2010 IEEE International Conference on Robotics and Automation
Anchorage Convention District
May 3-8, 2010, Anchorage, Alaska, USA

FAB-MAP + RatSLAM: Appearance-based SLAM for Multiple Times of Day

Arren J. Glover, William P. Maddern, Michael J. Milford, *Member*, *IEEE*, and Gordon F. Wyeth, *Member, IEEE*

***Abstract*—Appearance-based mapping and localisation is especially challenging when separate processes of mapping and localisation occur at different times of day. The problem is exacerbated in the outdoors where continuous change in sun angle can drastically affect the appearance of a scene. We confront this challenge by fusing the probabilistic local feature based data association method of FAB-MAP with the pose cell filtering and experience mapping of RatSLAM. We evaluate the effectiveness of our amalgamation of methods using five datasets captured throughout the day from a single camera driven through a network of suburban streets. We show further results when the streets are re-visited three weeks later, and draw conclusions on the value of the system for lifelong mapping.**

I. Introduction

Appearance-based SLAM matches locations based on similarity in camera images taken from each location. This challenge is difficult as an environment can change greatly in visual appearance over time due to both scene alterations and differing lighting conditions [1]. Many recent studies in visual SLAM have been done using robust local features, which are designed to be scale and illumination invariant [2],[3]. Outdoor loops of 5 km length have been mapped using local features and stereo cameras [4], while 250m loops have been mapped using only a single hand held camera and optical flow techniques [5]. However these experiments were only performed on datasets gathered in a single time period and hence did not include extensive visual change. Recent work has shown that building dynamic maps that adapt to changing conditions can be achieved [6], albeit with a SICK laser range finder and only indoors.

Outside of the SLAM domain, research has investigated the use of local visual features in data association over very large periods of time [7]. Images of the same location were gathered over nine months and included extreme weather conditions, such as snow and sun glare. The lack of feature matching between different times of year suggested that current local feature detection algorithms could not successfully match the same scene over different seasons.

Recent work on a probabilistic approach to appearance-based data association using recursive Bayes and Chow Liu dependency trees, dubbed FAB-MAP, has successful mapped a loop of over 1000km [8]. FAB-MAP uses local SURF features that are illumination invariant, with the potential for successful mapping over large time periods. Mapping is performed purely in appearance space however, and there is no representation of pose. Perhaps the largest appearance-based full SLAM experiment was that performed using RatSLAM [9], which successfully mapped a 66km suburb using only a single low-cost webcam. However, RatSLAM's current visual data association system performance is strongly dependent on lighting conditions, and is unlikely to be suitable for outdoor mapping over larger time scales.

This paper presents a hybrid SLAM system between RatSLAM and FAB-MAP. The hybrid system combines RatSLAM's filtering and mapping algorithms with FAB-MAP's lighting invariant data association, with the aim to create a robust system for SLAM over long time periods in visually varying environments. Section 2 describes both FAB-MAP and RatSLAM and details how the two systems are integrated. Section 3 presents the experimental setup, with results in Section 4. Discussion is provided in Section 5, as well as directions for future work.

II. Applying Probabilistic Appearance-based Place Recognition to RatSLAM

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 [10]. Visual scenes, and hence locations in the real world, can be associated from high probability matches in appearance space.

A. Location Representation

Each image is held as a set of visual features, known as 'words'. Words are created by quantizing each SURF descriptor to an *a-priori* generated list of common features in the environment. It is therefore necessary to create the database of common words, named a 'codebook', as a once off calculation from a set of training data [11]. Every feature

Manuscript received September 15, 2009. This work was supported in part by the Australian Research Council under a Discovery Project Grant and the Australian Research Council and National Health and Medical Research Council under a Thinking Systems Initiative.

A. J. Glover, W.P. Maddern, M.J. Milford and G.F. Wyeth are with the School of Information Technology and Electrical Engineering, The University of Queensland, St Lucia, Queensland 4072, Australia (e-mail: {arren, william, milford, wyeth}@itee.uq.edu.au).

M. J. Milford is also with the Queensland Brain Institute, The University of Queensland, St Lucia, Queensland 4072, Australia.

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.

$$Z_k = \{z_1, ..., z_{|v|}\} \quad (1)$$

Each unique location L_k is represented by the probability that the object e_i (that creates observation z_i) is present in the scene.

$$\{p(e_i = 1 \mid L_k), ..., p(e_{|v|} = 1 \mid L_k)\} \quad (2)$$

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

B. Probabilistic Data Association

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

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

where $\mathcal{Z}^k$ is a collection of previous observations up to time k.

The likelihood that an observation comes from location L_i, $p(Z_k \mid L_i, \mathcal{Z}^{k-1})$, is assumed to be independent from all past observations and is calculated using a Chow Liu approximation [12]. 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. It has been shown that this method improves performance over a straight naïve Bayes model.

$$p(Z_k \mid L_i) \approx p(z_r \mid L_i) \prod_{q=1}^{|v|} p(z_q \mid z_{p_q}, L_i) \quad (4)$$

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

The prior probability of matching a location $p(L_i \mid \mathcal{Z}^{k-1})$ is estimated using a naïve motion model. The probability of a new place $p(L_{new} \mid \mathcal{Z}^{k-1})$ is set to a constant and given a 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 3 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:

$$\begin{aligned} p(Z_k \mid \mathcal{Z}^{k-1}) = & \sum_{m \in M} p(Z_k \mid L_m) p(L_m \mid \mathcal{Z}^{k-1}) \\ + & \sum_{n \in \bar{M}} p(Z_k \mid L_n) p(L_n \mid \mathcal{Z}^{k-1}) \end{aligned} \quad (5)$$

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 [13], where the unmapped location is estimated by creating an 'average location' from training data.

$$\sum_{n \in \bar{M}} p(Z_k \mid L_n) p(L_n \mid \mathcal{Z}^{k-1}) \approx p(Z_k \mid L_{avg}) p(L_{new} \mid \mathcal{Z}^{k-1}) \quad (6)$$

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:

$$\sum_{n \in \bar{M}} p(Z_k \mid L_n) p(L_n \mid \mathcal{Z}^{k-1}) \approx p(L_{new} \mid \mathcal{Z}^{k-1}) \sum_{u=1}^{n_s} \frac{p(Z_k \mid L_u)}{n_s} \quad (7)$$

where L_u is a sampled location and n_s is the total number of samples. A probability density function over all previous locations and the probability of a new place can then be calculated to provide data association for the RatSLAM system.

C. Visual Template Injection

The RatSLAM 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:

$$V^i = \{a^i, P^i\} \quad (8)$$

Each local view cell uses the FAB-MAP calculated probability that a new visual scene matches the current template to set the activation level. Since RatSLAM does not require probabilistic inputs, the probabilistic match p^i generated by FAB-MAP can be transformed to provide finite injection even at low probabilities:

$$a^i = \frac{1}{1 - \ln(p(V^i \mid \mathcal{Z}^k))} \quad (9)$$

When the activation level a^i for a new, previously unseen location is above a threshold, a new local view cell is created using the current visual scene as a visual template.

D. Pose Cell Network

Each active Local View cell injects activity into the pose cell network as follows:

$$\Delta \mathbf{P}_{x',y',\theta'} = \delta \sum_i a^i P^i \quad (10)$$

where δ is the visual calibration strength constant. The pose cell network takes the form of a three-dimensional competitive attractor network $P_{x',y',\theta'}$, 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

$$\varepsilon_{a,b,c} = \frac{1}{k_p \sqrt{2\pi k_d}} e^{-(a^2+b^2)/k_p} e^{-c^2/k_d} \quad (11)$$

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:

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

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 modulo created

boundaries which ensures that the three-dimensional wraparound on the grid is enforced during local excitation and inhibition. Local inhibition is calculated similarly to local excitation, using a three-dimensional inhibitory weight matrix $\psi_{a,b,c}$ and global inhibition value φ.

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 [14],[15].

E. 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:

$$e_i = \{E^i, P^i, V^i, \mathbf{p}^i\} \tag{13}$$

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:

$$E^i = \begin{cases} 0 & \text{if } V^i \neq V^{curr} \\ 1 - \left| p^i - p^{curr} \right| / \mu_p & \text{if } V^i = V^{curr} \end{cases} \tag{14}$$

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:

$$\Delta\mathbf{p}^i = \alpha \left[\sum_{j=1}^{N_f} \left(\mathbf{p}^j - \mathbf{p}^i - \Delta\mathbf{p}^{ij} \right) + \sum_{k=1}^{N_t} \left(\mathbf{p}^k - \mathbf{p}^i - \Delta\mathbf{p}^{ki} \right) \right] \tag{15}$$

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 [9]. By plotting the experience map positions $\mathbf{p}^i$, as well as the links between experiences, a topological map of the environment is formed.

III. Experiment Setup

A. Datasets

The experiment uses 10 datasets from a selection of streets in the suburb of St. Lucia, Brisbane, shown in Fig. 1. The first 5 datasets were gathered over 4 days with similar, sunny weather conditions. The second 5 datasets were gathered 3 weeks later over a period of 2 days, again with similar weather conditions. Each dataset was gathered over 20-25 minutes starting at 8:45am, 10:00am, 12:10pm, 2:10pm and 3:45pm. Fig. 2 shows a sample of the variation in the environment's visual appearance, both during the day and across a 3 week period.

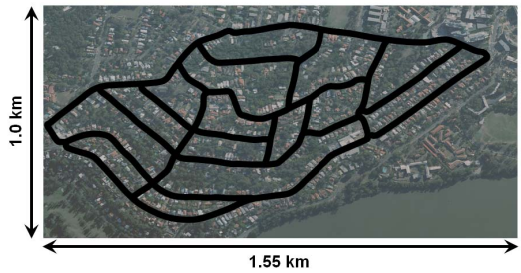


Fig. 1. Dataset route ground truth.

(a) (b) (c)

Fig. 2. The same location at (a) 8:45am (b) 3:45pm and (c) 8:45am approximately 3 weeks later

The video was captured using a Logitech QuickCam Pro 9000 web camera at 640x480 pixel resolution and an average of 15 frames per second. The camera was mounted facing forward at the top of the car windshield and has a field of view of approximately 62 degrees horizontally and 48 degrees vertically. GPS positions were also logged at 1 Hz during video capture.

The set-up is more consumer-viable than that used in [8] in which superior FAB-MAP results can be attributed to the use of a high-resolution, omni-directional camera, custom auto-exposure control and geometric matching of stereo image point clouds to eliminate false positives.

B. Algorithm Tuning

The codebook and Chow Liu tree was trained on a sub sampled version of an older RatSLAM experiment dataset of St. Lucia [9] in which all repeated sections of the video were removed, resulting in approximately 7000 non-overlapping images. The codebook was created using the modified sequential clustering algorithm [16] yielding 5730 words. To ensure performance was not affected by creating a codebook using a different camera and only at a single time of day a control codebook was created from the test dataset with negligible difference in results.

In contrast to other FAB-MAP experiments [10] in which disjoint locations are used, a new visual location is created when the calculated probability of a new location is > 0.99. To evaluate the probability of a new location the mean field approximation is used.

Visual odometry for this experiment, generated as in [9], resulted in the standard deviation of error against GPS ground truth to be 7.2m/s in the translational direction and 4.2 deg/sec in angular error. This large amount of error resulted in graphs, although correctly connected, to be

unrealistically shaped. We are aware that more sophisticated visual odometry methods are available (e.g. [17]) but as high quality visual odometry is not the focus of the study, more accurate visual odometry was simulated using a linear interpolation of the position differential of the GPS signal. Despite being GPS based ,this method introduces significant incremental error in accumulation of difference measurements to positions and is therefore suitable for simulating error prone dead reckoning.

TABLE 1
SUMMARY OF CONSTANTS USED IN ALGORITHMS FOR EXPERIMENTS.

FAB-MAP	
$p(z_i = 1 \mid e_i = 0)$	0
$p(z_i = 0 \mid e_i = 1)$	0.61
$p(L_{new} \mid Z^{k-1})$	0.9
RatSLAM	
Injection δ	0.005
Pose Cell Network Size	60 x 60 x 36 cells
Cell Size	10m x 10m x 10 degree
Graph Correction Constant	0.5
Graph Correction Repetition	25

IV. RESULTS

A. Mapping throughout the Day

The original RatSLAM system and the hybrid FAB-MAP/RatSLAM system were tested on the first five datasets to evaluate data association over the space of an entire day. Fig. 3 shows the percentages of frames that match over a threshold of 0.99 (as used in [10]) for FAB-MAP and a threshold of 0.1 (chosen to compensate for RatSLAM-profiles' lower precision) for RatSLAM-profiles. These results indicate the True Positives matched between datasets only, the self matches within a single dataset are not reported. For example, after running FAB-MAP through the 8:45am dataset, only 0.37% of frames in the 2:10pm dataset correctly matched back to the 8:45am dataset.

The recall rates when mapping a dataset to itself are not 100% as new visual locations are not created from every frame in the first pass through the dataset, so there is not always an exact frame to match to in the second run through. Recall generally reduces as the time difference between datasets increases, and is very low (0.32% to 3.4%) for matching between morning and afternoon. It can be seen that FAB-MAP data association outperforms the original RatSLAM-profile matching in both recall and precision.

Both algorithms, run independently, result in catastrophic failure over the full day dataset. The original RatSLAM fails due to low recall (Fig. 4) and FAB-MAP fails as any false positive causes an incorrect loop closure (Fig. 5).

The complete hybrid RatSLAM/FAB-MAP algorithm was used to map the five datasets sequentially, ordered chronologically in terms of time of day. The true positive loop closures to all *previous* datasets are shown as circles. A concise map can be seen in Fig. 6(a) demonstrating that the hybrid system can perform SLAM within loops of a single dataset. Each sequential map is initially started disjoint to the previous map but in every case enough data association occurs to allow RatSLAM to link the newly created map to the previous map. Fewer loop closures occur with greater

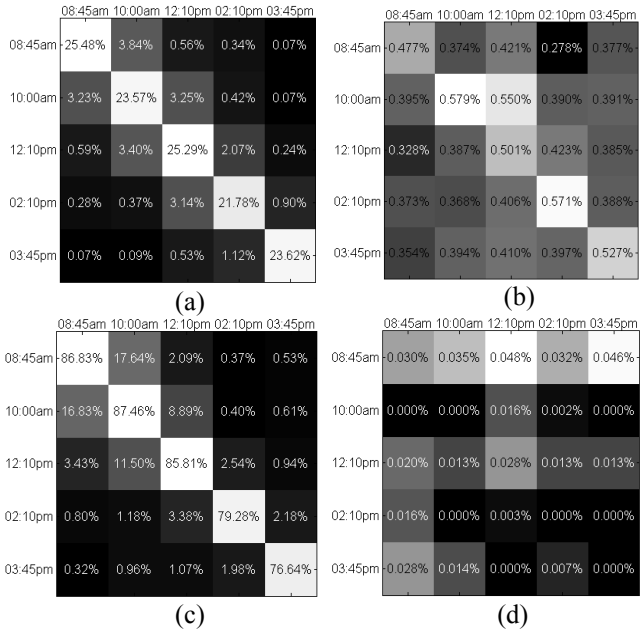


Fig. 3. True positive (TP) and false positive (FP) match percentages for data association techniques between different times of day (a) TP for RatSLAM-profiles (b) FP for RatSLAM-profiles (c) TP for FAB-MAP (d) FP for FAB-MAP.

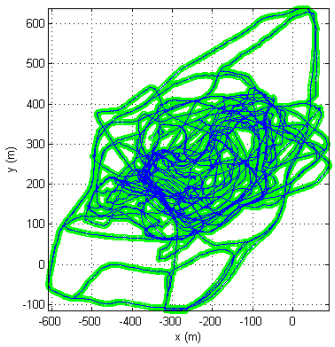


Fig. 4. Mapping using only RatSLAM-profiles, results in catastrophic failure over full day datasets.

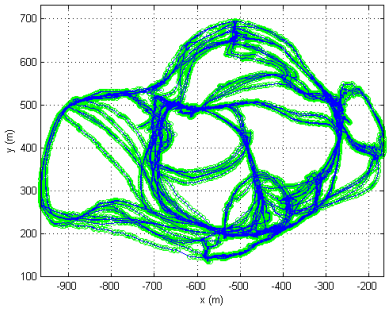
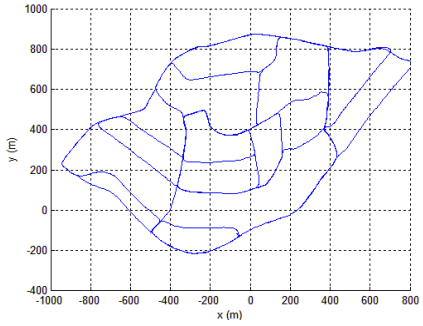
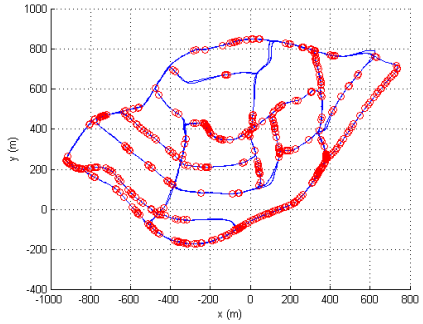


Fig. 5. Mapping using only FAB-MAP data association, results in catastrophic failure over full day datasets. This map was created using complete energy injection resulting in a loop closure on any reported match.

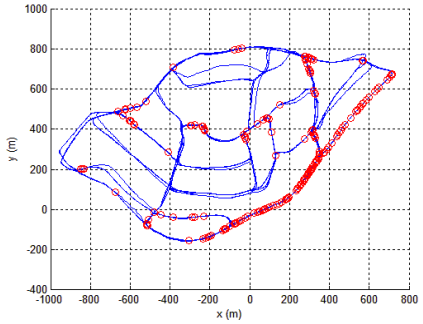
(a)



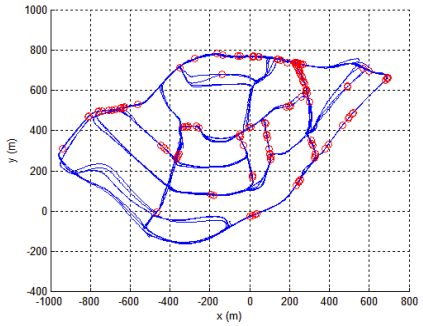
(b)



(c)



(d)



(e)

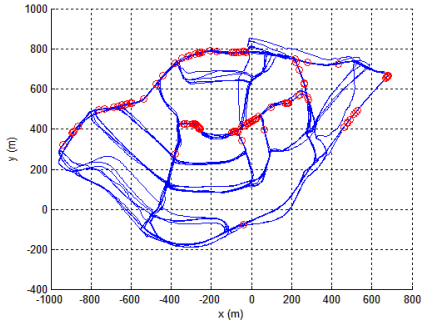


Fig. 6. RatSLAM-created experience maps and loop closure locations using FAB-MAP data association. Sequentially including (a) 8:45am (b) 10:00am (c) 12:10pm (d) 2:10pm and (e) 3:45pm.

time difference between visiting the location, indicated by fewer matches and consequently multiple paths for the same street. This has a direct correspondence to the true positives rates presented in Fig. 3. However, the false positive data association has not caused catastrophic failure (as seen in Fig. 5) because the RatSLAM pose filtering addresses the false positives produced by FAB-MAP. This is a necessary addition to any data association system, as false positives are inevitable when dealing with large long-term real-world datasets.

B. Mapping between weeks

The data association methods were also tested for precision and recall on datasets collected 3 weeks apart but at the same time of day. The true positive and false positive rates for FAB-MAP with a threshold of 0.99 and RatSLAM-profiles with a threshold of 0.1 are shown in Fig. 7. FAB-MAP again outperforms RatSLAM-profile matching with respect to recall and precision rates, but the presence of false positives again indicates FAB-MAP could not be used as the sole input to a mapping algorithm.

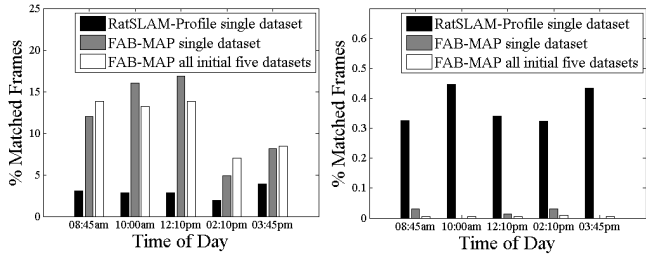


Fig. 7. Comparison of data association with a three week gap. The third series uses all five initial datasets in the location creation phase, (a) true positives (b) false positives.

The 12:10pm dataset from 3 weeks afterwards was then localised within the complete map from the first 5 datasets of the previous 3 weeks. The resulting map can be seen in Fig. 8. The relatively high true positive rate (16.85%) causes the new map to link to the previous 5 maps and consistent loop closures do not create any extra paths in the experience map.

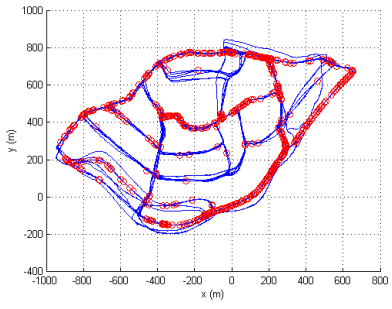


Fig. 8. RatSLAM-created experience map and loop closure locations using FAB-MAP from data captured at 12:10pm three weeks after having mapped the entire first five datasets.

C. Location Growth

An important aspect to note is the rate at which new locations are created as both space and computation time increase with the number of locations. Current FAB-MAP recall rates are not high enough to generalize completely both within and between datasets. A linear growth rate can be seen when mapping all six datasets in Fig. 9, even though the same locations are continually visited. When considering the aim is for long-term mapping on a mobile robot this continual growth rate would cause the current algorithm to be unsustainable without a map management or pruning algorithm [15].

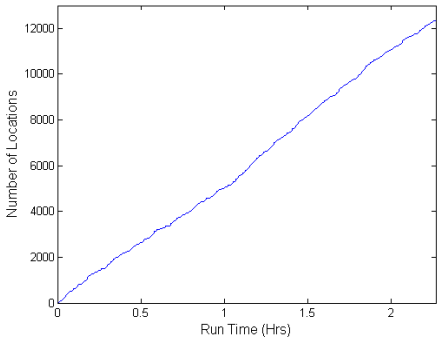


Fig. 9. FAB-MAP total number of locations vs. video frame number. The plot indicates near linear growth rate of locations over time.

V. Discussion

Individually, neither RatSLAM nor FAB-MAP can address the challenge of producing a coherent map across all times of day for the datasets presented in this paper. RatSLAM's lightweight heuristic approach to image-based location matching produces too small a true-positive-to-false-positive match ratio to create a useful map. FAB-MAP significantly improves the ratio of true-to-false positive matches, but does not provide sufficient filtering to remove the final few false matches, nor a system to combine pose estimation and re-localisation for performing full SLAM.

The hybrid RatSLAM/FAB-MAP system has shown that mapping can be performed even in difficult outdoor conditions when the environment's appearance varies due to changes in illumination and structure. However, the results clearly show that the map diverges when there are long sections of path where no matches occur. The algorithm also does not generalize well enough to reduce the rate at which new locations are produced when revisiting locations. This occurs in all cases; within a single dataset, across multiple datasets at different times of day and from week to week. This unchecked growth in locations would become computationally intractable in long-term real-time operation. We suspect that this is not a shortcoming of FAB-MAP, but rather the SURF features on which it is based, which are simply too variable over the course of a day to form a truly re-useable appearance-based map.

A. Future Work

A number of improvements to the FAB-MAP algorithm described here are presented in [8], including reducing computation time, improved robustness to false positives using the geometry of features in an image and updating location representations as they are revisited. A more robust set of features or multiple types of features could be incorporated to provide increased recall rates in variable lighting conditions addressing the problem of location growth. We are also working towards a new version of RatSLAM where the filtering and mapping components of the system can be more rigorously designed to suit the application at hand.

References

[1] Srinivasa Narasimhan, Chi Wang and Shree Nayar, "All the Images of an Outdoor Scene", pp. 3-13, in *Computer Vision — ECCV 2002*, 2002

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

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

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

[5] 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

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

[7] C. Valgren and A. Lilienthal, "Sift, surf, and seasons: Long-term outdoor localization using local features", In *Proc. of 3rd European Conference on Mobile Robots*, Freiburg, Germany, 2007

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

[9] 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

[10] 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, 2008

[11] 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

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

[13] 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

[14] 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

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

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

[17] S. Baker, D. Scharstein, J. P. Lewis, S. Roth, M. J. Black and R. Szeliski, "A Database and Evaluation Methodology for Optical Flow", I*EEE 11th International Conference on Computer Vision (ICCV)*, pages 1-8, 2007