Towards Hierarchical Place Recognition for Long-Term Autonomy

Kirk MacTavish and Timothy D. Barfoot

***Abstract*—Many state-of-the-art approaches to visual place recognition consider matches to each previously visited frame individually. This strategy is at least linear in computational complexity with respect to the number of previous image frames. In the context of long-term autonomy, this is not feasible as the map grows intractably large; the algorithm should operate with sub-linear complexity to handle extremely large scales.**

This paper takes the first steps in adapting the Fast Appearance-Based Mapping (FAB-MAP) algorithm to operate with sub-linear complexity, removing the reliance on the arbitrary image frame place discretization. In our naive preliminary approach, we achieve reasonable performance on several public datasets, grouping up to 256 frames. These hierarchical frame groupings are an enabling factor for a tree search with logarithmic complexity. The success of this naive method encourages further investigation into a purpose-designed hierarchical representation, accompanied by a logarthmic search algorithm.

Additionally, active imaging sensors such as lidar provide data that are robust to external lighting changes. This paper presents novel initial place recognition results on lidar intensity images, and shows that image groups increase performance for these low-quality images. Lidar data are acquired in a continuous scan, making it difficult to use traditional image-based place recognition. The proposed hierarchical place recognition does not depend on a single-image discretization, and is potentially compatible with unstructured lidar without major modification.

I. Introduction

Visual place recognition seeks to identify metrically close, but temporally distant places. The elapsed time between observations allows for challenges such as scene and lighting changes. Even in a mostly static outdoor scene, drastic lighting changes occur over a 24 hour period. Active lidar sensors provide a lighting invariant alternative to passive cameras, offering a solution to the lighting problem. However, data acquisition occurs continuously at a much lower rate, adding a new challenge. The slower data acquisition gives rise to the idea that a place is not simply the instantaneous field-of-view of an image. Rather, an image view is a convenient, but arbitrary, discretization of a metric region. This paper investigates the hypothesis that selecting different scales for this discretization could improve computational complexity. We also present preliminary results on lidar intensity images.

With the rise of relative Simultaneous Localization and Mapping (SLAM) techniques, such as adaptive relative bundle adjustment, some would argue that algorithms must have constant-time complexity to enable true long-term autonomy [1]. Current state-of-the-art appearance-based place recognition algorithms are at least linear in computational complexity [2, 3, 4, 5]. This paper presents the first steps towards a place recognition algorithm with logarithmic complexity. This is achieved by grouping images together at a hierarchy of scales, and in future work, by exploiting this structure to perform a tree search. Fig. 1 shows an example of this hierarchical structure and search process.

The authors are members of the Autonomous Space Robotics Lab at the Institute for Aerospace Studies, University of Toronto, 4925 Dufferin Street, Toronto, Ontario, Canada. kirk.mactavish@mail.utoronto.ca, tim.barfoot@utoronto.ca

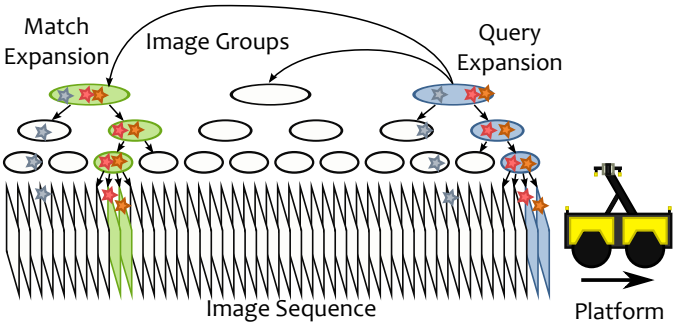


Fig. 1: This image shows an example of hierarchical query expansion. The query (blue) starts at a large grouping level, and finds a match (green). The query is expanded down until the image level is reached. This example shows that longer match sequences (pink and orange stars) are more likely to be expanded than shorter sequences (grey star).

II. Related Work

This paper uses the visual appearance-based place recognition algorithm FAB-MAP to evaluate the feasibility of a hierarchical scheme. FAB-MAP [6] uses a topological probabilistic technique, and has proven to be a seminal work in the field, inspiring community implementations [7]. FAB-MAP models the appearance of an image using a binary Bag-of-Words (BoW) descriptor composed of quantized feature descriptors. FAB-MAP considers matches between two image frames; a prior has a slight impact on weighting match sequences, but frames are primarily considered individually. FAB-MAP 2.0 is able to successfully process a 1000km experiment, the largest experiment of its kind at the time [2, 8]. While FAB-MAP 2.0 is very fast, the computational complexity is still linear with the number of previously visited places.

More recently, methods that represent a place with a sequence of images have been shown to out-perform single image representations. Guillaume et al. [9] use machine learning to encode a fixed-length sequence of images into a BoW using global image descriptors. Mei et al. [10] use word co-visibility to dynamically group images. Continuous Appearance-based Trajectory SLAM (CAT-SLAM) [4] uses a Rao-Blackwellized particle filter to represent loop closure likelihood. Sequential images impact each-other through a motion model and odometry. CAT-SLAM has been shown to out-perform FAB-MAP in recall when odometry is present.

The computational complexity of CAT-SLAM is proportional to the number of particles; it is assumed that the number of particles must increase linearly with the number of previously visited places to maintain a useful probability distribution.

Sequence SLAM (SeqSLAM) [3] performs appearance-based localization by comparing frames using global image descriptors. A line search to identify matching sequences in a smoothed confusion matrix. The computational complexity of SeqSLAM is linear with respect to the number of previous places. This method has shown success on a 3000km dataset across 4 seasons; however, the nature of the global image descriptor results in a significant sensitivity to camera misalignments between passes [11]. More recently, SeqSLAM has been shown to achieve impressive results with extremely low-resolution images [12].

Lidar-based place recognition algorithms have been proposed to meet the challenge of changing lighting conditions. Collier et al. [13] apply FAB-MAP simultaneously to features from 2D images and 3D range data. The addition of lidar results in a notable improvement in varying lighting conditions. Bosse and Zlot [5] introduce a lidar-based place recognition strategy with linear complexity based on keypoint voting. The keypoint voting matches against a subset of the previously observed places, but that subset grows linearly with the number of places. Lidar-based Normal Aligned Radial Features (NARFs) are used in a system similar to FAB-MAP with some additional geometric verification [14], resulting in excellent performance accompanied by slow run-times. The majority of these algorithms use a place description for a continuous lidar scan that is reliant on a fixed-size discretization.

Hierarchy has been used as a speed-up for appearance-based localization by Maohai et al. [15] with a two-scale (coarse and fine) search. The first stage uses global image descriptors integrated over a region to find coarse matches. The second stage uses feature matching between images to confirm matches. This algorithm is a localization system, and localizes against a map built off-line; it does not provide a probability for being in a new place (off-map). Lazebnik et al. show that a hierarchical spatial grouping of features can improve single-image description [16]. A bio-inspired place recognition algorithm using a hierarchical neural network is introduced by Zhang et al. [17]; however, the focus is placed more on scene categorization than long-term place recognition. Most recently, hierarchy has been used to improve recall performance by Chen et al. [18]. Support Vector Machines learn place appearance at multiple scales, and potential matches at a fine scale are cross-checked with larger scale match scores. A hierarchical vocabulary has been used in other work [19, 20, 21], showing that hierarchy has additional benefits for place recognition, beyond place grouping.

III. Methodology

This paper uses the OpenFABMAP implementation by Glover et al. [7]. The goal of this investigation was to avoid modifying the FAB-MAP algorithm. Instead, we seek to manipulate the inputs to achieve reasonable performance on hierarchical image groups. The details for the original FAB-MAP algorithm and implementation can be found in [6] and [7]; however, they will be summarized here for completeness. FAB-MAP requires a training dataset to train the vocabulary, the relationships between words in the vocabulary, and the typical appearance of unmapped places. FAB-MAP typically describes an image using Speeded Up Robust Features (SURF) descriptors, but any vector descriptor can be used. The SURF descriptors extracted from the training dataset are clustered to define the vocabulary. The SURF descriptors are quantized using the vocabulary, and BoW descriptors are constructed for each image. BoW descriptors are typically a histogram of quantized features, but FAB-MAP uses a binary BoW descriptor: each bit indicates word presence in the image. A Chow-Liu tree is used to encode word relationships, trained using word co-visibility in the training images. BoW descriptors from the training dataset are saved as representative unmapped places, concluding the training process.

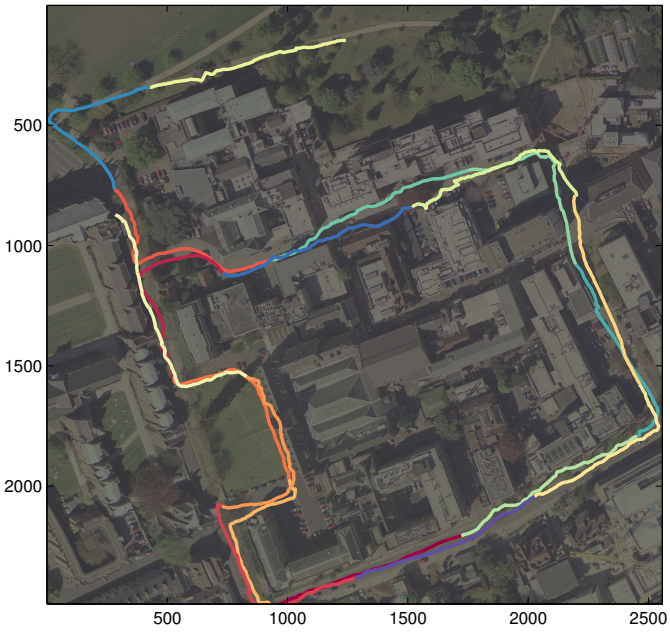


Fig. 2: The City Centre dataset [6] showing place sizes with 128 frames grouped together. We can detect matching groups of up to 256 frames with a single BoW descriptor.

On-line, FAB-MAP uses a probabilistic decision process to determine if the newest image frame matches a previous frame, or is a new place. This decision involves calculating the measurement likelihood at each previous place, the prior on the location, and a normalizing term computed with the aid of the trained representative unmapped places:

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

Where L_i is the i-th location, $\mathcal{Z}^k$ is all of the measurements up to the k-th image, Z_k is the measurement at the k-th image, and $k > i$ (equation (4) from [6]). A fixed threshold is used to determine if a match probability should be classified as a

loop closure. This threshold is swept to create the Receiver Operating Characteristics (ROC) and Precision-Recall (P-R) curves in Fig. 4, discussed further below.

Our hierarchical image groups are created by concatenating two consecutive lower level groups (or images). There are 2^n images in the n-th group level with level 0 containing a single image. Fig. 2 provides a sense of scale for these groupings. In order to run OpenFABMAP on places made from image groups, BoW descriptors for each image in the sequence were summed together.

$$Z_k^0 = [z_1, z_2, ..., z_{v-1}, z_v] \tag{2}$$

$$Z_k^n = \sum_{i=2k}^{2k+1} Z_i^{n-1} \tag{3}$$

Where Z_k^n is the k-th BoW histogram at the n-th level, composed of feature counts z_i for each word. This results in a BoW descriptor that contains the occurrence of every word seen in any of the images within the group. The trained vocabularies typically contained between two thousand and three thousand words. As the place sizes grew, the BoW descriptors became less sparse as more words were seen; increasing from a few hundred words to over half the vocabulary. This resulted in two problems: that the BoW descriptor eventually saturated, becoming less descriptive; and that the word relationships were trained on BoW descriptors which were much sparser, so the relationships were no longer valid.

These problems made groupings larger than a few images unusable. Boureau [22] proposes that sparsity is a key enabler of a binary BoW descriptor. Features seen many times are more important than infrequent features, so we sparsify the descriptor by setting a threshold for the number of times a word must appear to be present in the binary descriptor. This threshold was set so that average sparsity of the BoW descriptor is similar to that of the training images. The threshold value is calculated by increasing the threshold for the group until the group descriptor has similar sparsity to the training data. The thresholds used for each group size and dataset are shown in Fig. 3, and show a strong linear correlation with the number of images in the group. The ground-truth loop closures also need to be resampled for larger group sizes. If any two images match across two groups, those groups are a match.

In future work, the place recognition search will start with the largest group size, and the algorithm will decide which potential matches to expand. These expansions will flow down the tree until there are matches at the image level. Fig. 1 shows an example of this expansion. For this strategy to be effective, there should be as few false negatives as possible at the higher levels, even if it means allowing a few false positives. If the algorithm is required to run in a fixed time budget, the algorithm can choose to expand only the highest probability branches, enabling constant-time detection at the expense of recall.

IV. Experimental Results

This paper uses the New College and City Centre datasets [6], the odometry dataset from the KITTI vision

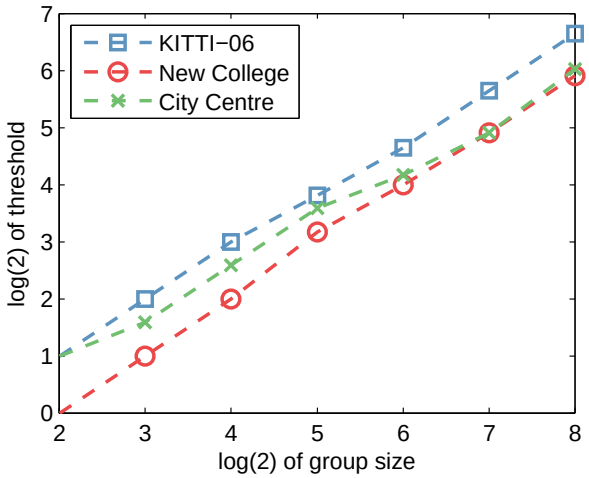


Fig. 3: BoW descriptors summed over a group of images need to be sparsified in order to be useful as binary descriptors. This figure shows the sparsification threshold that was used for each group size and dataset.

TABLE I: Dataset test and training configurations.

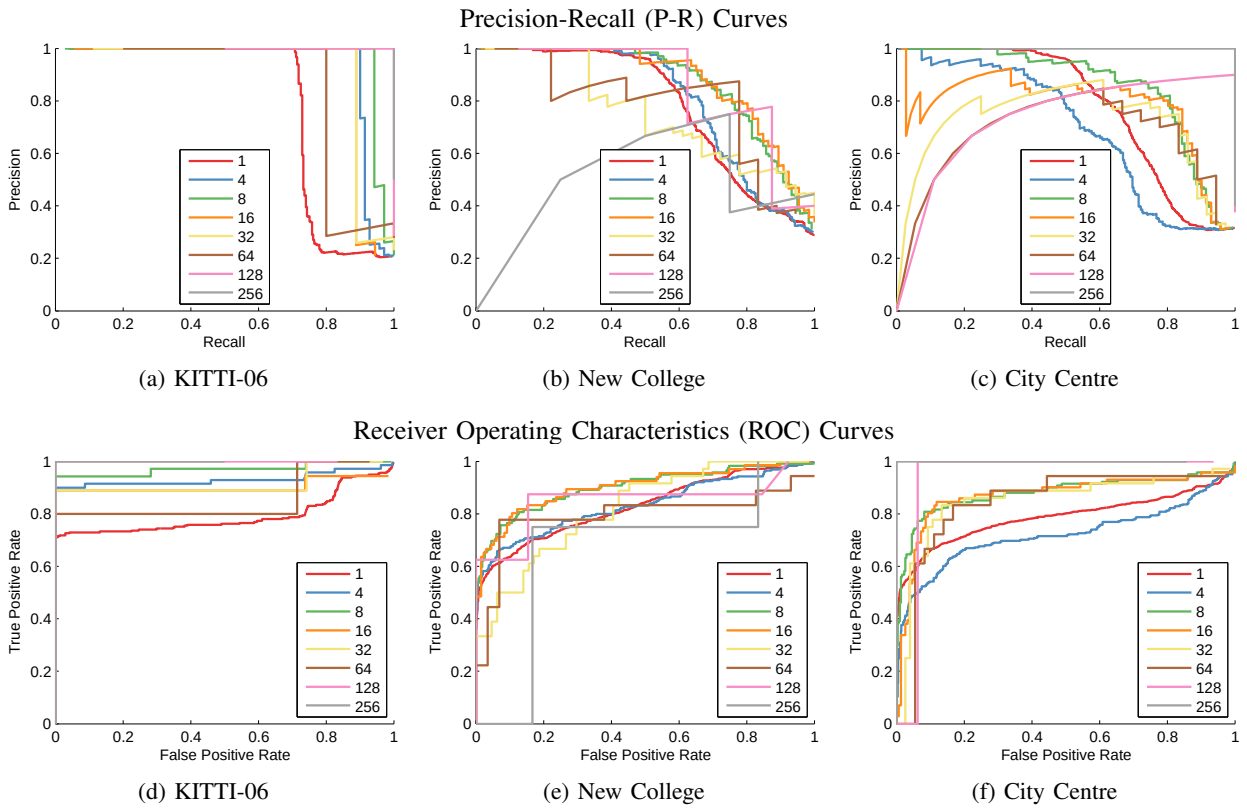
Test Dataset	# Images	Training Dataset(s)	# Images
New College	2146	City Centre	2474
City Centre	2474	New College	2146
KITTI-06	1101	KITTI-01,03,04,10	3374
ABL-T1	4800	ABL-T1	2000

benchmark suite [23], and the ABL Sudbury lidar dataset [25]. Table I shows the test and training dataset configurations that were used. The KITTI odometry 01, 03, 04 and 10 sequences were used for training as they contained no loops. The ABL dataset was broken geographically into a training region and a testing region as shown in Figure 6a. Approximate ground-truth loop-closure data are provided with the New College and City Centre datasets. The KITTI odometry and ABL datasets have ground-truth pose data; ground-truth loop closures were recorded if the images were taken within 15m and 3m of each other respectively.

The SURF detector threshold was set so that the image sequence generated approximately 100-300 features per image. The cluster size was set so that the vocabulary size was between two thousand and three thousand. For all datasets, the detector model was:

$$p(z_i|e_i) = 0.55$$
$$p(z_i|\neg e_i) = 0.15$$

It is important to recall that goal of this investigation is not to out-perform single image FAB-MAP; we are seeking to validate a hierarchical place definition that will enable lower computational complexity. Fig. 4 presents P-R and ROC curves for each of the image test datasets. The legend indicates the number of images grouped together as a place. A place size of 1 is traditional OpenFABMAP, and is used as a performance reference. At a grouping of 256, there are



(a) KITTI-06 (b) New College (c) City Centre

(d) KITTI-06 (e) New College (f) City Centre

Fig. 4: Performance for the three image datasets. Image group sizes are shown in the legend. The curves are created by sweeping the match probability threshold.

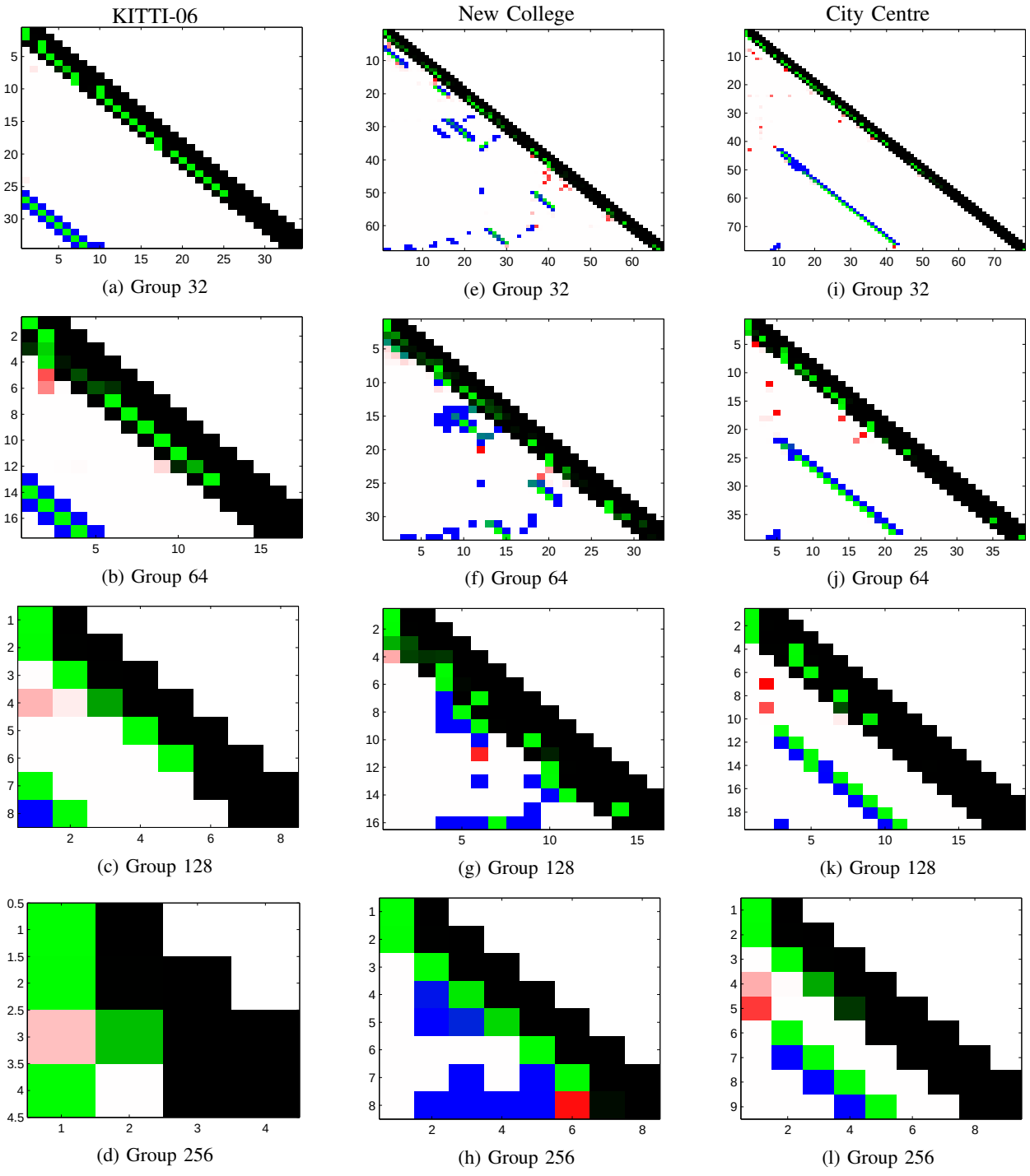
between 1 and 4 potential loop closures to identify. Fig. 5 shows the confusion matrices for the larger group sizes on each dataset. Ground-truth loop closures are shown in blue and green, green indicates that the loop closure was detected. Red indicates a false-positive loop closure.

Every group size out-performs single-image FAB-MAP on the KITTI-06 dataset. At around 70% recall, nearly all of the group sizes out-perform the single-image results on all datasets. This performance point is encouraging, as it fulfils our desire for high recall while sacrificing a few false-positives. The group size of 4 on City Centre had difficulty, as the BoW threshold did not provide enough resolution to achieve a sparsity comparable to the training data. We hypothesise that this issue could be resolved by adjusting the feature detection threshold, which provides much finer sparsity resolution. At 100% precision, single images tend to out-perform image groups, although this is not the performance point that holds our interest. The lidar dataset proved far more challenging. Figure 6e shows an example image from the dataset. It is low-resolution, has motion distortion, and there is tearing, making the features scarce and unstable. Figures 6c and 6b show that the grouping strategy was necessary to make the results at all useful, although the performance in general still needs improvement.

V. Conclusions and Future Work

This paper has presented a successful hierarchical place grouping for place recognition as well as novel results on lidar intensity images. The BoW sparsification strategy allows successful place recognition at a scale up to 256 grouped images. This hierarchical structure is a key enabler for logarithmic complexity place recognition, and place recognition with low-quality lidar intensity images. The initial results are promising, and are expected to improve with a description model that has been designed for a hierarchical structure.

Future work is required to verify that there is significant detection overlap at different group levels, so that the tree search is able to traverse the necessary branches. We also hypothesise that group sizes beyond those presented here will not be well described using this simple sparsification strategy. It is expected that a deeper descriptor will be needed to maintain descriptivity of these larger places. Other works [16, 24, 22] provide a good starting place for a deeper descriptor with the sparse binary characteristics FAB-MAP-like inferencing requires. We plan to design a hierarchical place description for feature vectors accompanied by a logarithmic-complexity probabilistic inferencing system. This system will be field tested as well as benchmarked on common datasets such as those used in this paper. We also aim to use this system with features extracted from unstructured 3D intensity and geometry data from lidar.



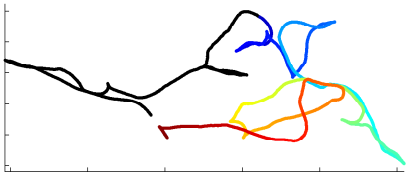
(a) Group 32 (e) Group 32 (i) Group 32

(b) Group 64 (f) Group 64 (j) Group 64

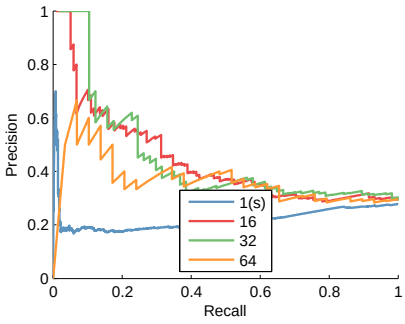
(c) Group 128 (g) Group 128 (k) Group 128

(d) Group 256 (h) Group 256 (l) Group 256

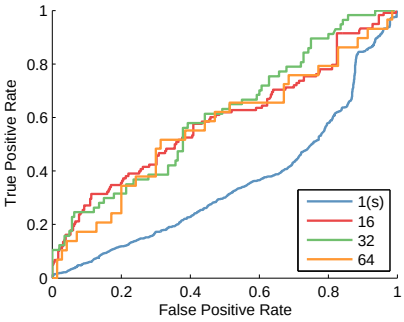
Fig. 5: Confusion matrices for the three image datasets at the larger group sizes. These groups provide intuition into the evolution of the performance at different levels of the hierarchy. Green represents true positives (or masked positives), red false positives, blue ground truth loop closures (potentially false negatives), black the masked region where no loop closures are considered. The strength of green or red indicates the estimated probability of the loop closure.



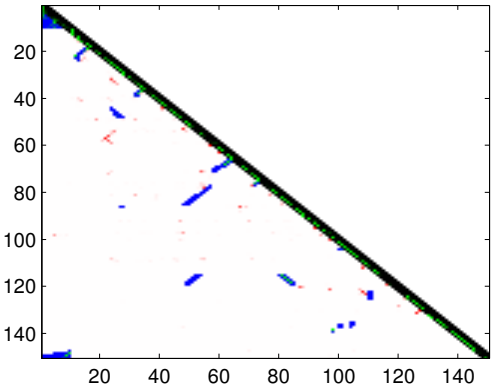
(a) Train sequence in black, test sequence is coloured. The test sequence is over twice as long as the train sequence.



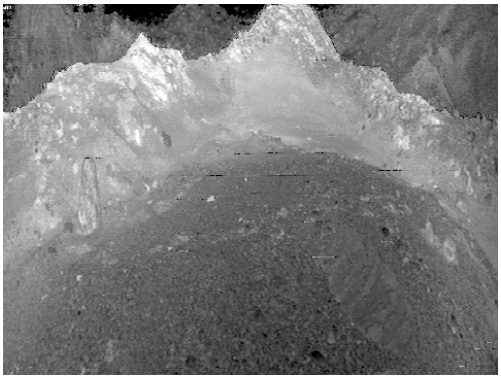
(b) P-R curve for alternating single images [1(s)], and three group sizes. Using image groups was necessary to make the results at all useful.



(c) ROC curve for alternating single images [1(s)], and three group sizes.



(d) Confusion matrix for a group size of 32. While there are many more false positives than the regular image datasets, we are still able to find some loop closures.



(e) An example lidar intensity image from the dataset. The images are low-resolution, have motion distortion, and tearing from missing pixels.

Fig. 6: Results for the Sudbury lidar dataset [25] using intensity images.

VI. Acknowledgment

We would like to extend our deepest thanks to the NSERC and the Canada Foundation for Innovation, DRDC Suffield, the Canadian Space Agency, and MDA Space Missions for providing us with the financial support necessary to conduct this research.

References

[1] G. Sibley, C. Mei, I. Reid, and P. Newman. Vast-scale Outdoor Navigation Using Adaptive Relative Bundle Adjustment. *The International Journal of Robotics Research*, 29(8):958–980, May 2010.

[2] M. Cummins and P. Newman. Highly scalable appearance-only SLAM-FAB-MAP 2.0. *Robotics: Science and Systems*, 2009.

[3] M. J. Milford and G. F. Wyeth. SeqSLAM: Visual route-based navigation for sunny summer days and stormy winter nights. In *2012 IEEE International Conference on Robotics and Automation*, pages 1643–1649. IEEE, May 2012.

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

[5] M. Bosse and R. Zlot. Place recognition using keypoint voting in large 3D lidar datasets. In *2013 IEEE International Conference on Robotics and Automation*, pages 2677–2684. IEEE, May 2013.

[6] M. Cummins and P. Newman. FAB-MAP: Probabilistic Localization and Mapping in the Space of Appearance. *The International Journal of Robotics Research*, 27(6): 647–665, June 2008.

[7] A. Glover, W. Maddern, M. Warren, S. Reid, M. Milford, and G. Wyeth. OpenFABMAP: An open source toolbox for appearance-based loop closure detection. In *2012 IEEE International Conference on Robotics and Automation*, pages 4730–4735. IEEE, May 2012.

[8] M. Cummins and P. Newman. Appearance-only SLAM at large scale with FAB-MAP 2.0. *The International Journal of Robotics Research*, 30(9):1100–1123, November 2010.

[9] H. Guillaume, M. Dubois, E. Frenoux, and P. Tarroux. Temporal Bag-of-Words-A Generative Model for Visual Place Recognition using Temporal Integration. In *VISAPP-International Conference on Computer Vision*

Theory and Applications, 2011.

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

[11] N. Sunderhauf, P. Neubert, and P. Protzel. Are We There Yet? Challenging SeqSLAM on a 3000 km Journey Across All Four Seasons. In *Proc. of Workshop on Long-Term Autonomy, International Conference on Robotics and Automation (ICRA)*, 2013.

[12] M. Milford. Vision-based place recognition: how low can you go? *The International Journal of Robotics Research*, 32(7):766–789, July 2013.

[13] J. Collier, S. Se, and V. Kotamraju. Multi-sensor Appearance-Based Place Recognition. *2013 International Conference on Computer and Robot Vision*, pages 128–135, May 2013.

[14] B. Steder, M. Ruhnke, S. Grzonka, and W. Burgard. Place Recognition in 3D Scans Using a Combination of Bag of Words and Point Feature based Relative Pose Estimation. In *Intelligent Robots and Systems (IROS), IEEE/RSJ International Conference on*, 2011.

[15] L. Maohai, S. Lining, H. Qingcheng, C. Zesu, and P. Songhao. Robust Omnidirectional Vision based Mobile Robot Hierarchical Localization and Autonomous Navigation. *Information Technology Journal*, 10(1):29–39, January 2011.

[16] S. Lazebnik, C. Schmid, and J. Ponce. Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2 (CVPR'06)*, volume 2, pages 2169–2178. IEEE, 2006.

[17] X. Zhang, J. Zhang, A. B. Rad, X. Mai, and Y. Jin. A novel mapping strategy based on neocortex model: Preliminary results by hierarchical temporal memory. In *2012 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, pages 476–481. IEEE, December 2012.

[18] Z. Chen, A. Jacobson, U. Erdem, M. Hasselmo, and M. Milford. Towards Bio-inspired Place Recognition over Multiple Spatial Scales. In *Australasian Conference on Robotics and Automation*, 2013.

[19] D. Nister and H. Stewenius. Scalable Recognition with a Vocabulary Tree. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2 (CVPR'06)*, volume 2, pages 2161–2168. IEEE, 2006.

[20] E. Fazl-Ersi, J. H. Elder, and J. K. Tsotsos. Hierarchical appearance-based classifiers for qualitative spatial localization. In *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 3987–3992. IEEE, October 2009.

[21] A. Kawewong, N. Tongprasit, and O. Hasegawa. A speeded-up online incremental vision-based loop-closure detection for long-term SLAM. *Advanced Robotics*, 27 (17):1325–1336, December 2013.

[22] Y.-l. Boureau. *Learning Hierarchical Feature Extractors For Image Recognition*. PhD thesis, New York University, 2012.

[23] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? The KITTI vision benchmark suite. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3354–3361. IEEE, June 2012.

[24] Y.-L. Boureau, F. Bach, Y. LeCun, and J. Ponce. Learning mid-level features for recognition. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2559–2566. IEEE, June 2010.

[25] S. Anderson, C. Mcmanus, H. Dong, E. Beerepoot, and T. D. Barfoot. The Gravel Pit Lidar-Intensity Imagery Dataset, 2013. URL http://asrl.utias.utoronto.ca/datasets/abl-sudbury/.