision–ECCV 2006*, pages 404–417, 2006. 17, 59, 111, 148

[8] N. Bellotto and H. Hu. Multisensor-based human detection and tracking for mobile service robots. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 39(1):167–181, 2009.

[9] P. Biber, H. Andreasson, T. Duckett, and A. Schilling. 3d modeling of indoor environments by a mobile robot with a laser scanner and panoramic camera. In *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems, 2004.(IROS 2004), 2004*, volume 4, pages 3430–3435. IEEE, 2004. 27

[10] P. Biber and T. Duckett. Dynamic maps for long-term operation of mobile service robots. In *Proc. of Robotics: Science and Systems (RSS)*, pages 17–24, 2005.

[11] C. Biernacki, G. Celeux, and G. Govaert. Assessing a mixture model for clustering with the integrated completed likelihood. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(7):719–725, 2000. 130

[12] C.M. Bishop and SpringerLink (Service en ligne). *Pattern recognition and machine learning*, volume 4. springer New York, 2006. 41

[13] J.L. Blanco, J.A. Fernández-Madrigal, and J. Gonzalez. Toward a unified bayesian approach to hybrid metric–topological slam. *IEEE Transactions on Robotics*, 24(2):259–270, 2008. 15

[14] D.M. Blei and J.D. Lafferty. Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*, pages 113–120. ACM, 2006. 79

[15] D.M. Blei and J.D. Lafferty. A correlated topic model of science. *The Annals of Applied Statistics*, pages 17–35, 2007. 79

[16] D.M. Blei and J.D. McAuliffe. Supervised topic models. *Arxiv preprint arXiv:1003.0783*, 2010. 80

[17] D.M. Blei, A.Y. Ng, and M.I. Jordan. Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022, 2003. 78, 82, 85, 86, 135

[18] O. Boiman and M. Irani. Detecting irregularities in images and in video. *International Journal of Computer Vision*, 74(1):17–31, 2007. 83

[19] O. Booij, Z. Zivkovic, and B. Krose. Sparse appearance based modeling for robot localization. In *IEEE/RSJ International Conference on Intelligent Robots and Systems, 2006*, pages 1510–1515. IEEE, 2006. 132

[20] J. Boyd-Graber and D.M. Blei. Syntactic topic models. *Arxiv preprint arXiv:1002.4665*, 2010. 79

[21] E. Brunskill, T. Kollar, and N. Roy. Topological mapping using spectral clustering and classification. In *IEEE/RSJ International Conference on Intelligent Robots and Systems, 2007. IROS 2007.*, pages 3491–3496. Ieee, 2007. 145

[22] M. Burl, M. Weber, and P. Perona. A probabilistic approach to object recognition using local photometry and global geometry. *Computer VisionÑECCV 1998*, pages 628–641, 1998. 24

[23] M.C. Burl, TK Leung, and P. Perona. Face localization via shape statistics. In *International Workshop on Automatic Face and Gesture Recognition*, pages 154–159. Citeseer, 1995. 25

[24] J. Chang and D. Blei. Relational topic models for document networks. In *Artificial Intelligence and Statistics*, pages 81–88, 2009. 79

[25] SC Choi and R. Wette. Maximum likelihood estimation of the parameters of the gamma distribution and their bias. *Technometrics*, pages 683–690, 1969. 46

[26] C. Chow and C. Liu. Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory*, 14(3):462–467, 1968. 18, 19, 47

[27] O. Chum, M. Perdoch, and J. Matas. Geometric min-hashing: Finding a (thick) needle in a haystack. In *IEEE Conference on Computer Vision and Pattern Recognition, 2009. CVPR 2009*, pages 17–24. IEEE, 2009. 24

[28] O. Chum, J. Philbin, J. Sivic, M. Isard, and A. Zisserman. Total recall: Automatic query expansion with a generative feature model for object retrieval. In *IEEE 11th International Conference on Computer Vision, 2007. ICCV 2007.*, pages 1–8. IEEE, 2007. 24, 179

[29] W. Churchill and P. Newman. Practice makes perfect? managing and leveraging visual experiences for lifelong navigation. In *Robotics and Automation*

(ICRA), 2012 IEEE International Conference on, pages 4525–4532. IEEE, 2012.

[30] P. Clough, H. Joho, and M. Sanderson. Automatically organising images using concept hierarchies. In *proceedings of the Multimedia Workshop running at ACM SIGIR conference*, 2005. 130

[31] A. Constantiniu, P. Steinmann, T. Bobach, G. Farin, and G. Umlauf. The adaptive delaunay tessellation: a neighborhood covering meshing technique. *Computational Mechanics*, 42(5):655–669, 2008. 177

[32] M. Cummins and P. Newman. Probabilistic appearance based navigation and loop closing. In *IEEE International Conference on Robotics and Automation, 2007*, pages 2042–2048. IEEE, 2007. 8, 15

[33] M. Cummins and P. Newman. Accelerated appearance-only slam. In *IEEE international conference on Robotics and automation, 2008. ICRA 2008.*, pages 1828–1833. IEEE, 2008. 8, 17, 21

[34] M. Cummins and P. Newman. Fab-map: Probabilistic localization and mapping in the space of appearance. *The International Journal of Robotics Research*, 27(6):647–665, 2008. 3, 5, 8, 13, 15, 16, 17, 19, 20, 21, 49, 50, 52, 53, 60, 61, 62, 76, 104, 105, 106, 118

[35] M. Cummins and P. Newman. Appearance-only slam at large scale with fab-map 2.0. *The International Journal of Robotics Research*, 30(9):1100–1123, 2011. 3, 8, 17, 22, 76, 77, 104

[36] J. Cussens. Bayes and pseudo-bayes estimates of conditional probabilities and their reliability. In *Machine Learning: ECML-93*, pages 136–152. Springer, 1993. 45, 157

[37] F. Dayoub and T. Duckett. An adaptive appearance-based map for long-term topological localization of mobile robots. In *Intelligent Robots and Systems, 2008. IROS 2008. IEEE/RSJ International Conference on*, pages 3364–3369. IEEE, 2008.

[38] B. Delaunay. Sur la sphere vide. *Izv. Akad. Nauk SSSR, Otdelenie Matematicheskii i Estestvennyka Nauk*, 7(793-800):1–2, 1934. 56

[39] T. Denton, J. Abrahamson, and A. Shokoufandeh. Approximation of canonical sets and their applications to 2d view simplification. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, volume 2, pages II–550. IEEE, 2004. 128

[40] T. Denton, M.F. Demirci, J. Abrahamson, A. Shokoufandeh, and S. Dickinson. Selecting canonical views for view-based 3-d object recognition. In *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, volume 2, pages 273–276. IEEE, 2004. 128

[41] Nikolas Engelhard Cyrill Stachniss Dominik Joho, Gian Diego Tipaldi and Wolfram Burgard. Nonparametric bayesian models for unsupervised scene analysis and reconstruction. In *Proceedings of Robotics: Science and Systems*, Sydney, Australia, July 2012. 81

[42] F. Endres, C. Plagemann, C. Stachniss, and W. Burgard. Unsupervised discovery of object classes from range data using latent dirichlet allocation. In *Proceedings of Robotics: Science and Systems*, 2009. 81, 180

[43] E. Fazl-Ersi and J.K. Tsotsos. Histogram of oriented uniform patterns for robust place recognition and categorization. *The International Journal of Robotics Research*, 2012.

[44] L. Fei-Fei and P. Perona. A bayesian hierarchical model for learning natural scene categories. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005. CVPR 2005.*, volume 2, pages 524–531. Ieee, 2005. 80

[45] M. Feigin, D. Feldman, and N. Sochen. From high definition image to low space optimization. *Scale Space and Variational Methods in Computer Vision*, pages 459–470, 2012. 180

[46] D. Feldman, M. Monemizadeh, C. Sohler, and D.P. Woodruff. Coresets and sketches for high dimensional subspace approximation problems. In *Proceedings of the Twenty-First Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 630–649. Society for Industrial and Applied Mathematics, 2010. 180

[47] Dan Feldman, Matthew Faulkner, and Andreas Krause. Scalable training of mixture models via coresets. In J. Shawe-Taylor, R.S. Zemel, P. Bartlett, F.C.N. Pereira, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 2142–2150. 2011. 180

[48] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2003*, volume 2, pages II–264. IEEE, 2003. 25

[49] A.M. Ferman and A.M. Tekalp. Multiscale content extraction and representation for video indexing. In *Proceedings of SPIE*, volume 3229, page 23, 1997. 131

[50] M.A. Fischler and R.A. Elschlager. The representation and matching of pictorial structures. *IEEE Transactions on Computers*, 100(1):67–92, 1973. 24

[51] A. Fitzgibbon and A. Zisserman. Automatic camera recovery for closed or open image sequences. pages 311–326. Springer, 1998. 132

[52] F. Fraundorfer, C. Engels, and D. Nistér. Topological mapping, localization and navigation using image collections. In *IEEE/RSJ International Conference on Intelligent Robots and Systems, 2007. IROS 2007.*, pages 3872–3877. IEEE, 2007. 14

[53] Y. Freund, H.S. Seung, E. Shamir, and N. Tishby. Selective sampling using the query by committee algorithm. *Machine learning*, 28(2):133–168, 1997. 85

[54] C. Galleguillos, A. Rabinovich, and S. Belongie. Object categorization using co-occurrence, location and appearance. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (VCVPR), 2008.*, pages 1–8. Ieee, 2008.

[55] D. Gallup, J.M. Frahm, P. Mordohai, and M. Pollefeys. Variable baseline/resolution stereo. In *IEEE Conference on Computer Vision and Pattern Recognition, 2008. CVPR 2008.*, pages 1–8. IEEE, 2008. 177

[56] Y. Girdhar and G. Dudek. Online navigation summaries. In *IEEE International Conference on Robotics and Automation (ICRA), 2010*, pages 5035–5040. IEEE, 2010. 132

[57] Y. Girdhar and G. Dudek. Onsum: A system for generating online navigation summaries. In *Proceedings of the IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROSÕ10)*, pages 746–751, 2010.

[58] Y. Girdhar and G. Dudek. Offline navigation summaries. In *IEEE International Conference on Robotics and Automation (ICRA), 2011*, pages 5769–5775. IEEE, 2011.

[59] Y. Girdhar and G. Dudek. Online visual vocabularies. In *IEEE Canadian Conference on Computer and Robot Vision (CRV), 2011*, pages 191–196, 2011.

[60] Y. Girdhar, P. Giguere, and G. Dudek. Autonomous adaptive underwater exploration using online topic modelling. In *International Symposium on Experimental Robotics*, 2012.

[61] T. Goedemé, T. Tuytelaars, and L.V. Gool. Visual topological map building in self-similar environments. *Informatics in Control Automation and Robotics*, pages 195–205, 2008. 16

[62] Y. Gong and X. Liu. Video summarization and retrieval using singular value decomposition. *Multimedia Systems*, 9(2):157–168, 2003. 131

[63] K. Granström, T.B. Schön, J.I. Nieto, and F.T. Ramos. Learning to close loops from range data. *The International Journal of Robotics Research*, 30(14):1728–1754, 2011.

[64] T.L. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*, 101(Suppl 1):5228, 2004. 79, 85, 86, 93, 94, 111, 112, 135, 136, 146, 148

[65] S. Guha and S. Khuller. Approximation algorithms for connected dominating sets. *Algorithmica*, 20(4):374–387, 1998. 133

[66] J.S. Gutmann and K. Konolige. Incremental mapping of large cyclic environments. In *Proceedings of the IEEE International Symposium on Computational Intelligence in Robotics and Automation, 1999. CIRA'99.*, pages 318–325. IEEE, 1999. 15

[67] PM Hall and MJ Owen. Simple canonical views. In *The British Machine Vision Conference.(BMVCÕ05*, volume 1, pages 7–16, 2005. 128

[68] B. Han, D. Comaniciu, and L. Davis. Sequential kernel density approximation through mode propagation: applications to background modeling. In *Proc. ACCV*, volume 2004, 2004. 46

[69] B. Han, D. Comaniciu, Y. Zhu, and L. Davis. Incremental density approximation and kernel-based bayesian filtering for object tracking. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1. IEEE Computer Society; 1999, 2004. 46

[70] R. Hartley, A. Zisserman, and Inc ebrary. *Multiple view geometry in computer vision*, volume 2. Cambridge Univ Press, 2003. 133

[71] KR Harvey and GJE Hill. Vegetation mapping of a tropical freshwater swamp in the northern territory, australia: a comparison of aerial photography, landsat tm and spot satellite imagery. *International Journal of Remote Sensing*, 22(15):2911–2925, 2001. 133

[72] D. Hayden, S. Chien, D.R. Thompson, and R. Castano. Onboard clustering of aerial data for improved science return. In *IJCAI Workshop on AI in Space*, 2009. 133

[73] D.S. Hayden, S. Chien, D.R. Thompson, and R. Castano. Using onboard clustering to summarize remotely sensed imagery. *Intelligent Systems, IEEE*, 25(3):86–91, 2010. 133

[74] G. Heinrich. Parameter estimation for text analysis. *Web: http://www. arbylon. net/publications/text-est. pdf*, 2005. 86, 102

[75] A. Hendel, D. Weinshall, and S. Peleg. Identifying surprising events in videos using bayesian topic models. *Computer Vision–ACCV 2010*, pages 448–459, 2011. 82

[76] P. Henry, M. Krainin, E. Herbst, X. Ren, and D. Fox. Rgb-d mapping: Using kinect-style depth cameras for dense 3d modeling of indoor environments. *The International Journal of Robotics Research*, 2012. 27

[77] K.L. Ho and P. Newman. Detecting loop closure with scene sequences. *International Journal of Computer Vision*, 74(3):261–286, 2007. 12

[78] M.D. Hoffman, D.M. Blei, and F. Bach. Online learning for latent dirichlet allocation. *Advances in Neural Information Processing Systems*, 23:856–864, 2010. 178

[79] T. Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 42(1):177–196, 2001. 79

[80] S.C.H. Hoi, R. Jin, J. Zhu, and M.R. Lyu. Batch mode active learning and its application to medical image classification. In *Proceedings of the 23rd international conference on Machine learning*, pages 417–424. ACM, 2006.

[81] A. Holub, P. Perona, and M.C. Burl. Entropy-based active learning for object recognition. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, 2008. CVPRW'08.*, pages 1–8. IEEE, 2008. 84

[82] E. Horster, R. Lienhart, and M. Slaney. Image retrieval on large-scale image databases. In *Proceedings of the 6th ACM international conference on Image and video retrieval*, pages 17–24. ACM, 2007. 97, 99, 100, 101

[83] T. Hospedales, S. Gong, and T. Xiang. Finding rare classes: Active learning with generative and discriminative models. *IEEE Transactions on Knowledge and Data Engineering*, (99):1–1, 2011. 84, 179

[84] T. Hospedales, S. Gong, and T. Xiang. Video behaviour mining using a dynamic topic model. *International Journal of Computer Vision*, pages 1–21, 2011. 83

[85] T. Hospedales, J. Li, S. Gong, and T. Xiang. Identifying rare and subtle behaviours: A weakly supervised joint topic model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (99):1–1, 2011. 83

[86] A.S. Huang, A. Bachrach, P. Henry, M. Krainin, D. Maturana, D. Fox, and N. Roy. Visual odometry and mapping for autonomous flight using an rgb-d

camera. In *Int. Symposium on Robotics Research (ISRR),(Flagstaff, Arizona, USA)*, 2011. 27

[87] L. Itti and P. Baldi. A principled approach to detecting surprising events in video. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005. CVPR 2005.*, volume 1, pages 631–637. IEEE, 2005. 82

[88] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1254–1259, 1998. 130

[89] A. Jaffe, M. Naaman, T. Tassa, and M. Davis. Generating summaries and visualization for large collections of geo-referenced photographs. In *Proceedings of the 8th ACM international workshop on Multimedia information retrieval*, pages 89–98. ACM, 2006. 129

[90] M.C. Jones, J.S. Marron, and S.J. Sheather. A brief survey of bandwidth selection for density estimation. *Journal of the American Statistical Association*, pages 401–407, 1996. 43, 45

[91] M.I. Jordan, Z. Ghahramani, T.S. Jaakkola, and L.K. Saul. An introduction to variational methods for graphical models. *Machine learning*, 37(2):183–233, 1999. 49

[92] J. Joseph, F. Doshi-Velez, A.S. Huang, and N. Roy. A bayesian nonparametric approach to modeling motion patterns. *Autonomous Robots*, pages 1–18, 2011. 81

[93] J. Joseph, F. Doshi-Velez, and N. Roy. A bayesian nonparametric approach to modeling battery health. 2012. 81

[94] A.J. Joshi, F. Porikli, and N. Papanikolopoulos. Multi-class active learning for image classification. In *IEEE Conference on Computer Vision and Pattern Recognition, 2009. CVPR 2009.*, pages 2372–2379. Ieee, 2009. 84

[95] A. Kapoor, K. Grauman, R. Urtasun, and T. Darrell. Gaussian processes for object categorization. *International Journal of Computer Vision*, 88(2):169–188, 2010. 84

[96] K. Konolige and J. Bowman. Towards lifelong visual maps. In *Intelligent Robots and Systems, 2009. IROS 2009. IEEE/RSJ International Conference on*, pages 1156–1163. IEEE, 2009.

[97] K. Konolige, J. Bowman, JD Chen, P. Mihelich, M. Calonder, V. Lepetit, and P. Fua. View-based maps. *The International Journal of Robotics Research*, 29(8):941–957, 2010. 14, 132

[98] J. Kosecka, L. Zhou, P. Barber, and Z. Duric. Qualitative image based localization in indoors environments. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003.*, volume 2, pages II–3. IEEE, 2003. 133

[99] A. Krause and C. Guestrin. Nonmyopic active learning of gaussian processes: an exploration-exploitation approach. In *Proceedings of the 24th international conference on Machine learning*, pages 449–456. ACM, 2007. 84

[100] A. Ladikos, E. Boyer, N. Navab, S. Ilic, et al. Region graphs for organizing image collections. *ECCV 2010 Workshop on Reconstruction and Modeling of Large-Scale 3D Virtual Environments*, 2010. 129

[101] P. Lamon, A. Tapus, E. Glauser, N. Tomatis, and R. Siegwart. Environmental modeling with fingerprint sequences for topological global localization. In *Proceedings of 2003 IEEE/RSJ International Conference on Intelligent Robots and Systems, 2003 (IROS 2003)*, volume 4, pages 3781–3786. IEEE, 2003.

[102] V. Lavrenko. *A generative theory of relevance*, volume 26. Springer-Verlag New York Inc, 2009. 102

[103] N.D. Lawrence, M. Seeger, and R. Herbrich. Fast sparse gaussian process methods: The informative vector machine. volume 15, pages 609–616, 2002. 84

[104] K.H. Lee and R. Ehsani. Comparison of two 2d laser scanners for sensing object distances, shapes, and surface patterns. *Computers and electronics in agriculture*, 60(2):250–262, 2008. 61

[105] A. Levin and R. Szeliski. Visual odometry and map correlation. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, volume 1, pages I–611. IEEE, 2004. 14

[106] X. Li, C. Wu, C. Zach, S. Lazebnik, and J.M. Frahm. Modeling and recognition of landmark image collections using iconic scene graphs. In *Proceedings of the European Conference on Computer Vision, ECCV*, volume 8, 2008. 128

[107] X. Liu and W.B. Croft. Cluster-based retrieval using language models. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 186–193. ACM, 2004. 81, 99

[108] Y. Liu and J. Snoeyink. A comparison of five implementations of 3d delaunay tessellation. *Combinatorial and Computational Geometry*, 52:439–458, 2005. 57

[109] D.G. Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004. 14

[110] W.L.D. Lui and R. Jarvis. A pure vision-based topological slam system. *The International Journal of Robotics Research*, 2012.

[111] C. Lund and M. Yannakakis. On the hardness of approximating minimization problems. *Journal of the ACM (JACM)*, 41(5):960–981, 1994. 143

[112] D.J.C. MacKay and L. Peto. A hierarchical dirichlet language model. *Natural language engineering*, 1(3):1–19, 1995. 79

[113] W. Maddern, M. Milford, and G. Wyeth. Cat-slam: probabilistic localisation and mapping using a continuous appearance-based trajectory. *The International Journal of Robotics Research*, 31(4):429–451, 2012. 15

[114] C.D. Manning, P. Raghavan, and H. Schutze. *Introduction to information retrieval*, volume 1. Cambridge University Press Cambridge, 2008. 62, 97

[115] A. McCallum and K. Nigam. Employing em in pool-based active learning for text classification. In *Proceedings of ICML-98, 15th International Conference on Machine Learning*, pages 350–358, 1998. 85

[116] C. Mei, G. Sibley, and P. Newman. Closing loops without places. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2010*, pages 3738–3744. IEEE, 2010. 15

[117] T. Mei, B. Yang, S.Q. Yang, and X.S. Hua. Video collage: presenting a video sequence using a single image. *The Visual Computer*, 25(1):39–51, 2009. 132

[118] B. Micusik and J. Kosecka. Semantic segmentation of street scenes by superpixel co-occurrence and 3d geometry. In *Proceedings of the IEEE 12th International Conference on Computer Vision Workshops (ICCV Workshops), 2009*, pages 625–632. IEEE, 2009.

[119] B. Micusik, J. Košecká, and G. Singh. Semantic parsing of street scenes from video. *The International Journal of Robotics Research*, 31(4):484–497, 2012.

[120] M. Milford and G. Wyeth. Persistent navigation and mapping using a biologically inspired slam system. *The International Journal of Robotics Research*, 29(9):1131–1153, 2010. 14

[121] Michael Milford. Visual route recognition with a handful of bits. In *Proceedings of Robotics: Science and Systems*, Sydney, Australia, July 2012. 14

[122] J. Miura, Y. Negishi, and Y. Shirai. Mobile robot map generation by integrating omnidirectional stereo and laser range finder. In *IEEE/RSJ International Conference on Intelligent Robots and Systems, 2002*, volume 1, pages 250–255. IEEE, 2002. 27

[123] AC Murillo, J. Kosecka, JJ Guerrero, and C. Sagüés. Visual door detection integrating appearance and shape cues. *Robotics and Autonomous Systems*, 56(6):512–521, 2008.

[124] D. Newman, A. Asuncion, P. Smyth, and M. Welling. Distributed inference for latent dirichlet allocation. *Advances in Neural Information Processing Systems*, 20(1081-1088):17–24, 2007.

[125] P. Newman, G. Sibley, M. Smith, M. Cummins, A. Harrison, C. Mei, I. Posner, R. Shade, D. Schroeter, L. Murphy, et al. Navigating, recognizing and describing urban spaces with vision and lasers. *The International Journal of Robotics Research*, 28(11-12):1406–1433, 2009. 51

[126] A.Y. Ng, M.I. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems*, 2:849–856, 2002. 145

[127] K. Nigam, A.K. McCallum, S. Thrun, and T. Mitchell. Text classification from labeled and unlabeled documents using em. *Machine learning*, 39(2):103–134, 2000. 79

[128] H. Ning, W. Xu, Y. Chi, Y. Gong, and T.S. Huang. Incremental spectral clustering by efficiently updating the eigen-system. *Pattern Recognition*, 43(1):113–127, 2010. 145

[129] D. Nister and H. Stewenius. Scalable recognition with a vocabulary tree. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2006*, volume 2, pages 2161–2168. Ieee, 2006. 14, 28

[130] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3):145–175, 2001. 129

[131] E.B. Olson. *Robust and efficient robotic mapping*. PhD thesis, Massachusetts Institute of Technology, Department of Electrical Engineering and Computer Science, 2008. 15

[132] Edwin Olson and Pratik Agarwal. Inference on networks of mixtures for robust robot mapping. In *Proceedings of Robotics: Science and Systems (RSS)*, Sydney, Australia, July 2012. 16, 52

[133] S.N. Parizi, J. Oberlin, and P.F. Felzenszwalb. Reconfigurable models for scene recognition. CVPR, 2012.

[134] R. Paul and P. Newman. Fab-map 3d: Topological mapping with spatial and visual appearance. In *IEEE International Conference on Robotics and Automation (ICRA), 2010*, pages 2649–2656. IEEE, 2010. 5

[135] R. Paul and P. Newman. Self help: Seeking out perplexing images for ever improving navigation. In *IEEE International Conference on Robotics and Automation (ICRA), 2011*, pages 445–451. IEEE, 2011. 5

[136] R. Paul and P. Newman. How was your day? online visual workspace summaries using incremental clustering in topic space. In *IEEE International Conference on Robotics and Automation (ICRA), 2012*. IEEE, 2012. 6

[137] Rohan Paul, Rudolph Triebel, Danieal Rus, and Paul Newman. Semantic categorization of outdoor scenes with uncertainty estimates using multi-class Gaussian process classification. In *Proc. of the International Conference on Intelligent Robots and Systems (IROS)*, 2012. to appear. 6

[138] M. Perd'och, O. Chum, and J. Matas. Efficient representation of local geometry for large scale object retrieval. In *IEEE Conference on Computer Vision and Pattern Recognition, 2009. CVPR 2009.*, pages 9–16. IEEE, 2009. 24

[139] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *IEEE Conference on*

Computer Vision and Pattern Recognition, 2007. CVPR'07., pages 1–8. Ieee, 2007. 24

[140] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Lost in quantization: Improving particular object retrieval in large scale image databases. In *IEEE Conference on Computer Vision and Pattern Recognition, 2008. CVPR 2008.*, pages 1–8. IEEE, 2008.

[141] J. Philbin, J. Sivic, and A. Zisserman. Geometric lda: A generative model for particular object discovery. In *In Proceedings of the British Machine Vision Conference (BMVC)*. Citeseer, 2008. 80

[142] A. Pigeau and M. Gelgon. Organizing a personal image collection with statistical model-based icl clustering on spatio-temporal camera phone meta-data. *Journal of Visual Communication and Image Representation*, 15(3):425–445, 2004. 130

[143] I. Posner, M. Cummins, and P. Newman. A generative framework for fast urban labeling using spatial and temporal context. *Autonomous Robots*, 26(2):153–170, 2009.

[144] Y. Pritch, A. Rav-Acha, A. Gutman, and S. Peleg. Webcam synopsis: Peeking around the world. In *IEEE 11th International Conference on Computer Vision, 2007. ICCV 2007.*, pages 1–8. IEEE, 2007. 131

[145] M. Quigley, S. Batra, S. Gould, E. Klingbeil, Q. Le, A. Wellman, and A.Y. Ng. High-accuracy 3d sensing for mobile manipulation: Improving object detection and door opening. In *IEEE International Conference on Robotics and Automation, 2009. ICRA'09.*, pages 2816–2822. IEEE, 2009.

[146] A. Ranganathan and F. Dellaert. A rao-blackwellized particle filter for topological mapping. In *Proceedings of the IEEE International Conference on Robotics and Automation, 2006. ICRA 2006.*, pages 810–817. IEEE, 2006. 16

[147] A. Ranganathan and F. Dellaert. Semantic modeling of places using objects. In *Proceedings of the 2007 Robotics Science and Systems Conference*, volume 3, pages 27–30, 2007. 26

[148] A. Ranganathan and F. Dellaert. Bayesian surprise and landmark detection. In *IEEE International Conference on Robotics and Automation (ICRA), 2009.*, pages 2017–2023. IEEE, 2009. 82, 132

[149] A. Ranganathan and F. Dellaert. Online probabilistic topological mapping. *The International Journal of Robotics Research*, 30(6):755–771, 2011. 52

[150] A. Ranganathan, E. Menegatti, and F. Dellaert. Bayesian inference in the space of topological maps. *IEEE Transactions on Robotics*, 22(1):92–107, 2006. 16, 52, 177

[151] A. Rav-Acha, Y. Pritch, and S. Peleg. Making a long video short: Dynamic video synopsis. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2006*, volume 1, pages 435–441. IEEE, 2006. 131

[152] V.C. Raykar and R. Duraiswami. Fast optimal bandwidth selection for kernel density estimation. In *Proceedings of the sixth SIAM International Conference on Data Mining*, pages 524–528. Citeseer, 2006. 45

[153] J. Repko and M. Pollefeys. 3d models from extended uncalibrated video sequences: Addressing key-frame selection and projective drift. In *Fifth Inter-*

national Conference on 3-D Digital Imaging and Modeling, 2005. 3DIM 2005., pages 150–157. IEEE, 2005. 132

[154] M. Rodriguez, S. Ali, and T. Kanade. Tracking in unstructured crowded scenes. pages 1389–1396, 2009. 81

[155] M. Rodriguez, J. Sivic, I. Laptev, and J.Y. Audibert. Data-driven crowd analysis in videos. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1235–1242. IEEE, 2011. 81

[156] C. Rother, L. Bordeaux, Y. Hamadi, and A. Blake. Autocollage. In *ACM Transactions on Graphics (TOG)*, volume 25, pages 847–852. ACM, 2006. 131

[157] C. Rother, S. Kumar, V. Kolmogorov, and A. Blake. Digital tapestry [automatic image synthesis]. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005.*, volume 1, pages 589–596. IEEE, 2005. 130

[158] P. Sala, R. Sim, A. Shokoufandeh, and S. Dickinson. Landmark selection for vision-based navigation. *IEEE Transactions on Robotics*, 22(2):334–349, 2006. 133

[159] A. Sanfeliu. The use of graph techniques for identifying objects and scenes in indoor building environments for mobile robots. *Structural, Syntactic, and Statistical Pattern Recognition*, pages 30–41, 2004. 25, 26

[160] F. Schaffalitzky and A. Zisserman. Multi-view matching for unordered image sets, or "how do i organize my holiday snaps?". In *Proceedings of the 7th European Conference on Computer Vision, Copenhagen, Denmark*, volume 1, pages 414–431. Springer-Verlag, 2002. 129

[161] P. Schmitz. Inducing ontology from flickr tags. In *Collaborative Web Tagging Workshop at WWW2006, Edinburgh, Scotland*, pages 210–214, 2006. 130

[162] T. Sebastian, B. Wolfram, and F. Dieter. Probabilistic robotics, 2005. 40, 43

[163] M. Seeger, C.K.I. Williams, and N.D. Lawrence. Fast forward selection to speed up sparse gaussian process regression. In *Workshop on AI and Statistics*, volume 9, page 2003, 2003. 84

[164] F. Serratosa, R. Alquézar, and A. Sanfeliu. Function-described graphs for modelling objects represented by sets of attributed graphs. *Pattern recognition*, 36(3):781–798, 2003. 26

[165] B. Settles. Active learning literature survey. *University of Wisconsin, Madison*, 2010. 84, 179

[166] C. Silpa-Anan and R. Hartley. Localisation using an image-map. In *Proceedings of the Australian Conference on Robotics and Automation*. Citeseer, 2004. 14

[167] C. Silpa-Anan and R. Hartley. Optimised kd-trees for fast image descriptor matching. In *IEEE Conference on Computer Vision and Pattern Recognition, 2008. CVPR 2008.*, pages 1–8. IEEE, 2008. 14

[168] D. Simakov, Y. Caspi, E. Shechtman, and M. Irani. Summarizing visual data using bidirectional similarity. In *IEEE Conference on Computer Vision and Pattern Recognition, 2008. CVPR 2008.*, pages 1–8. IEEE, 2008. 131

[169] I. Simon, N. Snavely, and S.M. Seitz. Scene summarization for online image collections. In *IEEE 11th International Conference on Computer Vision, 2007. ICCV 2007.*, pages 1–8. Ieee, 2007. 128

[170] G. Singh and J. Kosecka. Visual loop closing using gist descriptors in manhattan world. In *Omnidirectional Robot Vision workshop, with IEEE Int. Conf. on Robotics and Automation*, 2010. 14

[171] J. Sivic, B.C. Russell, A.A. Efros, A. Zisserman, and W.T. Freeman. Discovering object categories in image collections. In *In Proceedings of the Tenth International Conference on Computer Vision*. Citeseer, 2005.

[172] J. Sivic, B.C. Russell, A. Zisserman, W.T. Freeman, and A.A. Efros. Unsupervised discovery of visual object class hierarchies. In *IEEE Conference on Computer Vision and Pattern Recognition, 2008. CVPR 2008.*, pages 1–8. IEEE, 2008. 80

[173] J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *Proceedings of Ninth IEEE International Conference on Computer Vision, 2003.*, pages 1470–1477. Ieee, 2003. 17, 24, 34, 85, 111, 134, 146, 148

[174] M. Smith, I. Baldwin, W. Churchill, R. Paul, and P. Newman. The new college vision and laser data set. *The International Journal of Robotics Research*, 28(5):595–599, 2009. 5, 57, 116

[175] N. Snavely, S.M. Seitz, and R. Szeliski. Skeletal graphs for efficient structure from motion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, 2008. 129

[176] I. Stamos and P.K. Allen. Integration of range and image sensing for photo-realistic 3d modeling. In *IEEE International Conference on Robotics and Automation, 2000. ICRA'00.*, volume 2, pages 1435–1440. IEEE, 2000. 27

[177] Robin Deits Thomas Kollar Stefanie Tellex, Pratiksha Thaker and Nicholas Roy. Toward information theoretic human-robot dialog. In *Proceedings of Robotics: Science and Systems*, Sydney, Australia, July 2012. 85

[178] E.B. Sudderth, A. Torralba, W.T. Freeman, and A.S. Willsky. Describing visual scenes using transformed objects and parts. *International Journal of Computer Vision*, 77(1-3):291–330, 2008. 81

[179] R. Szeliski and H.Y. Shum. Creating full view panoramic image mosaics and environment maps. In *Proceedings of the 24th annual conference on Computer graphics and interactive techniques*, pages 251–258. ACM Press/Addison-Wesley Publishing Co., 1997. 131

[180] A. Tapus and R. Siegwart. Incremental robot mapping with fingerprints of places. In *IEEE/RSJ International Conference on Intelligent Robots and Systems, 2005.*, pages 2429–2434. IEEE, 2005. 27

[181] Y.W. Teh, M.I. Jordan, M.J. Beal, and D.M. Blei. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006. 80, 82, 178

[182] D.R. Thompson, T. Smith, and D. Wettergreen. Information-optimal selective data return for autonomous rover traverse science and survey. In *IEEE International Conference on Robotics and Automation, 2008. ICRA 2008.*, pages 968–973. IEEE, 2008. 133

[183] D.R. Thompson, D. Wettergreen, and R. Castano. Adaptive subsampling of image sequences for remote exploration. 133

[184] S. Thrun, W. Burgard, and D. Fox. A real-time algorithm for mobile robot mapping with applications to multi-robot and 3d mapping. In *IEEE International Conference on Robotics and Automation, 2000. ICRA'00.*, volume 1, pages 321–328. IEEE, 2000.

[185] S. Tong and D. Koller. Support vector machine active learning with applications to text classification. *The Journal of Machine Learning Research*, 2:45–66, 2002. 85

[186] R. Triebel, R. Paul, D. Rus, and P. Newman. Parsing outdoor scenes from streamed 3d laser data using online clustering and incremental belief updates. *Proc. AAAI'12*, 2012. 6

[187] W.H. Tsai and K.S. Fu. Error-correcting isomorphisms of attributed relational graphs for pattern analysis. *IEEE Transactions on Systems, Man and Cybernetics*, 9(12):757–768, 1979. 26

[188] H. Ueda, T. Miyatake, and S. Yoshizawa. Impact: an interactive natural-motion-picture dedicated multimedia authoring system. In *Proceedings of the SIGCHI conference on Human factors in computing systems: Reaching through technology*, pages 343–350. ACM, 1991. 131

[189] C. Valgren and A. Lilienthal. Sift, surf and seasons: Long-term outdoor localization using local features. In *Proceedings of the European conference on mobile robots (ECMR)*, pages 253–258, 2007.

[190] C. Valgren and A. Lilienthal. Incremental spectral clustering and seasons: Appearance-based localization in outdoor environments. In *Robotics and automation, 2008. ICRA 2008. IEEE international conference on*, pages 1856–1861. IEEE, 2008. 145

[191] J. Varadarajan and J.M. Odobez. Topic models for scene analysis and abnormality detection. In *IEEE 12th International Conference on Computer Vision Workshops (ICCV Workshops)*, pages 1338–1345. IEEE, 2009. 82

[192] H.M. Wallach. Topic modeling: beyond bag-of-words. In *Proceedings of the 23rd international conference on Machine learning*, pages 977–984. ACM, 2006. 79

[193] C. Wang, D. Blei, and F.F. Li. Simultaneous image classification and annotation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1903–1910. IEEE, 2009. 80

[194] C. Wang, J. Paisley, and D.M. Blei. Online variational inference for the hierarchical dirichlet process. In *Proceedings AISTATS*, 2011. 178

[195] J. Wang, L. Quan, J. Sun, X. Tang, and H.Y. Shum. Picture collage. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2006*, volume 1, pages 347–354. IEEE, 2006. 130

[196] T. Wang, T. Mei, X.S. Hua, X.L. Liu, and H.Q. Zhou. Video collage: A novel presentation of video sequence. In *IEEE International Conference on Multimedia and Expo, 2007*, pages 1479–1482. IEEE, 2007. 130

[197] X. Wang and E. Grimson. Spatial latent dirichlet allocation. *Advances in Neural Information Processing Systems*, 20:1577–1584, 2007. 80

[198] X. Wang, X. Ma, and W.E.L. Grimson. Unsupervised activity perception in crowded and complicated scenes using hierarchical bayesian models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(3):539–555, 2009. 82

[199] Y. Wang and G. Mori. Human action recognition by semilatent topic models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(10):1762–1774, 2009. 81

[200] X. Wei and W.B. Croft. Lda-based document models for ad-hoc retrieval. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 178–185. ACM, 2006. 81, 100

[201] B. Williams, M. Cummins, J. Neira, P. Newman, I. Reid, and J. Tardós. An image-to-map loop closing method for monocular slam. In *IEEE/RSJ International Conference on Intelligent Robots and Systems, 2008. IROS 2008.*, pages 2053–2059. IEEE, 2008.

[202] B. Williams, M. Cummins, J. Neira, P. Newman, I. Reid, and J. Tardós. A comparison of loop closing techniques in monocular slam. *Robotics and Autonomous Systems*, 57(12):1188–1197, 2009.

[203] A.K.C. Wong and M. You. Entropy and distance of random graphs with application to structural pattern recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (5):599–609, 1985. 26

[204] C. Yang, R. Duraiswami, N.A. Gumerov, and L. Davis. Improved fast gauss transform and efficient kernel density estimation. In *Ninth IEEE International Conference on Computer Vision, 2003.*, pages 664–671. Ieee, 2003. 45

[205] Y. Yang and S. Newsam. Spatial pyramid co-occurrence for image classification. In *Proceedings of the IEEE International Conference onComputer Vision (ICCV), 2011*, pages 1465–1472. IEEE, 2011.

[206] C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to ad hoc information retrieval. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 334–342. ACM, 2001. 99, 101

[207] S. Zhang, Q. Huang, Y. Lu, G. Wen, and Q. Tian. Building pair-wise visual word tree for efficent image re-ranking. In *Proceedings of the IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP), 2010*, pages 794–797. IEEE, 2010.

[208] Y. Zhang, J. Callan, and T. Minka. Novelty and redundancy detection in adaptive filtering. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 81–88. ACM, 2002. 83, 102

[209] Z. Zivkovic, B. Bakker, and B. Krose. Hierarchical map building using visual landmarks and geometric constraints. In *IEEE/RSJ International Conference on Intelligent Robots and Systems, 2005, (IROS 2005)*, pages 2480–2485. IEEE, 2005. 133