CAMAL: Context-Aware Multi-scale Attention framework for Lightweight Visual Place Recognition

Ahmad Khaliq[1], Shoaib Ehsan[1], Michael Milford[2] and Klaus McDonald-Maier[1]

Abstract— In the last few years, Deep Convolutional Neural Networks (D-CNNs) have shown state-of-the-art performances for Visual Place Recognition (VPR). Their prestigious generalization power has played a vital role in identifying persistent image regions under changing conditions and viewpoints. However, against the computation intensive D-CNNs based VPR algorithms, lightweight VPR techniques are preferred for resource-constraints mobile robots. This paper presents a lightweight CNN-based VPR technique that captures multilayer context-aware attentions robust under changing environment and viewpoints. Evaluation of challenging benchmark datasets reveals better performance at low memory and resources utilization over state-of-the-art contemporary VPR methodologies.

Index Terms— Convolutional Neural Network, Context-based Regional Attentions, Robot Localization, Visual Place Recognition.

I. INTRODUCTION

Using a visual sensor, an ability of an object to correctly localize itself within a well-known environment through image matching is Visual Place Recognition (VPR) and building maps of the surrounding environment is technically termed as Simultaneous Localization and Mapping (SLAM). Under significant visual changes experienced in day-night and summer-winter transitions, successful VPR employing image retrieval techniques quite hard to achieve [1]. Such a system takes a query image as an input and returns a closely matched database image after applying an image processing technique followed by the feature matching. Feature extraction technique can either be handcraft-based (such as SURF [2], SIFT [3] and HOG [4]) or deep neural networks with rich spatial information from middle and late convolutional layers [5]; both have made an attempt to solve the VPR problem.

Employing a CNN pre-trained on task dependent dataset for finding meaningful region-based attentions has been an area of ongoing research including image classification/retrieval [6] [7]. Likewise in VPR problems [5] [8] [9], employing such CNNs which are pre-trained on external task-based datasets for finding cues based on vital image regions under environment changes has been a great interest
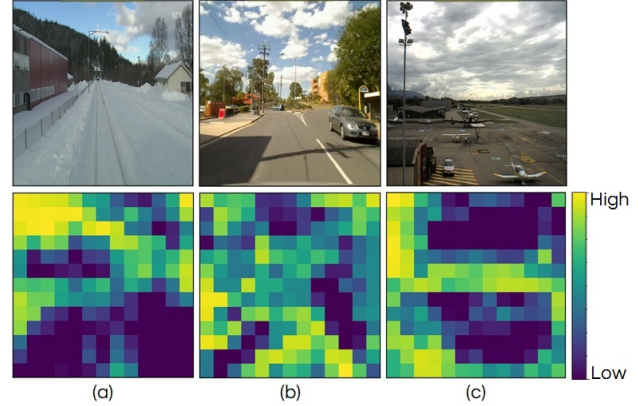
Fig. 1. In this paper, our proposed shallow CNN-based VPR technique identifies multi-semantic context-aware regional attentions under strong condition and appearance changes. Three exemplars are shown against which (a), (b) and (c) represent their identified novel multi-layer fused attentions. Our proposed framework concentrates on persistent image regions including building structures and road maps while filtering down the dynamic instances such as, vehicles, clouds and planes.

in robotics and computer vision communities [10]. However, such pre-trained CNNs are different in nature from recognizing the places under seasonal, lightning and viewpoint changes where activations are non-uniformly distributed over the convolution layers as compared to tasks [11] where a single object occupies the whole image.

Solving VPR problem in [12], Chen et al. employed a deep neural network VGG-16 [13] pre-trained on an object-centric ImageNet [11] and used middle convolutional layers for extracting features based on identified regions from late convolutional layers. Later in [14], Chen et al. fine-tuned the pre-trained object-centric VGG-16 [13] on VPR-centric SPED dataset [15]. A context-flexible block is integrated inside the fine-tuned deep feed forward neural network to learn context-based regions of interest (ROIs). Recently, Khaliq et al. in [16] have introduced a lightweight novel approach for extracting region-based CNN features from shallow CNN model. They have employed pre-trained scene-centric AlexNet365 [17] and used middle convolutional layer for regional features coupled with Vector of Locally Aggregated Descriptor (VLAD) [18] features encoding. The proposed Region-VLAD framework in [16] has shown boost in Area under Precision-Recall curves (AUC-PR) on several viewpoint- and condition-variant place-recognition datasets against the state-of-the-art contemporary VPR techniques including FAB-MAP [19], SeqSLAM [20], R-MAC [21], Cross-Region-BoW [12] and pooling methodologies like Sum- [22], Max- [21] and Cross-Pooling [23].

In VPR, focusing on dynamic entities other than static objects such as, road signs, buildings structures can instigate deceptive information in recognizing places. Despite the better real-time AUC-PR performance of the framework proposed in [12] [16], sometimes, it suffer with the inclusion of time-changing objects (such as pedestrians, vehicles, clouds) in the final region-based representations. To address this problem, we extend the idea of [16] to multi-layered region-based approach and integrated it within the shallow SPED-centric HybridNet [15]. The proposed framework captures powerful and rich semantic attentions where the attentions' areas vary with the image features. Employing D-CNNs, the authors in [15] [24] [25] [14] also attempt to learn fused multi-level context-aware regional features. However, improving VPR performance with D-CNNs does add computational and memory constraints in real-time robotic applications where response time is vital [10]. Figure 1 illustrates the novel multi-layered CNN-based regional attentions identified by our proposed lightweight M-Region-VLAD framework on three exemplars. Our main contributions of the work are as follow:

1) We have proposed a lightweight multi-scale context-aware attention approach for large scale environment-invariant VPR.

2) Taking precedence of shallow place recognition-centric HybridNet, the proposed M-Region-VLAD framework focuses on regional attentions which remain persistent under strong visual changes coupled with the confusing instances including cars and pedestrians.

3) A range of experiments on the challenging datasets exhibiting strong conditional and moderate viewpoint variations claim better performance against state-of-the-art D-CNNs based VPR approaches at low time and memory requirements.

The rest of the paper is organized as follows. Section II provides a literature review of both handcrafted and CNN-based VPR paradigms. In section III, we describe the proposed framework in detail. Section IV presents the experimental setup, detailed analysis and results obtained by evaluating the proposed framework on challenging benchmark datasets. Section V ends with the conclusions.

II. LITERATURE REVIEW

In a VPR system, Image processing is the first module involved in identifying and extracting distinguishing features. The early approaches consisted of human-made feature detection techniques [2] [3] [4], classified into local or global representations. Scale Invariant Feature Transform (SIFT) [3], a local feature detector that extracts and describes the keypoints using difference-of-gaussian and histogram-of-oriented-gradients. Other approaches includes HOG, SURF, FAST [26], GIST [27], FABMAP [19] and SEQSLAM [20]. FABMAP, a combination of SURF features with Bag-of-Words (BoW) [28] encoding scheme demonstrated robustness against viewpoint changes. SEQSLAM is another appearance-invariant VPR technique that subtracts patch-normalized frames captured in sequence, followed by a confusion matrix for best match retrieval.

The recent boom of deep learning in various computer vision applications inspires and opens up the research gate for the VPR community. Chen et al. in [5] for the first time used CNN-based features for the VPR problem. Later, [8] and [9] followed-up the work with a detailed analysis of middle and late convolutional layers' robustness for place recognition. Authors in [29] and [9] combined CNNs with external landmark-based approaches. All aforementioned techniques generally employed CNN models pre-trained on tasks other than place recognition. To compensate this research gap, Chen et al. in [15] have introduced and evaluated the performances of two place recognition-centric CNNs for VPR; AmosNet and HybridNet, pre-trained and fine-tuned the object-centric CaffeNet [11] on SPED dataset [15]. The results claimed that Spatial Pyramid Pooling (SPP) on late convolutional layers of HybridNet has shown better performance on several benchmark place-recognition datasets. Various feature pooling techniques including Sum-[22], Max- [21], Spatial Max- [15] and Cross-Pooling [23] for capturing the useful features from convolutional layers are proposed in image retrieval [7] and classification [6]. Arandjelovic et al. in [30] added a VLAD layer in the CNN architecture, named NetVLAD and trained the model end to end on severe condition- and viewpoint-variant datasets.

Recent VPR research work in [12] [14] [31] employed region-based features description of deep neural networks. Cross-Region-BoW [12] used a similar idea of cross-convolution [23] over late convolutional layers and employed 200 ROIs coupled with the 10k BoW dictionary. RMAC [21] also employed a regional approach based on maximum activations of convolution. Siagian et al. in [32] employed an attention-based regional approach for mobile robots. Authors in [33] and [6] further demonstrated that attention-based features can play an important role in improving vision-based robotics tasks. However, such attention capturing techniques require manually defined regional masks. Motivated from the work of [34] which overcomes the difficulty of manually employing a fixed regional attention mask over deep CNNs, Chen et al. in [14] have integrated a context-flexible attention block in a deep object-centric VGG-16 and fine-tuned it on condition-variant SPED dataset. The proposed system was trained end-to-end specifically for VPR under severe conditional changes. However, all the aforementioned contemporary VPR techniques used deep VGG-16 models and employed late convolutional layers for features extraction which means more memory and computational resources required at execution time.

Khaliq et al. in [16] bridged this research gap with a lightweight but manually designed CNN-based regional approach which can be incorporated within any CNN model. At cheap time-computation and resource utilization, the Region-VLAD approach in [16] employed middle convolutional layers of AlexNet365 and shown state-of-the-art performance over [12] in terms of AUC under PR-curves. Despite its better matching performance, at higher regional features,
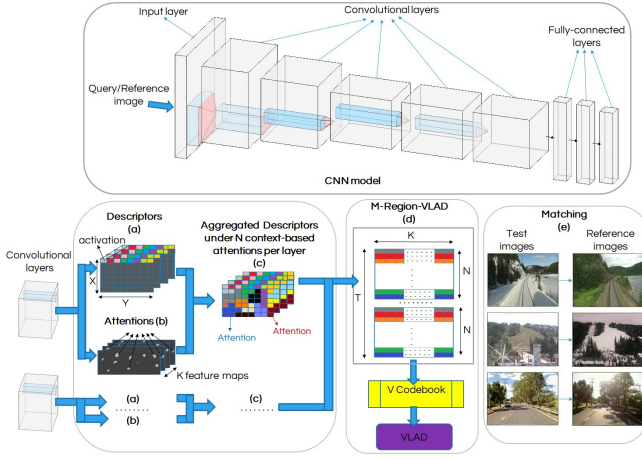
Fig. 2. Block diagram of the proposed VPR framework is shown here. N Regions of Interest (ROIs) per convolutional layer is identified, followed by local descriptors-based regional aggregation. T aggregated regional attentions across the multiple layers are used as features representation.

Region-VLAD [16] VPR framework sometimes encounters dynamic instances (such as, vehicles, clouds) in the captured regional features. Authors in [33] [15] [24] demonstrated that attentions based on multiple convolutional layer can provide more powerful feature representations. Taking inspiration from [16] [33], we have optimized the regional approach of [16] and integrated at multiple convolutional layers of shallow SPED-centric HybridNet. We have shown in the experiments that our proposed M-Region-VLAD framework captured meaningful and static structures invariant to strong conditional and appearance changes. At low time and memory utilization, evaluation of our proposed M-Region-VLAD framework on several benchmark place-recognition datasets have shown better and comparable matching performance in terms of Area under Precision-Recall (AUC-PR) curves over state-of-the-art D-CNNs based VPR approaches [21] [30] [12] [14] [16].

III. PROPOSED TECHNIQUE

This section describes the proposed framework in more detail. To subdivide an image into spatial regional representations, we first discuss the retrieval of local descriptors from the convolutional feature maps. We then demonstrate our approach of finding regional attentions from multiple convolutional layers, followed by the discussion on how to aggregate and map the regional local descriptors on a separate regional vocabulary to obtain a compact VLAD representation. The overall framework is shown in Figure 2.

A. Stacking of Convolutional Layers Activations for making Descriptors

In a neural network, $X \times Y \times K$ is the dimension of $3D$ convolutional layer tensor M, where X and Y represent the width and height of each channel and K is the number of channels, also termed as feature maps. In layman terms, each feature map $k = \{1, 2, ..., K\}$ corresponds to some filter being convolve on the input image I. At certain spatial location, we stack down the activations of K feature maps, and each

spatially stacked activations vector is termed as a local descriptor, visually shown in Figure 2(a). In (1), D^L denotes the K dimensional local descriptors at L^{th} convolutional layer of m_c model.

$$D^L = \{d_l \in M^K \quad \forall \, l \in \{(i,j) \mid i = 1, ..., X; j = 1, ..., Y\}\} \quad (1)$$

B. Identification of Context Aware Regional Attentions

Within the convolutional layer of a CNN, certain spatial regions of the feature maps do have more intensity mimicking the presence of certain visual patterns in the image such as road signal, buildings etc. For finding context-based most contributing regions in an image of the place, we employed a shallow pre-trained SPED-centric HybridNet [15]. Particularly, we process the feature maps of the convolutional layer and grouped the non-zero spatially connected activations such that two or more activations couple to represent a G_h salient region if roughly have similar responses, $\forall \, h \in \{1, ..., H\}$ where H is the total number of identified salient attentions from K feature maps at L^{th} convolution layer (visualized in Figure 2(b)). Similar to [16], energies of all the identified regional attentions are calculated by averaging over all the a_h activations lying under each G_h attention. In (2), a_h^s represents the sth activation lying under G_h region where E^L denotes the regional energies. In (3), with sorted E^L energies, R^L represents the top N energetic novel context-based ROIs..

$$E^L = \{\frac{1}{|G_h|} \sum_s a_h^s, \, \forall \, a_h^s \in G_h\} \quad (2)$$

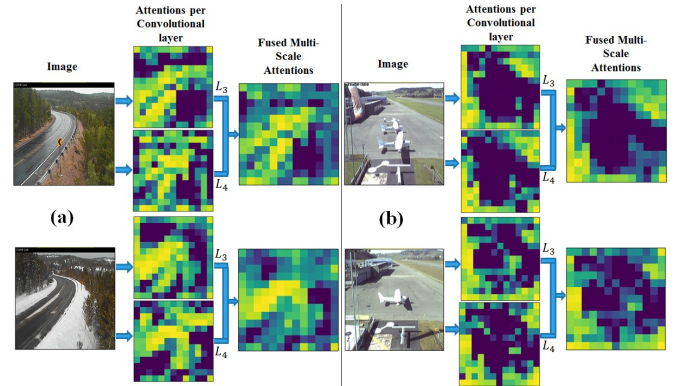$$R^L = \{G_t \, \forall \, t \in \{1, ..., N\}\} \quad (3)$$

Fig. 3. Fused multi-scale attentions captured under strong conditional changes coupled with dynamic instance experienced by the place (a) and (b) under different times of the year.

Considering real-time performance and to forbade the inclusion of time-varying objects in the final region-based features, $N = 300$ attentions per layer are captured. It is because with the inclusion of more but less energetic regions, activations concentrated on dynamic objects do get included. Experimentation at $N = 300$ confirms minimum dynamic instances in the captured regional representations. Under the q identified attention, D_q^L denotes the underlying regional local descriptors, aggregated in (4) to retrieve $N \times K$ dimensional f^L context-based regional features. Given an image I, F_I in (5) represents the concatenated $T \times K$ attentions captured

from L_3 and L_4 convolutional layers of the model, illustrated in Figure 2(d). Based on the energies, the fused attentions are sorted as illustrated in Figure 3; attentions captured from middle $L_3 = conv3$ and late $L_4 = conv4$ are fused. It is worth noting that the fused multi-scale attentions reduce down the impact of attentions focusing on dynamic instances (such as plane and tree) captured by the individual convolutional layers. With multi-scaling, we captured low and high level image regions that are persistent under changing conditions.

$$f^L = \{ \sum_{q \in R_t^L} D_q^L \ \forall \ t \in \{1,...,N\}\} \tag{4}$$

$$F_I = \{f^l \ \forall \ l \in \{L_3, L_4\}\} \tag{5}$$

C. Attention based Vocabulary and Extraction of VLAD for Image Matching

With smaller visual word vocabulary in tasks including image retrieval, recognition and object detection [21] [23], Vector of Locally Aggregated Descriptors (VLAD) [18] has shown state-of-the-art performance. Similar to [16], for attention-based dictionary, we have collected a separate dataset of $3K$ images which contains 1125 Query247 [35] images taken in day, evening and night times of 365 places. The other images consist of Garden Point [15], Eynsham [5] and multiple environment variant rural and urban road traverses captured from *Mapillary* [9] [12]. K-means is used to cluster $3000 \times T \times K$ dimensional context-aware attentions into $V = 128$ regions. For all the benchmark test and reference frames, their attentions are quantized to predict the dictionary clusters/labels. Similar to [16], the VLAD descriptor is obtained using the multi-layer context-aware attentions, predicted labels and attention-based pre-trained vocabulary.

IV. Datasets, Implementation details, Results and Analysis

This section discusses the benchmark datasets employed to determine the proposed framework efficiency of recognizing places under strong environment changes against the state-of-the-art VPR contemporary techniques. We first highlight the run-time implementation details followed with the discussion on the performance evaluation. We then compare the context-based attentions identified by our approach and state-of-the-art VPR techniques [14] [16].

A. Benchmark Place Recognition Datasets

For evaluating our proposed VPR system, we have targeted the three challenging place recognition benchmark datasets (please see Table I). All the datasets have captured two traverses along the same route taken at multiple times of the day/year under diverse range of environments which do exhibit scenarios experience by robots in real world. The first traverse is used for testing and the second traverse is served as reference frames. The St.Lucia dataset [15] was captured in the suburban route at multiple day times with sufficient viewpoint- and condition-variation. The original GPS annotation with the St.Lucia dataset employed to build place and frame level correspondence, used as Ground truth. The SPEDTest [14] is the newly introduced dataset which

contains very divergent scenarios captured with surveillance cameras in different year times (for more information, please see [14]). There is a strong illumination changes with mild viewpoint variance and for the ground truth, each test image in SPEDTest resembles with three known reference images provided with the dataset. The Synthesized Nordland dataset [14] is a modified version in which viewpoint variance is introduced by cropping frames to keep 75% resemblance. It's a train journey in winter and summer seasons, with frame and place level resemblance is used as ground truth.

TABLE I
BENCHMARK PLACE RECOGNITION DATASETS

Dataset	Traverse		Environment	Variation	
	Test	Reference		Viewpoint	Condition
St. Lucia	1249	1249	Suburban	Adequate	Significant
SPEDTest	607	1821	Diverse	Moderate	Strong
Synthesized Nordland	1622	1622	Train journey	Moderate	Very Strong

B. Setup and Implementation details

Deep learning techniques are computationally expensive which makes it indispensable to evaluate the run-time performance in order to realize the system's deployment in robotic VPR applications. The presented VPR framework is implemented in Python 3.6.4 and the system average run-time over 3 iterations with 3244 images is recorded. For all the baseline experiments, we employed HybridNet and used middle *conv3* and late *conv4* convolution layers to capture rich semantic context-aware regional features. For an image, the forward pass takes an average $M_f = 13.85 \ ms$ using Caffe on **Intel Xeon Gold 6134 @3.2GHz**. Other parameters including $N = 300$ attentions per layer with $V = 128$ clustered vocabulary for VLAD encoding. Extraction of T context-aware attentions per image takes around $M_e = 140.5 \ ms$ with VLAD encoding and (two VLADs) matching takes $M_v = 2.68 \ ms$ and $M_m = 0.07 \ ms$ [16]. Therefore, let say with $R = 1622$ reference VLAD representations, the total retrieval time M_q for a single query against R stored database VLADs can be calculated using (6), comes around $270.57 \ ms$. 128×384 dimensional VLAD representation per image consumes around $393KBytes$ memory.

$$M_q = M_f + M_e + M_v + M_m * R \tag{6}$$

In comparison, memory and time computation for NetVLAD, RMAC, Region-VLAD, Cross-Region-VLAD and SeqSLAM are higher than the proposed M-Region-VLAD framework, as reported in [36]. Employing **Titan X 1080 GPU**, state-of-the-art Context Flexible Attention [14] is evaluated on 1101 images and takes around $M_f + M_e = 14.1 \ ms$ $(M_v = 0)$ per image for features extraction. The $512 \times 14 \times 14$ dimensional feature vector consists of multi-scale fused attentions, consumes $401KBytes$ memory. Using Python 3.6.4, feature matching is performed by flattening the 3D vector, followed up with cosine distance matching further takes an average $M_v = 0.63 \ ms$ employing Intel Xeon Gold 6134 @3.2GHz. Therefore, an overall retrieval time for matching a single query against $R = 1622$ reference images takes around 1035.96 ms. It should be noted that

Context Flexible Attention [14] employed GPU for feature extraction using SPED-centric deep VGG-16 whereas our M-Region-VLAD framework employed SPED-centric shallow HybridNet and used CPUs for extracting context-aware regional features but with more resources, the retrieval time can further be improved, illustrated in Table II.

TABLE II
FEATURE ENCODING AND MATCHING TIMES OF THE VPR APROACHES.

Techniques	Feature Encoding (ms)	Feature Matching (ms)	Techniques	Feature Encoding (ms)	Feature Matching (ms)
Intel Xeon(R) Gold 6134 CPU @ 3.20GHz with 32 cores, 64GB RAM					
SeqSLAM	0	1.5	Cross-Region -BoW	830	160
NetVLAD	770	0.005	RMAC	470	0.04
Region-VLAD	460	0.12	M-Region-VLAD	157.03	0.07
Techniques	Feature Encoding (ms)		Feature Matching (ms)		
	Titan X 1080 GPU		Intel Xeon(R) Gold 6134 CPU @ 3.20GHz with 32 cores, 64GB RAM		
Context Flexible Attention	14.1		0.63		

C. Comparison Methods

To make a fair comparison, we also reported the performance of other VPR approaches evaluated in [14] that includes Attentive Attention [37], Cross-Pool, FABMAP, Fix-Context [25], Context Flexible Attention, Places365 [17] and SEQSLAM. Particularly, for state-of-the-art Attentive Attention approach and VPR-based Fix-Context framework, Chen et al. [14] have fine-tuned these models on SPED dataset while removing the geometric verification layer. For Cross-Pool [23], the late convolutional layer is employed to generate a fixed attention mask, used as features representations. For handcraft-based VPR approaches which include FABMAP and SEQSLAM, the authors employed their official implementations [38] [20]. Places365 is a CNN model pre-trained on *2 Million* diverse scenes. The authors used responses of the late fully-connected convolution layer as features representation.

Furthermore, other CNN-based VPR algorithms such as NetVLAD, RMAC, Cross-Region-BoW and Region-VLAD are also evaluated. For Region-VLAD, $N = 200$ regions are employed from *conv3* of AlexNet365 with $V = 128$ clustered vocabulary for VLAD retrieval [16]. All other approaches used VGG-16 pre-trained on object-centric ImageNet. Their layers configuration are kept same as in [36]; *conv5_2* is used for RMAC, with power- and l2-normalization on the regional features. For Cross-Region-BoW, *conv5_2* and *conv5_3* are employed with 10*k* BoW dictionary. For both the techniques, cosine matching is performed for filtering the mutual regions and their scores are summed and database image with highest score considered as matched place. Given an image, NetVLAD outputs a feature descriptor and cosine matching of the feature descriptors is performed with scores summation and reference image with highest score represents the currently encountered place.
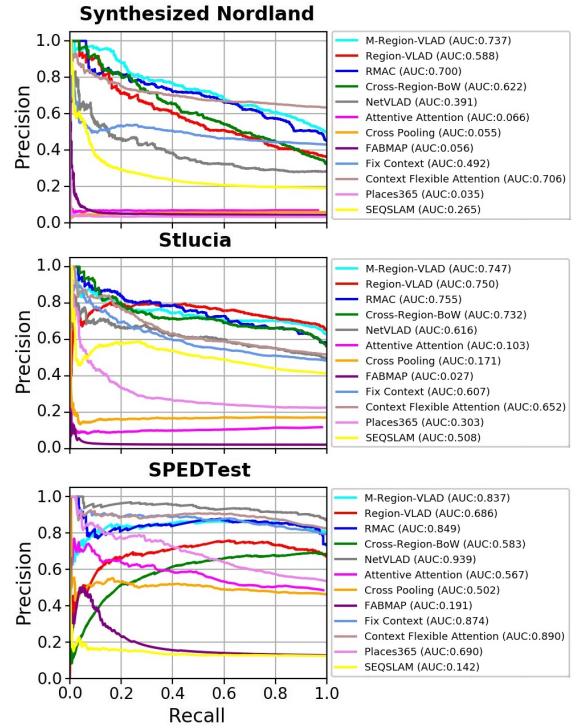

Fig. 4. Area under Precision-Recall curves for all the benchmark datasets on the contemporary VPR approaches are presented here.

D. Precision Recall Characteristics

For all the benchmark place recognition datasets, Area under Precision-Recall curves [39] (AUC-PR curves) is used for evaluating the proposed place recognition framework, state-of-the art image retrieval and VPR-based contemporary approaches (as mentioned in IV-C).

More area the PR-curve covers, better the performance of the technique. Figure 4 displays the AUC-PR curves for the benchmark datasets on the employed approaches. It is quite evident that for Synthesized Nordland and St.Lucia datasets, our proposed VPR approach has shown the best performance. Comparing from other datasets, St.Lucia exhibits moderate appearance change coupled with an appropriate viewpoint variation. A closer look at M-Region-VLAD results confirm that the system identified and captured context-based salient regions and boostup the overall retrieval performance, as illustrated in Figure 5. It mimics that employment of multiple convolutional layers is very productive under strong conditional changes.

For St.Lucia dataset, Region-VLAD, Cross-Region-BoW and RMAC have shown similar performance as M-Region-VLAD. However, their performance degrades for SPEDTest and Synthesized Nordland which experience strong seasonal and conditional changes. It suggests that under moderate conditions, all these regions-based techniques focus on place-recognition centric regions which results into better recognition performance. NetVLAD showcases nearly the similar PR-characteristic as Fixed Context and Context Flexible Attention. Better and comparable performance of M-Region-VLAD on all the dataset highlights the usefulness and generalization power of shallow attentions over deeply learned

context flexible salient representations [14].

In Figure 4, it is worth noticing that NetVLAD which underperformed under Synthesized Nordland, has shown state-of-the-art performance on SPEDTest. Although both the datasets exhibit severe condition-variation among the traverses. One of the reason could be the existence of perceptual aliasing in Synthesized Nordland i.e. much resemblance among the sequentially captured frames. For SPEDTest, the environment of the test images is very diverse and each has only three matched images in the reference traverse. Majority of the techniques perform well on this dataset. In comparison, our proposed M-Region-VLAD achieves comparable AUC under PR curve against deep Context Flexible Attention, RMAC and Fixed Context frameworks. Cross-Region-BoW has shown an average performance both on SPEDTest and Synthesized Nordland. It is observed that due to ImageNet-centric training of VGG-16, the cross-convolutional regions-based approach concentrates more on objects. As expected, Region-VLAD which is integrated with AlexNet365 exhibits a comparable performance for SPEDTest and Synthesized Nordland. It is probably because the model is pre-trained on scene-centric Place365 dataset and with novel region finding approach, it sometime considers dynamic instances e.g. sky as a valuable region for distinguishing the scene which leads to place mismatch, also shown in Figure 5.

SPEDTest dataset is a subset of SPED [15] but has not been used to train the models. A deep analysis suggests that although HybridNet [15] and Context Flexible Attention [14] models are fine-tuned on SPED dataset but training parameters such as, learning rates are kept different; dual learning rates approach was employed in [14]. Same goes for the weight decays and iterations which also differ from the values set for HyridNet and SPED-centric VGG-16. Also, employing three convolutional layers, deep multi-scale features of Context Flexible Attention [14] can be more robust against condition-invariance and hence, exhibits better performance for this datasets. However, under seasonal changes coupled with perceptual aliasing (Synthesized Nordland), the performance degrades. It should be noted that our proposed M-Region-VLAD approach employed only two convolutional layers of HybridNet and still delivers a comparable performance across all the datasets which mimics the generalization power at low computation and memory needs.

It is visible that the worst performance of FABMAP is consistent throughout the datasets. It is because FAPMAP used viewpoint-invariant SURF feature detector which is sensitive under condition and appearance changes. It is interesting that SEQSLAM with its better appearance tackling and whole image-based matching approach shown inferior performance under SPEDTest. It is probably due to the fact that the places exhibit diverse environment and sequence-based matching requirement is violated. Despite the better performances of Cross-Pool and Attention Attentive approach in other vision-based tasks, they under-performed in St.Lucia and Synthesized Nordland. This highlights the difference in other image retrieval/classification systems from
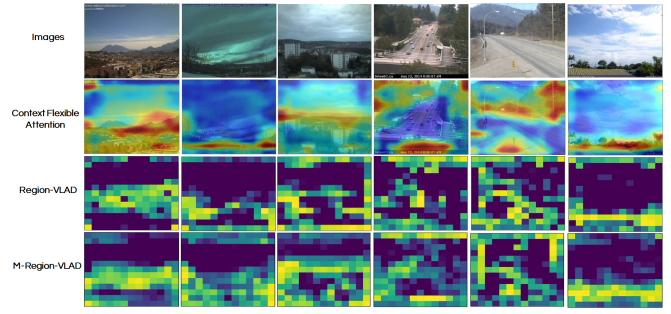

Fig. 5. Sample context-aware regional attentions identified by Context Flexible Attention [14], Region-VLAD [16] and our proposed M-Region-VLAD.

place recognition where convolutional layers' responses are non-uniformly distributed and the place is subdivided into multiple contributing salient regions. However, their better performances under SPEDTest point towards the importance of CNN training. Fixed Context and Places365 exhibit better results for St.Lucia and SPEDTest only. This implies that both the approaches are sensitive under perceptual aliasing experienced in Synthesized Nordland.

Furthermore, to analyse and differentiate the multi-semantic attentions captured by our proposed M-Region-VLAD framework against the state-of-the-art Context Flexible Attention [14] and Region-VLAD [16], Figure 5 shows some of the sample places with their corresponding salient regions. Both Context Flexible Attention and M-Region-VLAD find most distinguishing structures, such as, houses, street lights as novel attentions while filtering out confusing areas including clouds, vehicles etc. It is evident that Region-VLAD sometime includes sky and other dynamic instances as vital regions. It is worth noticing that M-Region-VLAD captures meaningful and place-centric spatial regions from a shallow CNN architecture against long-term condition and seasonal variations. Datasets and results are placed at [40].

V. CONCLUSION

Despite the recent state-of-the-art performances of D-CNNs for VPR, the high computation and memory cost limit their practical deployment for battery-operated mobile robots. Achieving superior performance with shallow CNN architectures is thus desirable, but a challenging problem. In this paper, a multi-scale context-aware attention approach is presented that combines salient regions from multiple convolutional layers of a light-weight CNN architecture. The proposed approach captures persistent regional features under changing conditions and viewpoints while filtering down the confusing instances including sky, moving objects etc. Evaluation on several challenging benchmark datasets confirms the dominance over state-of-the-art algorithms in terms of area under precision-recall curves.

In future, we will incorporate the proposed multi-scale attention block in a shallow feed forward neural network and fine-tune the CNN model on a large-scale place recognition dataset. It should reduce the feature encoding time and the system should learn image regions invariant to strong viewpoint and condition variations.

REFERENCES

[1] S. Lowry and et al., "Visual place recognition: A survey," *IEEE Transactions on Robotics*, vol. 32, no. 1, pp. 1–19, 2015.

[2] H. Bay and et al., "Surf: Speeded up robust features," in *ECCV*. Springer, 2006, pp. 404–417.

[3] Y. Ke, R. Sukthankar *et al.*, "Pca-sift: A more distinctive representation for local image descriptors," *CVPR (2)*, vol. 4, pp. 506–513, 2004.

[4] N. Dalal and et al., "Histograms of oriented gradients for human detection," in *CVPR*, vol. 1. IEEE Computer Society, 2005, pp. 886–893.

[5] Z. Chen and et al., "Convolutional neural network-based place recognition," arXiv:1411.1509 [cs.CV], 2014.

[6] F. Wang and et al., "Residual attention network for image classification," in *IEEE CVPR*, 2017, pp. 3156–3164.

[7] P. Wang and et al., "Multi-attention network for one shot learning," in *IEEE CVPR*, 2017, pp. 2721–2729.

[8] N. Sünderhauf and et al., "On the performance of convnet features for place recognition," arXiv:1501.04158 [cs.RO], 2015.

[9] N. Sünderhauf, S. Shirazi, and et al., "Place recognition with convnet landmarks: Viewpoint-robust, condition-robust, training-free," *Proceedings of RSS XII*, 2015.

[10] C. Cadena and et al., "Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age," *IEEE T-RO*, vol. 32, no. 6, pp. 1309–1332, 2016.

[11] A. Krizhevsky and et al., "Imagenet classification with deep convolutional neural networks," in *Advances in NIPS*, 2012, pp. 1097–1105.

[12] Z. Chen and et al., "Only look once, mining distinctive landmarks from convnet for visual place recognition," in *IROS*. IEEE, 2017, pp. 9–16.

[13] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv:1409.1556 [cs.CV], 2014.

[14] Z. Chen and et al., "Learning context flexible attention model for long-term visual place recognition," *IEEE RAL*, vol. 3, no. 4, pp. 4015–4022, 2018.

[15] Z. Chen, A. Jacobson, and et al., "Deep learning features at scale for visual place recognition," in *ICRA*. IEEE, 2017, pp. 3223–3230.

[16] A. Khaliq and et al., "A holistic visual place recognition approach using lightweight cnns for severe viewpoint and appearance changes," arXiv:1811.03032 [cs.RO], 2018.

[17] B. Zhou and et al., "Places: A 10 million image database for scene recognition," *IEEE transactions on pattern analysis and machine intelligence*, 2017.

[18] H. Jégou and et al., "Aggregating local descriptors into a compact image representation," in *CVPR*. IEEE, 2010, pp. 3304–3311.

[19] M. Cummins and P. Newman, "Fab-map: Probabilistic localization and mapping in the space of appearance," *The International Journal of Robotics Research*, vol. 27, no. 6, pp. 647–665, 2008.

[20] M. J. Milford and et al., "Seqslam: Visual route-based navigation for sunny summer days and stormy winter nights," in *ICRA*. IEEE, 2012, pp. 1643–1649.

[21] G. Tolias and et al., "Particular object retrieval with integral max-pooling of cnn activations," arXiv:1511.05879 [cs.CV], 2015.

[22] A. Babenko and et al., "Aggregating local deep features for image retrieval," in *ICCV*, 2015, pp. 1269–1277.

[23] L. Liu and et al., "Cross-convolutional-layer pooling for image recognition," *IEEE T-PAMI*, vol. 39, no. 11, pp. 2305–2313, 2017.

[24] X. Chu and et al., "Multi-context attention for human pose estimation," in *IEEE CVPR*, 2017, pp. 1831–1840.

[25] H. J. Kim and et al., "Learned contextual feature reweighting for image geo-localization," in *CVPR*, 2017, pp. 2136–2145.

[26] E. Rosten and T. Drummond, "Machine learning for high-speed corner detection," in *ECCV*. Springer, 2006, pp. 430–443.

[27] A. Oliva and et al., "Building the gist of a scene: The role of global image features in recognition," *Progress in brain research*, vol. 155, pp. 23–36, 2006.

[28] J. Sivic and et al., "Video google: A text retrieval approach to object matching in videos," in *null*. IEEE, 2003, p. 1470.

[29] P. Panphattarasap and et al., "Visual place recognition using landmark distribution descriptors," in *ACCV*. Springer, 2016, pp. 487–502.

[30] R. Arandjelovic and et al, "Netvlad: Cnn architecture for weakly supervised place recognition," in *CVPR*, 2016, pp. 5297–5307.

[31] J. M. Facil and et al., "Condition-invariant multi-view place recognition," arXiv:1902.09516 [cs.CV], 2019.

[32] C. Siagian and et al., "Biologically inspired mobile robot vision localization," *IEEE TRO*, vol. 25, no. 4, pp. 861–873, 2009.

[33] L.-C. Chen and et al., "Attention to scale: Scale-aware semantic image segmentation," in *IEEE CVPR*, 2016, pp. 3640–3649.

[34] H. J. Kim and et al., "Learned contextual feature reweighting for image geo-localization," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017, pp. 3251–3260.

[35] A. Torii and et al, "24/7 place recognition by view synthesis," in *CVPR*, 2015, pp. 1808–1817.

[36] M. Zaffar, A. Khaliq, and et al., "Levelling the playing field: A comprehensive comparison of visual place recognition approaches under changing conditions," *arXiv preprint arXiv:1903.09107*, 2019.

[37] D. Yu and et al., "Multi-level attention networks for visual question answering," in *CVPR*. IEEE, 2017, pp. 4187–4195.

[38] M. Cummins, "Highly scalable appearance-only slam-fab-map 2.0," *Proc. Robotics: Sciences and Systems (RSS), 2009*, 2009.

[39] J. A. Hanley and et al., "The meaning and use of the area under a receiver operating characteristic (roc) curve." *Radiology*, vol. 143, no. 1, pp. 29–36, 1982.

[40] "Results and datasets," https://github.com/Ahmedest61/M-Region-VLAD-VPR/.