Beyond a Shadow of a Doubt: Place Recognition with Colour-Constant Images

Kirk MacTavish, Michael Paton, and Timothy D. Barfoot

Abstract Colour-constant images have been shown to improve visual navigation taking place over extended periods of time. These images use a colour space that aims to be invariant to lighting conditions – a quality that makes them very attractive for place recognition, which tries to identify temporally distant image matches. Place recognition after extended periods of time is especially useful for SLAM algorithms, since it bounds growing odometry errors. We present results from the FAB-MAP 2.0 place recognition algorithm, using colour-constant images for the first time, tested with a robot driving a 1 km loop 11 times over the course of several days. Computation can be improved by grouping short sequences of images and describing them with a single descriptor. Colour-constant images are shown to improve performance without a significant impact on computation, and the grouping strategy greatly speeds up computation while improving some performance measures. These two simple additions contribute robustness and speed, without modifying FAB-MAP 2.0.

1 Introduction

Visual place recognition aims to recognize, from a stream of images, if the vehicle is revisiting a place it has previously seen. Since integrated odometry measurements drift over time, this information is especially useful if a long period of time has passed since the last visit. Over this period, lighting conditions will change, making it more difficult to recognize the matching image. To address this problem, colour-constant images transform an RGB image into a colour-space that changes less with lighting conditions than greyscale [19, 7, 3, 12, 10, 16]. This paper presents experimental results from a challenging multi-day dataset [16] where colour-constant images

e-mail: {kirk.mactavish, mike.paton}@mail.utoronto.ca, tim.barfoot@utoronto.ca
University of Toronto Institute for Aerospace Studies, Toronto, Ontario, M3H 5T6, Canada

improve place recognition performance with no modification to the underlying inference algorithm, Fast Appearance-Based Mapping (FAB-MAP) 2.0 [6].

The use of colour-constant images does add a small computational overhead, since these images are used alongside the original greyscale images, increasing vocabulary size and the average number of observed features. To recover this computation effort, we use the the image grouping strategy introduced by MacTavish and Barfoot [9]. This method is faster by an order of magnitude, improves some performance measures (see Section 4), and does not require modification or parameter tuning of the place recognition algorithm.

Similar work has been performed by Maddern and Vidas [11], who used FAB-MAP with a monochromatic and thermal camera, with a similar channel-concatenated Bag-of-Words (BoW). Collier et al. [2] address lighting change using lidar geometry and monochromatic images, running FAB-MAP separately on each sensor. MacTavish and Barfoot [9] use lidar intensity with FAB-MAP to achieve lighting invariance, requiring specialized hardware and introducing motion distortion due to a rolling shutter. Paul and Newman [17] augment visual features with spatial information using lidar. This paper focuses on improved lighting invariance without additional hardware beyond an RGB camera.

Sunderhauf et al. [20] use Sequence SLAM (SeqSLAM) [14] with monochromatic images to localize a train over 3000 km across seasons with impressive results, but do not perform full Simultaneous Localization and Mapping (SLAM) with the ability to add new places. Milford [13] shows how SeqSLAM can use very-low-resolution images to localize by leveraging sequence information. The FAB-MAP image-grouping strategy [9] used in this paper also makes use of sequence information by grouping local regions in a single descriptor.

In an effort to learn appearance change and proactively translate the image to different appearance conditions, Neubert et al. [15] introduce a super-pixel-based translation algorithm. This algorithm targets large seasonal change rather than lighting, and requires training data of the expected appearance domain. Pepperell et al. [18] blacken the sky in daytime images for better matching against those captured at night using a whole-image matching technique. Aiming to improve lighting invariance at the descriptor level, Carlevaris-Bianco and Eustice [1] train neural-net features using data from outdoor webcams. Colour-constant images improve lighting invariance without algorithm modification even at the descriptor level.

Corke et al. [3] compute image similarity scores across a small set of colour-constant images, and Maddern et al. [10] perform local metric localization; however, there has not been an evaluation of place recognition using colour-constant images. In this paper, we discuss this task and present the results of our approach on an 11 km dataset consisting of over 2000 images.

This paper presents novel results for place recognition using colour-constant images. This contribution goes beyond the simple image similarity scores that have been used in previous work to benchmark this image transform. In Section 2 we discuss the place recognition and image processing techniques that we have used. In Section 3 we discuss the field experiment, and in Section 4 we present and analyze the experimental results. Final conclusions and future work are discussed in Section 5.

2 Methodology

2.1 Place Recognition

The FAB-MAP algorithm and its extensions [5, 4, 6] have been extensively tested and widely used; in particular, FAB-MAP 2.0 has been tested on a 1000 km dataset. This paper examines the results of two input preprocessing techniques for place recognition: colour-constant images, and BoW image grouping. For place recognition itself, we use the OpenFABMAP implementation [8] of the FAB-MAP 2.0 algorithm which is summarized below.

FAB-MAP uses a BoW descriptor to describe images. To train the BoW vocabulary, Speeded Up Robust Features (SURF) descriptors are extracted from a training image dataset. These descriptors are clustered, and the BoW vocabulary is described by these cluster centers (words). An image can now be described by a BoW descriptor by quantizing its SURF features using the vocabulary, and listing which words were seen. A BoW descriptor can be represented as a binary vector of word presence, or as a list of which words were observed. To learn a factorized probability prior distribution over BoW descriptors, FAB-MAP trains a Chow-Liu Tree (CLT) using the BoW descriptors from the training dataset.

FAB-MAP represents a place as a vector of Bernoulli variables indicating the existence of the generator for each word in the vocabulary. The measurement model is given by the trained CLT and two user-specified parameters, and full Bayesian inference determines the posterior generator probabilities. The probability of being in a new place is determined using a Monte-Carlo approximation, sampling training images as representative new places. FAB-MAP 2.0 speeds up inference using an inverted index for each word in the vocabulary and slightly modified inference.

FAB-MAP 2.0 also uses geometric verification in the form of a 1-point Random Sample Consensus (RANSAC) test to improve precision. The results in this paper focus only on the recall task, and have not used any geometric verification, though they have used the FAB-MAP 2.0 simple motion model. Since the Visual Teach & Repeat (VT&R) algorithm [16] used to collect the dataset is already performing visual odometry, it would be straightforward to use only features that are stable over a short distance to verify geometric stability; we leave this as future work.

2.2 Colour-Constant Images

Colour-constant images were first developed in the optics community. Recent methods are based on the theory that a 1D colour space that is invariant to outdoor lighting conditions can be calculated from the channel responses of an RGB camera, given certain assumptions about the sensor and environment [19, 7]. The method presented by Ratnasingam and Collins [19] asserts that a colour-constant feature, F, can be extracted from a three-channel camera from the following:

$$F = \log(R_2) - \alpha \log(R_1) + \beta \log(R_3), \tag{1}$$

where R_i, is the approximated sensor response for channel i, and α and β are weights subject to the following constraints:

$$\frac{1}{\lambda_2} = \frac{\alpha}{\lambda_1} + \frac{\beta}{\lambda_3}, \quad \beta = (1-\alpha), \tag{2}$$

where $\lambda_1, \lambda_2, \lambda_3$ are the peak sensor responses numbered from highest to lowest. The result of equation (1) is a 1D feature with much of the effect of lighting removed.

Colour-constant images have appeared in various forms [3, 12, 10, 16] in the robotics and computer vision community. The approach taken in this paper is identical to that of Paton et al. [16], which uses experimentally trained coefficients of equation 1 to obtain two colour-constant images: $\{F'_v, F'_r\}$ that perform well in vegetation and rocks-and-sand, respectively. Examples of these images can be seen in Figure 1.

Fig. 1 This figure illustrates the transformation of an RGB image into a set of greyscale images. The top image is a typical greyscale image obtained from the green channel, and the bottom two are the colour-constant image pair $\{F'_v, F'_r\}$ [16] used in this paper to boost place recognition. By making assumptions about the sensor and environment, a weighted log-difference of the three camera channels can cancel the effect lighting has on the appearance of the scene. Credit: [16]

These images were used to great success in an autonomous route-following algorithm presented by Paton et al. [16], which was used to collect the dataset that is used in this paper. Details on the environment and route can be found in Section 3.

Since FAB-MAP requires a single BoW descriptor for each observation, we can create a unified place descriptor by concatenating the BoW descriptors from each channel [16, 11]: the green channel (greyscale), F'_v, and F'_r. A separate vocabulary is trained for each channel, and each is quantized into a separate BoW descriptor. These per-channel-BoW descriptors are concatenated into a stacked BoW that is used to train the CLT, and for online place recognition. We expect that there will be a strong

correlation between words in each of the channels, since the channels themselves are correlated. Luckily the CLT accounts for this correlation to the extent that it is apparent in the training dataset.

2.3 Image Grouping

MacTavish and Barfoot [9] show that sequences of images can be grouped together and described with a single BoW descriptor. This provides two benefits: temporal smoothing, which can improve robustness if features are somewhat unstable; and a theoretical speedup of n^2 for groups of n images. The major drawback is that matches are not established at an image level. Simply adding the BoW descriptors loses sparsity as group size increases. For the CLT training to be valid, these grouped BoWs must have similar sparsity to the single-image training descriptors. We can meet this requirement by increasing the binary BoW threshold, requiring multiple observations of a word before it is considered present. A detailed description and results for this method is available by MacTavish and Barfoot [9].

3 Field Experiment

A four day field trial was conducted at the Canadian Space Agency (CSA)'s Mars Emulation Terrain (MET) at Montreal, Quebec on May 12-15th, 2014, with the purpose of testing the colour-constant VT&R algorithm introduced by Paton et al. [16]. The MET, pictured in Figure 2, is a 60x120 m manicured environment emulating the surface of Mars. It consists primarily of rock and sand, with interesting features such as outcroppings and craters. The MET is surrounded by unstructured vegetation containing trees, marshland, open fields, a small stream, and a gravel roadway.

The field trial proceeded by teaching a 1 km path, marked as a yellow line in Figure 2, through the MET and its surrounding fields. This path was taught at approximately 11 am on the first day during sunny conditions with pronounced shadows. Over the course of the field trial, this path was autonomously traversed 26 times in varying lighting conditions. During this time the robot maintained an autonomy rate of 99.9% of distance travelled.

The hardware setup used during these experiments is pictured in Figure 3. The robot is the Clearpath Grizzly Robotic Utility vehicle. The VT&R algorithm ran on an on-board computer using a Point Grey Research Bumblebee XB3 stereo camera. GPS data was collected for the purpose of visualization only.

During the traversals of the 1 km path, the robot recorded rectified 512×384 stereo RGB images at 16 hz from the Grizzly's front PGR XB3 Camera. The result is close to 1TB of stereo data along the same path in many lighting conditions. In this paper we present results using 11 of these traversals, from dawn to dusk, selected with the intent of maximizing appearance variation. Additionally, a 247 image, 1.2

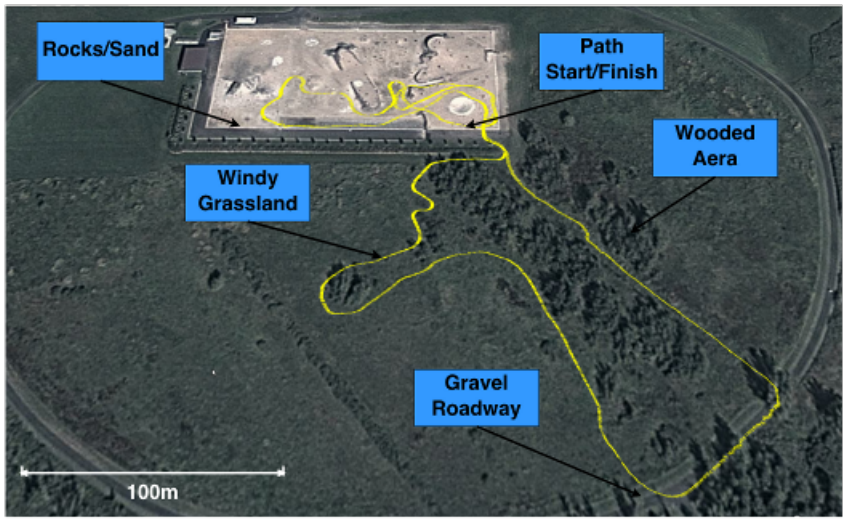


Fig. 2 Satellite Imagery of the CSA MET, with the teach pass from the 2014 field trials highlighted in yellow, and interesting environmental features annotated. Credit: [16]

Fig. 3 Grizzly Robotic Vehicle autonomously repeating a route during the CSA field trials, with applicable sensors highlighted. Credit: [16].

km dataset was collected in Ontario, Canada, which was used for training the place recognition algorithm.

4 Results

This section presents the place recognition results for the colour-constant and image grouping techniques. Parameter training for the FAB-MAP algorithm is covered by Cummins and Newman [5], the tuning process and results for the colour-constant images are detailed by Paton et al. [16], and the tuning process for image groups is explained by MacTavish and Barfoot [9].

The CSA dataset is quite challenging for several reasons. Over the course of the experiment, the terrain was significantly modified by the vehicle, as shown in Figures 4a,4c,4d. This dataset is collected by a single camera pointed forward and down, meaning a significant portion of the field of view is physically changing over the course of the experiment. As anticipated, the changing lighting conditions had a large effect – including the robot's own shadow being visible when the sun was behind (see Figure 4b), leading to similar features being seen in different places depending on the time of day. Natural environments also tend be more challenging than urban [6], and the geometric intricacy of vegetation leads to difficult shadows as lighting changes. Finally, at times the lighting conditions were so extreme that the auto exposure was unable to produce a usable image (see Figure 4e).

FAB-MAP is fairly sensitive to feature stability, and SURF detector thresholds had to be carefully selected for the colour-constant images, due to their far-lower dynamic range (see Figure 1), and limited intensity information (by design). Initial results used a detector threshold that would extract a similar number of features across all image channels. This resulted in poorer performance than greyscale on its own, since many of the colour-constant features turned out to be unstable. Since colour-constant images are deliberately removing intensity information from the image to provide invariance, there is less information remaining. This leads to a noisier image, and noisier feature descriptors. The final SURF thresholds for the greyscale, F_v', and F_r' images lead to an average of 83, 9, and 22 keypoints per training image, respectively. The clustering threshold was set so that the feature-to-vocabulary-size ratio was similar for the image channels, resulting in 1017, 85 and 244 words per image type, respectively. The performance for the greyscale-and-colour-constant stack is shown in Figure 5 as *Stack*, and for the greyscale only baseline as *Grey*. Colour-constant-only results have not been shown, as the low feature count and vocabulary size are unable support place recognition alone. For equivalent recall, the precision is strictly better using the colour-constant stack. The timing results in Table 1 show that there is a 22% increase in computation, due to a larger vocabulary. Figure 6a shows an example of a place that is correctly recognized by the colour-constant stack, but not by greyscale.

Image sequences were also grouped in sequences of 5 images, to illustrate the speed-up without introducing a large disparity in match specificity. MacTavish

(a) Tall grass that was flattened by the vehicle over the course of the experiment.

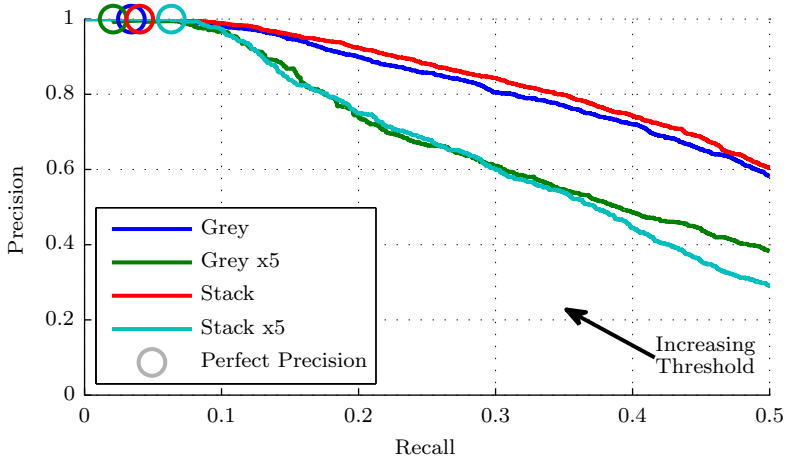
(b) The vehicle's shadow is seen in different places depending on the time of day.

(c) The same location during the first and last loop showing the terrain modification on sand.

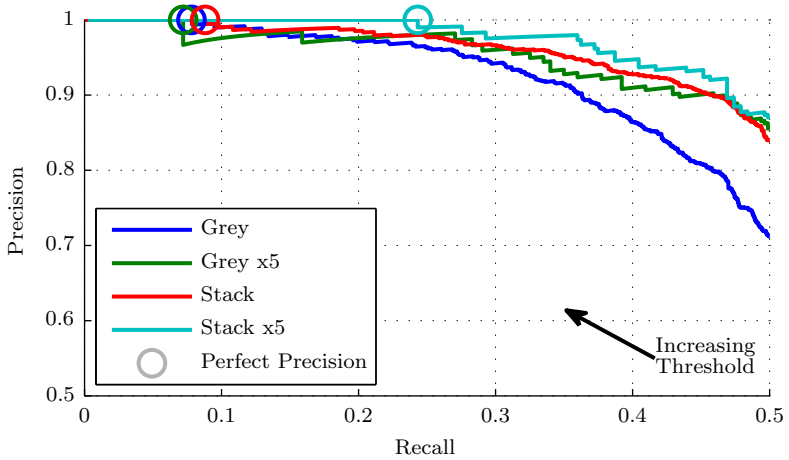
(d) The same location during the first and last loop showing the terrain modification on vegetation.

(e) The same location during the first and latest-in-the-day loop showing the auto exposure struggling with low light and a still-bright sky.

Fig. 4 Example images from the test dataset showing several of the challenging cases.



(a) Precision-Recall for *all* matches between loops. Unfortunately, the colour-constant stack shows only modest improvement, and the image grouping fares far worse. This measure is the most common, but is not necessarily representative of the desired output. The curve below presents an alternative measure that might represent a more realistic use case.



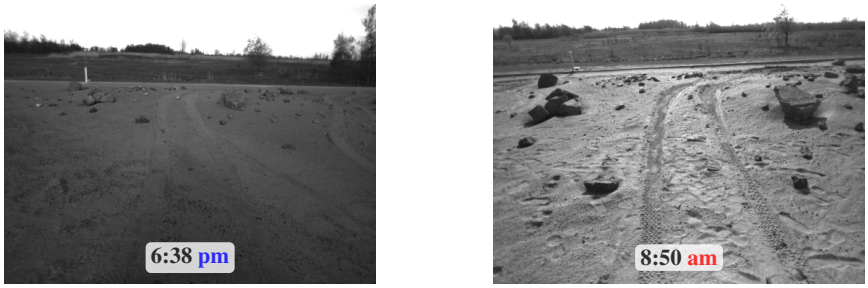
(b) Precision-Recall for *at least one* matches between loops (per query). This P-R curve represents how the system might actually be used; if every query has at least one match, the connected graph (chain of matches) will cover all of the loops even if they aren't explicit. For example, if query B matches place A and query C also matches place A, we can infer that C also matches B, without needing to explicitly label that match. Contrary to 5a, the image groupings show *improved* performance compared to their ungrouped counterparts, and the colour-constant stack is *significantly* improved over greyscale. Both techniques combined produce far better recall at 100% precision.

Fig. 5 Precision-Recall curves for the recall-only task (no geometric verification). *Grey* indicates only greyscale images were used, *Stack* consists of the greyscale as well as both colour-constant images. The *x5* indicates that sequences of 5 images were grouped and described with a single BoW descriptor. Matches are labelled as true if they are within 30 m of ground truth.

Table 1 Timing results show that the colour-constant stack only adds a small amount of overhead, and that the image grouping is faster by an order of magnitude.

Name	# Queries	Average time (s)
Grey	2189	1.18
Stack	2189	1.44
Grey x5	437	0.11
Stack x5	437	0.15

(a) A successful match at 95% precision with the colour-constant stack that was not found using only greyscale (no image grouping).



(b) A successful match at 95% precision with the image grouping that was not found using single images (no colour-constant channels).

Fig. 6 Interesting examples of successful match hypotheses with the two processing techniques.

and Barfoot [9] further investigate different sized image groups. The binary BoW threshold was chosen as 2 feature occurrences per group to maintain sparsity. The mean binary BoW density for single images were 0.1357 and 0.1227; after grouping and thresholding, they were 0.1149 and 0.1639, respectively. In both cases, the speedup is approximately an order of magnitude (see Table 1). The precision-recall curves shown in Figure 5 show that the grouping deflates the first measure, but improves the second. The first measure considers the precision-recall if the task is to identify *all* of the possible loop closures for each query. The second measure only aims to find *at least one* of the loop closures. Due to the temporal ordering of the queries, if every query has correctly identified at least one loop closure, all possible loop closures are connected without the match being explicitly identified; e.g., B matches A and C matches B, therefore C and A must be a match.

The training for FAB-MAP must be done prior to run-time and is fairly time-consuming compared to the online algorithm. Therefore, the place recognition algorithm is trained in a geographically separate but visually similar environment. Due to geographic limitations, and since this was the first major field deployment for this robotic platform, our training dataset was restricted to 247 images over 1.2 km. It consists of a dry-run for the CSA experiment that took place in Ontario, Canada, primarily in vegetation with a very small sand portion. The confusion matrices, showing the match probabilities for each query are shown in Figure 7. The difficult checkered square regions are the rocks-and-sand sections of the trajectory, the terrain type that was underrepresented in the training dataset.

5 Conclusion and Future Work

We can conclude that both colour-constant images and image grouping show value for place recognition in real outdoor environments. We have also shown reasonable system performance despite a very limited and not fully representative training dataset, and difficult lighting conditions that changed over the course of the day. Future work consists of improving the stability of the colour-constant image channels. By increasing the contrast of the images, the features descriptors may be less corrupted by quantization error, and the detector response may be more stable. A geometric consistency check such as the FAB-MAP 2.0 1-point RANSAC will also improve results by using more-stable features [6]. We can also verify geometric stability by only using features that have been tracked through several frames by VT&R [16].

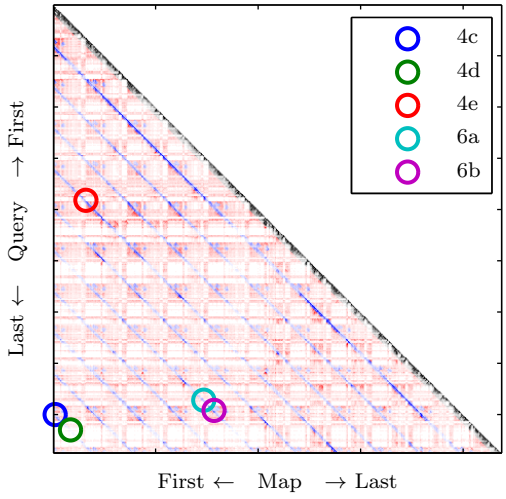
Acknowledgements

We would like to extend our deepest thanks to the Natural Sciences and Engineering Research Council (NSERC) through the NSERC Canadian Field Robotics Network (NCFRN), the Canada Foundation for Innovation, the Canadian Space Agency, and MDA Space Missions for providing us with the financial and in-kind support necessary to conduct this research.

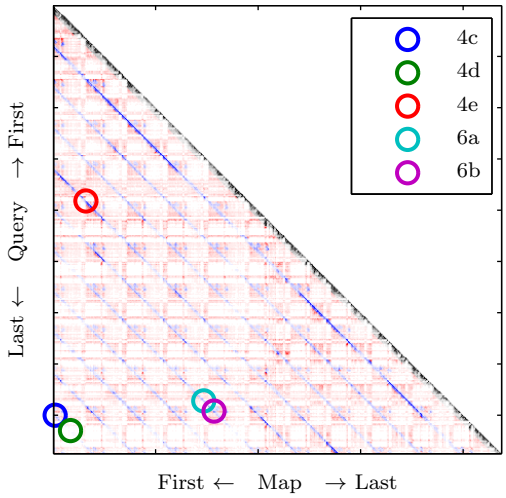
References

[1] Carlevaris-Bianco, N., Eustice, R.M.: Learning Visual Feature Descriptors for Dynamic Lighting Conditions. In: Robot. Autom. (ICRA), 2014 IEEE Int. Conf. (2014)

[2] Collier, J., Se, S., Kotamraju, V., Jasiobedzki, P.: Real-time lidar-based place recognition using distinctive shape descriptors. SPIE Defense, Security, and



(a) Greyscale.



(b) Colour-constant stack.

Fig. 7 Contrast-enhanced confusion matrices show the probability mass for each query (rows) over the mapped places (columns). Correct (true positive) match probability is shown in **blue**, incorrect (false positive) in **red**, and ignored matches (temporally close) in **grey**. The ground-truth for the confusion matrices is shown in Figure 8. The circled interest points correspond to the image examples in Figures 4 and 6. The darker red checkering of false positives show that the system struggled in the rocks-and-sand of the MET, which was underrepresented in the training dataset.

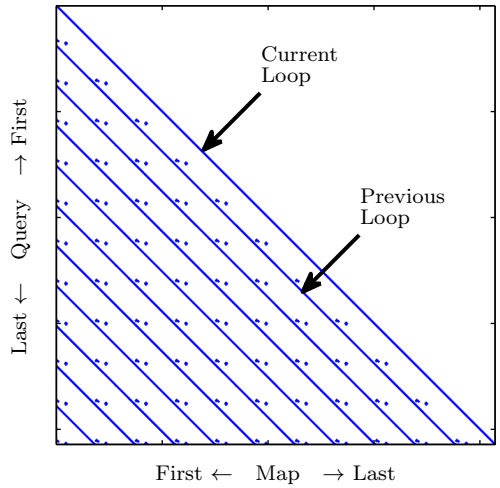


Fig. 8 Ground truth confusion matrix. Since the dataset is a repeated loop, there is diagonal banding, with the current loop on the diagonal, and previous loops on the off-diagonal bands. The smaller dots are regions of the MET that are re-observed toward the end of the loop.

Sensing. 83870P–83870P (May 2012),
[3] Corke, P., Paul, R., Churchill, W., Newman, P.: Dealing with shadows: Capturing intrinsic scene appearance for image-based outdoor localisation. IEEE Int. Conf. Intell. Robot. Syst. pp. 2085–2092 (2013)
[4] Cummins, M., Newman, P.: Accelerated Appearance-Only SLAM. ICRA pp. 1828–1833 (May 2008),
[5] Cummins, M., Newman, P.: FAB-MAP: Probabilistic Localization and Mapping in the Space of Appearance. Int. J. Rob. Res. 27(6), 647–665 (Jun 2008),
[6] Cummins, M., Newman, P.: Appearance-only SLAM at large scale with FAB-MAP 2.0. Int. J. Rob. Res. 30(9), 1100–1123 (Nov 2010),
[7] Finlayson, G., Hordley, S., Cheng, L., Drew, M.: On the removal of shadows from images. Pattern Analysis and Machine Intelligence, IEEE Transactions on 28(1), 59–68 (Jan 2006)
[8] Glover, A., Maddern, W., Warren, M., Reid, S., Milford, M., Wyeth, G.: OpenFABMAP: An open source toolbox for appearance-based loop closure detection. In: 2012 IEEE Int. Conf. Robot. Autom. pp. 4730–4735. IEEE (May 2012),
[9] MacTavish, K., Barfoot, T.D.: Towards hierarchical place recognition for long-term autonomy. In: ICRA Workshop on Visual Place Recognition in Changing Environments (2014)
[10] Maddern, W., Stewart, A.D., McManus, C., Upcroft, B., Churchill, W., Newman, P.: Illumination invariant imaging: Applications in robust vision-based

localisation, mapping and classification for autonomous vehicles. Proc. Work. Vis. Place Recognit. Chang. Environ. IEEE Int. Conf. Robot. Autom. (2014),

[11] Maddern, W., Vidas, S.: Towards Robust Night and Day Place Recognition Using Visible and Thermal Imaging. Proc. Robot. Sci. Syst. pp. 1–6 (2012)

[12] McManus, C., Upcroft, B., Newman, P.: Scene Signatures: Localised and Pointless Features for Localization. In: Proc. Robot. Sci. Syst. Berkely, USA (2014)

[13] Milford, M.: Vision-based place recognition: how low can you go? Int. J. Rob. Res. 32(7), 766–789 (Jul 2013),

[14] Milford, M.J., Wyeth, G.F.: SeqSLAM: Visual route-based navigation for sunny summer days and stormy winter nights. In: 2012 IEEE Int. Conf. Robot. Autom. IEEE (May 2012),

[15] Neubert, P., Sünderhauf, N., Protzel, P.: Superpixel-based appearance change prediction for long-term navigation across seasons. Rob. Auton. Syst. (Aug 2014),

[16] Paton, M., MacTavish, K., Ostafew, C.J., Barfoot, T.D.: Lighting-resistant stereo visual teach & repeat using color-constant images. In: IEEE International Conference on Robotics and Automation (ICRA) (2015)

[17] Paul, R., Newman, P.: FAB-MAP 3D: Topological mapping with spatial and visual appearance. 2010 IEEE Int. Conf. Robot. Autom. pp. 2649–2656 (May 2010),

[18] Pepperell, E., Corke, P.I., Milford, M.J.: Towards Vision-Based Pose- and Condition-Invariant Place Recognition along Routes. IEEE Int. Conf. Intell. Robot. Syst. (2014)

[19] Ratnasingam, S., Collins, S.: Study of the photodetector characteristics of a camera for color constancy in natural scenes. J. Opt. Soc. Am. A 27(2), 286–294 (Feb 2010),

[20] Sunderhauf, N., Neubert, P., Protzel, P.: Are We There Yet? Challenging SeqSLAM on a 3000 km Journey Across All Four Seasons. In: Proc. Work. Long-Term Auton. Int. Conf. Robot. Autom. (2013),