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

Sourav Garg[1], Adam Jacobson[1], Swagat Kumar[2] and Michael Milford[1]

***Abstract*— The problem of place recognition actually comprises two distinct subproblems; "traditional" 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 traditional 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 [1], 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 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 [2] and during difficult day and night traverses of an environment [3]. 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 categorization as an extension to the place recognition problem, where it is possible to use similar frameworks to solve both problems. We highlight the main differences 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 and are expected to be able to generalize

[1]The authors are with Australian Centre for Robotic Vision, Queensland University of Technology, 2 George St, Brisbane, Australia

[2]The author is with Innovation Labs, Tata Consultancy Services, New Delhi, India

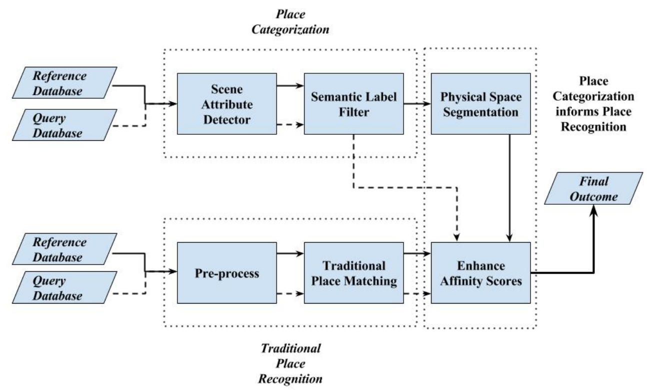


Fig. 1. A block diagram showing the flow of semantic information from the place categorization module to the place recognition module for better performance.

class labels to classify and extend to unseen environments. The place categorization framework typically utilizes 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.

The place recognition problem has historically been solved for ideal environmental conditions [4],[5],[2],[6] until recently with a significant interest towards all conditions place recognition [3], [7], like all weather [8], [9], all season [10], [11] and all times-of-day [12],[13],[14] for long-term localization and mapping. This has also led to a creation of massive 1000 km dataset [15] with repeated traversals over a time period of one year, therefore exhibiting variety of conditional and structural changes in the environment across the traversals.

The variations in environmental conditions for visual place recognition have been so far explored only for global changes in conditions across different traversals of the same route. For example, a single route traversal at night exhibits only night conditions throughout the traverse; similarly, route traversed during sunny weather doesn't exhibit other weather conditions in the same traverse. In many real world situations, a lots of local changes in environmental conditions occur, especially for long-term and seamless navigation systems. Such local changes can be either be moderate like moving from an indoor environment to an outdoor one, or these changes can be extreme like transiting from an artificially-lit underground car-parking to an open outdoor road-network at night; or entering a house during daytime and navigating path in dark before turning on the lights.

In this paper, we look at the effects of local condition variations in the environment on known condition-invariant place recognition method and present methodology to deal with it by leveraging semantic place categorization in order to effectively recognize places despite local and/or global changes in conditions. Our primary contribution is the development of a novel framework to incorporate semantic labels and place categorization results to inform and improve place recognition place estimates (as seen in Fig.1). Once a place is categorized, we leverage the SeqSLAM [3] framework to perform place recognition, implementing a novel dynamic weighting scheme, biasing place matches with similar place characteristics 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 indoor and outdoor campus environments and the CTA-Rail Dataset consists of a single camera mounted on a train traversing 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 the 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 the implementation of our CNN place categorization framework, outlines our place recognition framework and the proposed technique for combining the two frameworks to produce superior place recognition results. We present the experimental setup in Section 4, and results of multiple levels of evaluation in Section 5. Section 6 discusses the significance of the research and 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, with 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 [2], [16].

There have been many attempts to improve performance of place recognition systems. This has included the inclusion of temporal information, fusing multiple sensory modalities and implementing unique preprocessing steps to improve localization capabilities like shadow removal techniques.

Temporal information has been incorporated into the place recognition framework with the introduction of the SeqSLAM framework [3], integrating place hypotheses over small distances to accrue evidence and improve place recognition performance.

Multi-sensor fusion has been investigated in a number of works [17], [18], attempting to introduce unique sensory modalities which have different failure modes to produce robust place estimates.

Furthermore, the introduction of unique sensor preprocessing techniques to improve sensor data for place recognition has also been explored. Frameworks utilizing techniques for shadow removal [19] or the introduction of illumination invariant color spaces [20] to remove temporal or environmental changes from images to improve localization.

The work presented in place categorization attempts to develop the generic capability to identify types of places in the world, potentially enabling improvements in place recognition capabilities.

B. Place Categorization

Place categorization systems are an extension of the place recognition problem and attempt to attach semantic meaning to particular places in an environment; attempting to utilize labels from a training set like indoor, outdoors, kitchen, office and bedroom to categorize the location within which an image was taken. These frameworks are powerful as they facilitate generalization of room labels to different environments, for example identifying a bedroom within an unexplored house, potentially enabling robotic platforms to perform generic tasks in unknown environments by recognizing the type of place [21].

There have been a number of works which attempt to imbue traditional SLAM architectures with the ability to semantically label locations in an environment[1], [22]. These types of frameworks utilize the place categorization labels to provide information about a space, for example a location is mostly likely a kitchen, but these labels are not utilized in the process of generating the map or performing place recognition or localization.

In a recent work [23], authors develop a method to generate different categories of environments from a large available reference database for place recognition in order to reduce the search space for matching places. They basically segment the overall physical space into categories of similar environments within the place recognition system and do not use any semantic place categorization.

Place categorization systems have been leveraged in previous work to improve object detection and classification, enabling reduction of the object search space and improvement in object recognition performance [24].

However, there has been no prior work utilizing place categorization information to improve place recognition place estimates.

III. Proposed Method

The proposed approach has three main components: place categorization, physical space segmentation and place recognition as depicted in Fig. 1, with semantic information flowing from the former to the latter to generate the final place match estimate. Our core contribution is the development of a technique to utilize place categorization information to improve place recognition performance. In order to achieve this, we use the semantic labels to divide the physical space into different regions based on its appearance, that is, semantic scene attributes. These segmented regions are then used in the place recognition module for biasing place matches that lie in a paritcular semantic region. We use CNN model VGG16-places365 [25] pre-trained on the Places365 database [26] for labeling reference and query database frames with most probable scene attributes [27]. We use SeqSLAM [3] for showing improved place recognition using semantic information by appropriately enhancing the image matching scores.

A. Place Categorization

The pre-trained CNN model classifies an image with probabilities associated with each of the 365 place categories. It is also made to predict the most probable scene attributes (out of 102 attributes trained on the SUN database [27]) using one of its fully-connected layers ('fc7'). We use these scene attributes and their associated probabilities to post-process the image labels for semantic segmentation of 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 only on the reference database. The semantic labels as obtained are used to temporally divide the reference image sequence into different segments.

B. Physical Space Segmentation

The place categorization module provides semantic labels for each reference image ranked according to the probabilities associated with those labels. In order to segment the reference database, a unique label corresponding to each image is required while taking into consideration the temporal nature of the input and avoiding any transient errors produced by place categorization module. This is achieved using a Hidden Markov Model (HMM) ((**cite**), where we estimate the model parameters and the hidden state corresponding to each image in the reference image sequence using the semantic label probabilities as the observed variable for that image.

The sequence of semantic labels feature vector corresponding to the reference database having T number of images is represented as a random variable $X = (x_1, x_2 \dots x_T)$ and the hidden variables are denoted as a random variable $Z = (z_1, z_2 \dots z_T)$, where z_t at a given time instant t can belong to one of the N hidden states. It is assumed that given the z_{t-1}, z_t is independent of previous hidden variables and current observation x_t only depends on current hidden state z_t. Hence, the state transition probability matrix, represented as A and initial state distribution π_i is given as:

$$A = \{a_{ij}\} = p(z_t = j | z_{t-1} = i) \qquad \forall i, j \in [1, N] \tag{1}$$

$$\pi_i = p(z_1 = i) \tag{2}$$

The probability of an observation at time t for being in state i is defined as:

$$b_i(x(t)) = p(x(t) | z_t = i) \tag{3}$$

Our objective is to find the hidden state sequence of the model, that is, the desired unique labels for the reference image sequence. This is given by the posterior probability of the state sequence:

$$p(Z|X, \theta) = \frac{p(X, Z|\theta)}{P(X|\theta)} \tag{4}$$

where $\theta = (A, b_i(x(t)), \pi)$ are the parameters of the model and

$$p(X, Z|\theta) = \pi_{z_1} \prod_{t=1}^{T-1} a_{z_t z_{t+1}} \prod_{t=1}^{T} b_{z_t}(x(t)) \tag{5}$$

$$P(X|\theta) = \sum_Z p(X, Z|\theta) \tag{6}$$

The final labels for the reference images, represented as L_t, are obtained after estimating the parameters θ of the model:

$$L_t = \arg\max_i Z_t(i) \qquad \forall i \in [1, N] \tag{7}$$

The input feature vector, that is, the observation $x(t)$, is the output response vector of the place categorization module with size 1x102, where 102 dimensions represent the probability associated with each of the scene attributes. The feature vector is normalized to the range $[0, 1]$ before feeding into the HMM. The parameters θ of the model are determined using Baum-Welch algorithm (**cite**) and the most likely hidden state sequence is obtained. The implementation of HMM used for this work is available here (**cite**). The number of hidden states, that is, N is empirically determined for the datasets used in the paper, though there are ways to determine N using cross-validation (**cite**) or by using Infinite HMM (**cite**), and is not the focus of our work.Fig. 3 shows the images and their semantic labels at the segmentation transition points for one of the datasets used in this paper.

C. Place Recognition

In general, a place recognition system comprises of a preprocessing 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, as seen in Fig. 1.

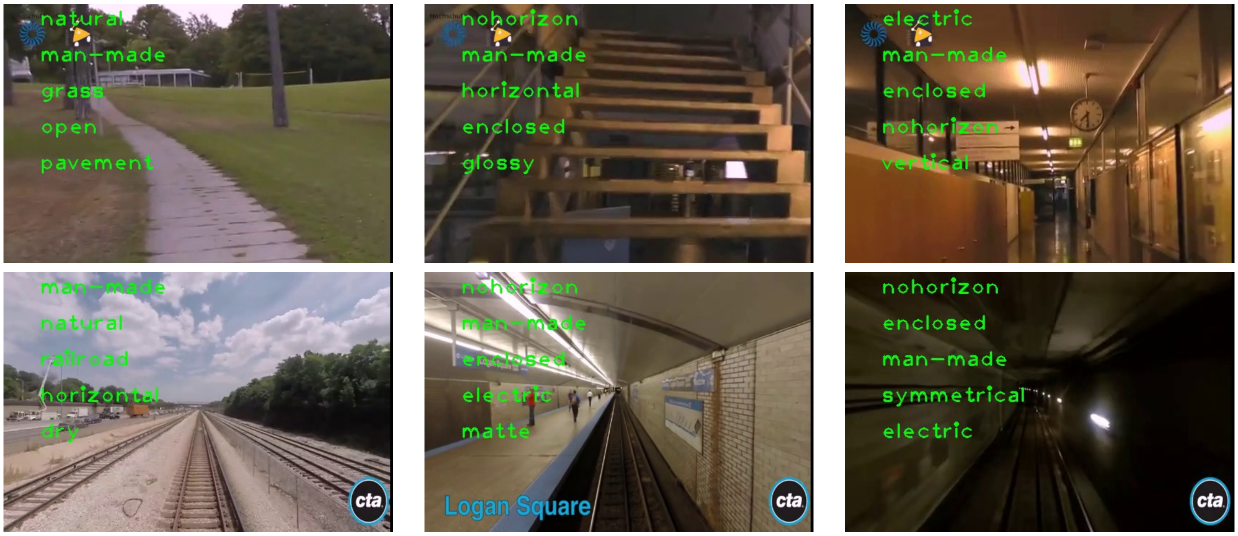


Fig. 2. Images from reference database with top-5 most probable semantic labels out of 102 scene attributes for Campus Dataset (top) and CTA Rail Dataset (bottom).

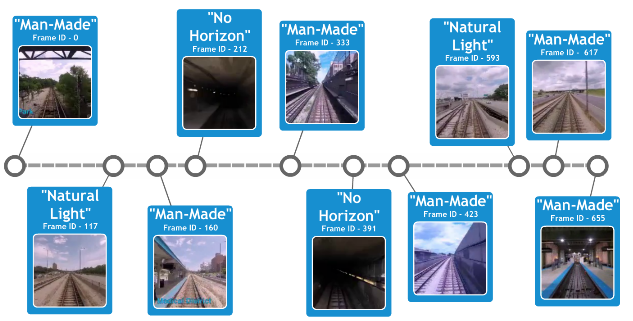


Fig. 3. The time-line of CTA-Rail dataset with semantically labeled images at the transition points of segmented reference database. (Time-line created using [28])

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 [3] 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 [29], [30], [31] 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 [3].

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.

$$D_i = \frac{1}{S_x S_y} \sum_{x=0}^{S_x} \sum_{y=0}^{S_y} |p^j_{x,y} - p^i_{x,y}| \tag{8}$$

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

The difference vector 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.

$$\hat{D_i}^R = \frac{D_i - \bar{D_i}^R}{\sigma_i^R} \tag{9}$$

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 the best score is then selected using a trajectory uniqueness threshold μ.

2) Localized and semantically-informed matching: The neighborhood normalization of place matching scores within the window R, as calculated in Eq. 8 and 9, reflects the emphasis on matching a local physical region of the environment, instead of finding a global minima. In general, finding a global match is prone to false noisy matching and doesn't take into consideration the temporal nature of reference image database, which can help in identifying similar patterns of matching scores in any local region of the database. Hence, the parameter R represents the span of environment, where the matching scores can be locally

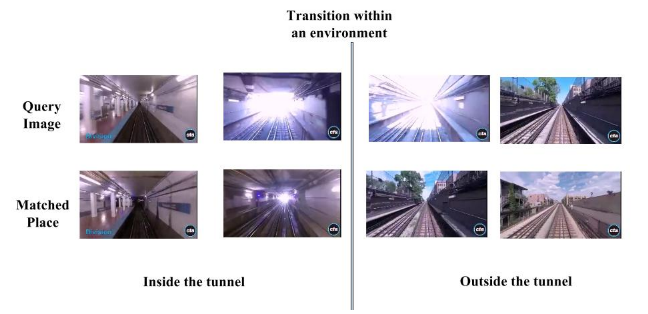


Fig. 4. Matched places at the point of transition within an environment for CTA-Rail dataset where environment changes from being inside the tunnel to outside the tunnel.

enhanced. This helps in preventing the dissimilar images from unnecessarily deviating the mean and increasing the variance of the matching scores in these regions, which otherwise beats the purpose of finding a local match. Our aim is to pre-define these physical regions of the environment that share similar semantic labels.

The segmentation of the dataset as described in earlier sections using HMM, 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. Instead of arbitrarily choosing the neighborhood for the reference image as in Vanilla SeqSLAM method, we propose to use the neighborhood regions obtained using labels L_t generated using HMM. The segmented regions are denoted as a set of pairs R'_t:

$$R'_t = \{(i,j) \mid t \in [i,j) \text{ and } L_k = L_{k+1} \quad \forall k \in [i,j) \\ \forall i,j \in [1,T]\} \tag{10}$$

where t iterates over all the reference images and a pair in R'_t defines the lower and upper limit of the segmented region within the reference database. The neighborhood normalization equation (9) is therefore updated as below:

$$\hat{D_i}^{R'_i} = \frac{D_i - \bar{D_i}^{R'_i}}{\sigma_i^{R'_i}} \tag{11}$$

Fig. 5 shows the method described in this section for choosing R. The incorporation of segmented regions based neighborhood normalization as shown above, makes sure that different environmental conditions encountered within a traversal are handled separately for finding a local best match.

IV. Experimental Setup

The experiments are performed using four different datasets described in following subsections. The image classification for place categorization is performed off-line as a preprocessing step and all the experiments are conducted using Dell Latitude E7450 Intel Core i7-5600 CPU @ 2.6 GHz x 4 processor having 16 GB RAM and running Ubuntu 14.04.

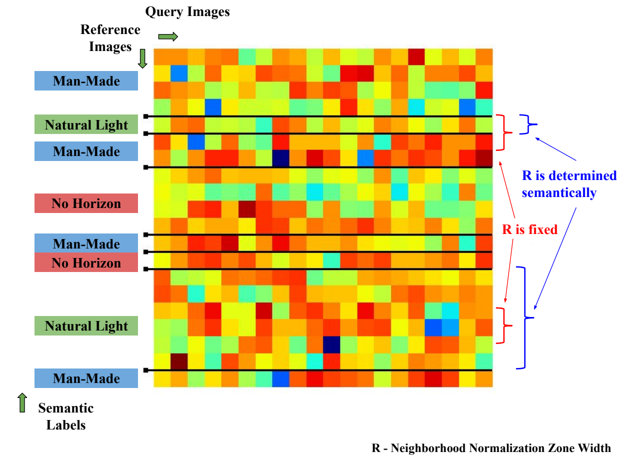


Fig. 5. The semantic segmentation of environment decides the normalization regions for better place recognition. The matrix represents the Sum of Absolute Difference score between reference and query images of CTA-Rail dataset. The black horizontal lines mark the transitions from one type of environment to the other. The red markers on the right show the fixed neighborhood normalization zone width (R) for SeqSLAM and blue markers refer to the proposed method for determining R.

Fig. 6. CTA Dataset Trajectory Aerial View with sample images. The path is traversed along a fixed rail route partially underground and partially through outdoor rail-tracks. (Marked Trajectory Source - [32])

A. Datasets

The four datasets used in the experiments are practical application scenarios for visual SLAM and seamless navigation within different types of environments exhibiting variations in environmental conditions *within* and/or *across* the traversals. These four different datasets have diversity with respect to the (a) size - varying from small campus traversals to large rail routes; (b) condition variations within the traversal - medium variations like transiting from open-space to enclosed ones, and extreme variations by making such transitions from bright to dark space or vice-versa; (c) condition variations across the traversals - from none to day and night; (d) amount of motion - varying from pedestrian motion, to motorbike and to trains; (e) viewpoint - from slight variations due to gait and jerks on bike, to deliberate lateral offsets along the footpath.

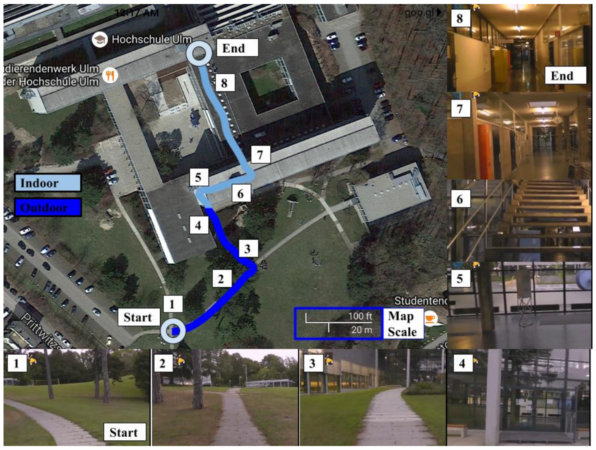


Fig. 7. Campus Indoor-Outdoor Dataset Trajectory Aerial View with sample images. The path is traversed from outside of the campus and then into the campus with both reference and query datasets belonging to daytime. (Source - Google Maps)

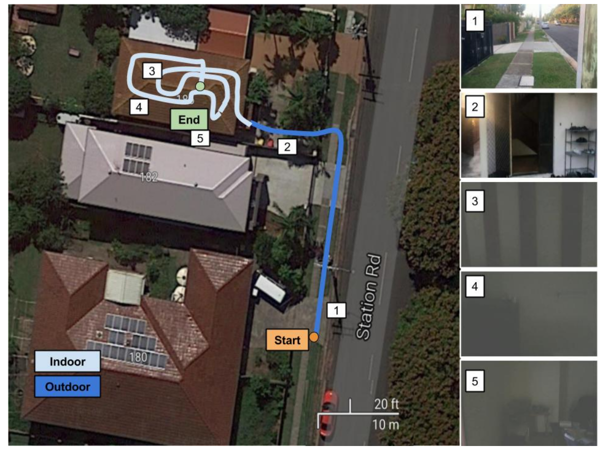


Fig. 9. Residence Indoor-Outdoor Dataset Trajectory Aerial View with sample reference images. The path is traversed from outside the house to indoor hall (ground floor) and to the bedroom (first floor). (Source - Google Maps)

Fig. 8. Parking Lot Dataset Trajectory Aerial View with sample query images. The path is traversed from the underground parking lot to outdoor area and then back to underground parking. The reference data was collected in day and query data was collected at night. (Source - Google Maps)

1) CTA-Rail: The CTA-Rail (Chicago Transit Authority) dataset (Fig. 6) comprises of two videos traversing a 23 km railway route (Blue Line, Forest Park to O'Hare), recorded once in 2014 [33] and then in 2015 [34], available online. A single camera is placed at the head of the train facing forward towards 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 an outside lawn to an inside corridor [35], [36]. The videos have been recorded using a hand-held device with single camera and exhibit 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.

3) Parking Lot: The Parking Lot dataset is captured inside a society of residential buildings, traversing through the underground as well as open parking area. It comprises of two videos traversing the same path, once during daytime and then at night. The dataset exhibit transition from underground parking area having artificial lighting to open naturally lit space during day time, and to the dark sky during night time with some street lights. This dataset hence possess variations in environmental conditions within and across the route traversals. The videos have been captured using hand-held mobile device while driving on a motor-bike and contain 6407 and 6396 image frames respectively for day and night traversal. We process every 20th frame, therefore processing 320 frames for each of reference and query database.

4) Residence Indoor-Outdoor: The Residence Indoor-Outdoor dataset comprises of two traversals from outside of a house and then entering inside the house via corridors to the common area and then to the bedroom via stairs. The reference database was captured during daytime with good natural lighting outside the house and with minimum lighting inside the house. The query database was captured at night with street lights lighting the way outside of house and adequate lighting inside the house. Therefore, in this dataset as well, there are variations in the environmental conditions within and across the traversals. Moreover, the path traversed outside the house also exhibits a change in viewpoint in the two traversals, due to lateral offset of around 1m while walking down the pavement. The videos are captured using hand-held camera. The reference and query database comprise of 2200 and 2180 image frames respectively and are processed by skipping 10 frames, therefore comparing 220 and 218 image frames.

B. Ground Truth

The place recognition ground truth for all the datasets was generated manually for intermittent frames and then interpolated for the 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.

C. SeqSLAM parameters

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

TABLE I
SEQSLAM PARAMETERS.

S_xxS_y	Image Down-sampling Size	64x32
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	Varies from 5 to 1280 (**sure?**)
V	Sequence Search Velocity Range	$(1 \pm 0.2)d_s$
μ	Trajectory Uniqueness Threshold	Varied

V. RESULTS AND DISCUSSION

We used maximum F1 score to measure changes in place recognition performance using the proposed approach. The trajectory uniqueness parameter (described in [3]), that is, the threshold for deciding a correctly matched place was varied to calculate precision-recall curve and maximum F1 score. The comparative results were generated between the proposed method and vanilla SeqSLAM for four real world datasets. In order to gain an in-depth understanding of the place recognition performance changes due to proposed approach, we used two parameters of SeqSLAM method, patch normalization window size (P) and neighborhood normalization zone width (R), to measure the trends in performance change. The results are as shown in Fig.10, and the significance and effect of these parameters is discussed in subsequent section.

In this work, we presented a framework which combines place categorization information to inform and improve place recognition results. The system was tested by using four 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 and the improvements achieved by the proposed approach.

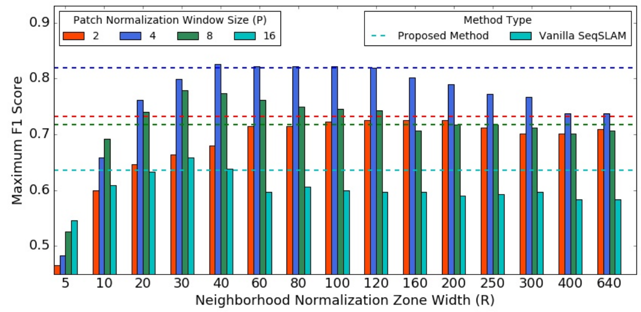
A. Neighborhood Normalization Zone Width (R)

The neighborhood normalization zone width R defines a temporal region around the reference image in order to find a local best match for the query image. As shown in Fig. 10, the proposed approach outperforms the vanilla method by adequately segmenting the reference image database and selecting the right temporal region for enhancing the place matching scores. The regions R' being determined using semantics are independent of the parameter R, hence the performance measure is always constant w.r.t. it.

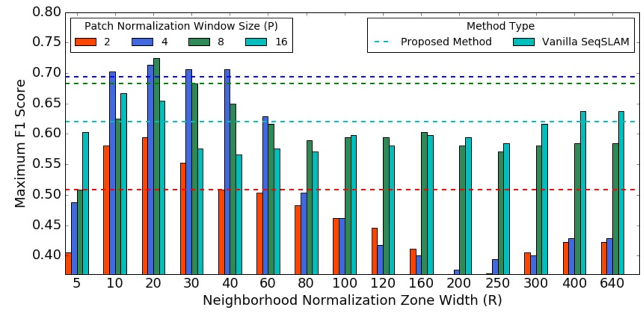
Our proposed method performs consistently better than the vanilla approach for smaller values of R for all the datasets. This is an expected outcome as a smaller temporal window around the reference image means spanning across very similar images, wherein normalization of matching scores creates local maximas and minimas within that region. As this process is repeated for all the reference images, the local extremas being very similar to each other cause inter-region redundancy and therefore add confusion, leading to false matches. Ideally, a temporal region around the reference image should be such that it spans across non-overlapping dissimilar images in the environment in order to correctly recognize a local match. This is achieved by our proposed method as it uses the semantically-segmented environment for creating adequate local temporal regions to effectively highlight the true matches.

It can be noted in the Fig. 10, that the performance of vanilla SeqSLAM for CTA-Rail and Campus Indoor-Outdoor dataset gets better with increasing the parameter R, but for the Parking Lot and Residence Indoor-Outdoor datasets, it achieves its peak performance and then starts to fall before becoming constant. This happens because of the fact that the former datasets exhibit moderate variations in conditions within the traversal and none across the traversals, whereas the latter exhibit extreme changes in condition both within and across the traversals.

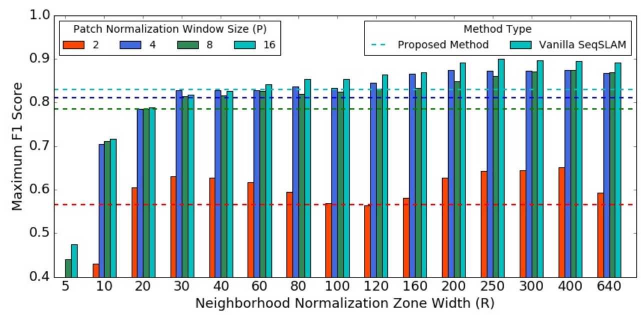
A large normalization zone essentially means finding a global match in the entire reference database which gives rise to false matches as variations in environmental conditions across traversals becomes extreme. For example, in Residence Indoor-Outdoor dataset's reference database, images in the beginning have been captured in broad daylight outdoor setting, which then transits to indoor environment with images captured in darkness of enclosed hall and bedroom. On the other hand, the query database images initially traverse the outdoor environment at night under street light and then transit into indoor areas of the residence brightened by lamps and bulbs. This is shown in Fig. 11, where 11(a) shows the query images, (b) shows the correct matches from the reference database using the proposed method and (c) shows false matches that occur using vanilla SeqSLAM with R value set to its maximum ($R = 640$). The false matches show that the vanilla approach finds a global minima that matches dark with dark and bright with bright rather than realizing that the condition-variant true match of query image lies in the other half of the reference database. This pitfall



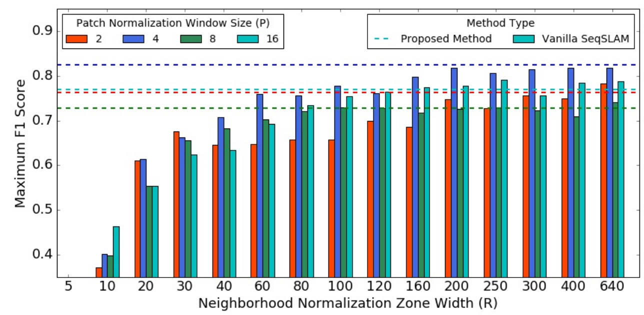
(a) Parking Lot



(b) Residence Indoor-Outdoor



(c) CTA-Rail



(d) Campus Indoor-Outdoor

Fig. 10. Performance charts showing maximum F1 Score w.r.t. different R (Neighborhood Normalization Zone Width) values. R is varied to the maximum value, that is, the size of reference database, after which maximum F1 score becomes constant. The patch normalization window size (P) parameter with value 4 happens to perform better as compared to others most of the times.

is avoided in SeqSLAM by using the local best match in an arbitrarily chosen fixed-size window R. We overcome the same in this paper by understanding the already-visited environment and semantically segmenting it to recognize places despite extreme changes in environmental conditions *within* or *across* the traversals.

B. Patch Normalization Window Size (P)

The images used for finding SAD score are preprocessed by down-sampling them to the size of 64x32 and then patch normalized in order to counter the variations in environmental conditions as proposed in SeqSLAM. Depending on the type of environment and corresponding imagery, the choice of patch normalization window size can lead to changes in performance. It is evident from Fig.10, that the performance trends are similar for different values of P for varying R, but there is no consistency for a value of P which could always give better results than the others. In the experiments performed for current work, we found that for our proposed method, patch normalization window size of 4 for the given down-sampling image size performs better in most of the cases.

C. Viewpoint Variation

The Residence Indoor-Outdoor dataset was captured with lateral offset between the camera position in its reference and query database of approximately $1m$. This was done in order to show that the performance of proposed approach improves upon the baseline approach with whatsoever capabilities the baseline approach has. The vanilla SeqSLAM can handle only limited viewpoint variations and so does the proposed method. The research problem pertaining to developing condition and viewpoint invariant place recognition system has been looked into widely **cite**, and is not the aim of current work.

VI. Conclusion and Future Work

In this paper, we proposed a method to The current work can be extended to incorporate labels of place categories or scene attributes defined at finer levels, as also discussed in previous section. Such fine level place categorization will be directed towards the traditional place recognition problem where each place, despite being from same semantic category, is treated as a separate place. It would also be worth exploring ways to dynamically determine the sequence length for matching places while using the semantic place information.

References

[1] 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.

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

[3] 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.

[4] W. B. Thompson and H. Pick Jr, "Vision-based navigation," in *Proc. ARPA Image Understanding Workshop*. Citeseer, 1993.

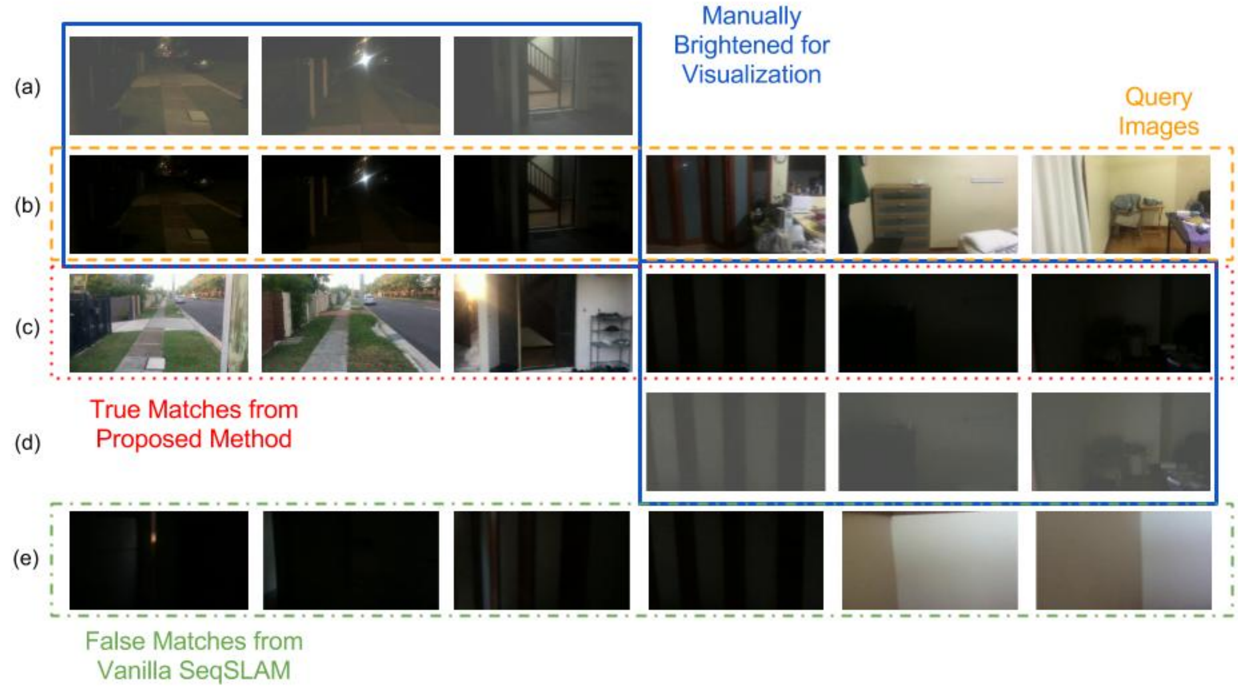


Fig. 11. Images from the Residence Indoor-Outdoor dataset. (a) Manually brightened query images for visualization purpose only. (b) Query images intermittently selected in temporal order show transition from night outdoor to bright indoor environment. (c) Correct matches from the reference database using proposed method. (d) Manually brightened reference images for visualization purpose. e) False matches that occurred using vanilla SeqSLAM.

[5] I. Ulrich and I. Nourbakhsh, "Appearance-based place recognition for topological localization," in *Robotics and Automation, 2000. Proceedings. ICRA'00. IEEE International Conference on*, vol. 2. Ieee, 2000, pp. 1023–1029.

[6] N. Kejriwal, S. Kumar, and T. Shibata, "High performance loop closure detection using bag of word pairs," *Robotics and Autonomous Systems*, vol. 77, pp. 55–65, 2016.

[7] S. Lowry and M. J. Milford, "Supervised and unsupervised linear learning techniques for visual place recognition in changing environments," *IEEE Transactions on Robotics*, vol. 32, no. 3, pp. 600–613, 2016.

[8] C. Linegar, W. Churchill, and P. Newman, "Made to measure: Bespoke landmarks for 24-hour, all-weather localisation with a camera," in *Robotics and Automation (ICRA), 2016 IEEE International Conference on*. IEEE, 2016, pp. 787–794.

[9] E. Pepperell, P. Corke, and M. Milford, "Routed roads: Probabilistic vision-based place recognition for changing conditions, split streets and varied viewpoints," *The International Journal of Robotics Research*, vol. 35, no. 9, pp. 1057–1179, 2016.

[10] C. McManus, B. Upcroft, and P. Newman, "Learning place-dependant features for long-term vision-based localisation," *Autonomous Robots*, vol. 39, no. 3, pp. 363–387, 2015. [Online]. Available: http://dx.doi.org/10.1007/s10514-015-9463-y

[11] Z. Chen, A. Jacobson, N. Sunderhauf, B. Upcroft, L. Liu, C. Shen, I. Reid, and M. Milford, "Deep Learning Features at Scale for Visual Place Recognition," *arXiv preprint arXiv:1701.05105*, 2017.

[12] C. Linegar, W. Churchill, and P. Newman, "Work smart, not hard: Recalling relevant experiences for vast-scale but time-constrained localisation," in *Robotics and Automation (ICRA), 2015 IEEE International Conference on*. IEEE, 2015, pp. 90–97.

[13] W. Maddern, A. D. Stewart, and P. Newman, "LAPS-II: 6-DoF Day and Night Visual Localisation with Prior 3D Structure for Autonomous Road Vehicles."

[14] P. Nelson, W. Churchill, I. Posner, and P. Newman, "From dusk till dawn: Localisation at night using artificial light sources," in *Robotics and Automation (ICRA), 2015 IEEE International Conference on*. IEEE, 2015, pp. 5245–5252.

[15] W. Maddern, G. Pascoe, C. Linegar, and P. Newman, "1 year, 1000 km: The Oxford RobotCar dataset," *The International Journal of Robotics Research*, p. 0278364916679498, 2016.

[16] D. Nister and H. Stewenius, "Scalable recognition with a vocabulary tree," in *Computer vision and pattern recognition, 2006 IEEE computer society conference on*, vol. 2. IEEE, 2006, pp. 2161–2168.

[17] A. Tapus and R. Siegwart, "A cognitive modeling of space using fingerprints of places for mobile robot navigation," in *Proceedings 2006 IEEE International Conference on Robotics and Automation, 2006. ICRA 2006.* IEEE, 2006, pp. 1188–1193.

[18] M. Milford and A. Jacobson, "Brain-inspired Sensor Fusion for Navigating Robots," in *International Conference on Robotics and Automation*. Karlsruhe, Germany: IEEE, 2013.

[19] P. Corke, R. Paul, W. Churchill, and P. Newman, "Dealing with shadows: Capturing intrinsic scene appearance for image-based outdoor localisation," in *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2013, pp. 2085–2092.

[20] C. McManus, W. Churchill, W. Maddern, A. D. Stewart, and P. Newman, "Shady dealings: Robust, long-term visual localisation using illumination invariance," in *2014 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2014, pp. 901–906.

[21] J. Wu, H. I. Christensen, and J. M. Rehg, "Visual place categorization: Problem, dataset, and algorithm," in *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2009, pp. 4763–4770.

[22] A. Ranganathan and J. Lim, "Visual place categorization in maps," in *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2011, pp. 3982–3989.

[23] M. Mohan, D. Gálvez-López, C. Monteleoni, and G. Sibley, "Environment selection and hierarchical place recognition," in *2015 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2015, pp. 5487–5494.

[24] A. Torralba, K. P. Murphy, W. T. Freeman, and M. A. Rubin, "Context-based vision system for place and object recognition," in *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*. IEEE, 2003, pp. 273–280.

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

[26] 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.

[27] 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.

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

[29] 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.

[30] 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.

[31] 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.

[32] 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].

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

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

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

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