Invited Applications Paper

FAB-MAP: Appearance-Based Place Recognition and Mapping using a Learned Visual Vocabulary Model

Mark Cummins MJC@ROBOTS.OX.AC.UK
Paul Newman PNEWMAN@ROBOTS.OX.AC.UK
University of Oxford, Parks Road, Oxford, OX1 3PJ, UK

Abstract

We present an overview of FAB-MAP, an algorithm for place recognition and mapping developed for infrastructure-free mobile robot navigation in large environments. The system allows a robot to identify when it is revisiting a previously seen location, on the basis of imagery captured by the robot's camera. We outline a complete probabilistic framework for the task, which is applicable even in visually repetitive environments where many locations may appear identical. Our work introduces a number of technical innovations - notably we demonstrate that place recognition performance can be improved by learning an approximation to the joint distribution over visual elements. We also investigate several principled approaches to making the system robust in visually repetitive environments, and define an efficient bail-out strategy for multi-hypothesis testing to improve system speed. Our model has been shown to substantially outperform standard tf-idf ranking on our task of interest. We demonstrate the system performing reliable online appearance mapping and loop closure detection over a 1,000 km trajectory, with mean filter update times of 14 ms.

1. Introduction

This paper reviews FAB-MAP (Fast Appearance Based Mapping), a technique for place recognition and mapping developed for mobile robotics applications. It addresses some key aspects of the navigation problem, which is a core task for autonomous robots. FAB-MAP is developed in detail in (Cummins & Newman, 2008a;b; 2009; Cummins, 2009; Newman et al., 2009), and is presented in overview here.

For many tasks, a robot must be able to reliably determine its position within its environment. It is often necessary to solve this navigation problem on the basis of the robot's internal sensors alone, without the aid of external infrastructure. This can dictated by considerations of flexibility and costs, or by necessity. For example, no infrastructure is available to planetary exploration robots, and GPS is unavailable in many terrestrial environments such as indoors, near tall buildings, under foliage, underground and underwater. In these situations the robotic system must solve the navigation problem unaided. This problem is known as Simultaneous Localization and Mapping (SLAM), and has been an active area of research in mobile robotics for several decades.

A broad class of techniques, which we will refer to as "metric SLAM", approach the problem by jointly maintaining an estimate of the pose of the robot and a set of map landmarks using an Extended Kalman Filter or particle filter (Durrant-Whyte & Bailey, 2006). These metric SLAM techniques have been very successful in small to moderately sized environments, but tend to experience problems at larger scales. In particular, it is common for these techniques to fail when a robot revisits a previously seen location after conducting a long traverse through unexplored terrain. This is known as the "loop closure problem". There are a number of reasons why this situation poses a challenge to SLAM algorithms; one of the most fundamental is that map landmarks are typically tracked locally, without any efficient method for recognising them when a location is revisited.

FAB-MAP and related approaches, which we term "appearance-only SLAM" or appearance-based navigation, have been developed as a solution to the problem of loop closure detection. In contrast to metric

Appearing in *Proceedings of the 27^{th} International Conference on Machine Learning*, Haifa, Israel, 2010. Copyright 2010 by the author(s)/owner(s).

SLAM techniques, we do not maintain any explicit estimate of vehicle position, but rather aim to recognise places via their appearance. The system navigates in "appearance space", assigning each new observation to either a new or previously visited location, without reference to metric position. The map of locations is constructed fully incrementally, even in repetitive environments. Because distinctive places can be recognized after unknown vehicle motion, appearance-only SLAM techniques provide a natural solution to loop-closure detection, multi-session mapping and kidnapped robot problems which are challenging for metric SLAM systems.

2. Outline of the Problem

In constructing our appearance-based navigation system, the core problem that we must address is the following: Given two observations or sequences of observations (typically images), what is the probability that these observations were collected at the same location?

This is a hard problem for a number of reasons. Firstly, the world is dynamic. Two images of the same location may look different due to changes in lighting or viewpoint, arrival or departure of vehicles, changes in weather, etc. Secondly, and more challengingly, the world is visually repetitive. Common textures such as brickwork and foliage are present everywhere. Mass produced objects and symbols such as road markings appear throughout the environment. These problems are illustrated in Figure 1. Under many reasonable metrics, the image pair shown in Figure 1(b) is more similar than the pair in 1(a). In our bag-of-words representation (introduced later), 1(b) share 28% of their visual words, whereas 1(a) share only 15%. However, we would like to build a system that asserts with high confidence that 1(a) shows two images of the same location, whereas 1(b) probably does not. To do this, we must go beyond a simple computation of image similarity, and build a model that allows us to evaluate the *distinctiveness* of images. This will involve learning a probability distribution over the space of images, which captures the fact that the scene in Figure 1(b) is common in the environment, so despite the similarity the images are unlikely to come from the same location.

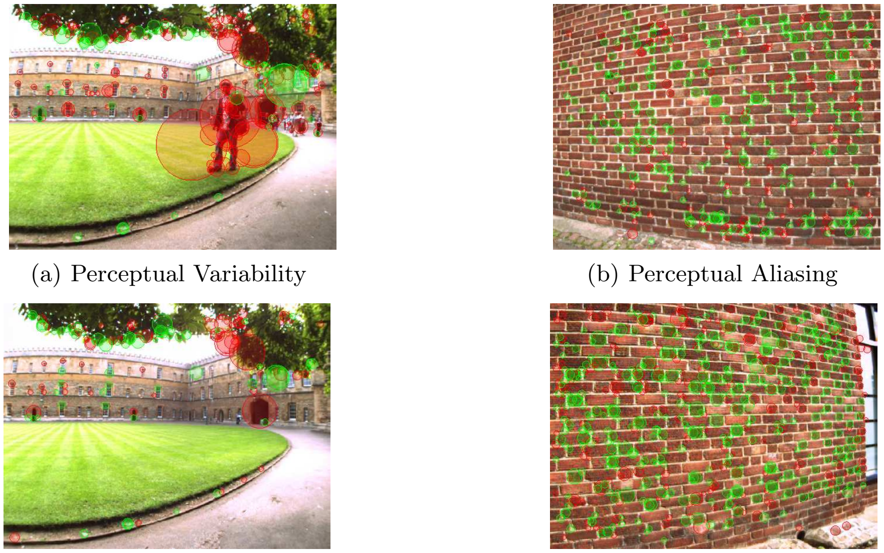
3. Related Work

While appearance-based navigation has a long history within robotics (Dudek & Jugessur, 2000), there has been considerable development in the field in the last five years. Appearance-based navigation and loop closure detection systems operating on trajectories on the order of a few kilometers in length are now commonplace. Indeed, place recognition systems similar in character to the one described here are now used even in single-camera SLAM systems designed for small-scale applications (Eade & Drummond, 2008).

Use of these systems on the scale of tens of kilometers or more has also begun to be feasible. For example, in (Milford & Wyeth, 2008) a system employing a set of biologically inspired approaches achieved successful loop closure detection and mapping in a collection of more than 12,000 images from a 66 km trajectory, with processing time of less than 100 ms per image. The appearance-recognition component of the system was based on direct template matching, so scaled linearly with the size of the environment. Operating at a similar scale, Bosse and Zlot describe a place recognition system based on distinctive keypoints extracted from 2D lidar data (Bosse & Zlot, 2008), and demonstrate good precision-recall performance over an 18 km suburban data set. In (Cummins & Newman, 2009), we demonstrated successful loop closure detection on a 1,000 km trajectory, using a version of FAB-MAP modified to work with an inverted index architecture.

Another recent research direction is the development of integrated systems which combine appearance and metric information. Olson described an approach to increasing the robustness of general loop closure detection systems by using both appearance and relative metric information to select a single consistent set of loop closures from a larger number of candidates (Olson, 2008). The method was evaluated over several kilometers of urban data and shown to recover high-precision loop closures even with the use of artificially poor image features. More loosely coupled systems have also recently described in (Konolige et al., 2009; Newman et al., 2009).

Considerable relevant work also exists on the more restricted problem of global localization. For example, Schindler *et al.* describe a city-scale location recognition system (Schindler et al., 2007) based on the vocabulary tree approach of (Nistér & Stewenius, 2006). The system was demonstrated on a 30,000 image data set from 20 km of urban streets, with retrieval times below 200 ms. Also of direct relevance is the research on content-based image retrieval systems in the computer vision community, where systems have been described that deal with more than a million images (Philbin et al., 2007; Nistér & Stewenius, 2006; Jégou et al., 2008) . However, the problem of retrieval from a fixed index is considerably easier than the full loop-closure



(a) Perceptual Variability

(b) Perceptual Aliasing

Figure 1. Appearance based navigation systems must assess the probability that two images come from the same place. This is difficult because place appearance can vary, (a), and more challengingly, because different places in the environment can appear identical, (b). The appearance based navigation problem is thus harder than typical content based image retrieval tasks.

problem, because it is possible to tune the system directly on the images to be recognized, and the difficult issue of new place detection does not arise. We believe the results presented in (Cummins & Newman, 2009) and summarized in this paper represent the largest scale system that fully addresses these issues of incrementality and perceptual aliasing.

4. Technical Overview

The FAB-MAP system is at heart a bag-of-words image retrieval system similar to those developed in the computer vision community (Sivic & Zisserman, 2003). While these systems employ many image-specific optimisations in parts of their architecture, the final ranking of images is almost universally based on the standard tf-idf relevance measure borrowed from text retrieval. We argue that tf-idf ranking is not particularly suitable for image data, and that by paying attention to the specific qualities of imagery, much better ranking measures can be derived.

For our FAB-MAP navigation system, we develop a probabilistic model on top of the bag-of-words representation which allows us to define appearance-based navigation as recursive Bayesian filtering problem. This framework is similar to many other established SLAM techniques. In defining this model, the question arises of how to account for the probabilistic nature of visual word occurrence, and how to treat the correlations between visual words. We show that by approximating the joint distribution using a Chow Liu tree, inference performance can be improved relative to a naive Bayes model. We also introduce a noisy detector model to account for the unreliable nature of visual feature detection. We show that for our task the model outperforms the standard tf-idf ranking. This work is outlined in (Cummins & Newman, 2008a; 2007; 2009; Cummins, 2009).

Our initial model is somewhat slow to evaluate, limiting application to a few thousand images or about two kilometers of robot exploration. In subsequent work we developed a number of techniques to improve the speed of the system. In (Cummins & Newman, 2008b), we in introduced a probabilistic bail-out test based on the use of concentration inequalities (specifically Bennett's inequality), which enabled the system to rapidly identify promising location hypotheses and exclude less likely locations from further processing. This yielded 25-50x speedup with minimal loss of accuracy. The general approach is applicable to many types of multi-hypothesis testing. In (Cummins & Newman, 2009; Cummins, 2009) we introduced a second approach to improving system speed. This subsequent work adapts our probabilistic model for use with an inverted index architecture similar to typical search engines. This approach allowed us to apply FAB-MAP to navigation problems on the very largest scale, investigating performance on robot navigation tasks on trajectories 1,000 km in length.

5. Probabilistic Model

Our probabilistic model outlined in detail in (Cummins & Newman, 2008a; Cummins, 2009). We review it briefly here.

The basic data representation used is the bag-of-words approach developed in the computer vision community (Sivic & Zisserman, 2003). Features are detected in raw sensory data, and these features are then quantized with respect to a *vocabulary*, yielding *visual words*. The vocabulary is learned by clustering all feature vectors from a set of training data. The Voronoi regions of the cluster centres then define the set of feature vectors that correspond to a particular visual word. The continuous space of feature vectors is thus mapped into the discrete space of visual words, which enables the use of efficient inference and retrieval techniques. In this paper, the raw sensor data of interest is imagery, processed with the SURF feature detector (Bay et al., 2006), though in principle the approach is applicable to any sensor or combination of sensors, and we have explored multi-sensory applications elsewhere.

FAB-MAP, our appearance-only SLAM system, defines a probabilistic model over the bag-of-words representation. An observation of local scene appearance captured at time k is denoted $Z_k = \{z_1, \ldots, z_{|v|}\}$, where $|v|$ is the number of words in the visual vocabulary. The binary variable z_q, which we refer to as an observation component, takes value 1 when the q^{th} word of the vocabulary is present in the observation. $\mathcal{Z}^k$ is used to denote the set of all observations up to time k.

At time k, our map of the environment is a collection 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, which we parameterize in terms of unobservable "scene elements", e_q. A detector yields visual word observations z_q, which are noisy measurements of the existence of the underlying scene element e_q. The appearance model of a location in the map is our belief about the existence of each scene element at that location:

$$L_i : \{p(e_1 = 1|L_i), \ldots, p(e_{|v|} = 1|L_i)\} \tag{1}$$

where each of the scene elements e_q are generated independently by the location. A detector model relates scene elements e_q to feature detection z_q. The detector is specified by

$$\mathcal{D} : \begin{cases} p(z_q = 1|e_q = 0), & \text{false positive probability.} \\ p(z_q = 0|e_q = 1), & \text{false negative probability.} \end{cases} \tag{2}$$

A further salient aspect of the data is that visual words do not occur independently – indeed, word occurrence tends to be highly correlated. For example, words associated with car wheels and car doors are likely to be observed simultaneously. We capture these dependencies by learning a tree-structured Bayesian network using the Chow Liu algorithm, which yields the optimal approximation to the joint distribution over word occurrence within the space of tree-structured networks. Importantly, tree-structured networks also permit efficient learning and inference even for very large visual vocabulary sizes.

Given our probabilistic appearance model, localization and mapping can be cast as a recursive Bayes estimation problem, closely analogous to metric SLAM. A pdf over location given the set of observations up to time k is given by:

$$p(L_i|\mathcal{Z}^k) = \frac{p(Z_k|L_i, \mathcal{Z}^{k-1})p(L_i|\mathcal{Z}^{k-1})}{p(Z_k|\mathcal{Z}^{k-1})} \tag{3}$$

Here $p(L_i|\mathcal{Z}^{k-1})$ is our prior belief about our location, $p(Z_k|L_i, \mathcal{Z}^{k-1})$ is the observation likelihood, and $p(Z_k|\mathcal{Z}^{k-1})$ is a normalizing term. We briefly discuss the evaluation of each of these terms below. For a detailed treatment we refer readers to (Cummins & Newman, 2008a; Cummins, 2009).

Observation Likelihood

To evaluate the observation likelihood, we assume independence between the current and past observations conditioned on the location, and make use the Chow Liu model of the joint distribution, yielding:

$$p(Z_k|L_i)=p(z_r|L_i)\prod_{q=2}^{|v|} p(z_q|z_{p_q}, L_i) \tag{4}$$

where z_r is the root of the Chow Liu tree and z_{p_q} is the parent of z_q in the tree. After some further manipulation (see (Cummins & Newman, 2008a)), each term in the product can be further expanded as:

$$p(z_q|z_{p_q}, L_i) = \sum_{s_{e_q} \in \{0,1\}} p(z_q|e_q = s_{e_q}, z_{p_q})p(e_q = s_{e_q}|L_i) \tag{5}$$

which can be evaluated explicitly.

In some configurations of the system we find that these likelihood are be too peaked, so we introduce an optional smoothing step which can be applied:

$$p(Z_k|L_i) \longrightarrow \sigma p(Z_k|L_i) + \frac{(1-\sigma)}{n_k} \tag{6}$$

where n_k is the number of places in the map and σ is the smoothing parameter, which we typically set to be slightly less than 1.

Location Prior

The location prior $p(L_i|\mathcal{Z}^{k-1})$ is obtained by transforming the previous position estimate via a simple motion model. The model assumes that if the vehicle is at location i at time $k-1$, it is likely to be at one of the topologically adjacent locations at time k.

Normalization

In contrast to a localization system, a SLAM system requires an explicit evaluation of the normalizing term $p(Z_k|\mathcal{Z}^{k-1})$. The normalizing term converts the appearance likelihood into a probability of loop closure, by accounting for the possibility that the current observation comes from a location not currently in the robot's map. Intuitively $p(Z_k|\mathcal{Z}^{k-1})$ is a measure of the distinctiveness of an observation, and thus directly related to the problem of perceptual aliasing.

To calculate $p(Z_k|\mathcal{Z}^{k-1})$, we divide the world into the set of places in our current map, $\mathcal{L}^k$, and the set of unmapped places $\overline{\mathcal{L}^k}$, so that

$$p(Z_k|\mathcal{Z}^{k-1}) = \sum_{m\in\mathcal{L}^k} p(Z_k|L_m)p(L_m|\mathcal{Z}^{k-1}) \quad (7)$$

$$+ \sum_{u\in\overline{\mathcal{L}^k}} p(Z_k|L_u)p(L_u|\mathcal{Z}^{k-1}) \quad (8)$$

The second summation cannot be evaluated directly because it involves all possible unknown locations. However, if we have a large set of randomly collected location models L_u, (readily available from previous runs of the robot or other suitable data sources such as, for our application, Google Street View), we can approximate the summation by Monte Carlo sampling. Assuming a uniform prior over the samples, this yields:

$$p(Z_k|\mathcal{Z}^{k-1}) \approx \sum_{m\in\mathcal{L}^k} p(Z_k|L_m)p(L_m|\mathcal{Z}^{k-1}) \quad (9)$$

$$+p(L_{new}|\mathcal{Z}^{k-1}) \sum_{u=1}^{n_s} \frac{p(Z_k|L_u)}{n_s} \quad (10)$$

where n_s is the number of samples used, and $p(L_{new}|\mathcal{Z}^{k-1})$ is our prior probability of being at a new location.

Data Association

Once the pdf over locations is computed, a data association decision is made. The observation Z_k is used either to initialize a new location, or update the appearance model of an existing location. Recall that an appearance model consists of a set of beliefs about the existence of scene elements at the location, $\{p(e_1=1|L_i),\ldots,p(e_{|v|}=1|L_i)\}$. Each component of the appearance model can be updated according to:

$$p(e_i=1|L_j,\mathcal{Z}^k) = \frac{p(Z_k|e_i=1,L_j)p(e_i=1|L_j,\mathcal{Z}^{k-1})}{p(Z_k|L_j)} \quad (11)$$

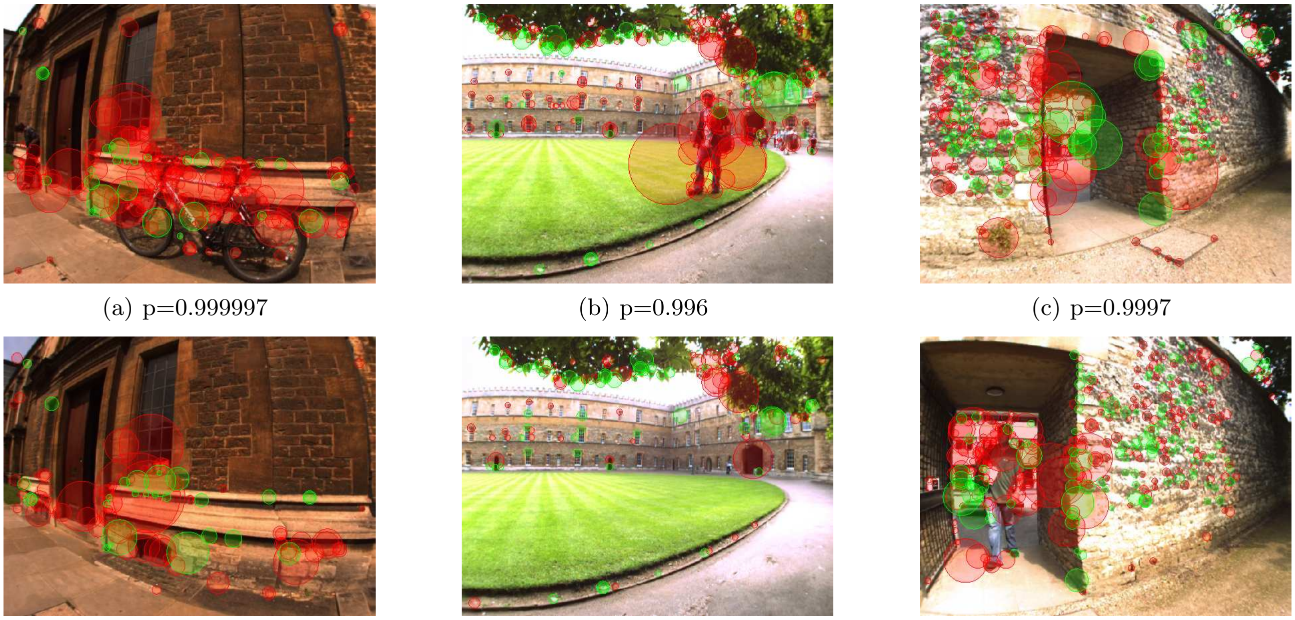
In the case of new locations, the values $p(e_i=1|L)$ are first initialized to the marginal probability $p(e_i=1)$ derived from training data, and then the update is applied.

6. Overview of Results

For a detailed evaluation of the system, we refer readers to our other publications. We briefly present typical results here.

Some examples of typical image matching results are presented in Figures 2 and 3. Figure 3 highlights robustness to perceptual aliasing. Here very similar images that originate from different locations are correctly assigned low probability of having come from the same place. We emphasize that these results are not outliers; they show typical system performance. The result is possible because most of the probability mass for the query image is captured by locations in the sampling set – effectively the system has learned that images like these are common in the environment. Of course, had these examples been genuine loop closures they might also have been assigned low probability values. We would argue that this is correct behaviour, modulo the fact that the probabilities in 3a and 3b seem too low. The very low probabilities for these examples are due to the fact that good matches for the query images are found in the sampling set, capturing almost all the probability mass. This is less likely in the case of a true but ambiguous loop closure.

Figure 2 shows matching performance in the presence of scene change. Many of these image pairs have far fewer visual words in common than the examples of perceptual aliasing, yet are assigned high probability of having come from the same place. To give a quantitative example, loop closures detected by the system sometimes have as few as 8% of their visual words in common, whereas perceptual aliasing examples often share most of their visual words – for example, the images shown in Figure 3b have 46% of their words in

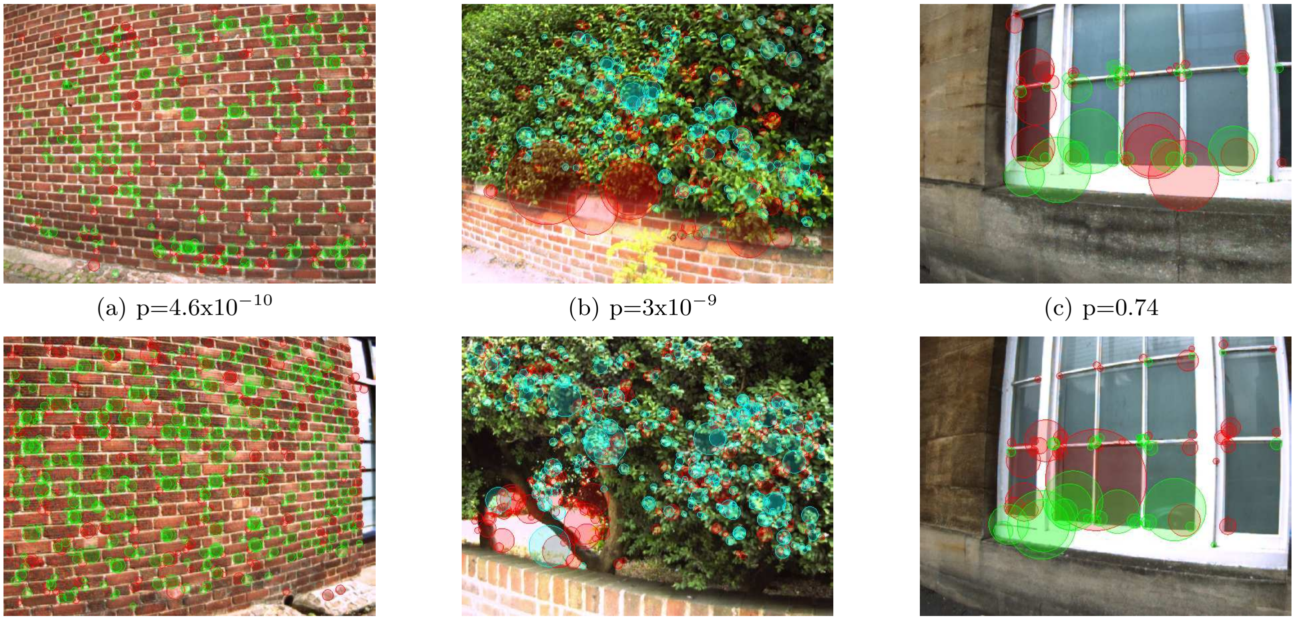


(a) p=0.999997

(b) p=0.996

(c) p=0.9997

Figure 2. Some images assigned high probability of having come from the same place, despite scene change. Words common to both images are shown in green, others in red. Loop closure probability indicated between the pairs.



(a) p=4.6x10^{-10}

(b) p=3x10^{-9}

(c) p=0.74

Figure 3. Some similar-looking images from different parts of the workspace correctly assigned low probability of having come from the same place. Words common to both images are shown in green (in blue for (b)), others in red.

common.

6.1. Comparing Approximations

This section highlights the effect of correctly accounting for the correlations between the words of the visual vocabulary. We compare the Naive Bayes vs. Chow Liu approximations to $p(Z|L_i)$, and also compare different approaches to the evaluation of the partition function $p(Z_k|\mathcal{Z}^{k-1})$. For the partition function we compare the Monte Carlo approximation outlined in Section 5 with a simpler mean field approximation outlined in (Cummins & Newman, 2008a). Figure 4 shows precision-recall curves from a 2 km data set for the four possible combinations.

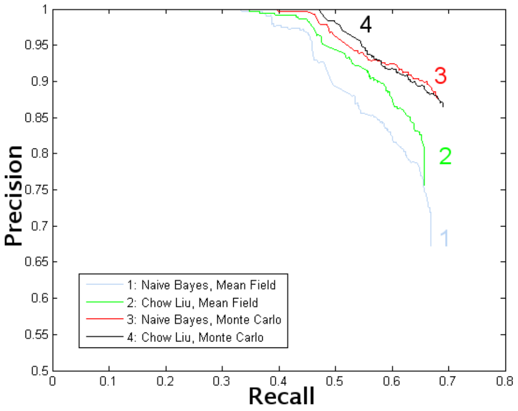
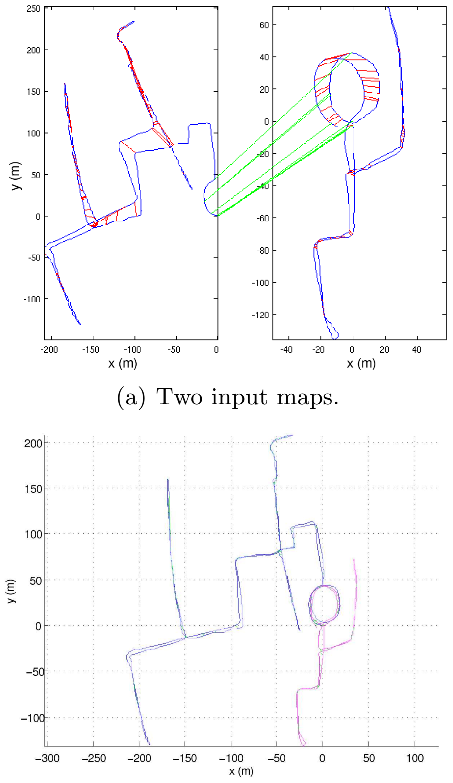


Figure 4. Precision-Recall curves for the four variant algorithms on the New College data set. Notice the scale. All results include the location prior which allows evidence to accumulate through a sequence of observations.

6.2. Use with Metric SLAM Systems

The motivation for the development of FAB-MAP was to solve the loop closure detection and multi-session mapping problems for metric SLAM systems. Typical results for this application are shown in Figure 5, where the alignment between two partially overlapping data sets collected on different days is recovered. The vehicle trajectory, consisting of a set of vehicle poses and the relative metric transformations between them (with associated uncertainty), is determined by a visual SLAM system developed in (Mei et al., 2009). Loop closure constraints between poses are determined by FAB-MAP. The relative metric transformation for these constraints is determined by the visual SLAM system using stereo image pairs collected by the robot at each location. The final trajectory estimate is obtained by relaxing the pose graph using the method described in (Newman et al., 2009).



(a) Two input maps.

(b) Single output map.

Figure 5. Use of FAB-MAP for multi-session mapping. Sub-figure (a) shows two robot trajectories collected on different days (blue), which have an unknown transformation between them. Place recognition constraints between poses in these trajectories are detected by FAB-MAP. Constraints within each trajectory are shown as red links, and between the two trajectories as green links. Pose graph relaxation yields the final output trajectory shown in (b). The two trajectories now share a common coordinate frame.

7. Summary

This paper has presented an overview of a new approach to appearance-only SLAM. The framework is fully probabilistic, and deals with challenging issues such as perceptual aliasing and new place detection. In other work (Cummins, 2009), we have shown that as a pure ranking function it considerably out-performs the baseline tf-idf approach, at least for our task. In (Cummins, 2009) we evaluated the system on two substantial data sets, of 70 km and 1,000 km. Both experiments are larger than any existing result we are aware of. Our approach shows very strong performance on the 70 km experiment, in conditions of challenging perceptual aliasing. The 1,000 km experiment is more challenging, and we do not consider it fully solved, nevertheless our system's performance is already sufficient to provide a useful competency for an autonomous vehicle operating at this scale. Our data sets are available to the research community, and we hope that they will serve as a benchmark for future systems.

Acknowledgments

The work reported in this paper was funded by the Systems Engineering for Autonomous Systems (SEAS) Defence Technology Centre established by the UK Ministry of Defence and by the EPSRC.

References

Bay, H., Tuytelaars, T., and Van Gool, L. SURF: Speeded Up Robust Features. In *Proc 9th European Conf on Computer Vision*, volume 13, pp. 404–417, Graz, Austria, May 7 2006.

Bosse, M. and Zlot, R. Keypoint design and evaluation for place recognition in 2D lidar maps. In *Robotics: Science and Systems Conference : Inside Data Association Workshop*, 2008.

Cummins, M. and Newman, P. Probabilistic appearance based navigation and loop closing. In *Proc. IEEE International Conference on Robotics and Automation (ICRA'07)*, Rome, April 2007.

Cummins, M. and Newman, P. FAB-MAP: Probabilistic Localization and Mapping in the Space of Appearance. *The International Journal of Robotics Research*, 27:647–665, 2008a.

Cummins, M. and Newman, P. Accelerated appearance-only SLAM. In *Proc. IEEE International Conference on Robotics and Automation (ICRA'08)*, Pasadena, California, April 2008b.

Cummins, M. and Newman, P. Highly scalable appearance-only SLAM - FAB-MAP 2.0. In *Proceedings of Robotics: Science and Systems*, Seattle, USA, June 2009.

Cummins, Mark. *Probabilistic Localization and Mapping in Appearance Space.* PhD thesis, University of Oxford, 2009.

Dudek, G. and Jugessur, D. Robust place recognition using local appearance based methods. In *Proceedings of IEEE International Conference on Robotics and Automation*, volume 2, 2000.

Durrant-Whyte, H. and Bailey, T. Simultaneous localization and mapping: Part I. *IEEE Robotics & Automation Magazine*, 13(2):99–110, 2006.

Eade, E. and Drummond, T. Unified loop closing and recovery for real time monocular slam. In *Proc. 19th British Machine Vision Conference*, 2008.

Jégou, H., Douze, M., and Schmid, C. Hamming embedding and weak geometric consistency for large scale image search. In David Forsyth, Philip Torr, Andrew Zisserman (ed.), *European Conference on Computer Vision*, volume I of *LNCS*, pp. 304–317. Springer, oct 2008.

Konolige, K., Bowman, J., Chen, J.D., Mihelich, P., Calonder, M., Lepetit, V., and Fua, P. View-based maps. In *Proceedings of Robotics: Science and Systems (RSS)*, 2009.

Mei, C., Sibley, G., Cummins, M., Newman, P., and Reid, I. A constant time efficient stereo slam system. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2009.

Milford, M.J. and Wyeth, G.F. Mapping a Suburb With a Single Camera Using a Biologically Inspired SLAM System. *IEEE Transactions on Robotics*, 24 (5):1038–1053, 2008.

Newman, P., Sibley, G., Smith, M., Cummins, M., Harrison, A., Mei, C., Posner, I., Shade, R., Schrter, D., Murphy, L., Churchill, W., Cole, D., and Reid, I. Navigating, recognising and describing urban spaces with vision and laser. *The International Journal of Robotics Research*, 2009.

Nistér, D. and Stewenius, H. Scalable recognition with a vocabulary tree. In *Conf. Computer Vision and Pattern Recognition*, volume 2, pp. 2161–2168, 2006.

Olson, E. *Robust and Efficient Robotic Mapping.* PhD thesis, Massachusetts Institute of Technology, June 2008.

Philbin, J., Chum, O., Isard, M., Sivic, J., and Zisserman, A. Object retrieval with large vocabularies and fast spatial matching. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007.

Schindler, G., Brown, M., and Szeliski, R. City-Scale Location Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–7, 2007.

Sivic, J. and Zisserman, A. Video Google: A text retrieval approach to object matching in videos. In *Proceedings of the International Conference on Computer Vision*, Nice, France, October 2003.