Augmenting RatSLAM using FAB-MAP-based Visual Data Association

Will Maddern[1,2], Arren Glover[1], Michael Milford[2,1] and Gordon Wyeth[1]
School of Information Technology and Electrical Engineering[1]
Queensland Brain Institute[2]
University of Queensland
{william, arren, milford, wyeth}@itee.uq.edu.au

Abstract

This paper investigates the use of the FAB-MAP appearance-only SLAM algorithm as a method for performing visual data association for RatSLAM, a semi-metric full SLAM system. While both systems have shown the ability to map large (60-70km) outdoor locations of approximately the same scale, for either larger areas or across longer time periods both algorithms encounter difficulties with false positive matches. By combining these algorithms using a mapping between appearance and pose space, both false positives and false negatives generated by FAB-MAP are significantly reduced during outdoor mapping using a forward-facing camera. The hybrid FAB-MAP-RatSLAM system developed demonstrates the potential for successful SLAM over large periods of time.

1 Introduction

In recent years, the proliferation of field and service robotics as well as low cost mobile computing devices has caused a renewal of interest in automatic mapping and localisation, for the purposes of robot navigation, location-based services and augmented reality (AR). The underlying problem of simultaneous localisation and mapping (SLAM), however, has not been solved to a point where it can be easily adapted to a wide range of platforms and environments [Hager, *et al.*, 2007; Neira, *et al.*, 2008]. This paper focuses specifically on the application of SLAM in outdoor environments, where both scenery and illumination are highly variable.

Traditionally, SLAM has been performed using scanning radar, ultrasonic or LIDAR sensors, such as the popular SICK laser scanner [Thrun, 2003]. While these sensors provide accurate range and bearing measurements and well-understood noise characteristics, their 'active' sensing nature means many are bulky, expensive and draw large amounts of power. Additionally, it is difficult to perform data association between features detected by these sensors without resorting to computationally expensive point cloud comparison techniques [Nuchter, *et al.*, 2005]. Stereo or multiple cameras have been used with varying degrees of success in an attempt to provide similar metric feature measurements as the active sensors listed above, but using a passive image sensor array [Konolige and Agrawal, 2008].

More recently, work by [Davison, *et al.*, 2007] has shown that SLAM can be successfully performed in outdoor environments using a single camera and a sophisticated combination of optical flow, feature-based data association and metric SLAM algorithms. Using this method, an outdoor loop of approximately 250m in length was successfully mapped [Clemente, *et al.*, 2007]. Further work has been done to improve robustness, including the use of high speed cameras [Gemeiner, *et al.*, 2008], however no results mapping larger scale environment have been presented.

By changing two key paradigms in the SLAM problem – that the data association does not have to be feature-based, and that the map does not have to be metric – an entire 66km suburban road network was successfully mapped in [Milford and Wyeth, 2008] using a single low-cost webcam. This was accomplished using a biologically-inspired SLAM system based off rodent hippocampal models, known as RatSLAM [Milford, *et al.*, 2004]. Because the data association component of RatSLAM does not rely on vehicle odometry information, nor does the pose filter component explicitly propagate uncertainty with increasing distance, it has the capability to perform loop closure irrespective of the size of the loop [Milford and Wyeth, 2008]. However, the current visual data association system used in the RatSLAM is strongly dependent on lighting conditions, and is therefore unsuitable for outdoor mapping over larger time scales.

A probabilistic approach to visual data association has been demonstrated in [Cummins and Newman, 2008b]. The system, dubbed Fast Appearance-based Mapping (FAB-MAP), uses a bag-of-words dimensional reduction on SURF-generated features from images. A recursive Bayes estimation incorporating a Chow-Liu dependency tree is used to infer the probability that two images were representative of the same place. While not a full SLAM system (since the mapping is performed purely in appearance space and there is no representation of pose), FAB-MAP has demonstrated successful data association on road loops as large as 1000km [Cummins and Newman, 2009].

Attempts have been made to combine visual feature-based data association with non-metric SLAM algorithms [Newman, *et al.*, 2009], but common to these were the reliance on a secondary, geometric-based matching system to ensure the data association produced

no false positives. However, the geometric information about the environment was generated using either a stereo camera pair or LIDAR sensor, reintroducing many of the physical drawbacks of traditional SLAM systems.

In this paper we investigate a combination of visual feature-based data association as implemented by FAB-MAP with the spatio-temporal pose filter and mapping system provided by RatSLAM. A brief review of both systems is provided in the next section, followed by details of their implementation as a combined system. The full St Lucia suburb dataset (as presented in [Milford and Wyeth, 2008]) is used to evaluate the performance of the hybrid SLAM system, and results and conclusions are presented in the final section.

2 Appearance-Based Place Recognition

Given a visual scene, the FAB-MAP system calculates the probability that the scene matches to any previously visited location, as well as the probability that the scene is from an unvisited location [Cummins and Newman, 2008b]. Visual scenes, and hence locations in the real world, can be associated from high probability matches in appearance space. When only selecting matches with high certainty, the FAB-MAP system is proposed to provide low numbers of false positive matches.

The system uses a visual bag-of-words created using SURF feature extraction [Bay*, et al.*, 2006] to represent an image. A recursive Bayes technique along with a Chow-Liu dependency tree is used to calculate probabilities.

2.1 Visual Bag of Words

Each image is represented as a set of visual features, known as 'words'. Words are created by quantising each surf descriptor to an a-priori generated list of common features in the environment [Sivic and Zisserman, 2003]. It is therefore necessary to create the collection of common words, known as a 'codebook', as a once off calculation from a set of training data [Lowe, 2004]. Every feature extracted from the image is converted to the closest word in the codebook, reducing each image to a vector of which words are present in the image.

$$Z_k = \{z_1, \dots, z_{|v|}\} \tag{1}$$

Using this representation, geometry information is not kept and the image simply becomes a list of which common features are observed. Each location L_i is represented by the probability that the object e_i (that creates observation z_i) is present in the scene.

$$L_k = \{p(e_i = 1|L_k), \dots, p(e_{|v|} = 1|L_k)\} \tag{2}$$

Also required is a detector model, which captures the sensors' ability to make an observation of an object given that the object is in the image. This method therefore takes into account errors arising from imperfect visual sensors.

$$\begin{matrix} p(z_i = 1|e_i = 0) \\ p(z_i = 0|e_i = 1) \end{matrix} \tag{3}$$

This location representation can be compared to other locations using Bayesian probability to determine their similarity.

2.2 Probabilistic Data Association

The probability of the new image coming from the same location as a previous image is estimated using recursive Bayes:

$$p(L_i|\mathbb{Z}^k) = \frac{p(Z_k|L_i, \mathbb{Z}^{k-1})p(L_i|\mathbb{Z}^{k-1})}{p(Z_k|\mathbb{Z}^{k-1})} \tag{4}$$

where $\mathbb{Z}^k$ as a collection of previous observations up to time *k*. The likelihood that an observation comes from location L_i, $p(Z_k|L_i, \mathbb{Z}^{k-1})$, is assumed to be independent from all past observations and is calculated using a Chow-Liu approximation [Chow and Liu, 1968]. The Chow-Liu tree is used to describe a full joint probability distribution as a product of second-order conditional and marginal distributions. The tree is constructed once as an offline process based on training data, from which the conditional probabilities between root and parent nodes are learned. It has been shown that this method improves performance over a straight naive Bayes models

$$p(Z_k|L_i) \approx p(z_r|L_i) \prod_{q=1}^{|v|} p\left(z_q \middle| z_{p_q}, L_i\right) \tag{5}$$

where *r* is the root node of the Chow-Liu tree and p_q is the parent of node q.

To create the full PDF the likelihood for each previous observation needs to be calculated. This process can be computationally expensive but can be accelerated using the fast bail-out method without losing accuracy [Cummins and Newman, 2008a]. Instead of evaluating $p(Z_k|L_i)$ for each observation in the mapped set consecutively, individual elements $p(z_q|z_{p_q}, L_i)$ can be calculated for all observations at once, effectively calculating all likelihoods in parallel. Observations which have a very low probability of being a good match according to the Bennett inequality can be discarded to reduce computation. This process has been shown to accelerate the matching time by a factor of up to 50.

The prior probability of matching a location, $p(L_i|\mathbb{Z}^{k-1})$, is estimated using a naive motion model. This is done by setting the probability of a new place $p(L_{new}|\mathbb{Z}^{k-1})$ to a constant and assuming that for location *i* at time *t*, the probability of matching to locations *i*-1, *i*, and *i*+1 are equal at time *t*+1.

The denominator of equation 4 incorporates the probability of matching to a new location in addition to localisation within known places. To estimate if a new observation comes from a previously unvisited location the model needs to consider all locations, not just visited locations. This can be split into mapped and unmapped locations

$$\begin{aligned} p(Z_k|\mathbb{Z}^{k-1}) = & \sum_{m \in M} p(Z_k|L_m)p(L_m|\mathbb{Z}^{k-1}) \\ & + \sum_{n \in \overline{M}} p(Z_k|L_n)p(L_n|\mathbb{Z}^{k-1}) \end{aligned} \tag{6}$$

where M is the set of mapped locations. Since the second term cannot be evaluated directly (as it would require information on all unknown locations), an estimation must be used. Two calculations for this estimation are presented. The first is a mean field approximation [Jordan, *et al.*, 1999], where the unmapped location is estimated by creating an 'average location' from training data.

$$\sum_{n\in M} p(Z_k|L_n)p(L_n|\mathbb{Z}^{k-1}) \approx p(Z_k|L_{avg})p(L_{new}|\mathbb{Z}^{k-1}) \quad (7)$$

The second method is a sampling technique, where a random selection of scenes from training data is used to evaluate the unmapped location according to:

$$\sum_{n\in M} p(Z_k|L_n)p(L_n|\mathbb{Z}^{k-1}) \approx p(L_{new}|\mathbb{Z}^{k-1}) \sum_{u=1}^{n_s} \frac{p(Z_k|L_u)}{n_s} \quad (8)$$

where L_u is a sampled location and n_s is the total number of samples.

3 RatSLAM

While originally closely derived from neural models of the rat hippocampus [Milford and Wyeth, 2003; Milford, *et al.*, 2004], RatSLAM has undergone a number of enhancements based on increasing its performance and utility for real-time SLAM systems. The primary component of RatSLAM is the pose cell network, which performs the spatio-temporal filtering of the odometry and data association in pose space [Milford, *et al.*, 2004]. Loop closure is realised by injecting energy into the pose cell network until the dominant energy packet forms at the location the loop was last traversed. Finally, the experience map takes input from the visual system, odometry and pose cell network to form a semi-metric topological map, corrected for loop closures using graph relaxation techniques [Milford, *et al.*, 2005].

3.1 Pose Cell Network

The pose cell network takes the form of a three-dimensional competitive attractor network $\boldsymbol{P}_{x',y',\theta'}$, where each neuron in the grid simultaneously excites and inhibits its neighbours. The excitatory weight matrix $\varepsilon_{a,b,c}$ takes the form of a normalised spherical Gaussian, which is calculated by

$$\varepsilon_{a,b,c} = \frac{1}{k_p\sqrt{2\pi k_d}} e^{-(a^2+b^2)/k_p} e^{-c^2/k_d} \quad (9)$$

where k_p and k_d are the directional constants in the *x'-y'* and *θ'* directions respectively. The update cycle for the pose cell network is as follows:

$$\Delta P_{x',y',\theta'} = \sum_{i=0}^{(n_{x'}-1)} \sum_{j=0}^{(n_{y'}-1)} \sum_{k=0}^{(n_{\theta'}-1)} P_{i,j,k}\varepsilon_{a,b,c} \quad (10)$$

where $n_{x'}$, $n_{y'}$ and $n_{\theta'}$ are the sizes of each dimension of the pose cell network, and a, b, and c are found by

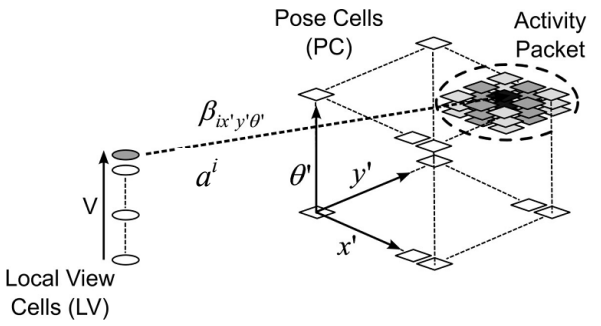

Figure 1 - Local view cell to pose cell injection diagram. Each local view cell links to a location in the pose cell network, and activity in the local view cells causes corresponding excitation in the pose cell network. Image © Springer 2008

$$\begin{aligned} a &= (x' - i) \bmod n_{x'} \\ b &= (y' - j) \bmod n_{y'} \\ c &= (\theta' - k) \bmod n_{\theta'} \end{aligned} \quad (11)$$

The modulo arithmetic above ensures that the three-dimensional wraparound on the grid is enforced during local excitation and inhibition. Similar to local excitation, given a three-dimensional inhibitory weight matrix $\psi_{a,b,c}$ and global inhibition value φ, the local inhibition is calculated as follows:

$$\Delta \boldsymbol{P}_{x',y',\theta'} = \sum_{i=0}^{(n_{x'}-1)} \sum_{j=0}^{(n_{y'}-1)} \sum_{k=0}^{(n_{\theta'}-1)} -\boldsymbol{P}_{i,j,k}\psi_{a,b,c} - \boldsymbol{\varphi} \quad (12)$$

Path integration in the pose cell network is accomplished by shifting all activity packets in the direction of vehicle motion, such that identical trajectories of forward and angular velocity result in identical paths through *x',y',θ'* space. Further details of path integration are provided in [Milford, *et al.*, 2004; Milford and Wyeth, 2009].

3.2 Visual Template Injection

The local view cells store each unique visual location as a 'template', an activation level a^i that corresponds to its match to the current visual scene, and the peak activity location in the pose cell network P^i when each template was generated:

$$V_i = \{a^i, P^i\} \quad (13)$$

When no activation level a^i is sufficient to match to the current visual scene, a new local view cell is creating using the current visual scene as a visual template. As shown in Figure 1, each active Local View cell injects activity into the pose cell network as follows:

$$\Delta \boldsymbol{P}_{x',y',\theta'} = \boldsymbol{\delta} \sum_i a^i P^i \quad (14)$$

where δ is the visual calibration strength constant.

3.3 Experience Mapping

The experience map forms the useful output of RatSLAM; it combines outputs from both local view cells and pose

cells, as well as odometry information, to form a topological map of the path taken by the SLAM system. Each experience e_i encodes an activation level E^i, pose cell location P^i, visual template V_i and position $\mathbf{p}^i$ in experience space:

$$e_i = \{E^i, P^i, V_i, \mathbf{p}^i\} \tag{15}$$

The activation level of each experience is based upon how well it matches the current pose cell location and current visual template, and is calculated as follows:

$$E^i = \begin{cases} 0 & \text{if } V^i \neq V^{\text{curr}} \\ 1 - \left|P^i - P^{\text{curr}}\right| / \mu_p & \text{if } V^i = V^{\text{curr}} \end{cases} \tag{16}$$

where P^{curr} and V^{curr} are the current pose cell activity location and visual template numbers respectively, and μ_p is a zone constant for pose cell location. If all activation levels are less than or equal to 0, a new experience is created using the current pose cell location and visual template number.

As the experience map develops it is necessary to correct locations in experience space $\mathbf{p}^i$ to account for errors in odometry found during loop closure. The following function is applied iteratively to each experience to update all associated positions:

$$\Delta\mathbf{p}^i = \alpha \left[\sum_{j=1}^{N_f} (\mathbf{p}^j - \mathbf{p}^i - \Delta\mathbf{p}^{ij}) + \sum_{k=1}^{N_t} (\mathbf{p}^k - \mathbf{p}^i - \Delta\mathbf{p}^{ki}) \right] \tag{17}$$

where N_f is the number of links from experience e^i to other experiences, N_t is the number of links from other experiences to e^i, and α is a correction constant, typically equal to 0.5 for maximum map correction without causing instability [Milford, 2008]. By plotting the experience map positions $\mathbf{p}^i$, as well as the links between experiences, a topological map of the environment is formed.

4 Experimental Procedure

The following section details the configuration of the algorithms presented in the above section, as well as other details required to implement a full SLAM system capable of mapping using a single forward-facing camera.

4.1 Experimental Setup

The mapping experiment in this section was performed using the same dataset presented in [Milford and Wyeth, 2008], and all following RatSLAM-only results refer to this publication. The route was a 66km path traversing every street in the suburb of St Lucia, Brisbane (as pictured in Figure 2). The path taken through the suburb forms 51 inner loops, and traverses over 80 intersections. Since a forward facing camera was used, all repeated sections of road were driven in the same direction.

The video was captured using an Apple *iSight* webcam built into an Apple *MacBook* at 10 frames per second and 640 x 480 resolution. The *iSight* camera has a field of view of approximately 50 degrees horizontal by 40 degrees vertical. The final video consisted of approximately 60,000 frames.

The dataset differs from original FAB-MAP experiments , which involved the use of two wide-angle stereo cameras [Cummins and Newman, 2008b], or a single omnidirectional camera [Cummins and Newman, 2009]. All the visual data association is performed using the single forward-facing camera. The video captured by the *iSight* webcam forms the sole sensory input to the SLAM algorithm, for both odometry information and visual data association.

Figure 2 – St Lucia test environment, consisting of a 66km road network covering approximately 6 square kilometres. Image © IEEE Transactions on Robotics 2008.

4.2 Algorithm Details

FAB-MAP Training Data

For this experiment, both the 'bag of words' codebook and the Chow-Liu dependency tree were generated using a modified version of the full 60,000 frame video (where all repeated sections were manually removed, and the video was subsampled to approximately 7000 frames). The primary goal was not to demonstrate the use of FAB-MAP as an online system but to test the performance based upon the ideal set of training data.

The codebook was generated using the modified sequential cluster algorithm [Teynor and Burkhardt, 2007] yielding 5730 unique words.

Vehicle Odometry

The odometry data used for this algorithm was generated in an identical fashion to that presented in [Milford and Wyeth, 2008], using scanline intensity profile comparisons. Although this is not sufficiently accurate to produce a metric map, it is more than sufficient for use with the RatSLAM algorithm to produce an accurate experience map of the suburb.

Visual Data Association

In contrast to original FAB-MAP implementations where a location was assigned to every video frame, a clustering approach was taken for this implementation. Similar to the visual templates used in the RatSLAM-only system, a new location 'template' is only created when no existing locations sufficiently describe the current location. For this case, a new location template is created if the probability that the current image comes from a new location $p_{\bar{M}}$ is above 0.99, in line with the data association threshold in [Cummins and Newman, 2008b]. To evaluate the probability that an observation comes from a new location the mean field approximation is used. Currently the sampling method is unsuitable for use as our training data is built from the test dataset; therefore using this method would create exact matches for new location probabilities. This would cause some locations in the dataset to always have a high probability of being a new location even after having observed that location in testing.

The detector probabilities were set at

$p(z_i = 1|e_i = 0) = 0$ and $p(z_i = 0|e_i = 1) = 0.61$, and the probability of a new location $p(L_{new}|\mathbb{Z}^{k-1}) = 0.9$.

The interface between FAB-MAP and RatSLAM takes the form of equating FAB-MAP locations L_i with RatSLAM visual templates V_i, and calculating an appropriate injection value a^i. Since RatSLAM does not require probabilistic inputs, the probabilistic match p_i generated by FAB-MAP can be 'smoothed' to provide finite injection even at low probabilities:

$$a^i = \frac{1}{1 - \ln\left(p(V_i|\mathbb{Z}^k)\right)} \tag{18}$$

The choice of the visual calibration constant δ proved critical to the success or failure of the visual data association system. If this constant was too high, the system would instantly relocalise to any spurious match returned by FAB-MAP, and conversely too low a value would cause a failure to successfully close loops under difficult perceptual matching conditions. A value of 0.05 was found to provide no false positive experience map loop closures, while ensuring sufficient injection to perform loop closures from partial FAB-MAP matches.

Pose Cell Configuration

The pose cell network was tuned as in [Milford and Wyeth, 2008], as this set of parameters has provided consistently good results in previous RatSLAM experiments. To reduce the chance of perceptual aliasing due to hash collisions in pose cell space [Milford, *et al.*, 2006], the size of the network was increased to 60 x 60 x 36 cells, each cell representing a 10m x 10m x 10 degree location.

Experience Mapping

Since this algorithm was not required to run in real-time, the timing constraints on experience map corrections were relaxed. For each iteration of the algorithm the experience map correction was applied 25 times, with the correction constant α equal to 0.5.

5 Results

The following section details a number of results from applying the hybrid FAB-MAP-RatSLAM algorithm to the St Lucia dataset in different configurations, as well as comparisons to the original RatSLAM-only results.

5.1 FAB-MAP-Only Results

Figure 3 (a) shows the experience map obtained for the entire dataset overlaid with colour-coded FAB-MAP connections. For every section of road that was repeatedly driven along, FAB-MAP returns significant numbers of correct matches with $p > 0.99$ (shown in green). However, there are also 4 false positives with $p > 0.99$ in various locations along the map, as shown by red markers with a connecting line indicating which positions have matched.

Figure 3 (b) and (c) show detail of the two false positives at locations L1 and L2. FAB-MAP has assigned a probability of greater than 99.8% that these pairs of frames have originated from the same location, however upon inspection the images are clearly from different parts of the map. In Figure 3 (b) FAB-MAP has matched the distinctive pattern on the pedestrian crossing to the same pattern on the following speed-bump. As these are the only two occurrences of this particular pattern in the entire dataset, the Chow-Liu dependency tree ranks them highly as unique identifiers of a particular place. The cause of the false match in Figure 3 (c) is somewhat more difficult to observe, but the particular combination of buildings, street lamps and clouds caused FAB-MAP to assign a high probability of the two images matching to the same location.

Figure 3 (d) presents two images from location L3 (with approximately an hour time difference) that FAB-MAP assigned a very low probability of matching. To a human observer the match is apparent; however due to the somewhat different lighting conditions and prevalence of foliage and trees in the image (which are very common throughout the dataset and thus have a low probability of causing a match) no frames along this

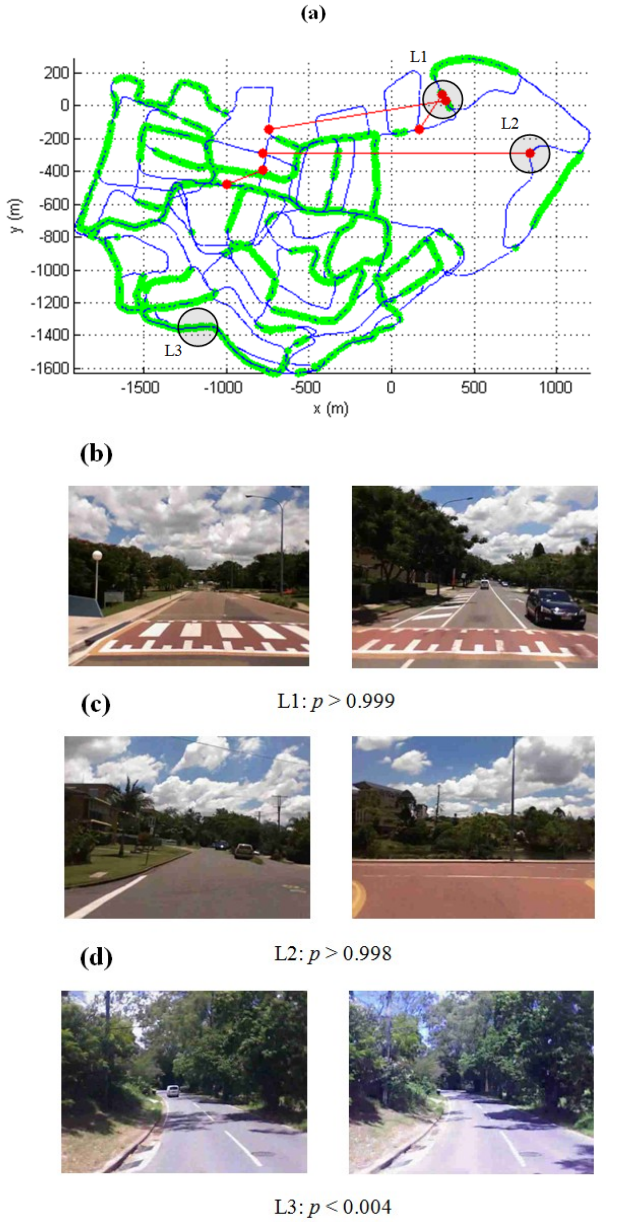

Figure 3 – FAB-MAP results on St Lucia dataset. (a) shows the FAB-MAP connection diagram for $p > 0.99$, with true positives shown in green and false positives in red. (b) and (c) show video frames that FAB-MAP falsely identified as matching, while (d) shows frames that FAB-MAP did not match.

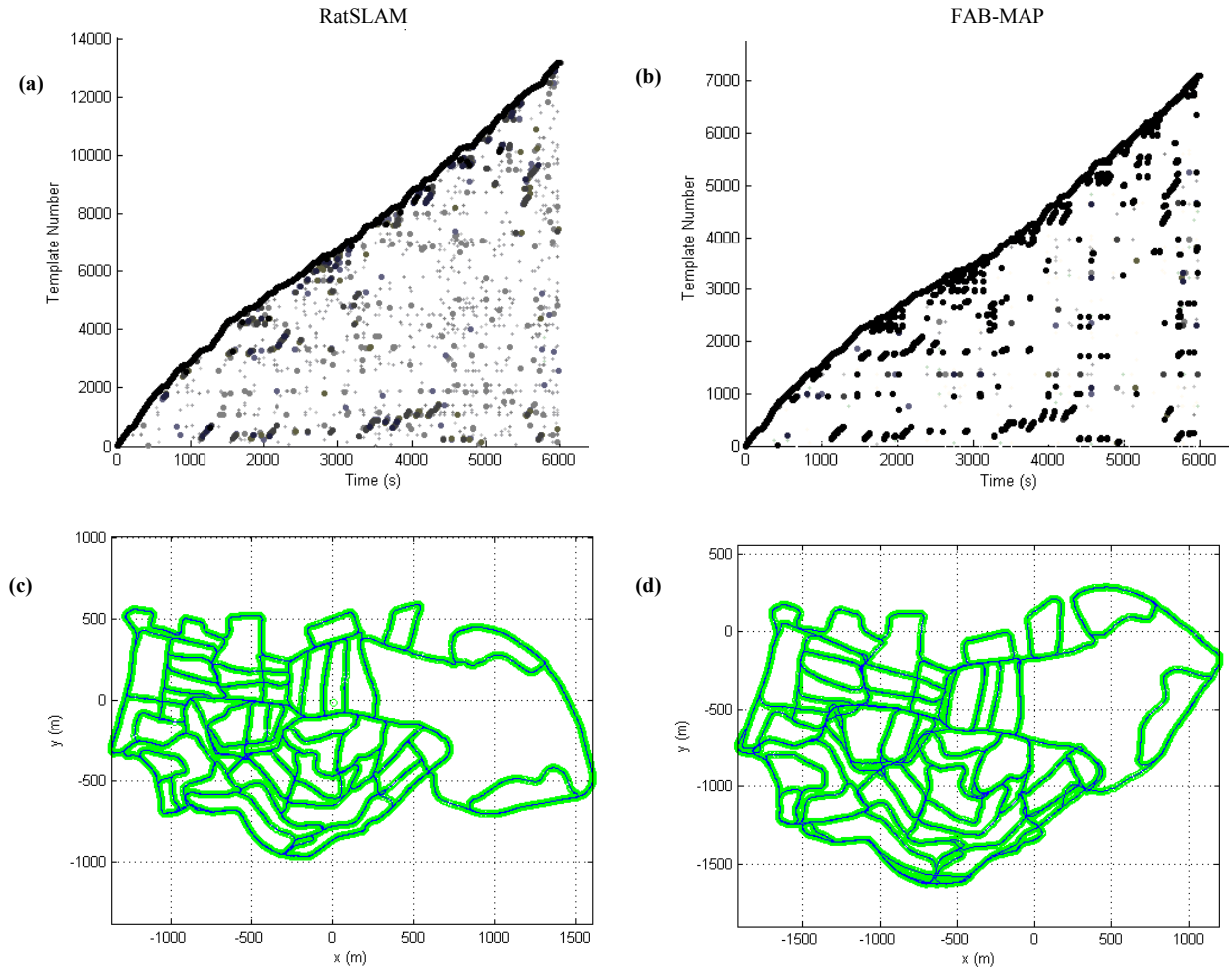

Figure 4 - Comparison between visual template match graphs and experience maps for RatSLAM and the FAB-MAP-RatSLAM hybrid. (a) and (b) graph template matches a^i against the current frame number for all 60,000 frames in the video, with stronger matches represented by darker points in the plot. (c) and (d) show the two experience maps generated.

section of road were correctly matched. While false negatives are difficult to quantitatively measure in a semi-metric system, they were reasonably common in this dataset and were the primary cause of the experience map distortion (as discussed in the following sections).

These results indicate that while FAB-MAP alone can map the St Lucia dataset in the space of appearance with a very low error rate, it is still not sufficient as the sole method of data association for this application, due to the false positives and negatives still present in the final result.

5.2 Template Generation

Figure 4 (a) and (b) show the results of both the RatSLAM-only template matching system and the template list generated by FAB-MAP. While RatSLAM produces approximately 13000 distinct visual templates, the method implemented to create new FAB-MAP templates only generates half as many. However, the correct matches obtained by FAB-MAP tend to be stronger, and fewer partial matches are generated.

While the RatSLAM-only template matching system had to be carefully tuned to provide acceptable results, the FAB-MAP method of automatically generating new templates required almost no parameter tuning to provide good results. Attempts to manually introduce a threshold or create a new template with every video frame (as in [Cummins and Newman, 2008b]) generated greater numbers of false positives in this experiment. However, since FAB-MAP generates half as many templates each visual location is approximately twice the size, which leads to weaker loop closures (as the strength of a loop closure is only dependent on the number of local experience map links).

Due to changing illumination levels throughout the day and the inability of the RatSLAM-only visual template system to incorporate differing visual representations of the same place (as a template is never modified after it has been created), the matches become progressively weaker with increased time, as evidenced in the latter part of Figure 4 (a). In contrast, the FAB-MAP based system consistently returns strong matches throughout the entire dataset, even though it too does not modify visual templates to accommodate for scene changes. Although false negatives are still produced, this is only when the scene does not contain sufficiently unique features (as in Figure 3 (d)).

5.3 Experience Map Generation

Figure 4 (c) shows the experience map generated by the RatSLAM-only system and (d) the hybrid FAB-MAP-RatSLAM system developed in this paper. While both maps are topologically correct, the map

generated by the original RatSLAM-only system more accurately reflects the route taken in Figure 2. The FAB-MAP-based system fails to capture many of the finer details of the route, particularly in the lower left-hand corner of the map. Additionally, there are multiple locations such as (-750,-1750) where FAB-MAP failed to relocalise and thus multiple representations of the same route exist in the experience map.

Although the primary cause of the experience map degradation is the number of false negatives returned by FAB-MAP, the pose filtering properties of RatSLAM and the injection mapping in equation 18 served to reduce their influence on the final map. While FAB-MAP did not produce any correct relocalisations with $p > 0.99$ along the southmost section of road at (-750, -1750), it produced many partial matches. These partial matches caused a continuous set of injection into the pose cell network, allowing it to build up a hypothesis over time until a correct loop closure was made at (-250, -1500).

6 Conclusions

The results presented in this paper have demonstrated a successful combination of a FAB-MAP based visual data association scheme with the pose filtering properties of RatSLAM. Using this combined algorithm, the entire suburb of St Lucia was mapped using a semi-metric method.

Given the size of the dataset and the narrow field of view of the camera, FAB-MAP has produced very few false positives, which generally cause catastrophic failure for SLAM systems. While [Cummins and Newman, 2009] demonstrated successful mapping of a dataset of approximately equal scale with zero false positives, the dataset presented here has significantly more inner loops of varying size, and was captured using only a low-cost forward-facing webcam.

By replacing the intensity-based visual data with FAB-MAP, some independence from illumination changes and dynamic scenes was gained. However, while this improved performance when matching to unique environments over large periods of time, it generated false negatives in other locations which did not contain sufficiently unique features. In contrast, although the scanline-intensity based method used in [Milford and Wyeth, 2008] generates matches over both unique and non-unique sections, it produces progressively weaker matches over the period of the dataset. While this method produced superior results for this particular dataset in terms of experience map generation, it is unlikely to perform better than a FAB-MAP based method over larger periods of time.

The addition of the RatSLAM pose filter allows FAB-MAP to function as a full SLAM algorithm in pose space. Its enhancement to FAB-MAP is twofold: firstly, it allows spurious false positives to be filtered using odometry data so as not to cause a false loop closure, and it allows multiple correct partial matches (which would otherwise be classed as false negatives) to build up over time, eventually causing a correct loop closure. By incorporating all the information available from the single camera in terms of visual data association and vehicle odometry, a robust SLAM system has been developed that outperforms localisation in appearance space alone.

6.1 Future Work

Although many visual SLAM algorithms are significantly improved by using wider field of view, multiple or omnidirectional cameras, using only low-cost forward facing cameras provides a greater challenge to the algorithm, but can be applied to low-end consumer systems where these cameras are commonplace (such as domestic robotics and augmented reality systems).

A number of improvements to the FAB-MAP algorithm are described in [Cummins and Newman, 2009]. These include a sparse likelihood update which reduces computation time by a factor of up to 80, a location clustering scheme (similar to the one implemented in this paper) that also incorporates new visual data into templates to 'average' location representations, and a geometric post-verification stage that compares the structure of known visual words in an image (rather than simply determining whether they are present or not). Implementing all these additional features will increase the performance of the FAB-MAP visual data association system in terms of computation time, and reduce the number of false positives found for a given dataset.

With respect to experience map generation, the appearance of the map can be improved by using a more robust method of determining vehicle motion. Rather than inferring translational and rotational velocity based upon changes in image intensity, the motion of SURF features (already generated by FAB-MAP) could be used to provide a better estimate of true vehicle motion.

Although this paper did not present a significant advantage of the combination of FAB-MAP and RatSLAM over the RatSLAM-only system, the results showed some promise of FAB-MAP's ability to relocalise across larger differences in scene and lighting conditions. This will be investigated in subsequent experiments by applying this algorithm to datasets generated at different times of day and across different days, to determine whether FAB-MAP can provide superior results for persistent outdoor visual data association and SLAM.

References

[Bay, *et al.*, 2006] Herbert Bay, Tinne Tuytelaars and Luc Van Gool. *SURF: Speeded Up Robust Features*, pp. 404-417, in *Computer Vision – ECCV 2006,* 2006

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

[Clemente, *et al.*, 2007] L. Clemente, A. J. Davison, I. D. Reid, J. Neira and J. D. Tardos. Mapping large loops with a single hand-held camera. In *Robotics: Science and Systems*, Atlanta, GA, USA, 2007

[Cummins and Newman, 2008a] M. Cummins and P. Newman. Accelerated appearance-only SLAM. *IEEE International Conference on Robotics and Automation, 2008 (ICRA 2008)*, pages 1828-1833, 2008a

[Cummins and Newman, 2008b] Mark Cummins and Paul Newman. FAB-MAP: Probabilistic Localization and Mapping in the Space of Appearance. *International Journal of Robotics Research*, 27(6):647-665, 2008b

[Cummins and Newman, 2009] Mark Cummins and Paul Newman. Highly Scalable appearance-only SLAM - FAB-MAP 2.0. In *Robotics Science and Systems*, Seattle, 2009

[Davison, *et al.*, 2007] A. J. Davison, I. D. Reid, N. D. Molton and O. Stasse. MonoSLAM: Real-Time Single Camera SLAM. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(6):1052-1067, 2007

[Gemeiner, *et al.*, 2008] P. Gemeiner, A. J. Davison and M. Vincze. Improving localization robustness in monocular SLAM using a high-speed camera In *Proceedings of Robotics: Science and Systems (RSS)*, 2008

[Hager, *et al.*, 2007] Greg Hager, Martial Hebert and Seth Hutchinson. Editorial: Special Issue on Vision and Robotics, Parts I and II. *International Journal of Computer Vision*, 74(3):217-218, 2007

[Jordan, *et al.*, 1999] Michael I. Jordan, Zoubin Ghahramani, Tommi S. Jaakkola and Lawrence K. Saul. An Introduction to Variational Methods for Graphical Models. *Machine Learning*, 37(2):183-233, 1999

[Konolige and Agrawal, 2008] K. Konolige and M. Agrawal. FrameSLAM: From Bundle Adjustment to Real-Time Visual Mapping. *IEEE Transactions on Robotics*, 24(5):1066-1077, 2008

[Lowe, 2004] David G. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, 60(2):91-110, 2004

[Milford and Wyeth, 2003] M. J. Milford and G. Wyeth. Hippocampal Models for Simultaneous Localisation and Mapping on an Autonomous Robot. In *Australasian Conference on Robotics and Automation*, Brisbane, Australia, 2003

[Milford, *et al.*, 2004] M. J. Milford, G. F. Wyeth and D. Prasser. RatSLAM: a hippocampal model for simultaneous localization and mapping. *Proceedings of the IEEE International Conference on Robotics and Automation, 2004 (ICRA '04)*, pages 403-408 Vol.401, 2004

[Milford, *et al.*, 2005] M. J. Milford, D. Prasser and G. Wyeth. Experience mapping: Producing spatially continuous environment representations using RatSLAM. In *Australasian Conference on Robotics and Automation*, Sydney, Australia, 2005

[Milford, *et al.*, 2006] M. J. Milford, G. Wyeth and D. Prasser. RatSLAM on the Edge: Revealing a Coherent Representation from an Overloaded Rat Brain. *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 4060-4065, 2006

[Milford, 2008] M. J. Milford. *Robot Navigation from Nature*. Springer-Verlag, 2008

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

[Milford and Wyeth, 2009] M. J. Milford and G. Wyeth. Persistent navigation and Mapping using a Biologically Inspired SLAM System. *International Journal of Robotics Research*, 2009

[Neira, *et al.*, 2008] J. Neira, A. J. Davison and J. J. Leonard. Guest Editorial Special Issue on Visual SLAM. *IEEE Transactions on Robotics*, 24(5):929-931, 2008

[Newman, *et al.*, 2009] Paul Newman, Gabe Sibley, Mike Smith, Mark Cummins, Alastair Harrison, Chris Mei, Ingmar Posner, Robbie Shade, Derik Schroeter, Liz Murphy, Winston Churchill, Dave Cole and Ian Reid. Navigating, Recognizing and Describing Urban Spaces With Vision and Lasers. *The International Journal of Robotics Research*, 2009

[Nuchter, *et al.*, 2005] A. Nuchter, K. Lingemann, J. Hertzberg and H. Surmann. 6D SLAM with approximate data association. *Proceedings of the 12th International Conference on Advanced Robotics, 2005 (ICAR '05)*, pages 242-249, 2005

[Sivic and Zisserman, 2003] J. Sivic and A. Zisserman. Video Google: a text retrieval approach to object matching in videos. *Proceedings of the Ninth IEEE International Conference on Computer Vision, 2003*, pages 1470-1477 vol.1472, 2003

[Teynor and Burkhardt, 2007] Alexandra Teynor and Hans Burkhardt. *Fast Codebook Generation by Sequential Data Analysis for Object Classification*, pp. 610-620, in *Advances in Visual Computing,* 2007

[Thrun, 2003] S. Thrun. *Robotic mapping: a survey*, pp. 1-35, in *Exploring artificial intelligence in the new millennium,* Morgan Kaufmann Publishers Inc., 2003