Memorable Maps: A Framework for Re-Defining Places in Visual Place Recognition

Mubariz Zaffar, Shoaib Ehsan, *Senior Member, IEEE*, Michael Milford, *Senior Member, IEEE*, and Klaus D. McDonald-Maier, *Senior Member, IEEE*

***Abstract*— This paper presents a cognition-inspired agnostic framework for building a map for Visual Place Recognition. This framework draws inspiration from human-memorability, utilizes the traditional image entropy concept and computes the static content in an image; thereby presenting a tri-folded criteria to assess the 'memorability' of an image for visual place recognition. A dataset namely 'ESSEX3IN1' is created, composed of highly confusing images from indoor, outdoor and natural scenes for analysis. When used in conjunction with state-of-the-art visual place recognition methods, the proposed framework provides significant performance boost to these techniques, as evidenced by results on ESSEX3IN1 and other public datasets.**

***Index Terms*— Visual Place Recognition, memorable maps, ESSEX3IN1, memorability, staticity.**

I. Introduction

VISUAL Place Recognition (VPR) is a well-defined, albeit a highly challenging module of a Visual-SLAM (Simultaneous Localization and Mapping) based autonomous system [1]. It represents the ability of a robot to 'remember' a previously visited place in the world map and thus subsequently generating a belief about the robot's location in the world. VPR can either be used as a stand-alone vehicle localization system in an appearance-only topological and/or topometric map or it can be combined with metric SLAM techniques to perform loop closure [2]. The scope of this work and our evaluations are limited to the former, however, it is possible to adopt the combination of our work and VPR within SLAM systems for loop-closure. Some key advances in SLAM research can be broken down into semantic mapping (surveyed in [3]) and visual place recognition (surveyed in [1]), where the latter can be annexed into the former [3].

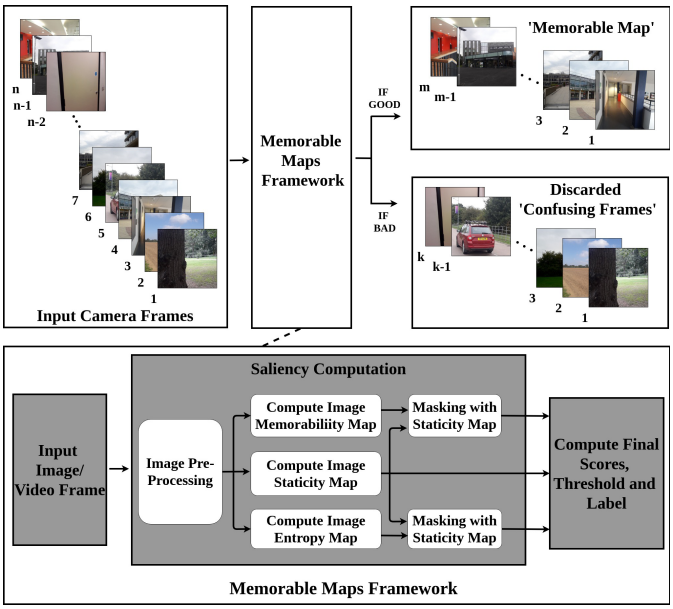

Fig. 1. A block-level overview of the proposed memorable maps framework is shown here.

Traditionally, for visual place recognition, 'Places' have been selected/sampled based on time-interval [4], distance [5] or distinctiveness [6] in different approaches. These approaches are discussed in depth in the next section. Most of these methods attempt to reduce the size of robot's map and do not quantify if a sampled/sub-sampled image is a good representation of a place; thereby has a greater chance of matching upon revisiting. The quality of image selection mechanism restricts the performance of a VPR system, both in the short-term and long-term. Due to limited number of images being stored in the map, it is critical to select those images that can be matched successfully upon repeated traversal-the motivation for this research.

In this work, we look at image selection from a semantic point of view and draw inspiration from images memorable to a human-cognition system. We use a Convolutional Neural Network [7] to compute the memorability of an incoming camera frame. However, while objects like vehicles and pedestrians in an image are subjectively-memorable; they are intrinsically not good for VPR as these dynamic objects

Manuscript received March 1, 2019; revised October 21, 2019 and March 13, 2020; accepted June 5, 2020. This work was supported in part by the U.K. Engineering and Physical Sciences Research Council under Grant EP/R02572X/1 and Grant EP/P017487/1 and in part by the RICE Project funded by the National Centre for Nuclear Robotics Flexible Partnership Fund. The work of Michael Milford was supported in part by the Australian Research Council (ARC) under Grant FT140101229 and Grant CE140100016 and in part by the QUT Centre for Robotics. The Associate Editor for this article was K. Wang. *(Corresponding author: Mubariz Zaffar.)*

Mubariz Zaffar, Shoaib Ehsan, and Klaus D. McDonald-Maier are with the School of Computer Science and Electronic Engineering, University of Essex, Colchester CO4 3SQ, U.K. (e-mail: mubariz.zaffar@essex.ac.uk; sehsan@essex.ac.uk; kdm@essex.ac.uk; mubarizzaffar@gmail.com).

Michael Milford is with the QUT Centre for Robotics, School of Electrical Engineering and Robotics, Queensland University of Technology, Brisbane, QLD 4000, Australia, and also with the Australian Centre for Robotic Vision, Queensland University of Technology, Brisbane, QLD 4000, Australia (e-mail: michael.milford@qut.edu.au).

Digital Object Identifier 10.1109/TITS.2020.3001228

are rarely re-observed. We thus perform object detection to compute the staticity of an image and mask memorability of dynamic content. In addition to being memorable and static, an image should be content-rich thus we calculate the entropy map.

The contribution of this work is a semantically coherent framework (Fig. 1) that filters an input image through a tri-folded criteria. Hence, ensuring that every image to be inserted against a place in robot's map is a good representation of the said place and highly recognizable. To analyze the effectiveness of this framework, we created a dataset 'ESSEX3IN1' from indoor, outdoor and natural environments. Unlike existing VPR datasets, ESSEX3IN1 mimics a robot exploring an environment instead of traditional path-following and is thus composed of highly confusing images from all three environments. We show how these confusing images lead to poor performance of current visual place recognition systems. The final results show the effectiveness of proposed framework in segregating these 'confusing' images from 'good' images, thereby increasing VPR precision and reducing database size. We also evaluate our framework on other public VPR datasets to show that this performance enhancement can be generalized. Due to its agnostic nature, any VPR technique can obtain a performance boost by stacking the presented framework as an additional layer in the VPR pipeline.

The remainder of the paper is organized as follows. In Section II, a comprehensive literature review regarding VPR state-of-the-art is presented with focus on image selection and semantic-mapping. Section III presents the motivation, design and implementation details of the 'memorable maps' framework developed in this work. Section IV is dedicated to the experimental setup for evaluating and analyzing state-of-the-art VPR techniques with and without proposed framework. Following-up on Section IV, Section V puts forth the results/analysis obtained by combining memorable maps and contemporary VPR techniques on multiple public datasets and ESSEX3IN1. Finally, conclusions and future directions are presented in Section VI.

II. Related Work

VPR and SLAM have seen major developments through different cognitive, intuitive or semantic approaches to the problem. A comprehensive review of these techniques is performed by Lowry *et al.* [1]. An earlier work on probabilistic implementation of SLAM in visual-appearance domain, called 'FAB-MAP', is presented by Cummins *et al.* [8]. This work was combined in [9] with a biologically inspired SLAM technique 'RAT-SLAM' [10], mimicking the Rat's hippocampus. Milford *et al.* [11] utilize sequence of images instead of individual frames to successfully match previously visited places under significant environmental variations. Similar to other fields, Convolutional Neural Networks (CNNs) have been a game-changer for VPR. The application of CNN for VPR was first studied by Chen *et al.* [12]. Authors in [13] trained two dedicated Neural Networks for VPR on Specific Places Dataset (SPED) containing images from different seasons and times of day. Unlike previous implementations where image descriptors were manually formed from CNN layer activations, Arandjelovic *et al.* [14] trained a new VLAD (Vector of Locally Aggregated Descriptors) layer for an end-to-end CNN-based-VPR. For images containing repetitive structures, Torii *et al.* [15] proposed a robust mechanism for collecting visual words into descriptors. Synthetic views are utilized for enhanced illumination invariant VPR in [16], which shows that highly condition variant images can still be matched if they are from the same viewpoint. In [17], authors try to extract local features from convolutional layers corresponding to salient Regions of Interest (ROI), thus providing significant viewpoint invariance. State-of-the-art performance is shown by authors in [18], by combining VLAD descriptors with ROI-extraction to show immunity to appearance and viewpoint variation. Fang *et al.* [19] presented a ground texture-based localisation technique for illumination-controlled localisation estimation.

Traditionally, places have been described by camera frames, where a place is selected from multiple video frames based on either time-step, distance or distinctiveness. Most of the VPR datasets [4], [9], [11], [17], [20]–[23] are time-based, as frames are selected given a fixed FPS (frames per second) rate of a video camera. However, time-based place selection assumes a constant non-zero speed of the robotic platform and is thus impractical in many situations. To cater for variable speed, distance-based frame selection is used where a frame is picked every few metres to represent a new place [5], [24]. Both time- and distance-based approaches lead to huge database sizes and frequently sample visually identical frames as different places; thus leading to inaccuracies and impracticality for long-term autonomy.

Different research works have tried to overcome these intrinsic limitations of image sampling by proposing image selection based on visual distinctiveness. Chapoulie *et al.* [6] use a customised algorithm that detects change point for segmentation between different topological places in both indoor and outdoor scenes. Image sequence partitioning for creating sparse topological maps is presented by Korrapati *et al.* [25], where sequences of images are divided into nodes/places using four descriptors namely GIST, Optical Flow, Local Feature Mapping and Common-Important Words. In [26], a thematic approach is adapted to evaluate the novelty of an incoming image by co-relating it with the redundancy of visual features/topics. Bayesian surprise is adapted with immunity to sensor type, for extracting landmarks to create a sparse topological map in [27]. Online topic modeling with visual surprise calculation is done by Girdhar *et al.* [28] for under-water explorations. An incremental unsupervised place discovery scheme is adopted by Murphy *et al.* [2] which fuses information over time to find visually distinct places.

Semantic mapping techniques for summarizing a robot's experience are surveyed by Kostavelis *et al.* [3]. Authors in [29] present both offline and online solutions for finding images that best summarize a given sequence. The score for every incoming image is related to the difference of posterior distribution from prior distribution using bayesian surprise or set theoretic surprise. In [30], coresets are used to pre-cluster input image stream and then topic-based image representation

is used followed with graph-based incremental clustering. A place detection scheme is proposed by Karaoguz *et al.* [31] based on bubble-space representation. A new place is checked for informativeness based on surface deformation and variance in a time-window of coherent images. The authors in [32] use region proposals in spatio-temporal context instead of low-level features to represent input frames and then based on region-adjacency-graph detect visually distinct places. A human-augmented change point detection scheme is presented by Topp *et al.* [33] where a change stimuli could either be pointed out by the robot or its operator. The authors propose the change as a structural ambiguity, which can be pointed out either by the robot or a human operator during a guided tour. Detection of change point is also targeted by Ranganathan [34] with a Bayesian probabilistic model. One common element to all these works is that they focus on map compression, video segmentation or experience summarisation, but do not discuss if the resulting compressed/summarized map is actually composed of good matchable images of places. These methods define the distinctive nature of images based on their visual difference from previously seen images. Resultingly, such visually different images may come from grassy plains, natural scenery, dynamic objects or low-textured places leading to poor VPR performance. Drawing inspiration from the said, we in this work, define distinctiveness based on a place's memorability (cognitive), static-content and information-richness leading to highly matchable compressed maps dubbed as 'memorable maps'. The human-memorability of an image is a well-known research domain and has drawn significant interest of the research community over the past many years. This work on predicting the memorability of an image was pioneered by Isola *et al.* [35] and deep learning was used to achieve state-of-the-art co-relation with human-memorability by Khosla *et al.* [7]. There are 3 key tracks that have been adopted in literature for improving the memorability models: 1) Introducing the role of emotions into the concept of memorability [36], [37]., 2) Introducing regional attention-based models [38]–[40]., 3) Studying the memorability of outdoor natural scenes [41], [42].

In addition to the above literature review, we discuss two works that have similar motivation to our approach. The interesting work by Hartmann *et al.* [43] proposes a random forest classifier of 5 decision trees trained on a dataset of 455 outdoor images. The objective of this random forest is to find keypoints in an image with low matchability and subsequently discarding them. This technique is computationally intensive in comparison to our methodology as we compute a single matchability (memorability) score against an image instead of scores against each image keypoint. Moreover, in VPR, features coming from dynamic objects and low-textured scenes are usually not re-observable/matchable (as shown later in our paper) but have not been examined in [43]. Although vegetation is considered to belong to non-matchable category, results show features coming from trees as being classified as matchable in [43]; which usually in VPR contribute negatively to the distinctiveness of a place (as shown in Fig. 3). More recently, a CNN able to classify input frames as stable/unstable is trained by Dymczyk *et al.* [44] for long term visual place recognition. Similar to [43], this work also proposes that vegetation in outdoor scenes is not suitable, but does not consider outdoor dynamic objects like cars, pedestrians, animals etc. Also, informativeness of stable frames in terms of extracted features and predicted matchability is not inspected given that walls are selected as stable elements of an image. Therefore, to the best of author's knowledge, our work combines for the first time all three of these criteria namely memorability, staticity and entropy to create memorable maps. We show utility of our proposed framework by reporting results on multiple public datasets. The agnostic nature of our framework is presented by using multiple state-of-the-art VPR techniques in combination with memorable maps.

III. Methodology

This section presents in depth the framework developed in our work. A sub-section is dedicated to each of the three criteria (i.e., memorability, staticity and entropy) adopted by the framework. We also discuss the integration of our framework with VPR techniques as a final sub-section.

For the purpose of evaluation and analysis, we have used AMOS-Net [13], Hybrid-Net [13] and Region-VLAD [18] as our VPR techniques throughout this paper. The details of these techniques are given in Section IV-B.

A. Memorability

1) Why Memorability?: The human-cognition system is powerful in evaluating what images are useful to be stored in the brain's memory fragments [45], [46]. We usually remember concrete structures like buildings, streets, squares etc. However, more natural scenes like fields, forests, grassy plains and far out sceneries are less memorable. This 'memorability' concept is also intuitive as it is easy to confuse different natural scenes with each other compared to concrete structures. In reference to VPR, there are two further reasons for the non-salience of trees, vegetation and natural scenery: 1) They are highly appearance-variant compared to concrete structures, 2) Local features coming from trees and vegetation etc. are usually mismatched, as explored in the works of [47]–[49]. In order to explain (1), we have shown samples of appearance changes in Fig. 2, along with the memorability maps created (methodology explained in sub-section III-A.2) by our framework. In Fig. 3, we have shown how non-memorable scenes are mismatched by state-of-the-art VPR systems leading to false-positives.

2) Memorability Implementation: Inspired from human-memorability, we apply the work done originally for marketing and advertising in [7] to VPR problem. A Convolutional Neural Network namely Hybrid-CNN [50] which was originally trained on Places365 dataset [50] for deep learning-based scene recognition, has been fine-tuned on LaMem dataset by Khosla et al in [7]. The authors in [7] have named this fine-tuned Hybrid-CNN as 'MemNet'. The LaMem dataset (introduced by [7]) is composed of 60,000 images covering multiple scenarios ranging from natural scenery, indoor scenes, outdoor scenes and distinctive objects. The ground-truth human-memorability provided in

Fig. 2. Concrete structures appear similar under seasonal changes while non-memorable elements like trees, vegetation and natural scenery appear very different. The memorability maps (in last row) show the effectiveness of our memorability implementation (sub-section III-A.2) in segregating concrete structures from these appearance variant regions. All examples images have been utilised from the Nordland dataset [21] and SPEDTest dataset [12] to ensure consistency with the evaluation mechanism.

Fig. 3. Natural places mismatched by VPR methods due to confusing features coming from trees, grass and plains are shown here. Red boundary represents an incorrect match while green boundary represents a correct match. All images in this figure were found static and information-rich, i.e., human-memorability is the only criterion that can discard these images.

LaMem dataset has been computed for each of the images using an interactive game played by multiple human subjects. Images are shown to players in a sequence and are repeated after a random interval where a human has to identify/recall a previously seen image. By using this LaMem dataset, the authors [7] fine-tuned the Hybrid-CNN achieving a high co-relation (0.64) with human memorability. Resultingly, the output of this fine-tuned Hybrid-CNN (MemNet) is a human-memorability score m for each input image in the range of $0-1$, with $m = 1$ being the most memorable. However, for our framework, we require a memorability map (as in Fig. 6) against every image instead of a single memorability score as output by MemNet [7]. The motivation for this memorability map is to cater for highly memorable but dynamic objects as discussed later in this sub-section and utilised in sub-section III-D.

The CNN input layer size is set to $W1 \times H1$. We re-size every incoming image to $W2 \times H2$.

$$where; \quad W2 = a \times W1, \;\; H2 = b \times H1$$

We then split this rescaled-image into C ($where$, $C = a \times b$) non-overlapping crops of size $W1 \times H1$ each and sequentially feed them as inputs to CNN. This in turn gives us the memorability matrix M as shown below.

$$M = \begin{bmatrix} m_{11} & m_{12} & \dots & m_{1b} \\ m_{21} & m_{22} & \dots & m_{2b} \\ \vdots & \vdots & \ddots & \vdots \\ m_{a1} & m_{a2} & \dots & m_{ab} \end{bmatrix}$$

where, m_{ij} is the memorability of each $W1 \times H1$ cropped image. To create a memorability map, we rescale the matrix M from $a \times b$ to $W2 \times H2$ with bilinear interpolation. Some examples of memorability maps overlayed on images are shown in Fig. 2 and Fig. 6. We have employed $C = 5 \times 5$ through-out this work and a parametric variation of this is shown later in sub-section V-F. It can be seen (in Fig. 2 and Fig. 6) that vegetation, natural scenery and trees are identified as less-memorable which is consistent with our motivation in sub-section III-A.1. However, for human cognition (and therefore for [7]), objects such as faces, animals and vehicles are memorable. But, such dynamic objects are not re-observable and therefore, they are not salient for VPR; we cater for this in the following sub-section.

B. Staticity

1) Why Staticity?: The previous sub-section shows how memorability is a good evaluation criterion for a camera frame to be used in VPR. However, one limitation is the fact that objects like cars, pedestrians, buses, animals and bicycles in an image are all classified as highly memorable but are not re-observable (for VPR problem). Resultingly, images that may be memorable but have high dynamic content will fail to match upon repeated traversal. Fig. 4 shows some of these images mismatched by VPR techniques [13], [18].

2) Staticity Implementation: To cater for highly dynamic images, we perform image segmentation into static and dynamic pixels. We re-size all input images to $W2 \times H2$. We use an object detector [51] that can detect 80 different classes of objects in an image. Out of these 80 classes, 21 correspond to highly-dynamic, commonly-observed objects. These dynamic objects include cars, pedestrians, buses, trucks, animals etc. We, therefore, only consider proposals of bounding boxes coming from objects of interest, i.e., dynamic objects. Please note that we have used the default parameters of YOLO [51] in our work.

Since the staticity map is computed for each pixel in the image, it can be represented as a staticity-matrix S of size

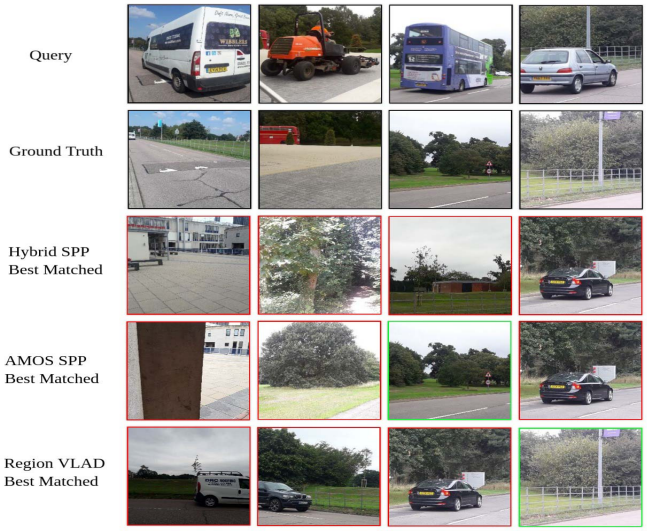

Fig. 4. Dynamic places mismatched by state-of-the-art VPR systems. Features coming from vehicles are not re-observable in addition to the occlusion caused by them in different scenes. Red boundary represents an incorrect match while green boundary represents a correct match.

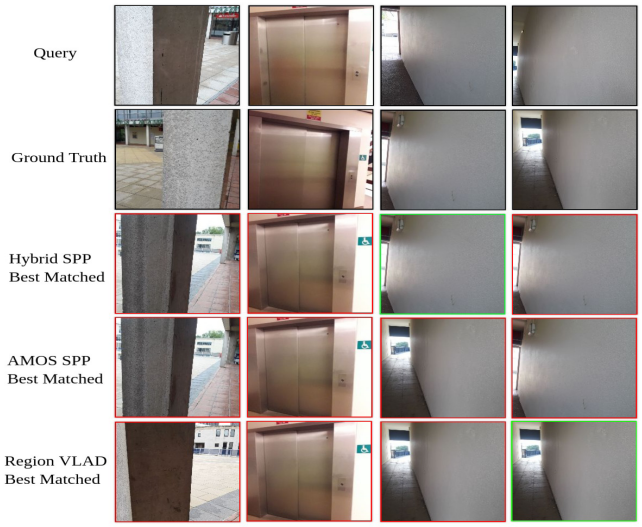

Fig. 5. Low-entropy places mismatched by state-of-the-art VPR methods can be commonly observed in indoor robot navigation datasets. Along with intrinsically less-informative images of doors/walls, static occlusion can also lead to poorly defined places. Red boundary represents an incorrect match while green boundary represents a correct match.

$W2 \times H2$ as below.

$$S = \begin{bmatrix} s_{11} & s_{12} & \dots & s_{1W_2} \\ s_{21} & s_{22} & \dots & s_{2W_2} \\ \vdots & \vdots & \ddots & \vdots \\ s_{H_2W_2} & s_{H_2W_2} & \dots & s_{H_2W_2} \end{bmatrix}$$

$$where; \quad \{s_{ij} \in \mathbb{Z}_2 \mid \mathbb{Z}_2 = [0,1]\}$$
$$s_{ij} = 1 | Pixel = Static$$
$$s_{ij} = 0 | Pixel = Dynamic$$

Fig. 6 shows the typical staticity map computed in our framework. However, although an image containing low-textured scenes (walls/door/pillars) can be classified as concrete (memorable) and static but it does not have distinguishable features and hence, it is not distinct. We accommodate this limitation in the following sub-section.

C. Entropy

1) Why Entropy?: An input camera frame containing a room/lift door is commonly observed by a robot navigating indoors. Such a frame is classified as memorable and static, but has little to no information differentiating it from other doors in the building, thus leading to false positives. The same can be extended to any other frame with occlusion resulting from walls, pillars etc. Examples of such confusing frames are shown in Fig. 5.

2) Entropy Implementation: To avoid less informative or occluded frames, we evaluate the information content of an image by computing its local entropy against every image pixel. This local entropy corresponds to the number of bits required to encode the local gray-scale distribution in an image. Based on standard boolean algebra, the number of bits required to represent any positive integral value can be computed by $log_2(Numerical\_Value)$. We use a circular window of r pixels radius as our local neighbourhood to get the entropy map of an incoming camera frame against each pixel. The total number of histogram bins used for entropy computation are 256 corresponding to $0-255$ gray-scale intensity values. The generic algorithm for entropy map computation is shown below and adapted from [52].

Algorithm 1 Computing Entropy Map

```
Create a Histogram of 256 Bins
for all Local Neighbourhoods in Image do
  for all Pixels in Current Neighbourhood do
    if Current_Pixel lies in BinX then
      Items_in_BinX = Items_in_BinX + 1
    end if
  end for
  Local_Entropy = log2(No.of Filled Histogram Bins)
  Clear all Histogram Bins
end for
```

This algorithm gives us an entropy map represented as matrix E of size $W2 \times H2$. Local circular regions of images containing texture-less doors/walls have a small range of intensity gradients within the region and thereby have lower entropy value. The maximum value of entropy is computed from equation (1) and equals 8, given that the maximum number of filled histogram bins is 256. Fig. 6 shows examples of entropy maps computed in our framework. We have used $r = 5$ in our work, where the reasons for this selection and parametric variation are shown in sub-section V-F.

$$Max\ Entropy = log_2\,(No.\ of\ Histogram\ Bins) \quad (1)$$

$$E = \begin{bmatrix} e_{11} & e_{12} & \dots & e_{1W_2} \\ e_{21} & e_{22} & \dots & e_{2W_2} \\ \vdots & \vdots & \ddots & \vdots \\ e_{H_2W_2} & e_{H_2W_2} & \dots & e_{H_2W_2} \end{bmatrix}$$

$$where; \quad \{e_{ij} \in K \mid K \subseteq \mathbb{R} \wedge K = \{0, \dots, 8\}\}$$

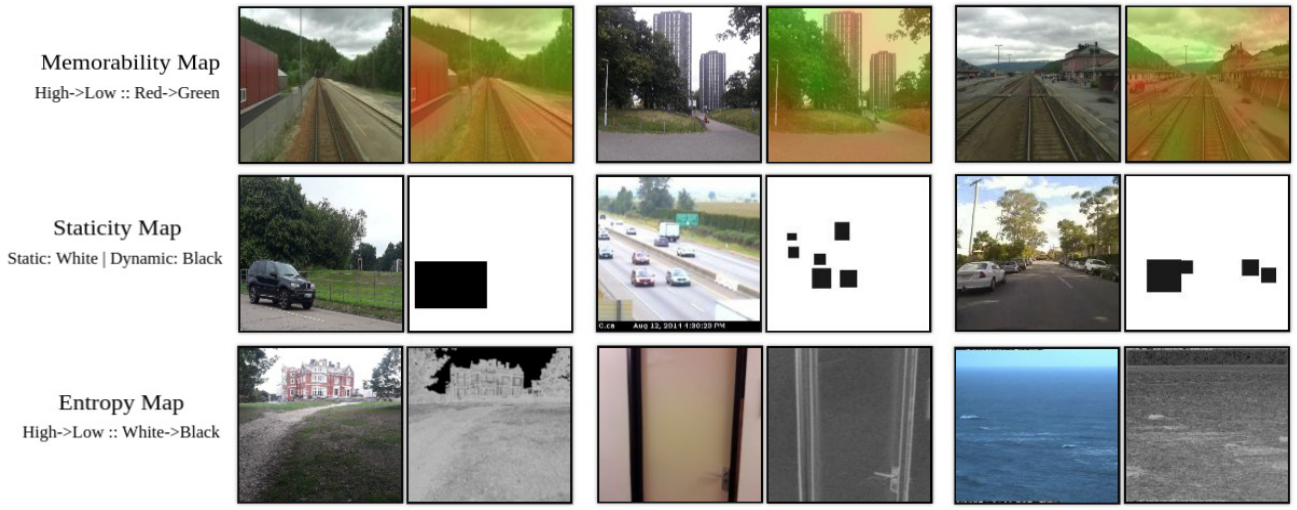

Fig. 6. The three types of image maps created by proposed framework for evaluating the content of an input image. Concrete structures like buildings and roads are memorable in comparison to grassy plains and trees [Top]. Cars, pedestrians and other dynamic objects are detected and evaluated for the amount (approximate) of pixels they occupy [Middle]. Uniform and texture-less scenes, sky portions have low-entropy compared to feature rich structures [Bottom].

D. Computing Scores and Thresholding

After acquiring all three maps of an image, we mask memorability map and entropy map with staticity map. This ensures that our decision to select an image based upon memorability and entropy is immune to the information coming from dynamic objects. Next, we compute the memorability score (MS) of an image as the average value of memorability map and compare it with a memorability threshold (MT), to evaluate if this image/frame is memorable enough for use in VPR. Secondly, we compute the percentage of static pixels in our staticity map to get a staticity score (SS). This is then contrasted with staticity threshold (ST) to decide if an incoming frame has enough static content to be inserted into the map. Thirdly, we calculate the average value of entropy map and scale it with the maximum value of entropy to get the percentage of information content. This percentage dubbed as the entropy score (ES), is compared with the entropy threshold (ET) to settle if an input frame has enough information.

Finally, we use a tri-input AND criteria to select images that are memorable, static and information-rich to be inserted into the memorable map.

Algorithm 2 Image Selection for Memorable Map

for all *Incoming Images* **do**
Compute All Three Image Maps

$$MS = \frac{1}{W2 \times H2} \sum_{i,j=1,1}^{W2,H2} m_{ij}$$

$$SS = \frac{1}{W2 \times H2} \sum_{i,j=1,1}^{W2,H2} s_{ij}$$

$$ES = \frac{1}{W2 \times H2 \times 8} \sum_{i,j=1,1}^{W2,H2} e_{ij}$$

if $MS \geq MT \;\&\; SS \geq ST \;\&\; ES \geq ET$ **then**
Insert into Map
else
Dicard Image
end if
end for

E. Integration of Memorable Maps and VPR Techniques

The integration of our framework with VPR techniques is seamless and straight-forward. The core component that all VPR techniques require to operate is a reference image database, using which the VPR techniques propose a place-match (image-retrieval) given an input query image. The creation of this reference database by employing the memorable maps framework instead of the traditional time-based or distance-based approaches is what brings our framework together with state-of-the-art VPR techniques. This integration of memorable maps framework with the VPR methods can be in an online or an offline fashion.

In an offline approach, where *a priori* knowledge of the environment is available in the form of images, memorable maps framework can take this knowledge (images) and output a memorable map as depicted in Fig. 1. In this case, the 'Input Camera Frames' block of Fig. 1 represents the input knowledge where each image is indexed in a sequential manner and evaluated by our framework yielding a memorable map. The contemporary VPR techniques can then use this memorable map instead of the original time-based, distance-based or distinctiveness-based reference image database, achieving place matching performance boost and map-size reduction as reported later in Section V.

Before discussing the integration of our framework in an online manner, it is important to understand that every query image in an online VPR system becomes a reference image at the next time step and is stored in the reference image database. Thus, for every input query image two operations are traditionally performed: 1) It is input to a VPR technique to search for a prospective place match, 2) If it matches to a previously known place, it is stored as an additional representation of the place and if it does not match to a previously known place, it is stored in the map as a 'new place'. Given this understanding, the memorable maps framework can easily be integrated into an online VPR system, where the input query image is first evaluated for its salience by our framework. If it is largely memorable, static and information-rich, only then it is used for VPR and subsequent storage in the reference map. For the online case, images in Fig. 1 would represent query images such that their indices represent time-stamps.

IV. Experimental Setup

This section discusses the datasets, VPR techniques and evaluation metric used in our analysis. We present a new dataset ESSEX3IN1, which is publicly available.[1] Additionally, we briefly discuss three pre-existing public datasets used for reporting our framework's performance. The VPR techniques used for our results and analysis are then summarized. We utilise area-under-the Precision-Recall curve (AUC) which is a well-established performance metric for VPR techniques.

A. Evaluation Datasets

This sub-section introduces the 4 datasets that we have used in our work to discuss and analyse the performance of memorable maps framework. Please note that none of these datasets were used for training the 3 VPR techniques employed in our work.

1) ESSEX3IN1 Dataset: Most of the Visual Place Recognition datasets have been created from a pre-planned path traversal. Thus, these datasets do not contain confusing images that an exploration robot may come across. Also, these datasets focus on a single type of environment either indoor or outdoor. To evaluate and challenge our framework, we have created a new dataset ESSEX3IN1 which is composed of images from indoor, outdoor and natural scenes.

The dataset was created in two stages using a human-held mobile phone camera at the University of Essex (Colchester Campus) and contains 210 query images and 210 reference images with viewpoint variations. In the first stage, the objective was to take images from all sorts of environments that were either 'confusing' or didn't qualify the definition of a 'distinct Place', where this indistinctness of a place refers to perceptual aliasing. Two-third of the images in ESSEX3IN1 are from this first stage. The second stage, consists of images that were not confusing and could be defined as 'distinct places'. One-third of the total images are from this second stage. Some images from these stages are shown in Fig. 7. The ground-truth data provides information about a single correct reference image against every query image. This ground-truth is created manually by looking at individual images such that the ground-truth pair of query and reference image represent the same geographic location in the world.

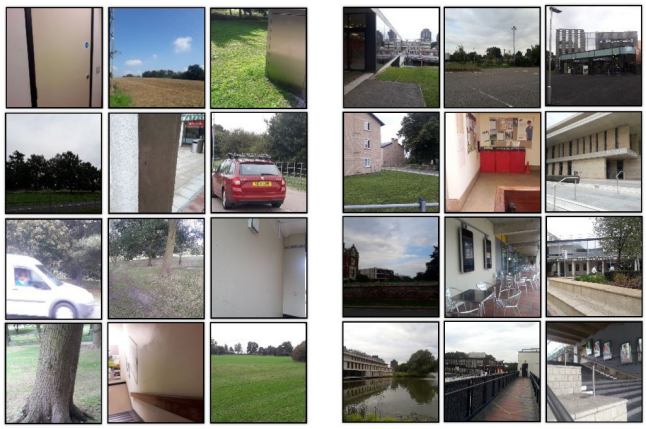

Fig. 7. Sample images from ESSEX3IN1 dataset. The first stage [on the left hand side] images contain occlusions, dynamic objects, information-less frames and non-memorable content like plains, natural scenery, vegetation and trees. In contrast, the second stage [on the right hand side] contains semantically identifiable and distinguishable images of various places from University of Essex (Colchester campus).

It is important to note that none of these images were used in tuning our three thresholds and were not seen prior by the proposed framework. The collection of dataset in this two-staged manner was useful for analysis in Section V.

2) Nordland Dataset: The Nordland dataset [21] comprises of a train journey through Norway and is collected in four different seasons with frame-to-frame ground-truth correspondence. We use a subset of this dataset which consists of 1622 query images and 1622 reference images. The query images are from the traversal performed in summer where as the reference images are from winter. Although this dataset does not provide any viewpoint variation, but has significant conditional variation. A retrieved image n is considered true-positive if the original ground-truth is between $n-1$ to $n+1$, i.e., each query image has 3 ground-truth references.

3) St. Lucia Dataset: The St. Lucia dataset was first introduced in [9]. It was recorded in the surroundings of University of Queensland's St. Lucia campus during multiple times of the day. This dataset consists of moderate viewpoint and illumination variation. The dataset also contains dynamic objects and scene variation. The ground-truth is derived manually from GPS data such that each query image has three reference images as true-positives. The total number of query images is 1261 and the total number of reference images is 1317.

4) SPEDTest Dataset: The SPEDTest dataset was introduced in [12] and is a sub-set of the original Specific Places Dataset [13]. It consists of 607 query images coming from a variety of scenes and environments. Frame-to-frame correspondence is available as the ground-truth.

[1] https://github.com/MubarizZaffar/ESSEX3IN1-Dataset

B. VPR Techniques

We have used three state-of-the-art VPR techniques (namely AMOS-SPP, Hybrid-SPP and Region-VLAD) [13], [18] that have shown promising results in recent research. AMOS-Net is a modified Caffe-Net [53] with all parameters trained on SPED dataset [13]. Hybrid-Net is another modified version of Caffe-Net with weights for top 5 convolutional layers initialized from Caffe-Net [53]. We have used Spatial Pyramidal Pooling as a feature descriptor for both AMOS-Net and Hybrid-Net since it shows excellent results as compared to other feature encoding methods. Features are extracted from 'conv5' layer in case of both AMOS-Net and Hybrid-Net. The third VPR technique, Region-VLAD, uses features extracted from selected/interesting regions of an AlexNet pre-trained on Places365 dataset [50]. Vector-of-Locally-Aggregated-Descriptors [54] is subsequently used for encoding the extracted features. In case of Region-VLAD, we use features from 'conv4', number of regions-of-interest as 400 and a visual dictionary size of 128. Evaluation of VPR techniques on existing datasets is an offline process, therefore the integration of our memorable maps framework with these techniques is in accordance to the discussion for an offline VPR system in sub-section III-E.

C. Evaluation Metric

For evaluating the performance of different VPR techniques, Area-Under-the-precision-recall-Curve (AUC) has been repeatedly used by the VPR research community [1], [12], [14], [17], [18], [55]–[57]. AUC acts as a good metric to assess the performance of a system based on true-positives (TP), false-positives (FP) and false-negatives (FN). For a given VPR technique, if a matched-place is the same as the ground-truth, it is labelled as a TP. If a matched-place does not match the ground-truth, it is labelled as a FP. Additionally, if a matched-place that was discarded due to a lower confidence-score but was actually a correct match according to the ground-truth is labelled as a FN. These are then used to compute the precision and recall at different thresholds of confidence-score, which are then plotted on y-axis and x-axis respectively. Area under this Precision-Recall curve is computed and named as AUC. The below 3 formulae are used for computing the Precision, Recall and AUC.

$$Precision = TP/(TP + FP) \tag{2}$$

$$Recall = TP/(TP + FN) \tag{3}$$

$$AUC = \sum_{i=1}^{N-1} \frac{(p_i + p_{i+1})}{2} \times (r_{i+1} - r_i) \tag{4}$$

$$where; \quad N = No.\ of\ Query\ Images$$

$$p_i = Precision\ at\ point\ i$$

$$r_i = Recall\ at\ point\ i$$

The extensive review of VPR research performed by Lowry et al in [1] and the VPR research community [12], [14], [17], [18], [55]–[57] in general agree that a highly precise VPR system with high recallability is required, which serves as our motivation to adopt AUC as an evaluation metric. This paper ensures consistency and fair comparison of AUC scores for different VPR methods on all datasets by computing and reporting results only using equation (4).

Fig. 8. VPR false positives upon evaluation on ESSEX3IN1 stage: 1. Images with cars, trees and natural scenes are mismatched. Additionally, images with low information and memorability are almost indistinguishable for even human cognition.

V. Results and Analysis

This section presents the results and analysis in a sequential manner. We first show that images collected from the first stage of ESSEX3IN1 actually lead to poor performance of VPR systems and are not good for insertion into a robot map. Secondly, we show the segregation performance of proposed framework on these 'confusing' images and 'good' images. Thirdly, we present the AUC improvement of different VPR systems when plugged with our framework on all datasets discussed in sub-section IV-A. This is followed-up with a sub-section dedicated to qualitative analysis showing sample images selected and discarded from all datasets. We then highlight the contribution of each framework criterion qualitatively and quantitatively. Next, we report the effect on VPR performance by sweeping framework parameters within possible range. We show in the next sub-section, how our framework leads to reduced map size and place matching time. Finally, we show the integration of Spatio-Temporal filtering with our framework to avoid large image gaps for localization.

A. Contemporary VPR Systems on ESSEX3IN1 Stage 1

The majority of VPR false positives against ESSEX3IN1 are from the first stage of dataset collection. This is due to the confusing images of fields, trees, doors, cars etc that lead to perceptual aliasing. Some of these false positives are shown in Fig. 8.

We show the AUC performance of VPR systems separately on Stage 1 and Stage 2 in Fig. 9.

B. Segregation Performance of Proposed Framework

For this sub-section, we apply the proposed memorable maps framework on complete and randomized

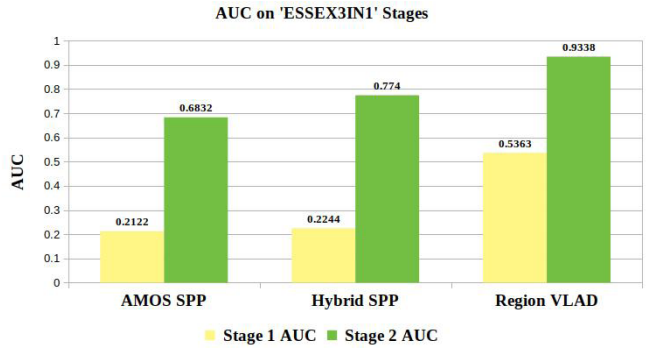

Fig. 9. Separate evaluation of VPR methods on each of ESSEX3IN1 stages reveals the challenge faced by contemporary VPR techniques for matching low-entropy, low-memorability and dynamic images.

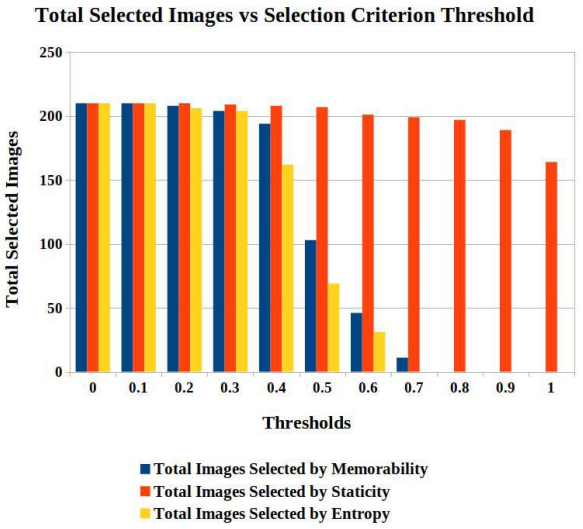

Fig. 10. The decrease in total selected images as each selection criterion is increased can be observed here for the whole ESSEX3IN1 dataset. For each threshold variation, the other two thresholds were set equal to zero, i.e., inactive. Majority of the ET/MT based image selection is done between $0.4 - 0.7$. Purely static images (without vehicles, human and animals) exist in the dataset which is why $ST = 1$ does not reduce map size to zero.

ESSEX3IN1 dataset. We use the below thresholds to segregate and discard 'confusing' images from 'good' images.

$$\text{Memorability-threshold} = 0.5$$
$$\text{Staticity-threshold} = 0.6$$
$$\text{Entropy-threshold} = 0.4$$

These values for the thresholds were selected from our analysis on pre-existing public VPR datasets. Increasing these thresholds reduces the number of images inserted into the memorable map. This is shown in Fig. 10 by varying each threshold from $0 - 1$, while setting the other two equal to 0. The manual selection of these particular values is based on the detailed analysis provided in sub-section V-F. Briefly, these particular values were employed for 3 reasons: 1) Agnostic performance boost across all 4 datasets (refer sub-section V-C)., 2) Reasonable number of 'good' images are left in the database as loop-closure candidates (refer sub-section V-G)., 3) Deviating

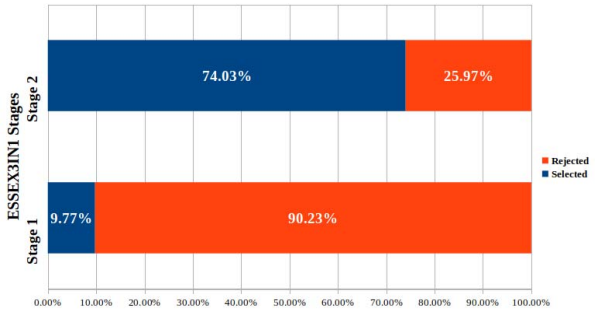

Fig. 11. The objective of memorable maps framework is to sample good frames and discard confusing frames. This objective achievement is presented by showing the contribution in memorable map from each ESSEX3IN1 stage.

significantly from these values could lead to zero or negative changes in the AUC (refer sub-section V-F). Setting any of the 3 thresholds equal to 0 will disable the corresponding criterion, e.g., in a continuous highly crowded scene, the ST can be disabled or the value of MT can be decreased for a continuous agricultural/natural environment. Increasing the thresholds towards 1 will result in decreased no. of images in the database, which will have higher salience.

The new database created by presented framework consists of memorable, static and informative images, thus dubbed as a memorable map. We show in Fig. 11, how many of the total images selected by presented framework are from which stage of the dataset.

C. AUC Improvement of VPR Systems

By selecting images that are memorable, static and have a higher entropy, the memorable maps framework gives performance boost to state-of-the-art VPR techniques. Here, we use fixed thresholds, as in previous sub-section V-B, but an AUC sweep across these thresholds is presented later in sub-section V-F. AUC evaluation is performed on the entire (both stages combined randomly) ESSEX3IN1 dataset along with the three public VPR datasets. It is important to note that bad/confusing images found by our framework are not removed from the reference database when evaluating AUC, but are treated as true negatives. This ensures that AUC boost reported here is not due to reduction of reference database size. For comparison with our framework, we also show the AUC performance for each technique by only employing static images on all datasets. Please note that because SPEDTest, Stlucia and Nordland datasets are largely static, the performance boost by just employing static images is only evident for ESSEX3IN1 dataset. This further validates the utility of our new proposed dataset ESSEX3IN1 for VPR, while simultaneously advocating for the efficacy of memorability and entropy criteria.

Fig. 12 depicts the AUC increase by employing our framework on ESSEX3IN1, St. Lucia, Nordland and SPEDTest dataset, respectively. We use the same values for MT, ST and ET as in sub-section V-B for ESSEX3IN1, Nordland and SPEDTest dataset. However, for St. Lucia we reduce each of the 3 selection thresholds by 0.05 to get a non-zero map size.

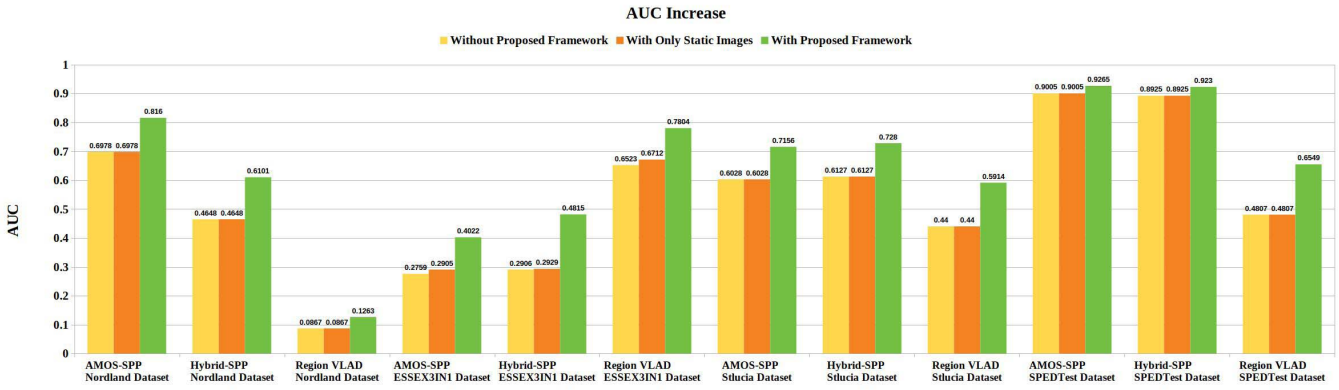

Fig. 12. Increase in AUC by using the proposed framework in combination with VPR techniques on all 4 datasets employed in our work is presented here. This performance increase for all techniques on all datasets advocates for the utility, generalisability and agnostic nature of our framework. Reference database size remained the same for all AUC evaluations by treating confusing images as true-negatives. Please note that ESSEX3IN1 is the only dataset with highly dynamic content and therefore the AUC boost for employing only static images is not evident on other datasets.

Fig. 13. Examples of images selected and discarded by the memorable maps framework from all 4 datasets are shown here. Top-left pairs of 4 images in each selected and discarded division are from ESSEX3IN1 dataset, followed-up with pairs from Stlucia dataset, Nordland dataset and SPEDTest dataset in clockwise manner. Selected images from ESSEX3IN1 are pre-dominantly of buildings with distinctive patterns and are largely static, while discarded images consist of far out natural scenes, dynamic objects or have low-entropy. Selected images in Stlucia dataset contain road signs, squares and houses. On the other hand, discarded images comprise of far out road scenes with trees and large portions of sky. Selected images from Nordland dataset consist of either appearing tunnels or bridges which contribute to their distinctiveness, while discarded images consist of vegetation or have low information. Staticity does not play any role in Nordland dataset due to the absence of dynamic objects. Selected images in SPEDTest dataset are from CCTVs covering buildings or distinctive locations, while discarded images consist of far out natural scenes and dynamic objects.

D. Selected vs Discarded Images

In this sub-section, we show some images from all 4 datasets that were selected or discarded by our framework. This gives a qualitative insight into the working of our framework in different environments/datasets. Since the memorable maps framework evaluates both the query images and reference images, the images in Fig. 13 are impartial to such distinction.

We also report the distribution of memorable images over the trajectories of Stlucia and Nordland datasets in Fig. 14. Because the ground-truth information for these datasets does not contain the exact inter-frame distance/time, the distribution in Fig. 14 is shown over image indices, which is very close to a constant distance-based distribution, as the speed of camera platform is mostly constant over the respective trajectories. ESSEX3IN1 and SPEDTest datasets are not trajectory-based, therefore, this distribution of memorable images over trajectory is not shown for these datasets.

E. Criterion Contribution Analysis

Each criterion in the memorable maps framework contributes to AUC boost. This subsection is dedicated to giving an insight into this individual contribution. We use ESSEX3IN1 for this purpose as it contains confusing images from all three (memorability, staticity and entropy) paradigms. For our AUC evaluation on ESSEX3IN1, we show the contribution of each criterion in Fig. 15. The analysis is performed based on the number of images that were mismatched by a VPR technique and were also discarded by atleast one of the

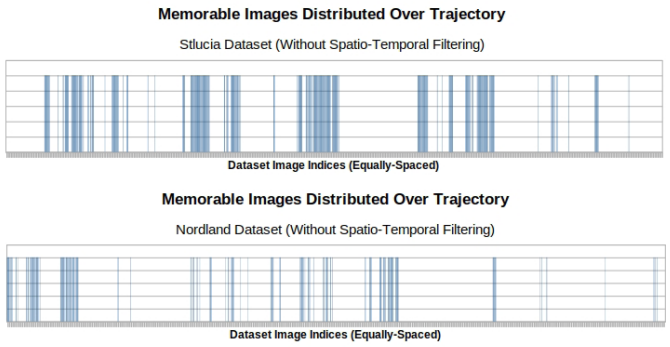

Fig. 14. Images selected as memorable over the trajectories of Stlucia [9] and Nordland datasets [21] are shown here. The horizontal axis represents the discrete, positive and equally-spaced indices of all the images in respective dataset. Each vertical bar represents an image selected as memorable by our framework. Because Spatio-Temporal filtering has not been utilized for this analysis, the selection of images is not uniform.

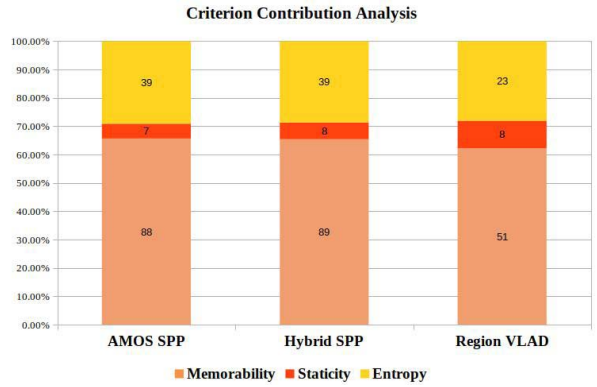

Fig. 15. Percentage contribution of each criterion into AUC increase is shown for ESSEX3IN1. This contribution is directly linked with the type of environment being explored. In a highly dynamic environment, the contribution of staticity will be more significant than suggested by this chart and such.

memorable maps framework criterion. Additionally, we also show in Fig. 16, a qualitative holistic view into cases where each criterion fails and others are used to cater for this failure, thereby, explaining intuitively why each of the criterion in our framework has its individual significance.

While Fig. 15 suggests that each of the three criteria are useful; the % contribution is linked to (and can vary with) the number of non-memorable, dynamic and information-less images in the dataset. (refer Fig. 10)

F. Parametric Variation

In this subsection, we present the variation in Visual Place Recognition performance with stricting framework criteria on ESSEX3IN1. We sweep each of the three criteria from 0-1 (Step size: 0.1) while keeping the other two inactive (i.e., set equal to zero). The data points for memorability and entropy thresholds have an upper-bound after which the total number of selected images equals to 0 (refer Fig. 10).

Fig. 17 shows that increasing entropy-threshold and memorability-threshold increases the AUC Performance for all three VPR techniques and follows a direct-relationship. On the other hand, the variation in AUC with increasing staticity-threshold follows a different trend. Firstly, the increase in AUC with ST is comparatively lower compared to MT/ET; which is due to the less number of dynamic images in the dataset compared to non-memorable and low-entropy images. Secondly, the variation in AUC with ST for Region-VLAD is higher compared to AMOS-SPP/Hybrid-SPP. We associate this with the fact that AMOS-SPP/Hybrid-SPP have been trained on SPED (Specific Places Dataset) and discourage features coming from vehicles. While our analysis/results reveal that Region-VLAD extracts and positively matches features coming from cars in different places (See Fig. 4). Thirdly, there is an evident decrease in AUC as ST goes above 0.9. This decrease is expected as images with very low dynamic content can still be matched by contemporary VPR-techniques and discarding such images leads to the observed decline in VPR-performance. Please note that the best AUC results in Fig. 17 are higher than the results reported in Fig. 12. This trend needs to be seen in co-relation with the reduction in map size as reported in Fig. 10. Increasing the three thresholds results in highly salient images stored in the map leading to higher AUC, however, it also reduces the absolute number of place recognition (loop-closure) candidates in the map and therefore, the framework thresholds need to be selected accordingly. The presented trends give a general idea for setting thresholds, thus to maintain a good balance between VPR performance and a salient representation of the world in a metric/topological/topo-metric map.

We also show the effect of varying the value of C from sub-section III-A.2 in Fig. 18 for the reader's understanding. Changing this parameter within the range shown in Fig. 18 does not have any effect on the AUC performance of all techniques on SPEDTest dataset, suggesting that our framework in not sensitive to this parameter. The effect on entropy map and entropy score (ES) by varying the local circular neighbourhood (r) in sub-section III-C.2 is reported in Fig. 19. The entropy score (ES) is dependent on this local circular neighbourhood r, such that increasing the value of r reduces the resolution of entropy map and increases the entropy score ES. This effect is similar to low-pass filtering and is explained as: Increasing the value of r increases the number of pixels to be added to the histogram, where the larger the radius of the circle, the greater will be pixel intensity divergence and hence higher is the log_2 score, leading to higher ES. This therefore, requires us to affix the value of r to a value where coupled with ET, we can successfully distinguish between low and high informative images. We are also interested in having high-resolution entropy maps instead of low-resolution entropy maps due to the salience of low-level features (like edges, corners etc) to the VPR problem.

G. Reduced Map Size and Computational Time

In addition to the increase in AUC, the developed framework helps in reducing the robot's map size which has been the motivation for semantic mapping research reviewed in this paper. This size reduction also leads to lesser computational overhead for VPR. The reduction in map size for the thresholds

Fig. 16. Examples of images that are selected/discarded based on various combinations of memorable maps framework criteria are shown here. Please note that the understanding for 'informative' and 'memorable' nature of an image is subjective and in our work, it is expressed by the practical/implementation nature of the term. For-example, an image of a bush (top-left) is called informative because it has several edges, corners and contours for computer-vision feature descriptors and thereby has a high entropy. Similarly, memorability is explained by its cognitive perception, i.e., the work of [7].

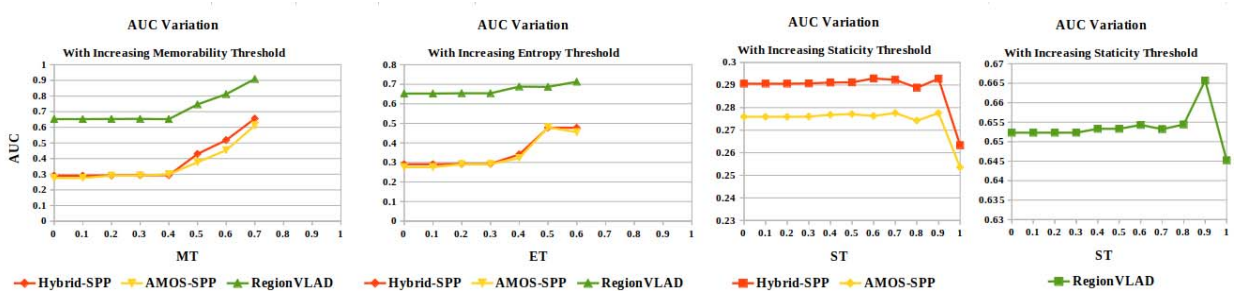

Fig. 17. Variation in VPR AUC performance by changing each of the memorable maps framework thresholds within their full range on ESSEX3IN1 is presented. For each threshold variation, the other two thresholds were set equal to zero, i.e., inactive. Memorability and entropy continuously increase AUC until the total number of selected images equals to zero; suggesting that images with higher memorability and entropy are well-matched by VPR methods. On the contrary, since images with low dynamic content should/can still be matched, variation in staticity threshold does not lead to a continuous AUC increase. AUC change with ST is not at the same scale as MT/ET so it is shown separately.

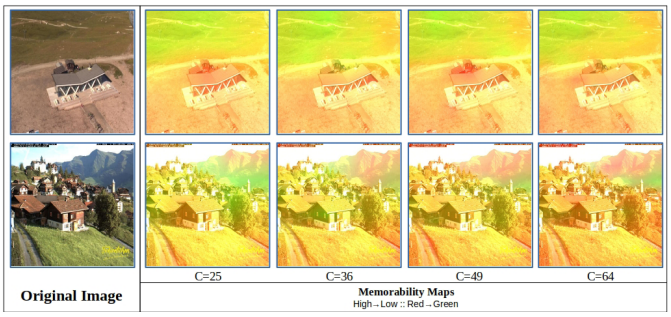

Fig. 18. Changing the value of C for computing memorability map does not result in any major change, as shown here. Images employed are from the SPEDTest [12] dataset and no change in AUC for this dataset was observed for the range of C used in this figure.

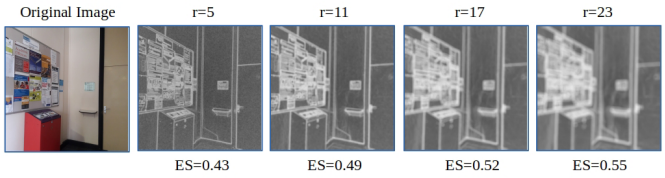

Fig. 19. Variation in the entropy map and the entropy-score (ES) are shown here for different values of local circular neighbourhood (r) given fixed image size ($W2 \times H2$). The larger the radius, the lower the resolution of entropy map. Increasing r also increases the value of ES due to increased no. of pixels for grayscale histogram that results in higher pixel intensity deviation.

presented in Section V-B is shown in Fig. 20. Because the map-size reduction by discarding non-memorable images can also lead to the reduction of absolute number of true-positives, we show this trend in Fig. 21. It can be seen in Fig. 21 that using the memorable maps framework does result in the decrease of true-positives, however, the proportion of discarded false-positives is greater than true-positives, which leads to the AUC boost reported previously in sub-section V-C.

The computational performance is reported by calculating the time required to match a query image with all the reference images (having pre-computed feature descriptors) in both a conventional map and a memorable map. This offline matching time is elaborated in Table I, where a memorable map having lesser number of reference images (see Fig. 20) achieves better matching time. The end-to-end time required in our implementation to compute the salience of an image for memorable map is around 5 sec. Because our current implementation utilizes a sequential combination of different research works, i.e., YOLO and MemNet, the timing is bottle-necked by the sum of individual timings of each of these works. We believe that there is room to improve the time required to compute these maps by employing a different suit of CNNs (object detectors and memorability maps), improving software implementation, utilizing hardware advances and by parallelizing the map computation by exploiting the independence of the three maps from each other.

TABLE I
MATCHING TIME PER QUERY IMAGE

System Specifications	Intel(R) Xeon(R) Gold 6134 CPU @ 3.20GHz, 64GB Physical Memory					
Framework	Without Memorable Maps			With Memorable Maps		
VPR Methods	AMOS-SPP	Hybrid-SPP	Region-VLAD	AMOS-SPP	Hybrid-SPP	Region-VLAD
ESSEX3IN1 (sec)	10.2	9.9	0.14	4.1	3.9	0.05
Nordland (sec)	78.7	76.4	1.1	9.1	8.7	0.12
St. Lucia (sec)	63.9	62.1	0.88	14.7	14.2	0.21
SPEDTest (sec)	29.5	28.6	0.41	7.7	7.5	0.11

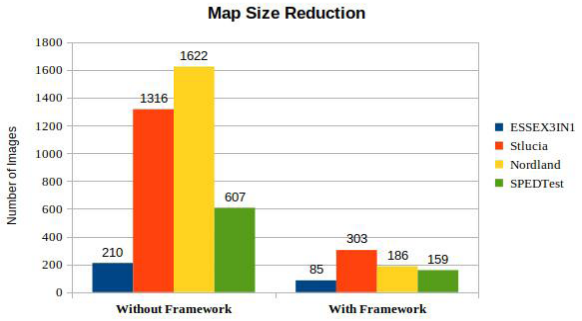

Fig. 20. Reduction in topological map size given similar or better VPR results is of prime importance for an autonomous robot to efficiently map/explore an environment. As depicted here, memorable maps framework intrinsically reduces map size while giving AUC boost to contemporary VPR systems.

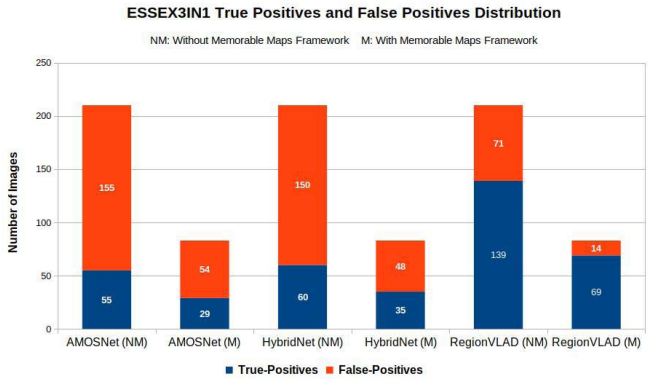

Fig. 21. The absolute decrease in true-positives and false-positives by using the memorable maps framework is shown here for all techniques on the ESSEX3IN1 dataset.

H. Spatio-Temporal Filtering With Proposed Framework

A natural extension to the memorable maps framework is to define an upper bound on the maximum distance and/or time travelled within which a best image (or Top-N images) from the traversed sequence should be be selected, even if the said image does not fully satisfy the proposed criteria. This can also be accommodated using a hysteresis-mechanism, where if a scene is continuously non-salient, the values for thresholds can be reduced to select relatively-less salient images. Depending on the size of employed hysteresis, it can be ensured that salient images distributed through-out the trajectory are selected. Thus, in a long traversal where the depicted scenery may not be memorable, static and/or highly-informative through-out the sequence, spatio-temporal filtering will ensure that the most memorable, static and informative image within the sequence is selected. This image

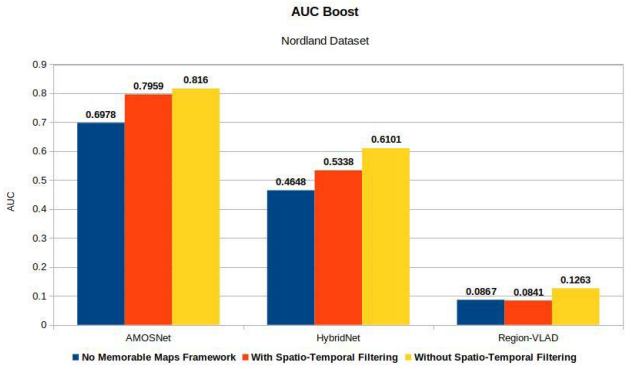

Fig. 22. Employing spatio-temporal filtering within the memorable maps framework to avoid large geographical gaps between salient images leads to lesser AUC boost as less salient images are added to the map. Using our framework without spatio-temporal filtering leads to the highest AUC, followed by our framework with spatio-temporal filtering and lastly without the memorable maps framework. Region-VLAD has significantly less number of true-positives through-out the trajectory, therefore AUC boost with spatio-temporal filtering is not evident for this technique.

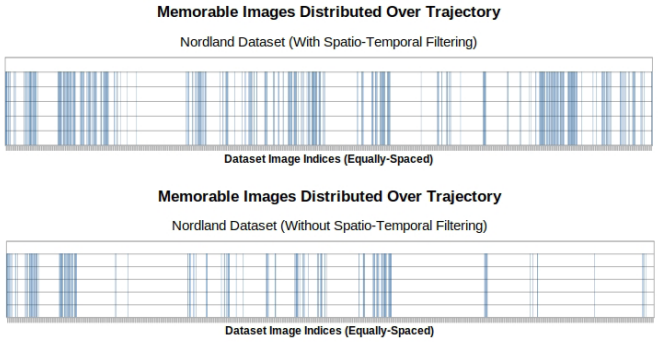

Fig. 23. Changes in memorable images selected over the Nordland trajectory by employing hysteresis-based spatio-temporal filtering within the memorable maps framework are shown here. Depending on the width of hysteresis, image gaps can be further reduced at the cost of reduced map salience.

can then be flagged as a 'low-quality' image in the memorable map and depending on the under-lying VPR system can either be treated differently (e.g., use longer matching sequences in [11]), avoided for use in loop-closure or treated as a potential false-positive prediction [58].

Because employing such a mechanism can lead to changes in AUC, we have reported this analysis of AUC boost with and without the spatio-temporal filtering in Fig. 22 for Nordland dataset. Our selection of filtering methodology is hysteresis-based, such that if in a sequence of 20 consecutive frames, none of the images satisfy the criteria thresholds of sub-section V-B, we reduce these thresholds by 0.03 for the respective sequence. It can be clearly seen in Fig. 22 that

allowing less-salient images into the map does lead to lesser AUC boost. We also show the changes in distribution of a total of 412 memorable images over the Nordland trajectory by employing such hysteresis-based spatio-temporal filtering in Fig. 23.

VI. Conclusion and Future Work

We proposed a cognition-inspired generalized framework for creating 'memorable maps'. This framework evaluates an incoming camera frame for its memorability, staticity and entropy to decide a frame's insertion into the robot's map. By using 'ESSEX3IN1', we show how images that are confusing and indistinct lead to perceptual aliasing and are also mismatched by contemporary VPR systems. The application of proposed framework in detecting these confusing images and subsequently improving VPR performance is presented. We generalise the applicability of our framework by reporting results on multiple public datasets. Due to its agnostic nature, memorable maps framework can be plugged into any VPR technique giving performance boost.

While presented thresholds are suitable for different indoor, outdoor and natural environments, they are not illumination invariant. In future work, it will be useful to integrate [59] to this work, thus making these thresholds as illumination-dependent variables. We acknowledge that there is room for a geography-based supervisory mechanism that determines the selection of salience thresholds, e.g, discarding images in a highly crowded urban environment may not be a desirable trait and such. The memorable maps framework coupled with different VPR techniques also enables the creation of a large-scale dataset containing 'good' and 'confusing' images. Such a dataset could subsequently help in training an end-to-end neural network for classifying an image as good/bad for map-insertion. Another important area of improvement is adopting the natural outdoor scenery-focused memorability computation, as in [41], [42], into the memorable maps framework. These natural scenery focused memorability prediction methods can further help to distinguish between distinct and indistinct outdoor scenes. We hope that our work draws attention of VPR community towards further research in segregation between confusing and good images. Thus, moving closer to practical deployment of VPR systems.

References

[1] S. Lowry *et al.*, "Visual place recognition: A survey," *IEEE Trans. Robot.*, vol. 32, no. 1, pp. 1–19, Feb. 2016.

[2] L. Murphy and G. Sibley, "Incremental unsupervised topological place discovery," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2014, pp. 1312–1318.

[3] I. Kostavelis and A. Gasteratos, "Semantic mapping for mobile robotics tasks: A survey," *Robot. Auto. Syst.*, vol. 66, pp. 86–103, Apr. 2015.

[4] M. Warren, D. McKinnon, H. He, and B. Upcroft, "Unaided stereo vision based pose estimation," in *Proc. ACRA*, 2010.

[5] S. Garg and M. Milford, "Straightening sequence-search for appearance-invariant place recognition using robust motion estimation," in *Proc. ACRA*, 2017.

[6] A. Chapoulie, P. Rives, and D. Filliat, "Topological segmentation of indoors/outdoors sequences of spherical views," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Oct. 2012, pp. 4288–4295.

[7] A. Khosla, A. S. Raju, A. Torralba, and A. Oliva, "Understanding and predicting image memorability at a large scale," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2390–2398.

[8] M. Cummins and P. Newman, "FAB-MAP: Probabilistic localization and mapping in the space of appearance," *Int. J. Robot. Res.*, vol. 27, no. 6, pp. 647–665, Jun. 2008.

[9] A. J. Glover, W. P. Maddern, M. J. Milford, and G. F. Wyeth, "FAB-MAP + RatSLAM: Appearance-based SLAM for multiple times of day," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2010, pp. 3507–3512.

[10] M. J. Milford, G. F. Wyeth, and D. Prasser, "RatSLAM: A hippocampal model for simultaneous localization and mapping," in *Proc. IEEE Int. Conf. Robot. Autom.*, vol. 1, May 2004, pp. 403–408.

[11] M. J. Milford and G. F. Wyeth, "SeqSLAM: Visual route-based navigation for sunny summer days and stormy winter nights," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2012, pp. 1643–1649.

[12] Z. Chen, O. Lam, A. Jacobson, and M. Milford, "Convolutional neural network-based place recognition," in *Proc. ACRA*, 2014.

[13] Z. Chen *et al.*, "Deep learning features at scale for visual place recognition," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2017, pp. 3223–3230.

[14] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "NetVLAD: CNN architecture for weakly supervised place recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 5297–5307.

[15] A. Torii, J. Sivic, T. Pajdla, and M. Okutomi, "Visual place recognition with repetitive structures," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 883–890.

[16] A. Torii, R. Arandjelovic, J. Sivic, M. Okutomi, and T. Pajdla, "24/7 place recognition by view synthesis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1808–1817.

[17] Z. Chen, F. Maffra, I. Sa, and M. Chli, "Only look once, mining distinctive landmarks from ConvNet for visual place recognition," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Sep. 2017, pp. 9–16.

[18] A. Khaliq, S. Ehsan, Z. Chen, M. Milford, and K. McDonald-Maier, "A holistic visual place recognition approach using lightweight CNNs for significant ViewPoint and appearance changes," *IEEE Trans. Robot.*, vol. 36, no. 2, pp. 561–569, Apr. 2020.

[19] H. Fang, C. Wang, M. Yang, and R. Yang, "Ground-texture-based localization for intelligent vehicles," *IEEE Trans. Intell. Transp. Syst.*, vol. 10, no. 3, pp. 463–468, Sep. 2009.

[20] M. Milford *et al.*, "Sequence searching with deep-learnt depth for condition- and viewpoint-invariant route-based place recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2015, pp. 18–25.

[21] S. Skrede. (2013). *Nordland Dataset*. [Online]. Available: https://bit.ly/2QVBOym

[22] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *Int. J. Robot. Res.*, vol. 32, pp. 1231–1237, Sep. 2013.

[23] T. Naseer, W. Burgard, and C. Stachniss, "Robust visual localization across seasons," *IEEE Trans. Robot.*, vol. 34, no. 2, pp. 289–302, Apr. 2018.

[24] E. Pepperell, P. Corke, and M. Milford, "Towards persistent visual navigation using smart," in *Proc. ACRA*, 2013.

[25] H. Korrapati, J. Courbon, Y. Mezouar, and P. Martinet, "Image sequence partitioning for outdoor mapping," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2012, pp. 1650–1655.

[26] R. Paul and P. Newman, "Self help: Seeking out perplexing images for ever improving navigation," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2011, pp. 445–451.

[27] A. Ranganathan and F. Dellaert, "Bayesian surprise and landmark detection," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2009, pp. 2017–2023.

[28] Y. Girdhar, P. Giguere, and G. Dudek, "Autonomous adaptive underwater exploration using online topic modeling," in *Experimental Robotics*. Springer, 2013, pp. 789–802.

[29] Y. Girdhar and G. Dudek, "Online navigation summaries," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2010, pp. 5035–5040.

[30] R. Paul, D. Feldman, D. Rus, and P. Newman, "Visual precis generation using coresets," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2014, pp. 1304–1311.

[31] H. Karaoguz and H. I. Bozma, "Reliable topological place detection in bubble space," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2014, pp. 697–702.

[32] M. Demir and H. I. Bozma, "Automated place detection based on coherent segments," in *Proc. IEEE 12th Int. Conf. Semantic Comput. (ICSC)*, Jan. 2018, pp. 71–76.

[33] E. A. Topp and H. I. Christensen, "Detecting structural ambiguities and transitions during a guided tour," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2008, pp. 2564–2570.

[34] A. Ranganathan, "Detecting and labeling places using runtime change-point detection and place labeling classifiers," U.S. Patent 8 559 717, Oct. 15, 2013.

[35] P. Isola, J. Xiao, D. Parikh, A. Torralba, and A. Oliva, "What makes a photograph memorable?" *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 7, pp. 1469–1482, Jul. 2014.

[36] Y. Baveye, R. Cohendet, M. Perreira Da Silva, and P. Le Callet, "Deep learning for image memorability prediction: The emotional bias," in *Proc. ACM Multimedia Conf.*, 2016, pp. 491–495.

[37] Z. Bylinskii, P. Isola, C. Bainbridge, A. Torralba, and A. Oliva, "Intrinsic and extrinsic effects on image memorability," *Vis. Res.*, vol. 116, pp. 165–178, Nov. 2015.

[38] B. Celikkale, A. Erdem, and E. Erdem, "Visual attention-driven spatial pooling for image memorability," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2013, pp. 976–983.

[39] M. Mancas and O. Le Meur, "Memorability of natural scenes: The role of attention," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2013, pp. 196–200.

[40] M. A. Cohen, G. A. Alvarez, and K. Nakayama, "Natural-scene perception requires attention," *Psychol. Sci.*, vol. 22, no. 9, pp. 1165–1172, Sep. 2011.

[41] J. Lu, M. Xu, and Z. Wang, "Predicting the memorability of natural-scene images," in *Proc. Vis. Commun. Image Process. (VCIP)*, Nov. 2016, pp. 1–4.

[42] J. Lu, M. Xu, R. Yang, and Z. Wang, "Understanding and predicting the memorability of outdoor natural scenes," 2018, *arXiv:1810.06679*. [Online]. Available: http://arxiv.org/abs/1810.06679

[43] W. Hartmann, M. Havlena, and K. Schindler, "Predicting matchability," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 9–16.

[44] M. Dymczyk, E. Stumm, J. Nieto, R. Siegwart, and I. Gilitschenski, "Will it last? Learning stable features for long-term visual localization," in *Proc. 4th Int. Conf. 3D Vis. (3DV)*, Oct. 2016, pp. 572–581.

[45] T. F. Brady, T. Konkle, G. A. Alvarez, and A. Oliva, "Visual long-term memory has a massive storage capacity for object details," *Proc. Nat. Acad. Sci. USA*, vol. 105, no. 38, pp. 14325–14329, Sep. 2008.

[46] T. Konkle, T. F. Brady, G. A. Alvarez, and A. Oliva, "Scene memory is more detailed than you think: The role of categories in visual long-term memory," *Psychol. Sci.*, vol. 21, no. 11, pp. 1551–1556, Nov. 2010.

[47] A. Mousavian, J. Kosecka, and J.-M. Lien, "Semantically guided location recognition for outdoors scenes," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2015, pp. 4882–4889.

[48] J. Knopp, J. Sivic, and T. Pajdla, "Avoiding confusing features in place recognition," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2010, pp. 748–761.

[49] P. Zhang, J. Wang, A. Farhadi, M. Hebert, and D. Parikh, "Predicting failures of vision systems," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 3566–3573.

[50] R. Raturi, "Adapting deep features for scene recognition utilizing places database," in *Proc. 2nd Int. Conf. Inventive Commun. Comput. Technol. (ICICCT)*, Apr. 2018, pp. 487–495.

[51] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7263–7271.

[52] S. van der Walt *et al.*, "Scikit-image: Image processing in Python," *PeerJ*, vol. 2, p. e453, Jun. 2014.

[53] A. Krizhevsky and *et al.*, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[54] H. Jegou, M. Douze, C. Schmid, and P. Perez, "Aggregating local descriptors into a compact image representation," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 3304–3311.

[55] Z. Chen, L. Liu, I. Sa, Z. Ge, and M. Chli, "Learning context flexible attention model for long-term visual place recognition," *IEEE Robot. Autom. Lett.*, vol. 3, no. 4, pp. 4015–4022, Oct. 2018.

[56] M. Zaffar, A. Khaliq, S. Ehsan, M. Milford, and K. McDonald-Maier, "Levelling the playing field: A comprehensive comparison of visual place recognition approaches under changing conditions," in *Proc. IEEE Int. Conf. Robot. Automat., Workshop Database Gener. Benchmarking*, 2019.

[57] M. Zaffar, A. Khaliq, S. Ehsan, M. Milford, K. Alexis, and K. McDonald-Maier, "Are State-of-the-art visual place recognition techniques any good for aerial robotics?" in *Proc. IEEE Int. Conf. Robot. Automat., Workshop Aerial Robot.*, 2019.

[58] E. Olson and P. Agarwal, "Inference on networks of mixtures for robust robot mapping," *Int. J. Robot. Res.*, vol. 32, no. 7, pp. 826–840, 2013.

[59] G. Finlayson, C. Fredembach, and M. S. Drew, "Detecting illumination in images," in *Proc. IEEE 11th Int. Conf. Comput. Vis.*, Dec. 2007, pp. 1–8.

Mubariz Zaffar received the B.E. degree in electrical engineering from the National University of Sciences and Technology (NUST), Pakistan, in 2016. He is currently pursuing the M.Sc. degree in computer science and electronic engineering with the University of Essex. He is working as a Research Officer with the National Centre for Nuclear Robotics, U.K. His research interests include computer vision and deep learning for autonomous robotics, visual place recognition and robot navigation, SLAM, radiation effects on embedded systems, and resilience techniques for nuclear robotics. He was a recipient of the South-Asian Helix Innovation Award, the DICE Foundation Innovation Award, the IET Present-Around-The-World Regional Awards, and the NUST High-Achiever's Award.

Shoaib Ehsan (Senior Member, IEEE) received the B.Sc. degree in electrical engineering from the University of Engineering and Technology, Taxila, Pakistan, in 2003, and the Ph.D. degree in computing and electronic systems with a specialization in computer vision from the University of Essex, Colchester, U.K., in 2012. He is currently a Senior Research Fellow with the University of Essex. He has extensive industrial and academic experience in the areas of embedded systems, embedded software design, computer vision, and image processing. His current research interests include intrusion detection for embedded systems, local feature detection and description techniques, image feature matching, and performance analysis of vision systems. He was a recipient of the University of Essex Post Graduate Research Scholarship and the Overseas Research Student Scholarship. He is the Winner of the Prestigious Sullivan Doctoral Thesis Prize by the British Machine Vision Association.

Michael Milford (Senior Member, IEEE) received the bachelor's degree in mechanical and space engineering and the Ph.D. degree in electrical engineering from the University of Queensland, Brisbane, QLD, Australia. He is currently a Professor and an Australian Research Council Future Fellow with the Queensland University of Technology (QUT), Brisbane, and a Chief Investigator with the Australian Centre of Excellence for Robotic Vision. He was a Research Fellow on the Thinking Systems Project with the Queensland Brain Institute until 2010, where he became a Lecturer with QUT. He conducts interdisciplinary research into navigation across the fields of robotics, neuroscience, and computer vision. He received the Inaugural Australian Research Council Discovery Early Career Researcher Award in 2012 and became a Microsoft Research Faculty Fellow in 2013.

Klaus D. McDonald-Maier (Senior Member, IEEE) received the Dipl.-Ing. degree in electrical engineering from the University of Ulm, Germany, the M.S. degree in electrical engineering from the École Supérieure de Chimie Physique Électronique de Lyon, France, in 1995, and the Ph.D. degree in computer science from Friedrich Schiller University, Jena, Germany, in 1999. He worked as a Systems Architect on reusable microcontroller cores and modules with the Infineon Technologies AG's Cores and Modules Division, Munich, Germany, and a Lecturer in electronics engineering with the University of Kent, Canterbury, U.K. In 2005, he joined the University of Essex, Colchester, U.K., where he is currently a Professor with the School of Computer Science and Electronic Engineering. His current research interests include embedded systems and system-on-a-chip design, security, development support and technology, parallel and energy efficient architectures, and the application of soft computing and image processing techniques for real-world problems. He is a member of the Verband der Elektrotechnik Elektronik Informationstechnik and the British Computer Society, and a fellow of the Institution of Engineering and Technology.