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

The problem of "traditional?" place recognition (find new word for ordinary) typically focuses on recognizing specific locations in the world stored within a database of "places". This form of place recognition is very powerful, enabling localization on very large scales [1] and during difficult day and night traverses of an environment [2]. The problem of place categorization is similar to the place recognition problem, where environments are evaluated to determine the type of place from a database of place types.

We see the problem of place recognition as an extension to the place categorization problem, where it is possible to use similar frameworks to solve both problems. The place categorization framework has less 'labels' than place recognition and more training examples of what represents a particular place category whereas place recognition has a "label" within a database representing every location in the environment. We also highlight difference in application between the two approaches, noting that within place recognition frameworks, the goal is to identify and utilize differences between locations within the dataset to enable unique localization. Place categorization algorithms highlight the similarity between intra-class samples to create a comprehensive representation of a particular place type.

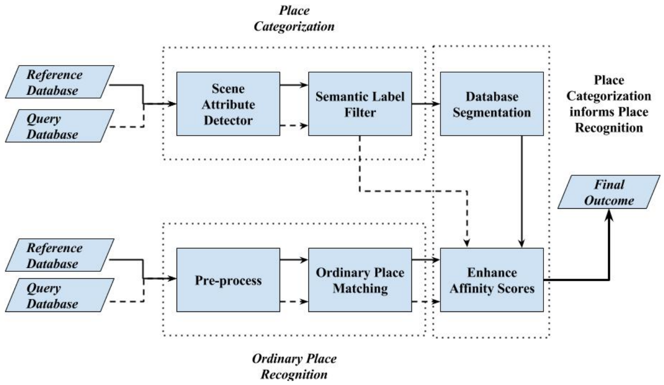


Fig. 1. A block diagram showing the flow of semantic information from the place categorization module to the place recognition module for generating the final outcome.

In this work, we combine the two frameworks of place recognition and place categorization to improve place recognition localization performance. Our primary contribution is the utilization of a place categorization framework to inform place recognition place matches (as seen in Fig. 1). We utilize the Convolutional Neural Network (CNN) model VGG16-places365 [3] pre-trained on Places365 database [4] and scene attributes [5] for performing place categorization. Once a place is categorized, we leverage the SeqSLAM framework to perform place recognition, implementing a dynamic weighting scheme biasing place matches with similar place characteristic and place categorization results.

We evaluate our proposed approach within the two real world datasets, the Campus Dataset and the CTA-Rail Dataset. The Campus Dataset utilizes a single camera traversing the indoors and outdoor campus environments and the CTA-Rail Dataset consists of a single camera mounted to a train traversing of indoor, outdoor scenes with subway station platforms, subway tunnels and railroad tracks. The proposed approach, incorporating place categorization information, outperforms a standalone state-of-the-art place recognition system in both environments.

The paper proceeds as follows. In Section 2, we review literature with a focus on place recognition and place categorization. Section 3 presents our approach describing

*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

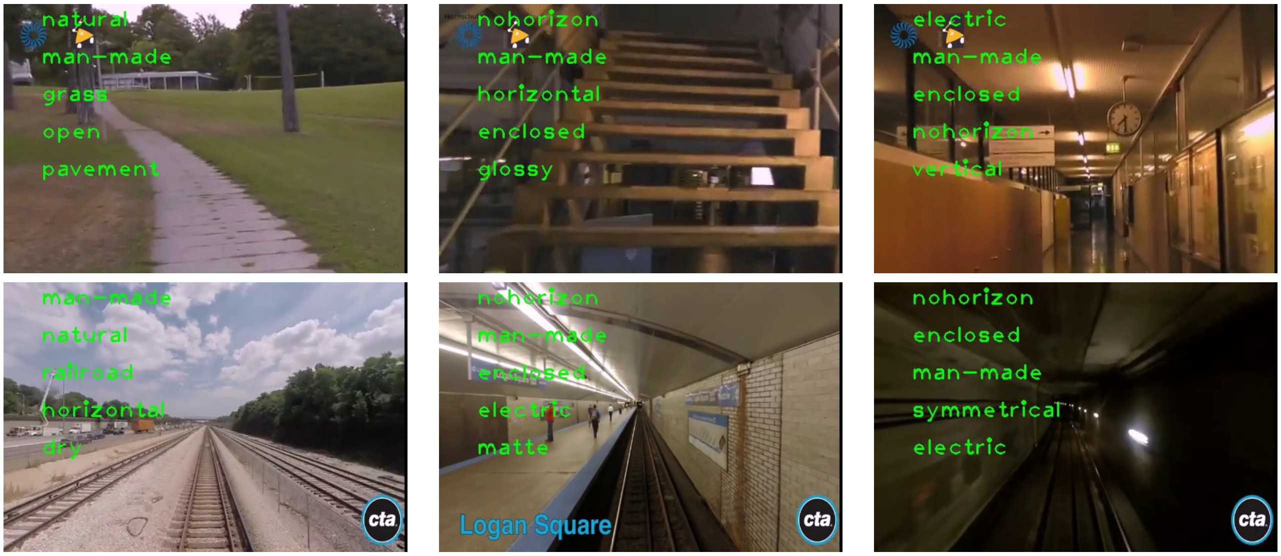


Fig. 2. Images from reference database with top-5 most probable semantic labels out of 102 scene attributes for Campus Dataset (top) and CTA Rail Dataset (bottom).

in the implementation of our CNN place categorization framework, outlines our place recognition framework and outlines the proposed fusion technique for combining the place recognition and place categorization results to produced superior place recognition results. In Section 4, we present the experimental setup, and present the results of multiple levels of evaluation in Section 5. Section 6 discusses the significance of the research and discuss areas of future work.

II. Related Work

In this section, we review current research in the areas of place categorization and place recognition. We specifically focus on place recognition, semantic mapping and place categorization frameworks.

A. Place Recognition

Visual place recognition leverages a visual map of the environment and compares visual information, typically from a camera sensor, to the map data to determine the current location of the camera within the map. There are many techniques which have been proposed to solve this problem of determining where an image has been taken within an environment. Typically, these approaches leverage single frame matching to determine the location of the camera in the environment. The key goal of place recognition frameworks is to separate places in the environment and highlight the unique attributes or features which uniquely describe individual locations in the environment [ref - fabmap, vocabulary trees].

Temporal information has been incorporated into the place recognition framework with the introduction of the SeqSLAM framework, integrating place hypotheses over small distances to accrue evidence and improve place recognition performance.

B. Place Categorization

[6]

Emphasize the generalization capability of categorization over place rec - Maybe echo earlier.

I dont have any background in this. If you have a particular message you want to get across here, let me know and I can try to write something.

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 [3] pre-trained on Places365 database [4] for labeling reference and query database with most probable scene attributes [5]. We use SeqSLAM [2] for showing improved place recognition using semantic information by appropriately enhancing the image matching scores.

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 [5]) 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 predicted scene attributes for some of the images from datasets used in this paper are shown in Fig. 2. 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 local temporal consistency. 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 component $c_k = \{N_{c_k}, N_{c_{k+1}}\}$ is obtained whenever an edge between consecutive nodes ceases to exist, where N_{c_k} and $N_{c_{k+1}}$ are the nodes marking the beginning and end of the connected component. The most probable prediction label L_i of each node in a connected component is used to find the most frequently occurring label L'_i within that component.

$$L'_i = mode(\biguplus_{i=a}^{b} L_i) \qquad \forall k \in [1, m] \tag{2}$$

where m is the total number of connected components and $mode(X)$ for a set X gives its statistical mode.

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 and merge some of the consecutive components. We use a sliding window of size s that passes through the entire image sequence and replaces the label L'_i by the mode of labels within the sliding window as given in Eq. 3.

$$\hat{L}_i = mode(\biguplus_{i-s/2}^{i+s/2} L'_i) \tag{3}$$

For sake of clearance, it can be noted that labels L'_i are obtained by looking at the labels of individual nodes of a connected component and are then assigned to that component and its participating nodes, whereas, the labels $\hat{L}_i$ are obtained by looking at the individual nodes of the entire database and then assigned to that particular node itself where the sliding window is centered. The purpose of the former is intra-component label consistency while the latter is for inter-component label consistency.

The filtered image labels $\hat{L}_i$ are finally used to segment the database into M chunks represented as $C_K = \{N_{C_K}, N_{C_{K+1}}\}$, where N_{C_K} and $N_{C_{K+1}}$ are the nodes that mark the beginning and end of the data chunk. The label corresponding to each chunk is same as the label for all its nodes and is represented as $\hat{L}_{C_K}$. 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 Fig. 2. Fig. 3 shows the images and their semantic labels at the segmentation transition points for one of the datasets used in this paper.

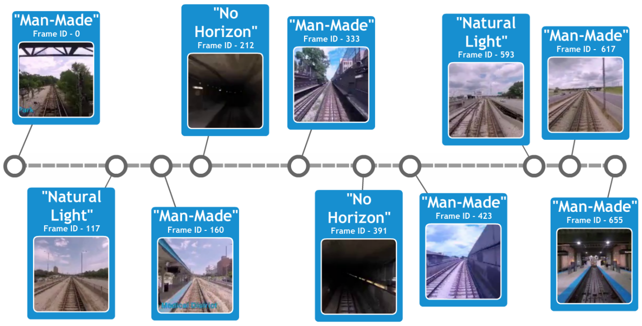


Fig. 3. The time-line of CTA-Rail dataset with semantically labeled images at the transition points of segmented reference database. (Time-line created using [7])

B. Place Recognition

In general, a place recognition system comprises of a pre-processing stage, then a method to calculate affinity scores between database places and the query, and finally a decision module for generating the best matching pairs. This is also depicted in Fig. 1.

1) Sequence-based place matching: In addition to the above mentioned place recognition pipeline, a sequence-based recognition method exploits the temporal information inherent in this problem. Therefore, searching for a matching sequence of places is a better approach than deciding a match based only on single matching template from reference imagery. SeqSLAM [2] is a sequence-based place recognition method developed on similar principle. Moreover, it is known to work remarkably well in challenging environmental conditions and is able to recognize places despite seasonal, weather or time of day variations. The recent advanced methods [8], [9], [10] etc. inspired from SeqSLAM further improve the state-of-the-art for place recognition. In this paper, we use the vanilla approach to show the performance improvement of a place recognition system, under the influence of variations in the surrounding environment, with the help of semantic information associated with those places. The detailed methodology of SeqSLAM can be referred to in [2].

SeqSLAM performs place recognition using Sum of Absolute Difference (SAD) scores represented as D between preprocessed reference and query images. The preprocessing step involves down-sampling of image to size S_x and S_y and patch normalizing it with a fixed square window of side length P.

$$D_i = \frac{1}{S_x S_y} \sum_{x=0}^{S_x} \sum_{y=0}^{S_y} |p^j_{x,y} - p^i_{x,y}| \tag{4}$$

where $p^i_{x,y}$ and $p^j_{x,y}$ are the pixel intensities of patch normalized reference and query images.

The difference vector so obtained for each query image undergoes neighborhood normalization within a sliding window of size R, also termed as neighborhood normalization zone width. The neighborhood normalized difference for a given query image, $\hat{D_i}^R$ is calculated using the local mean difference $\bar{D_i}^R$ and local standard deviation σ_i^R.

$$\hat{D_i}^R = \frac{D_i - \bar{D_i}^R}{\sigma_i^R} \tag{5}$$

The neighborhood normalized SAD matrix is then searched for local image sequence trajectories of length d_s, within a limited range of velocities, originating from each of the reference image. The sequence trajectory with best score is then selected using a trajectory uniqueness threshold μ.

2) Implicit dataset segmentation: The neighborhood normalization of place matching scores within the window R, as calculated in Eq. 4 and 5, reflects the emphasis on matching a local physical region of the environment. The parameter R represents the span of environment, where the matching scores are locally enhanced. Our aim is to pre-define these physical regions of the environment for effectively matching the places that share similar semantic labels. The subsequent section shows how this can be achieved using the place recognition method chosen above.

3) Semantically-informed place matching: The segmentation of dataset as described in earlier section using the consistency of semantic labels, is a way of separating the physical space into regions with similar environmental conditions. As shown in Fig. 1, in general, a place recognition system can use the semantic information from the place categorization module to enhance its affinity scores for matching places. We show the effect of this by implementing it for the place recognition method - SeqSLAM as described below.

The first step towards exploiting the segmented regions of the reference database and semantic labeling is to find the region candidates that have the same label as the query image. This is achieved by simply matching the labels of the segmented regions, $\hat{L}_{C_K}$ and the query image label $\hat{L}_j$. For a given query image N_j, these region candidates are used to define the neighborhood ranges in the reference database for normalization purpose:

$$\mathbf{R}'_i = \{\{N_{C_K}, N_{C_{K+1}}\} \mid \hat{L}_{C_K} = \hat{L}_j, i \in \{N_{C_K}, N_{C_{K+1}}\}\} \forall i, K \tag{6}$$

where i iterates over all the reference images, K iterates over all the segmented regions, and $\mathbf{R}'_i$ is the set of pairs of nodes that define the range for neighborhood normalization. If $\mathbf{R}'_i$ happens to be a null set, then the vanilla method is used, otherwise:

$$\hat{D_i}^{R'} = \frac{D_i - \bar{D_i}^{R'}}{\sigma_i^{R'}} \;\; \forall \;\; R' \in \mathbf{R}'_i \tag{7}$$

The incorporation of segmented regions based neighborhood normalization as shown above, makes the appropriate use of semantic information by handling the different physical environments separately.

This needs a picture, something as simple as 2 confusion matrices or difference vectors, one vanilla seqslam and one weighted version, with some labels.

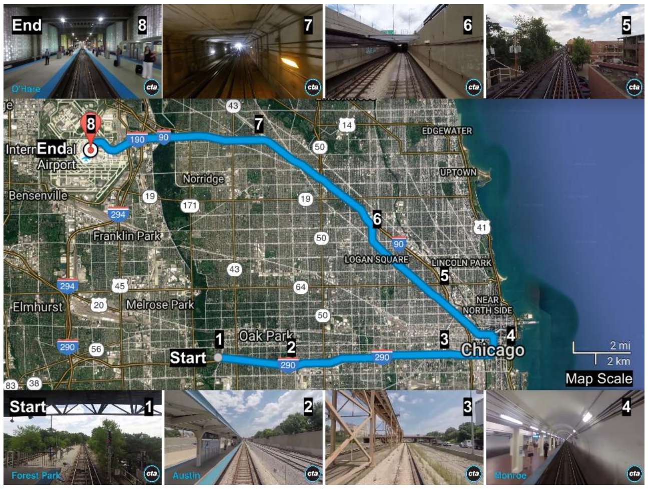


Fig. 4. CTA Dataset Trajectory Aerial View with sample images. (Marked Trajectory Source - [11])

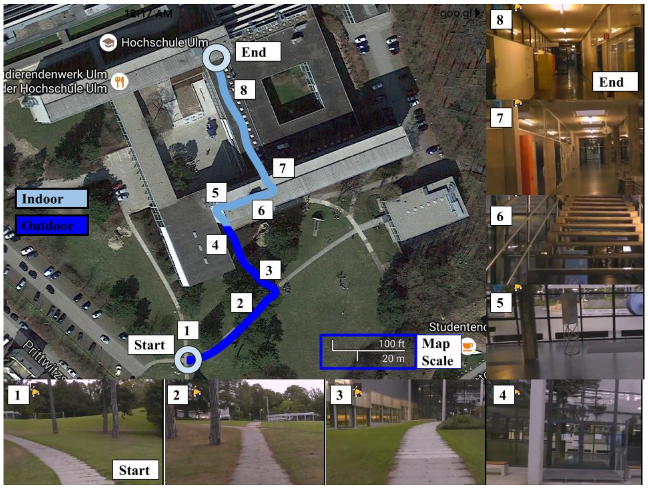


Fig. 5. Campus Indoor-Outdoor Dataset Trajectory Aerial View with sample images. (Marked Trajectory Source - [11])

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 image classification part is done as an off-line preprocessing step for reference as well as query database.

A. Datasets

The two datasets used in the experiments exhibit variations in environmental conditions as the route is traversed.

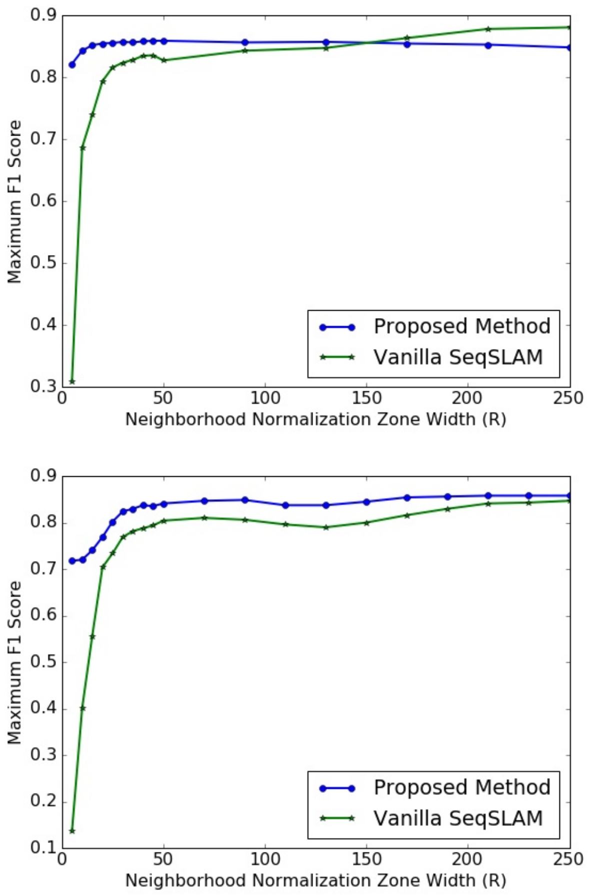


Fig. 6. Performance comparison of proposed method and vanilla SeqSLAM with respect to parameter R for CTA (top) and Campus (bottom) datasets.

1) CTA-Rail: The CTA-Rail (Chicago Transit Authority) dataset (Fig. 4) comprises of two videos traversing a 23 km railway route (Blue Line, Forest Park to O'Hare), recorded once in 2014 [12] and then in 2015 [13], available online. The camera is placed at the head of the train facing forward the railway track. The videos comprise of scenes from train stations platforms, subway station platforms, subway tunnels, and railroad tracks within highways and urban areas. The raw videos are approximately 73 and 84 minutes in duration with 132670 and 149090 frames respectively. We used the 480p version of the video and processed every 200th frame for all the experiments. The resultant reference and query databases have therefore 656 and 738 image frames respectively.

2) Campus Indoor-Outdoor: The Campus Indoor-Outdoor dataset comprises of two videos with repeated traversal of a part of Ulm University of Applied Sciences' campus from outside lawn to inside corridor [14], [15]. The videos have been recorded using a hand-held device and exhibits 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 datasets. The datasets comprise of scenes from outside the campus, with trees, grass and pavement, 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.

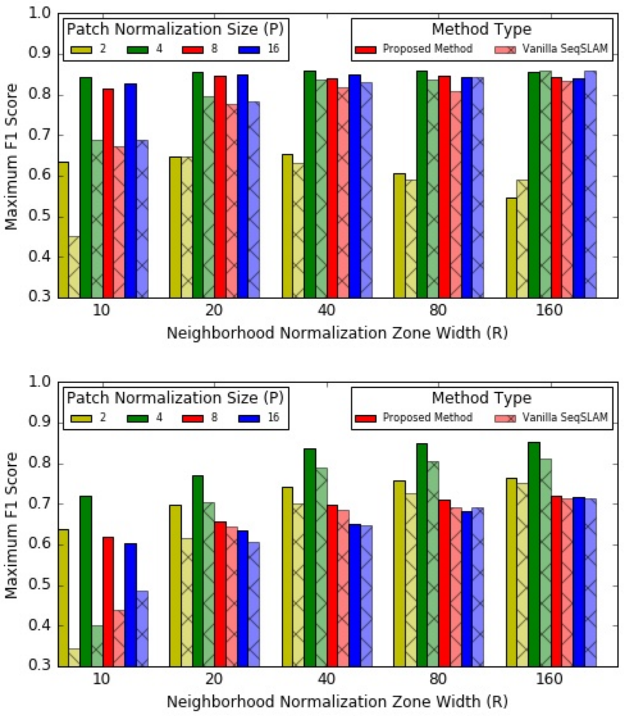


Fig. 8. Performance chart showing maximum F1 Score. The first row corresponds to the CTA-Rail dataset and second row to the Campus Indoor-Outdoor dataset. The performance improves with increasing the parameter value R, where vanilla SeqSLAM performs equally good as the proposed approach. The patch normalization Size (P) parameter with value 4 happens to perform better as compared to others.

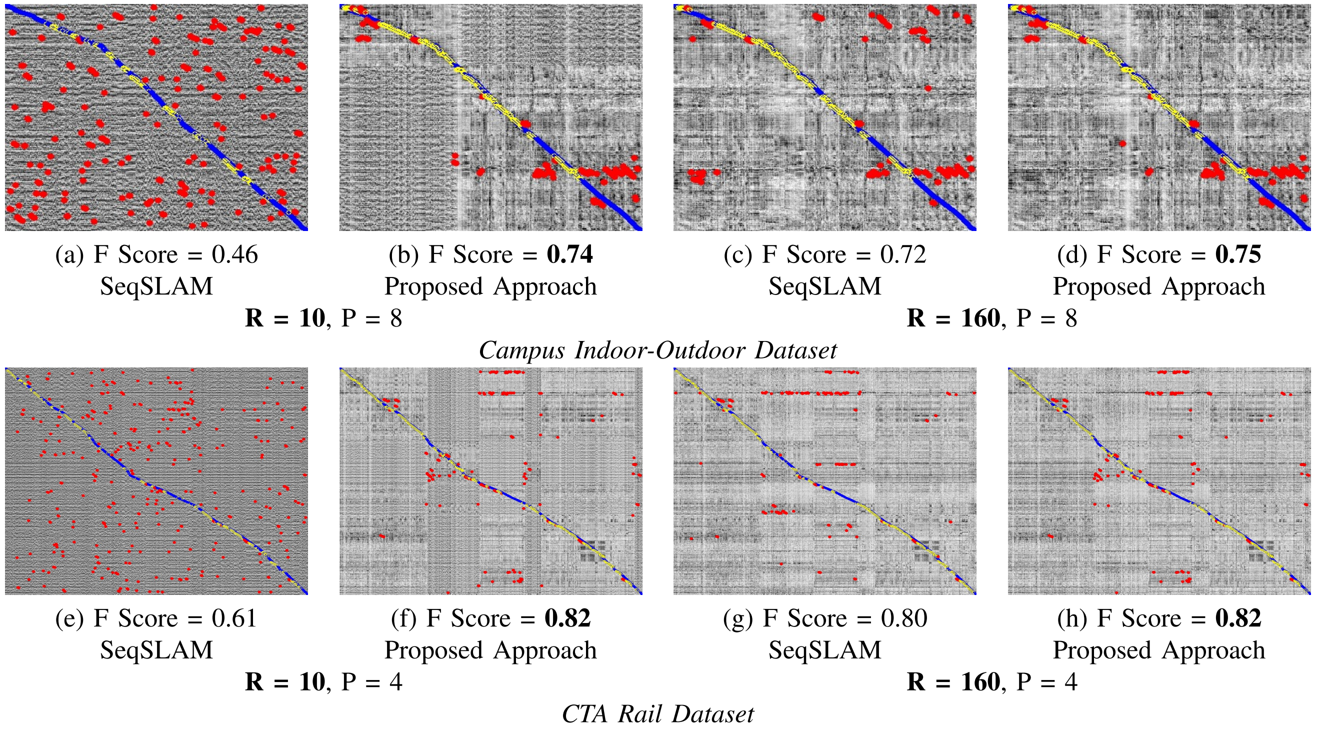
3) Ground Truth: The place recognition ground truth for both the datasets was generated manually for intermittent frames and then interpolated for rest of the image sequence. A query image is considered to be a true positive match for the reference image if its index lies within a range of 5 image frames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

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

V. Results

We used maximum F1 Score to measure place recognition performance of SeqSLAM by varying the trajectory uniqueness parameter (described in [2]), that is, the threshold for deciding a correctly matched place. The comparative results were generated between the proposed method and vanilla SeqSLAM for two different datasets. We used two parameters of SeqSLAM method that are known to affect its performance, to measure the trends in performance change. The performance improvement is as shown in Figure 8 and the effect of parameters is discussed in subsequent section.

Figure 7 shows the ground truth and the place recognition matches (without thresholding) corresponding to different parameter settings for both vanilla SeqSLAM and proposed



(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. 7. 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 significant for smaller values of R, whereas, with larger R values, performance of vanilla SeqSLAM method approaches to that of proposed method.

method. The performance improvement is large for both the datasets for smaller values of R (neighborhood normalization zone width). The large R values help the vanilla method to attain performance equivalent to the proposed approach as shown in Figure 7. The effect of dataset segmentation can be easily seen and understood in the these images in form of rectangular dark and light patches in the SAD matrix.

VI. Discussion and Future Work

In this work, we presented a system which combines place categorization information to inform and improve place recognition results. The system was tested by using two real world datasets, highlighting the proposed system's superiority over a state-of-the-art place recognition system. Here, we discuss the effects of system parameters, the improvements achieved by the proposed approach and potential extensions and future work for this approach.

A. Effect of SeqSLAM parameters

1) Neighborhood Normalization Zone Width (R): The neighborhood normalization parameter R is used to set the window size for locally enhancing the match scores. As shown in Figure 8, performance of vanilla SeqSLAM gets better with an increase in R. This can be expected because normalizing over a larger image sequence balances the overall variation in scores for an environment with varying conditions, but it leads to suppressing of correct matching pairs corresponding to images having low matching scores, especially in the smaller zones. The same argument is also applicable to smaller zone width at transition points in the environment. On the other hand, using segmented regions to set the normalization zone width, effectively highlights the matching scores in appropriate regions and generates more true positives. The performance is, therefore, most of the times better than the best achieved using vanilla approach with any value of R. Moreover, wider neighborhood zones for normalization, are not appropriate for large datasets and long time navigation due to computational burden.

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. In the experiments performed for current work, we found that patch normalization window size of 4 for the given down-sampling image size performs better.

B. Physical Space Segmentation

Fig. 6 shows a performance comparison for both the datasets with respect to the parameter R alone. The curves here give us an insight to understand how an optimal value for segmenting the physical space can be chosen. For both the datasets, the performance of vanilla SeqSLAM and the proposed approach becomes almost similar at a certain point, beyond which no significant improvement occurs. This happens when R is approximately between 20 and 25, which is almost 1 km of journey in the CTA Rail dataset captured at speed of train and about 50 m for Campus dataset captured at normal human speed. The visual data captured in both the cases is therefore sufficient enough to correctly segment its physical space. If the physical region spanned were to be smaller than this, it would have created an inter-region redundancy of visual data which would mean more confusing matching places. Hence, the optimal value for segmenting the environment would be based on the average rate of persistence of particular environmental conditions. However, it would still fail for the cases where there is a large variance in the span of different conditions existing within the environment, and a proper segmented environment based on semantic information will be the key to perform better.

TABLE I
SeqSLAM parameters.

$S_x\mathbf{x}S_y$	Image Down-sampling Size	32x32
P	Patch Normalization Window Size	2,4,8,16
O	Image Matching Offset Range	±10
d_s	Sequence Length	15
R	Neighborhood Normalization Zone Width	10,20,40,80,160
V	Sequence Search Velocity Range	$(1 \pm 0.2)*d_s$
μ	Trajectory Uniqueness Threshold	Varied

C. Future Work

Investigating spatio-temporal relationship between places with similar visual appearances, dynamically varying the sequence length to improve localization results?

Grid Cells?

3D object proposals to further inform place rec?

Others?

References

[1] M. Cummins and P. Newman, "Highly scalable appearance-only SLAM - FAB-MAP 2.0," in *Robotics: Science and Systems*, Seattle, United States, 2009.

[2] 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.

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

[4] 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.

[5] 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.

[6] N. Sunderhauf, F. Dayoub, S. McMahon, B. Talbot, R. Schulz, P. Corke, G. Wyeth, B. Upcroft, and M. Milford, "Place categorization and semantic mapping on a mobile robot," in *Proceedings of the International Conference on Robotics and Automation*. IEEE, 2016.

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

[8] 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.

[9] 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.

[10] 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.

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

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

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

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