Improving Condition- and Environment-Invariant Place Recognition with Semantic Place Categorization

Albert Author[1] and Bernard D. Researcher[2]

***Abstract*— The problem of place recognition actually comprises two distinct subproblems; "ordinary" place recognition which is recognizing a specific location in the world, and place "categorization", which involves recognizing the type of place. Both components of place recognition are competencies for robotic navigation systems and hence have each in isolation received significant attention in the robotics and computer vision community. In this paper, we leverage the powerful complementary nature of the place recognition and place categorization processes to create a new state-of-the-art ordinary place recognition system that uses place context to inform place recognition. We show that semantic place categorization creates a more informative natural segmenting of physical space than the blindly applied fixed segmentation used in algorithms such as SeqSLAM, which enables significantly better place recognition performance. In particular, where existing condition-invariant algorithms enable robustness to globally consistent change (such as day to night cycles), this new semantically informed approach adds robustness to significant changes within the environment, such as transitioning from indoor to outdoor environments. We perform a number of experiments using benchmark and new datasets and show that semantically-informed place recognition outperforms the previous state of the art systems. Like it does for object recognition [ref Niko IROS2015], we believe that semantics can play a key role in boosting conventional place recognition and navigation performance for robotic systems.**

I. INTRODUCTION

II. RELATED WORK

III. PROPOSED METHOD

The proposed approach has two main components: place categorization and place recognition as depicted in Fig. 1, with semantic information flowing from the former to the latter to generate the final outcome. The semantic labels divide the physical space into different regions based on its appearance, that is, scene attributes. These segmented regions are then used for improving the place recognition performance. We use CNN model VGG16-places365 [1] pre-trained on Places365 database [2] for labeling reference and query database images as indoor or outdoor along with scene attributes [3] prediction. We use SeqSLAM [4] for showing improved place recognition using semantic information by appropriately enhancing the image matching scores.

*This work was not supported by any organization

[1]Albert Author is with Faculty of Electrical Engineering, Mathematics and Computer Science, University of Twente, 7500 AE Enschede, The Netherlands albert.author@papercept.net

[2]Bernard D. Researcheris with the Department of Electrical Engineering, Wright State University, Dayton, OH 45435, USA b.d.researcher@ieee.org

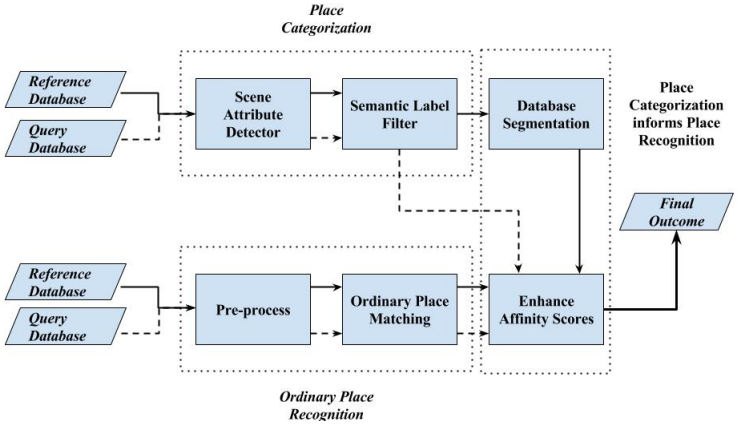


Fig. 1. A block diagram showing flow of semantic information from the place categorization module to the place recognition module for generating the final outcome.

A. Place Categorization

1) Image Classification: The pre-trained CNN model classifies an image with probabilities associated with each of the 365 place categories. It also predicts the most probable scene attributes (out of 102 attributes trained on SUN database [3]) using one of its fully-connected layers. We use the top-5 most probable attribute predictions for post-processing the image labels to semantically segment datasets. The classification is performed on the entire reference and query database. The reference database labels are used for temporally dividing the image sequence into different chunks based on its scene attributes. The query database labels are used later during place recognition for identifying the correct semantic segment of reference database and effectively matching the image sequence.

2) Dataset Segmentation: The image labels for reference database obtained from the classifier do not usually exhibit local temporal persistence of their labels. In order to achieve an adequate dataset segmentation, we find temporal connected components in the database. Each image in the database is represented by a node N_i defined as a set containing the top-5 predicted labels and the most probable prediction label L_i. A consecutive pair of nodes is considered to be connected if an edge $E_{i,i+1}$ exists between them as per Eq. 1.

$$E_{i,i+1} = \begin{cases} 1, & \text{if } |N_i \cap N_{i+1}| \geq 3 \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

The edges between the nodes can be determined by a single pass over the entire image sequence. A new connected

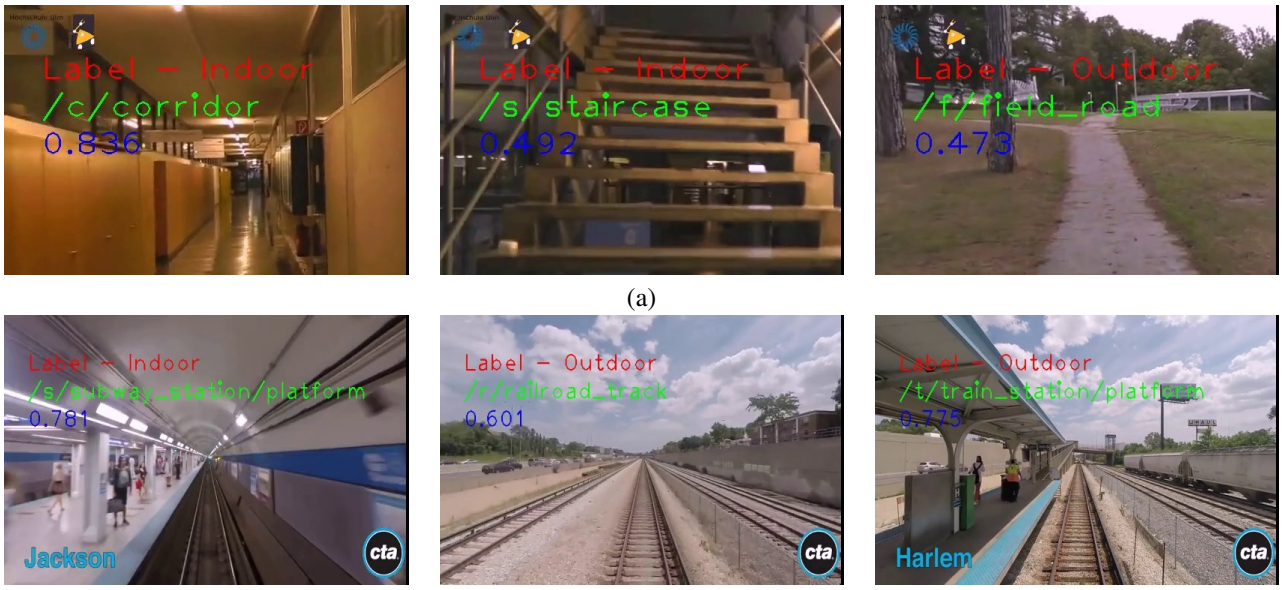


Fig. 2. Images from reference database with indoor/outdoor labels, top predicted label from 365 place categories and its associated probability. (a) Campus Dataset and (b) CTA Rail Dataset

component is obtained whenever an edge between consecutive nodes ceases to exist. The most probable prediction label L_i of each node in a connected component is used to find the most frequently occuring label L_i' within that component. This label L_i' is used for representing the connected component as well as each of its nodes. These newly obtained labels are further filtered to get rid of transient errors. We use a sliding window of size s that passes throught the entire image sequence and replaces the label L_i' by the mode of labels in the sliding window as given in Eq. 2, where $M(X)$ for a set X gives the statistical mode of the set.

$$\hat{L}_i = M(\bigcup_{i-s/2}^{i+s/2} L_i') \tag{2}$$

The filtered image labels $\hat{L}_i$ are finally used to segment the database into different chunks which possess a consistent label. These chunks represent the variation in appearance of the environment while traversing the route. For example, a train running underground as opposed to over the ground will have different appearance of its environment. Similarly, a person walking indoor or outdoor of a campus will witness different surrounding environment as shown in figure2. Fig. 3 shows the images and their semantic labels at the transition points of one of the segmented datasets used in this paper.

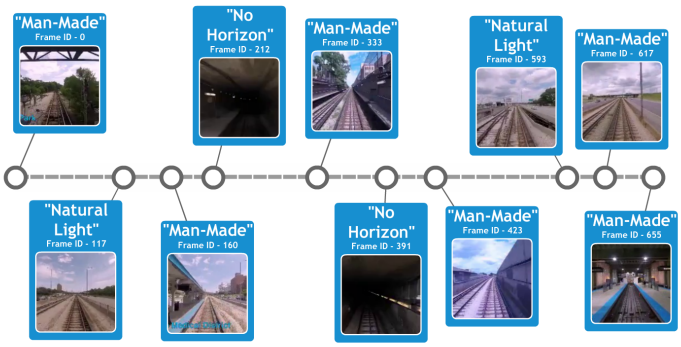


Fig. 3. The time-line of CTA-Rail dataset with semantically labeled images at the transition points of segmented reference database. Time-line created using [5]

B. Place Recognition

1) SeqSLAM: SeqSLAM [4] is a place recognition method known to work remarkably well in challenging environmental conditions. The recent advanced methods [6], [7], [8] etc. based on vanilla SeqSLAM further improve the state-of-the-art for place recognition. In this paper, we use the vanilla approach to show the performance improvement using semantic scene information. The detailed methodology of SeqSLAM can be referred to in [4] and is briefly explained here. The method uses Sum of Absolute Difference (SAD) scores between preprocessed reference and query images. The preprocessing step involves down-sampling of image to 32x32 size and patch normalizing the images with a fixed window size P. The SAD matrix formed between the reference and query database undergoes contrast enhancement for each query image. The contrast enhancement refers to the normalization of SAD scores within a sliding window of size R. The contrast enhanced SAD matrix is then searched for image sequence trajectories of size D_s within a limited range of velocities originating from each of the reference image. The sequence trajectory with best SAD score is then selected using a trajectory uniqueness threshold μ.

2) Segmented Region Normalization: The contrast enhancement step in SeqSLAM uses an arbitrary window size for locally normalizing the matching scores as described above. We use the segmented regions of the dataset to

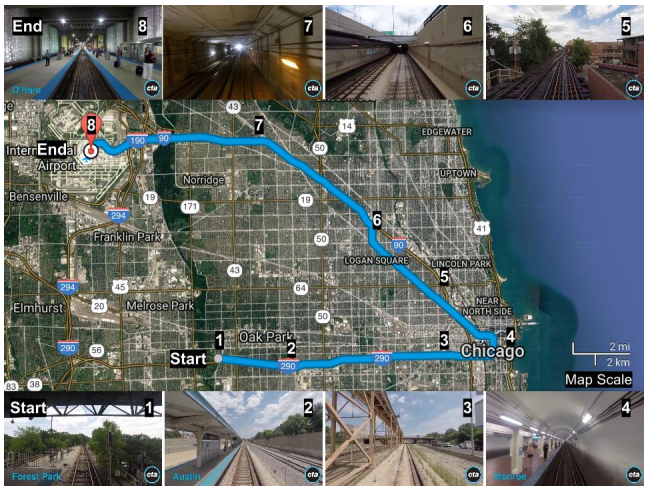


Fig. 4. CTA Dataset Trajectory Aerial View with sample images. Marked Trajectory Source - [11]

set the normalization window of each query image. The label associated with the query image is used to identify the matching regions of dataset where the normalization is performed with window size equivalent to the size of the matching region. The unmatched regions use the vanilla method of contrast enhancement.

IV. Experimental Setup

The experiments are performed using two datasets described in subsequent section. The image classification is performed using Dell M4800 Intel Core i7, 3.1 GHz processor with NVIDIA Quadro K2100M graphics card. The place recognition is performed using Dell E7450 Intel Core i7, 2.6 GHz processor. The classification is done as a preprocessing step for reference as well as query database off-line.

A. Dataset

The two datasets used in the experiments have instances of different environmental appearances as the route is traversed.

1) CTA-Rail: The CTA-Rail (Chicago Transit Authority) dataset (Fig. 4) comprises of two videos traversing a 23 km railway route (Blue Line to O'Hare) recorded once in 2014 [9] and then in 2015 [10] available online. The video comprises of scenes from train stations platforms, subway station platforms, subway tunnels, railroad tracks within highways and urban areas. The entire video sequence captured at a high frame-rate is used for current experiments processing every 200th frame. The resultant reference and query databases have 656 and 738 image frames respectively.

2) Campus Indoor Outdoor: The Campus Indoor Outdoor dataset comprises of two videos with repeated traversal of campus from outside lawn to inside corridor [12], [13]. The videos are recorded using a hand-held device and has jerky motion with huge motion blur. The raw videos are cropped to remove the comments at the bottom and an overlaid navigation display on the right side. The videos are also snipped from beginning so that the starting point is aligned in both the sequences. The dataset therefore comprises of scenes from outside the campus with trees, grass and field road, and from inside the campus traversing through entrance hall, staircase, lobby and corridor. The reference and query database is processed by using every 10th image and therefore uses 355 and 300 frames respectively.

3) Ground Truth: The place recognition ground truth for both the datasets is generated manually for intermittent frames and then interpolated for rest of the image sequence. A query image is considered to be a true positive match for the reference image if its index lies within a range of 5 image frames from the ground truth index.

4) SeqSLAM parameters: The parameters for SeqSLAM used for all the experiments are shown in Table I.

TABLE I
SeqSLAM parameters.

S	Image Down-sampling Size	32x32
P	Patch Normalization Window Size	2,4,8,16
O	Image Matching Offset Range	± 10
D_s	Sequence Length	10
R	Contrast Enhancement Window Size	10,20,40,80,160
V	Sequence Search Velocity Range	1 ± 0.4
μ	Trajectory Uniqueness Threshold	Varied from 0 to 255

V. Results

The place recognition performance is measured using maximum F1 Score by varying the trajectory uniqueness parameter (described in [4]), that is, the threshold for deciding a correctly matched place. We also varied two of the parameters of SeqSLAM method that are known to affect the performance, in order to understand the performance changes by using the proposed approach. The performance improvement is as shown in Figure 5 and the effect of parameters is discussed in subsequent section.

Figure 6 shows the ground truth and the place recognition matches (without thresholding) corresponding to different parameter settings for both vanilla SeqSLAM and proposed method. The results in first row correspond to Campus Indoor-Outdoor dataset and second row to CTA Rail dataset. The performance improvement is large for both the datasets for smaller values of R (contrast enhancement window size). The large R values help the vanilla method to attain performance equivalent to the proposed approach as shown in Figure 6. The effect of dataset segmentation can be easily seen and understood in the result images in form of rectangular dark and light patches in the SAD matrix.

VI. Discussion

A. Effect of SeqSLAM parameters

1) Sequence Normalization Zone Width (R): This parameter is used to set the window size for locally enhancing the contrast of the match scores, that is, normalizing the score values in a local window. As shown in Figure 5, for vanilla SeqSLAM method, the performance improves with increasing the value of this parameter. This is also expected because normalizing over a larger image sequence balances

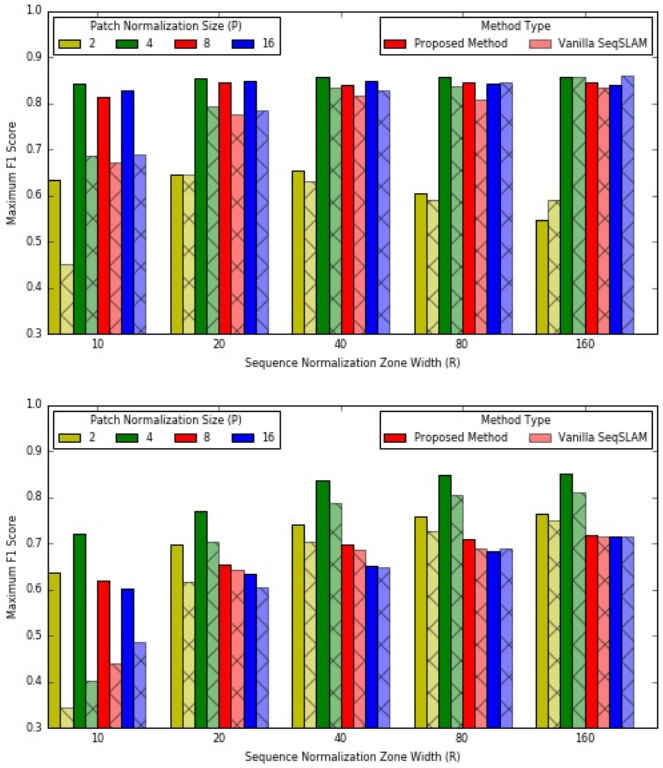


Fig. 5. Performance chart showing maximum F1 Score. The first row corresponds to the CTA-Rail dataset and second row to the Campus Indoor-Outdoor dataset. The performance improves with increasing the parameter value R, where vanilla SeqSLAM performs equally good as the proposed approach. The patch normlization Size (P) parameter with value 4 tends to give maximum score as compared to others.

the overall variation in scores, but it leads to suppressing of place recognition corresponding to images which have low matching scores. On the other hand, using segmented regions to set the window size, effectively enhances the contrast of matching scores in appropriate regions and yields correct matching pairs. The performance is, therefore, most of the times better than the best achieved using vanilla approach. Moreover, setting the window size large is not appropriate for large datasets and long time navigation. Fig. 7 shows a performance comparison for both the datasets with respect to the parameter R.

2) Patch Normalization Window Size (P): The images used for finding SAD score are preprocessed by down-sampling them to the size of 32x32 and patch normalizing them in order to counter the change in appearance of their matching counterparts. Depending on the type of environment and corresponding imagery, the patch normalization window size can give different performance, but a higher value is usually recommended. In the experiments conducted here, we found that patch normalization window size of 4 and 8 for the given down-sampling image size gives a better performance.

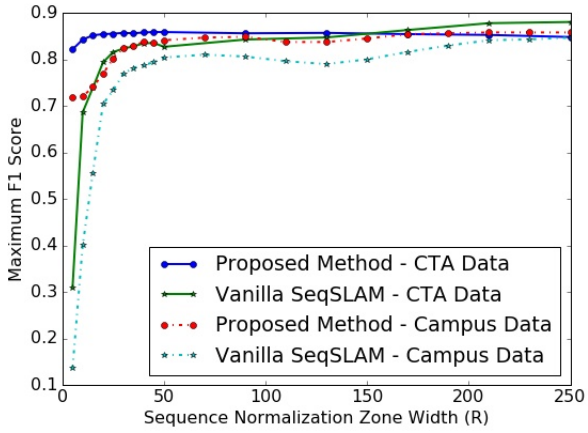


Fig. 7. Performance comparison of proposed method and vanilla SeqSLAM with respect to parameter R for CTA and Campus datasets.

VII. Conclusion

ACKNOWLEDGMENT

The preferred spelling of the word acknowledgment in America is without an e after the g. Avoid the stilted expression, One of us (R. B. G.) thanks . . . Instead, try R. B. G. thanks. Put sponsor acknowledgments in the unnumbered footnote on the first page.

References

[1] B. Zhou, "The Places365-CNNs," https://github.com/metalbubble/places365, 2016, [Online; accessed 15-August-2016].

[2] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning deep features for scene recognition using places database," in *Advances in neural information processing systems*, 2014, pp. 487–495.

[3] G. Patterson and J. Hays, "Sun attribute database: Discovering, annotating, and recognizing scene attributes," in *Proceeding of the 25th Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

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

[5] ReadWriteThink, "ReadWriteThink - http://www.readwritethink.org/," http://www.readwritethink.org/files/resources/interactives/timeline_2/, 2016, [Online; accessed 15-August-2016].

[6] M. Milford, C. Shen, S. Lowry, N. Suenderhauf, S. Shirazi, G. Lin, F. Liu, E. Pepperell, C. Lerma, B. Upcroft, *et al.*, "Sequence searching with deep-learnt depth for condition-and viewpoint-invariant route-based place recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2015, pp. 18–25.

[7] Y. Wang, X. Hu, J. Lian, L. Zhang, and X. Kong, "Improved seq slam for real-time place recognition and navigation error correction," in *Intelligent Human-Machine Systems and Cybernetics (IHMSC), 2015 7th International Conference on*, vol. 1. IEEE, 2015, pp. 260–264.

[8] M. Milford, H. Kim, M. Mangan, S. Leutenegger, T. Stone, B. Webb, and A. Davison, "Place recognition with event-based cameras and a neural implementation of seqslam," *arXiv preprint arXiv:1505.04548*, 2015.

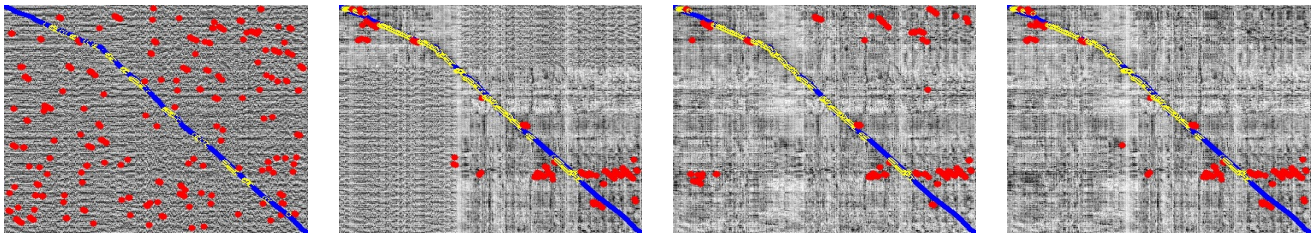
[9] CTAConnections, "CTA Ride the Rails: Blue Line to O'Hare in Real Time," https://youtu.be/n6xJFpPY_7s, 2014, [Online; accessed 15-August-2016].

[10] ——, "CTA Ride the Rails: Blue Line to O'Hare in Real Time (2015)," https://youtu.be/Kw_BbQoDv8o, 2015, [Online; accessed 15-August-2016].

[11] G. Maps, "Google Maps: Directions from Forest Park, IL, USA to O'Hare International Airport, IL, USA," https://goo.gl/maps/fX6aSDMnJpC2, 2016, [Online; accessed 15-August-2016].

[12] RoboticsAtHsUlm, "Seamless Indoor and Outdoor Navigation based on OpenStreetMap," https://youtu.be/_ConuKUOXH4, 2016, [Online; accessed 15-August-2016].

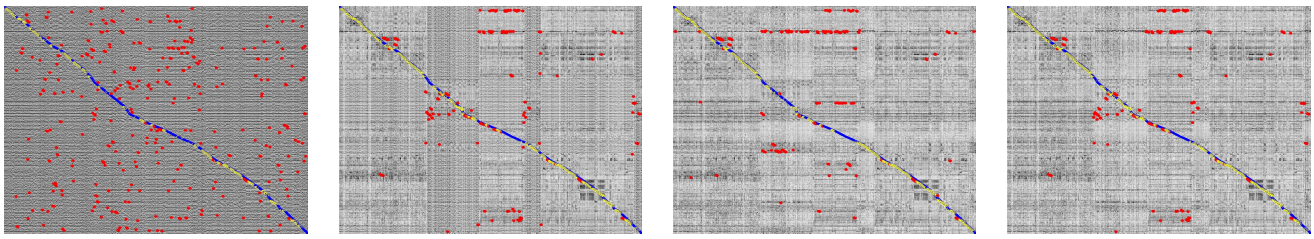
[13] ——, "Demonstrating System Integration by Composition: Seamless Indoor and Outdoor Navigation," https://youtu.be/fS3PJMswlH4, 2015, [Online; accessed 15-August-2016].



(a) F Score = 0.46 SeqSLAM | (b) F Score = **0.74** Proposed Approach | (c) F Score = 0.72 SeqSLAM | (d) F Score = **0.75** Proposed Approach

R = 10, P = 8 | **R = 160**, P = 8

Campus Indoor-Outdoor Dataset



(e) F Score = 0.61 SeqSLAM | (f) F Score = **0.82** Proposed Approach | (g) F Score = 0.80 SeqSLAM | (h) F Score = **0.82** Proposed Approach

R = 10, P = 4 | **R = 160**, P = 4

CTA Rail Dataset

Fig. 6. The SAD (Sum of absolute difference) matrix with reference database as rows and query database as columns. The ground truth is marked in blue, loop closures in red and true positives in yellow. The first and second row corresponds to Campus Indoor-Outdoor and CTA-Rail dataset respectively. Results in (a)-(b) and (e)-(f) correspond to smaller normalization window for vanilla SeqSLAM and proposed approach respectively while (c)-(d) and (g)-(h) correspond to larger normalization window. The dataset segmentation can be seen as rectangular dark and light patches. The images here show that performance improvement is large for smaller R values and with larger R values, performance of vanilla SeqSLAM method approaches to that of proposed method.