A Holistic Visual Place Recognition Approach using Lightweight CNNs for Severe ViewPoint and Appearance Changes

Preprint · November 2018

CITATIONS
0

READS
363

4 authors:

Ahmad Khaliq
University of Essex
9 PUBLICATIONS **48** CITATIONS
SEE PROFILE

Shoaib Ehsan
University of Essex
94 PUBLICATIONS **564** CITATIONS
SEE PROFILE

Michael Milford
Queensland University of Technology
288 PUBLICATIONS **6,838** CITATIONS
SEE PROFILE

Klaus McDonald-Maier
University of Essex
257 PUBLICATIONS **2,448** CITATIONS
SEE PROFILE

A Holistic Visual Place Recognition Approach using Lightweight CNNs for Severe ViewPoint and Appearance Changes

Ahmad Khaliq[1], Shoaib Ehsan[1], Zetao Chen[2], Michael Milford[3] and Klaus McDonald-Maier[1]

***Abstract*—This paper presents a lightweight visual place recognition approach, capable of achieving high performance with low computational cost, and feasible for mobile robotics under severe viewpoint and appearance changes. Results on several benchmark datasets confirm an average boost of 10% in accuracy, and 5% average speedup relative to state-of-the-art methods.**

***Index Terms*—Convolutional Neural Network, Feature Encoding, Robot Localization, Vector of Locally Aggregated Descriptors, Visual Place Recognition.**

I. INTRODUCTION

Given a query image, an image retrieval system aims to retrieve all images within a large database that contain similar objects as in the query image. Visual Place Recognition (VPR) can also be interpreted as an image retrieval system that tries to recognize a place by matching it with the places from the stored database [1]. A place database is a simplest way to represent a particular environment where appearance based information is stored with no pose related data. However, other VPR techniques use topological maps which contain relative information about the places in an environment (can be an ordered collection of images) and metric maps which are even more accurate in terms of absolute scale of the environment (such as distance, landmark position) but difficult to build and maintain. Sequence of images- and single image-matching are the two techniques employed by the VPR community. In this paper, we focus on database-centric place remembering approach with single image matching, thus, place recognition is solely based on appearance similarity and image retrieval techniques are applicable [2].

As with a range of other computer vision applications, deep learned CNNs have shown promising results for VPR and managed to shift the focus from traditional handmade features techniques [3][4] to CNNs [5]. Using a pre-trained CNN for VPR, there are three standard approaches to produce a compact image representation: (a) the entire image is directly fed into the CNN and its layers responses are extracted [5]; (b) the CNN is applied on the user-defined regions of the image and prominent activations are pooled from the layers representing those regions [6]; (c) the entire image is fed into the CNN and salient regions are identified by directly extracting distinguishing patterns based on convolutional layers responses [7][8]. Generally, category (a) results in global image representations which are not robust against severe viewpoint variations and partial occlusion. Image representations emerging from category (b) usually handle viewpoint changes better but are computation intensive. On the other hand, image representations resulting from category (c) address both the appearance and viewpoint variations. In this paper, we focus on category (c).

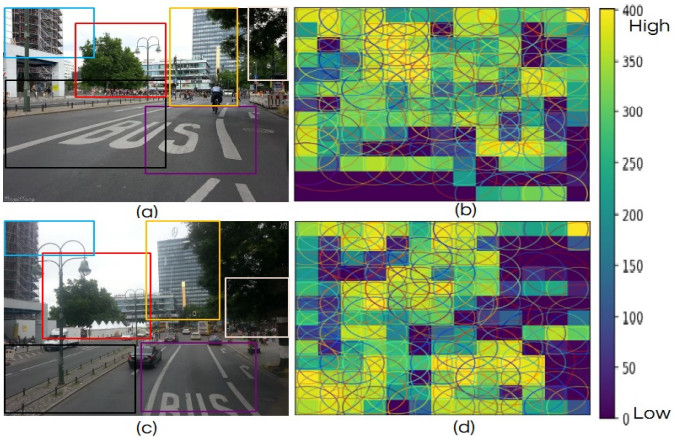

Fig. 1. For a query image (a), the proposed Region-VLAD approach successfully retrieves the correct image (c) from a stored image database under severe condition- and viewpoint-variation. (b) and (d) represent their CNN based meaningful regions identified by our proposed methodology.

The work by [7] and [8] are considered as state-of-the-arts in identifying prominent regions by directly extracting unique patterns based on convolutional layers' responses. In [7], the authors used VGG-16 network [9] pre-trained on ImageNet [10] and used late convolutional activations for regions identification. For regional features encoding, 10k bag-of-words (BoW)[11] codebook is employed. The system is tested on five benchmark place recognition datasets with AUC-PR curves [12] as the evaluation metric. It claims to outperform the FABMAP [13], SEQSLAM [14] and other image retrieval pooling techniques including Cross-Pooling [15], Sum/Average-pooling [16] and Max-pooling [17].

Despite its good AUC-PR performance, the method proposed in [7] has some shortcomings. A common strategy for improving CNN accuracy is to make it deep by adding more layers (provided sufficient data and strong regularization). However, increasing network size means more computation and using more memory both at training and test time (such as for storing outputs of intermediate layers and for storing parameters) which is not ideal for resource-constrained robots that are usually battery-operated. Utilizing late convolutional layers of deep VGG-16 for features extraction followed by regions' cross-matching while employing 10k BoW dictionary degrades the realtime performance. On the other hand, the employment of the CNN model pre-trained on an object-centric database in [7] results in CNN trying to put more emphasis on objects rather than the place itself. This reflects on the regional pooled features and leads to failure cases. Also, the regional approach proposed in [7] hinders the identification of individual static place-centric regions that can be more effective under condition and viewpoint variations.

To bridge these research gaps, this paper proposes a holistic approach targeted for a CNN architecture comprising a small number of layers pre-trained on a scene-centric [18] image databases to reduce the memory and computational costs. The proposed method detects novel CNN-based regional features and combines them with VLAD [19] adapted specifically for localization based VPR problem. The motivation behind employing VLAD comes from its better performance in various CNN-based image retrieval tasks utilizing a smaller visual word dictionary [19][20] compared to BoW [11]. To the best of our knowledge, this is the first work that combines novel lightweight CNN-based regional features with VLAD encoding adapted for computation-efficient VPR.

This work is supported by the UK Engineering and Physical Sciences Research Council through grants EP/R02572X/1 and EP/P017487/1

[1]Authors are with Embedded and Intelligent System Lab in Computer Science and Electronic Engineering department, University of Essex, Colchester, United Kingdom {ahmad.khaliq,sehsan,kdm}@essex.ac.uk

[2]Zetao Chen is with the Vision for Robotics Lab, ETH Zurich, Zurich 8092, Switzerland chenze@ethz.ch

[3]Michael Milford is with Electrical Engineering Science and Engineering Faculty, Electrical Engineering, Computer Science, Robotics and Autonomous Systems, Queensland University of Technology, Australia michael.milford@qut.edu.au

As opposed to [7] which uses object-centric VGG-16 architecture and employed a cross-convolution based regional extraction approach (resembles with [15]), the proposed VPR technique is particularly different both in identifying and extracting regional features (will discuss more in detail in section III-B). The presented approach in this paper showcases enhanced accuracy by employing middle convolutional layer of the CNN architecture comprising with small number of layers. Evaluation on several viewpoint- and condition-variant benchmark place recognition datasets shows an average performance boost of 10% over state-of-the-art VPR algorithms in-terms of AUC computed under Precision-Recall curves. In Figure 1, for a query image (a), our proposed system retrieved image (c) from the stored database. (b) and (d) highlight the salient regions which our proposed methodology identified under severe viewpoint- and condition-variation.

The rest of the paper is organized as follows. Section II provides related work for VPR and other image retrieval tasks. In Section III, the proposed methodology is presented in detail. Section IV illustrates the implementation details and the results achieved on several benchmark datasets. Section V presents the conclusion.

II. Literature Review

This section provides an overview of major developments in VPR under simultaneous viewpoint and appearance changes using hand-crafted features and CNN-based features. Other image retrieval tasks with their feature extracting and encoding approaches are further discussed and differentiated from VPR based image retrieval tasks.

FAB-MAP [13] is the first work that used handcrafted features (more specifically, SURF features) combined with BoW encoding for VPR. It demonstrated robustness under viewpoint changes by taking advantage of the in-variance properties of SURF. Another work based on sequence matching of the images named SEQSLAM [14] achieved remarkable results under severe appearance changes. However, it is unable to deal with simultaneous condition- and viewpoint-variation.

The first CNN-based VPR system is introduced in [5], which is followed by [21], [6] and [22]. In [5], the authors used Overfeat [23] trained on ImageNet. Eynsham [13] and QUT datasets with multiple traverses of the same route under environmental changes are used as benchmark datasets. Using the Euclidean distance on the pooled layers responses, test images are matched against the reference images. On the other hand, [22] and [6] used landmark-based approaches coupled with the pre-trained CNN models. In [24], the authors introduced two CNN models for the specific task of VPR (named AMOSNet and HybridNet) which are trained and fine-tuned the original object-centric CaffeNet [10] on place-recognition centric SPED dataset (2.5 million images). SPED dataset consists of thousands of places with severe-condition variance among the same places over different times of the years. The results showed that with Spatial Pyramidal Pooling (SPP) employed on middle and late convolutional layers, HybridNet outperformed AMOSNet, CaffeNet and PlaceNet on four publicly available datasets exhibiting strong appearance and moderate viewpoint changes [24]. [7] presented an approach that identifies pivotal landmarks by directly extracting prominent patterns based on responses of the later convolutional layers of a deep object-centric VGG-16 neural network for VPR. Recently, [8] introduced a context-flexible attention model and combines it with a pre-trained object-centric deep VGG-16 model fine-tuned on SPED dataset [24] to learn more powerful condition-invariant regional features. The system has shown state-of-the-art performance on severe condition-variant datasets. However, the efficiency of the proposed approach may be compromised if there be a simultaneous severe viewpoint and condition variations. Moreover, performance and efficient resource usage become two important aspects to be looked upon in real-life robotic VPR applications.

Image retrieval tasks which either rely on handcrafted features, such as, local SIFT and SURF features [3][4] or combining these with convolutional and fully connected layers of deep/shallow CNNs [2][25][5], Bag-of-Words (BoW) or Support Vector Machine (SVM) [26] are employed for classification, detection and recognition [17][15] purposes. As an alternative for BoW feature encoding scheme, several other approaches including Fisher vector [27] and Vector of Locally aggregated descriptor (VLAD) have shown promising results with smaller visual words vocabularies [19]. To perform instance level image retrieval where objects from the same category to be separated, [25] suggested to combine the rich spatial middle convolutional layers' features with VLAD encoding. [28] employed MSER [29] for regions identification, followed by the detection of SIFT features within the identified regions and describe each region/bundle as a fix sized VLAD, named as PBVLAD. 2D-based localization methods generally offer efficient database management at low accuracy cost whereas 3D-based techniques are computationally complex but more reliable in localization. Authors in [30] refute this notation by combining 2D-based approaches with SfM-based post-processing and shown better performances then structure-based methods. However, such post processing takes significant longer runtimes which is out of scope of this work since our proposed VPR system works like a 2D-based framework with an aim to improve the retrieval performance while reducing the computation complexities.

With the advent of several feature pooling techniques including Sum-Pooling [16], Max-Pooling [17], Spatial Max-Pooling [31] and Cross-Pooling [15] employed in deep CNNs have demonstrated performance boost in tasks requiring image classification/recognition and object detection/retrieval [17][15]. All these pooling approaches use the convolutional layers' feature maps as a whole to pick the regions, and with images focus on fewer objects make feature maps sparse in nature and finding single region of interest becomes relatively easier. However, such image retrieval tasks are different in nature from the VPR systems where recognizing a place which undergoes diverse changes due to illumination, winter to summer transitions or viewpoint variance added by different capturing angles is quite challenging because the place appears differently, thus, making it harder to identify the common regions. Even when external tasks based pre-trained CNNs are integrated with the above mentioned image retrieval pooling techniques for the VPR problem, the convolutional layers feature maps focus on the trained objects such as vehicles, pedestrians and other time varying objects which are not suitable for place recognition [7]. Therefore, it is still questionable for a generic Visual Place Recognition system to efficiently deal with simultaneous viewpoint and condition variations when employing CNN-based local features pre-trained on other image retrieval tasks.

Recently, authors in [32] trained the landmark detectors [33][17] with a newly introduced $1.2M$ Google Landmark dataset (GLD) containing $15k$ landmark categories including buildings, monuments and bridges annotated by human. They have proposed a technique which retrieves the normalized regional residuals, termed as R-VLAD. Thus, down-weights all the regional residuals and stores a single aggregated regional descriptor per image. Custom landmark detectors including ASMK [34], RMACB [33], RMAC [17] and selective search [35] are incorporated for the regional search and coupled with the proposed R-VLAD on deep CNNs. We can expect further boost in our proposed VPR framework with the integration of R-VLAD [32]. Authors in [8] have shown that the state-of-the-art region-based image retrieval techniques like Attentive Attention [36], Fixed Context [37] are not generally efficient in finding regions efficient for VPR under viewpoint and appearance changes.

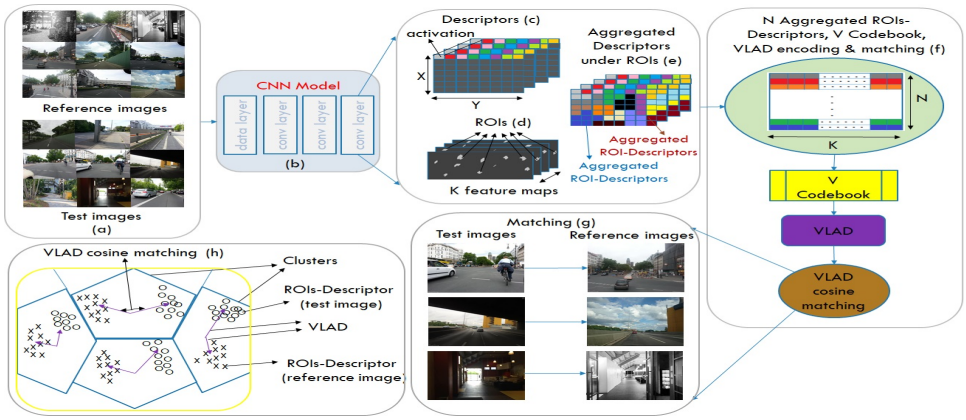

Fig. 2. Workflow of the proposed VPR framework is shown here; the test images are timely fed into the CNN model. Region-of-Interests (ROIs) are identified across all the feature maps of the convolution layer. Aggregated local descriptors under those identified ROIs are pooled for compact image representation. VLAD representations are retrieved and matched by mapping the aggregated regional features on a pre-trained regional vocabulary. The system is designed to store the reference VLAD representations, thus, the query VLAD run-time matching with the pre-stored reference VLADs reduces the retrieval time.

III. Proposed Technique

In this section, the key steps of the proposed methodology are described in detail. It starts from the idea of stacking feature maps' activations and the identification of distinguishing regional patterns. It then illustrates the aggregation of the stacked feature descriptors lying under those identified salient regions. Finally, it shows how to retrieve the VLAD representation from the extracted CNN-based regional representations to determine the match between two images. The workflow of the proposed methodology is shown in Figure 2.

A. Stacking of Convolutional Layer Activations for making Descriptors

For an image I in a convolutional layer of the CNN model, the output is $3D$ tensor M of $X \times Y \times K$ dimensions where K denotes the number of feature maps and X and Y are the width and height of each feature map / channel. We can also interpret M^k as a set of $X \times Y$ activations/responses for k^{th} feature map where $k = \{1,2,....,K\}$. For K feature maps in the convolution layer, we stack each activation at some certain spatial location into K dimensional feature representations as shown with different colours in Figure 2 (c). In (1), D^L represents the K dimensional d_l feature descriptors at L^{th} convolutional layer of m_c model.

$$D^L = \{d_l \in M^K \quad \forall\ l \in \{1,...,X \times Y\}\}\ ,\ L \in m_c \tag{1}$$

B. Identification of Regions of Interest

To extract region-based CNN features, the most prominent regions are need to be identified. Two or more activations are considered to be connected if they are neighbours and have approximately the same value. Thus, for K feature maps, H connected activations representing the regions visualized in Figure 2(d), each region is denoted by G_h^L, $\forall\ h \in \{1,...,H\}$ where H is the total number of identified regions at L^{th} convolution layer.

The mean energy of each of the G_h^L region is calculated by averaging over all the a_h activations lying under that region. In (2), a_h^f represents the f^{th} activation lying under the G_h^L region and E^L denotes the calculated mean energies of H regions. Based on the sorted E^L energies, top N energetic G_t^L ROIs (with their bounding boxes) are picked in (3), denoted as R^L novel regions at L^{th} convolution layer.

$$E^L = \{\frac{1}{|G_h^L|}\sum_f a_h^f,\ \forall\ a_h^f \in G_h^L\} \tag{2}$$

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

Figure 3 illustrates the identified top $N = \{50,200,400\}$ novel R^L regions in an image of the place. It is worth noticing that our novel CNN based identified regions strongly concentrate on the static objects including buildings, trees and road signals. To pool the regional features, D^L descriptors in (1) which fall under the bounding boxes of the R^L regions in (3) are aggregated in (4), each aggregated regional feature is denoted by $1 \times K$ dimensional f_t vector where q be the R_t^L region under which D_q^L descriptors fall. For N novel regions, (5) represents the $N \times K$ dimensional F^L region-based CNN features representing an image at L^{th} convolutional layer (intuitively shown in Figure 2 (e) and Figure 2 (f)).

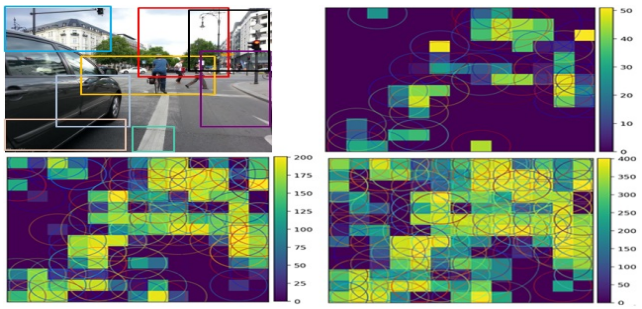

Fig. 3. Sample images of top 50, 200 and 400 Regions-Of-Interest (ROIs) identified by our proposed approach; CNN based identified regions put emphasis on static objects including buildings, trees and road signals.

$$f_t = \sum_{q \in R_t^L} D_q^L \tag{4}$$

$$F^L = \{f_t\ \forall\ t \in \{1,...,N\}\} \tag{5}$$

In comparison with [7] where the authors first identified distinguishing patterns, calculated their mean energies and picked $N = 200$ top energetic regions. Precisely, N identified regional activations of L^{th} convolution layer were mapped onto the $L-1^{th}$ convolutional feature maps, thus, aggregation of the modified regional local descriptors of $L-1^{th}$ convolution layer (lying under mapped ROIs) was carried out to retrieve the regional features. Figure 4 illustrates the sample ROIs identified from two feature maps with our approach and Cross-Region-BoW [7]. Note that depending upon the activations per ROI(s) and receptive fields of the convolutional layers' filters, bounding box per extracted region varies for [7]. We also observed that bigger sized regional patterns hinder the identification of the static place-centric instances in distinguishing the places under environment variations, thus, small sized regional features with our proposed approach improves the VPR performance. Few sample images exhibiting the novel identified regions by Cross-Region-BoW [7] and with our proposed Region-VLAD framework are shown in Figure 5.

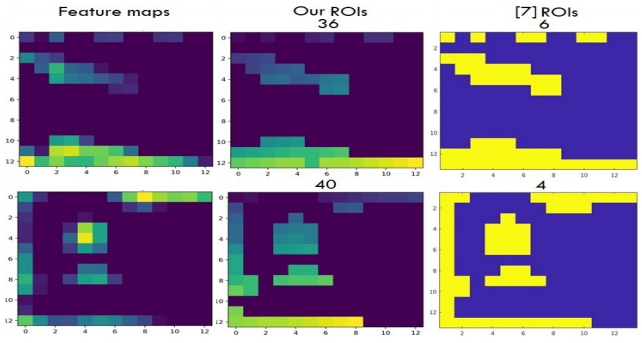

Fig. 4. Sample images of the identified regions using our proposed Region-VLAD and Cross-Region-BoW [7] are shown here. Our approach identifies large number of ROIs per feature map(s).

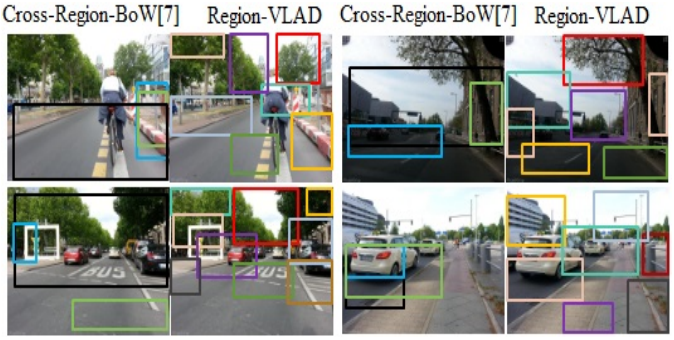

Fig. 5. Sample images of ROIs identified with Cross-Region-BoW [7] and Region-VLAD are shown here. Our regional approach subdivides each image into large number of most contributing regional blocks which helps in distinguishing the places efficiently.

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

Vector of Locally Aggregated Descriptors (VLAD) adopts K-means [11] based vector quantization and accumulates the quantization residues for features quantized to each dictionary cluster and concatenates those accumulated vectors into a single vector representation. A separate dataset of 2.6k images is collected and afore-described region-based aggregation is employed on it for generating a regional vocabulary. To learn a diverse vocabulary, we employed 1125 place-recognition centric images of 365 places from Query247 [38] (taken at day, evening and night times). Other images include a benchmark place recognition dataset St.lucia [24] with 1k frames of two traverses captured in suburban environment at multiple times of the day. The left over images consist of multiple viewpoint- and condition-variant traverses of urban and suburban routes collected from *Mapillary*[1](previously employed by [6] and [7] for capturing place recognition datasets). K-means is employed for clustering the $2600 \times N$ aggregated ROIs-Descriptors into V regions such that o_u in (6) represents the u^{th} region centre in the regional codebook C^L.

$$C^L = \{o_u \ \forall \ u \in \{1,...,V\}\}, \ V \in \{64,128,256\} \tag{6}$$

Using the learned codebook, the F^L regions of the benchmark test and reference traverses are quantized to predict the clusters/labels Z^L in (7), where α is the quantization function. Using the original region-based features F^L, predicted labels Z^L and the regional codebook C^L, summed residue v corresponding each u^{th} region can be retrieved using (8).

$$Z^L = \alpha(F^L) \tag{7}$$

In (8), for all the F^L regional features that fall in u^{th} region of the C^L codebook, the residues of the F^L_u regions and C^L_u codebook's region center are summed. Sometimes, few regions/words appear more frequently in an image than the statistical expectation known as visual word burstiness [39]. Standard techniques include power normalization [40] is performed in (9) to avoid it where each summed residue v_u undergoes non-linear transformation γ. In (10), power normalization is followed by l_2 normalization. For each image, V number of power and l_2 normalized summed residues each with dimension $1 \times K$ gets stored in (11) to get final $V \times K$ dimensional VLAD representation G^L.

$$v_u = \sum_{F^L_u : Z^L_u = C^L_u} F^L_u - C^L_u \tag{8}$$

$$v_u := sign(v_u)\|v_u\|^\gamma \tag{9}$$

$$v_u := \frac{v_u}{\sqrt{v_u^T v_u}} \tag{10}$$

$$G^L = \{v_u \ \forall \ u \in \{1,...,V\}\} \tag{11}$$

To match a test image "A" against the reference image "B", in (12), the dot/scalar product of their u^{th} regional VLAD components $G^{L^A}_u$ and $G^{L^B}_u$, each with dimension $1 \times K$ reaches to an individual regional matching score $j^{A,B}_u$ visualized in Figure 2 (h).

$$j^{A,B}_u = \frac{(G^{L^A}_u).(G^{L^B}_u)}{\|(G^{L^A}_u)\|\|(G^{L^B}_u)\|} \tag{12}$$

All the scalar $j^{A,B}_u$ scores for all V regions are summed up in (13) to get final single $J^{A,B}$ matching score. For each test image "A", the cosine matching in (12) is performed against all the reference images and at the end, reference image "X" with the highest similarity score is picked as a matched image using (14).

$$J^{A,B} = \sum_{u=1}^{V} j^{A,B}_u \tag{13}$$

$$P^A = \arg\max_X J^{A,X} \tag{14}$$

IV. Datasets, Implementation details, Results and Analysis

This section presents the implementation details of our proposed system which will attempt to evaluate its run-time performance for real-time robotic VPR applications. Comparison of the proposed method with state-of-the-art VPR and image retrieval algorithms has been conducted over several benchmark datasets and the obtained results are stated. The section ends by displaying the results on correctly matched and mismatched scenarios of our proposed Region-VLAD framework along with a discussion on the same.

A. Benchmark place recognition datasets

More specifically, challenging benchmark VPR datasets *Berlin A100*, *Berlin Halenseestrasse* and *Berlin Kudamm* (see [7] for detailed introduction), collected from crowd-sourced geotagged photo-mapping platform *Mapillary* are used to evaluate the proposed VPR framework. Each dataset covers two traverses of the same route uploaded by different users. One traverse is used as R reference database and the other traverse is employed as T test database (see TABLE I) where R' represents the reduced reference traverse which contains T' matched test images (will discuss in section IV-E). Another dataset, *Garden Point* was captured at QUT campus with one traverse taken in daytime on left side walk and the other traverse was recorded in right side walk at night time [24]. The *Synthesized Nordland* dataset was recorded on a train with one traverse taken in winter and other the traverse was recorded in spring. Viewpoint variance was added by cropping frames of summer traverse to keep 75% resemblance [8]. For *Berlin A100*, *Berlin Halenseestrasse* and *Berlin Kudamm*, geotagged information is used for ground truth. For *Garden Point* and *Synthesized Nordland*, the ground truth data is obtained by parsing the frames and maintaining place level resemblance.

[1]https://www.mapillary.com/

TABLE I
BENCHMARK PLACE RECOGNITION DATASETS

Dataset	Environment	Variation		T	R	T'	R'
		Viewpoint	Condition				
Berlin A100	urban	moderate	moderate	81	85	70	64
Berlin Haleenseetrasse	urban, suburban	very strong	moderate	67	157	50	138
Berlin Kudamm	urban	very strong	moderate	222	201	166	151
Garden Point	campus	strong	strong	200	200	152	150
Synthesized Nordland	train	moderate	very strong	1622	1622	1221	1217

B. Setup, Implementation details and Scalability

The proposed VPR framework is implemented in Python 3.6.4 and the system average runtime over 5 iterations is recorded with 1125 images. AlexNet pre-trained on Places365 dataset is employed as a CNN model for region-based features extraction with 256×256 input image size. For all the baseline experiments, we utilize middle *conv3* convolutional layer only due to its better performance in various VPR approaches [6][22].

For a single image, a forward pass takes around an average $0.305190ms$ and $15.57ms$ using Caffe on NVIDIA P100 and Intel Xeon Gold 6134 @3.2GHz. We extract and aggregate N ROIs-Descriptors with total time comparable with the state of the art methods [7] (see Table II). The VLAD representations are retrieved and matched using N ROIs-Descriptors mapped on V clustered dictionary C^L (trained on N ROIs-Descriptors of the $2.6k$ dataset). For direct comparison with [7], we use $N = 200$ with $V = 128$. The results are also reported for $N = 400$ with $V = 256$. Table II shows that for both the regional settings, our average VLAD matching times are $100x$ and $58x$ faster than [7].

In real-time robotic vision applications which include robotic agricultural devices, autonomous infrastructure, environmental monitoring equipment or other agriculture based use-cases, with exploration of new places, the size of the SLAM can grow unbounded, therefore, scalability is one of the important factor to be considered [41]. For both the regional settings, employing GPU for forward pass and CPU for features extraction and VLAD encoding, the overall times til VLAD encoding per test image are $396ms$ and $447ms$. Whereas, Titan X Pascal GPU in [7] takes $408ms$ per query for forward pass til regional features encoding. Figure 6 (a) further confirms that the proposed system consumes an average $0.07ms$ and $0.12ms$ for matching one query VLAD against one reference VLAD. Therefore, the total retrieval times per query against $R = 750$ reference images will be approximately around $446.405ms$ and $533.245ms$. In comparison, Cross-Region-BoW [7] takes $7ms$ per reference image features matching, so, the retrieval time against $R = 750$ reference images will be $5.658s$ which is $12x$ and $10x$ more than our proposed approaches and practically inappropriate for real-time applications. It is worth noticing our Region-VLAD VPR technique can store the encoded VLAD representations of all the reference frames whereas Cross-Region-BoW has to perform runtime cross matching of given query's regions against all the reference frames' regions, and mutually matched regional features are picked.

Furthermore, Figure 6 (b) evaluates our proposed system run-time performance when reference images are timely increased. For each PR-curve, we employed T test images and R reference images and retrieved their VLAD representations followed up by their cosine matching and in parallel, we record down the system's performance. We can see that as the test and reference traverses get bigger, the AUC under PR curves (will discuss later in section IV-D) approximately remains the same with "Time" represents the overall matching time of each test image. This mimics that the system is capable enough to handle large number of reference images while maintaining performance both in terms of accuracy and matching time. It is interesting to note that [7] used MATLAB implementation which is practically slower than Python but it will worth mentioning we employ CPUs in comparison to [7] which used GPU.

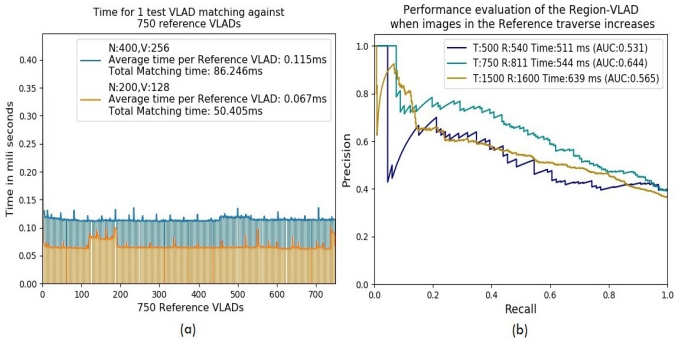

Fig. 6. Matching times for 1 test image VLAD against 750 reference VLADs are presented in (a). (b) illustrates the performance evaluation of the Region-VLAD with timely increase in test and reference images.

C. Comparing Methods

To show the dominance of our novel place-centric regions finding approach combined with the VLAD encoding, we integrated AlexNet with [7] (open-source MATLAB code can be found at [42]), and combined the regional features with the VLAD and BoW encodings, named as Cross-Region-VLAD and Cross-Region-BoW. For a fair comparison, using $2.6k$ dataset, we trained a separate regional vocabulary employing *conv4* for regions identification and *conv3* for features extraction. Keeping $N = 200$, we used $V = 128$ and $V = 2.6k$ for Cross-Region-VLAD and Cross-Region-BoW. Furthermore, results are also reported for HybridNet with Spatial Pyramidal Pooling (SPP) [24] employed on *conv5* and AlexNet integrated with RMAC [17] while performing power- and l2-normalization on the retrieved regions, followed by mutual regions filtering using cosine matching and aggregation of the mutual regions' scores.

PR curves across all other image retrieval approaches including Cross-, Max-, Sum-Pool, Whole and state-of-the-art VPR approaches FABMAP and SEQSLAM are taken from [7] which employed *conv5_2* of deep object centric VGG-16 as features representation. However, Cross-Region-BoW [7] with deep VGG-16 model used *conv5_3* for landmarks identification and *conv5_2* for features extraction. [7] employed standard FABMAP implementation [43] and three sequential frames configuration for SEQSLAM.

D. Precision Recall curves

In image retrieval tasks where there is a moderate to large class imbalance which means the positive class samples are quite rare as compare to negative classes, Precision Recall curves are usually employed as evaluation parameter [12]. Since, our aim is to localize a robot using its previous experience so, for any query place experienced by the robot, with maximum matching score approach (section III-C), it has to either match with the correct place from the previous experience or matches incorrectly with no third option.

1) Berlin Halenseestrasse: In Figure 7 (a), the proposed Region-VLAD PR-curves for *Berlin Halenseestrasse* dataset significantly outperforms all other state-of-the-art methods. Surprisingly, Cross-Region-VLAD PR-curve underperformed with a big margin. This mimics the boost up in our approach is encouraged with the use of our novel regional features. Cross-Region-BoW only considers the mutually matched regions, thus, exhibits better results. Further investigations for Cross-Region-VLAD suggest that under strong viewpoint variations and due to different regions finding approach [7] cause multiple regions to cover the subset of areas of other regions. Hence, the mapping of the cross-convoluted regional features over the vocabulary results into non-uniform features distribution. Although, normalization is carried out but still, many zero regional residues get stored in the VLAD which reflects on the PR-curves. RMAC which

TABLE II
COMPARISON OF THE PROPOSED METHOD WITH CROSS-REGION-BOW[7]

Methodology		Our Region-VLAD																		Cross-Region-BoW [7]
Model		AlexNet365																		VGG16
Images		1125																		1000
GPU/CPU		NVIDIA P100	Intel Xeon Gold 6134 @3.2GHz																	Titan X Pascal GPU
Forward pass time (ms)		0.305190	15.574639																	59
ROIs-Descriptors "N"		50			100			200			300			400			500			200
Extraction and Aggregation time (s)		0.328			0.361			0.394			0.402			0.443			0.452			0.349
Regions "V"		64	128	256	64	128	256	64	128	256	64	128	256	64	128	256	64	128	256	10k Visual words
Matching time (ms)	**VLAD encoding**	1.33	2.05	3.58	1.55	2.28	3.79	1.91	2.4	4.03	1.99	2.68	4.28	2.13	2.96	4.54	2.36	3.16	4.75	7
	VLAD matching	0.05	0.06	0.12	0.05	0.08	0.13	0.05	0.07	0.12	0.04	0.07	0.12	0.05	0.08	0.12	0.05	0.07	0.12	

is state-of-the-art in other image retrieval techniques and SPP, both under-performed on this dataset.

In Figure 7 (b), FABMAP despite its better viewpoint variations tackling and SEQSLAM with its sequence-based whole image approach have not perform well. Cross-Pool employs a similar idea of pooling as Cross-Region-BoW, so both have achieved nearly the similar PR-curves whereas other pooling techniques under-performed. It is worth noticing even with smaller regional dictionaries, our proposed Region-VLAD framework still manages to achieve better results than VGG-16 Cross-Region-BoW [7] and other methodologies which also indicates the potential of our shallow CNN based regional features robustness against strong viewpoint variations.

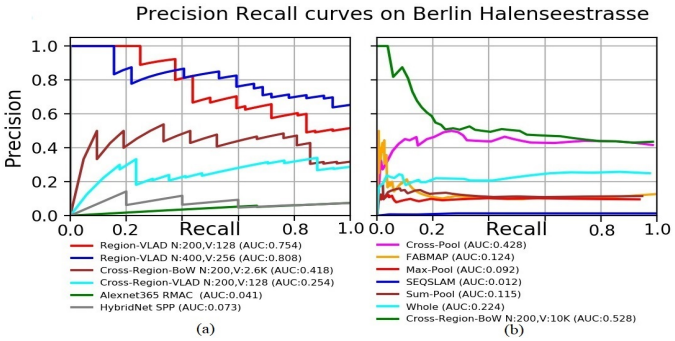

Fig. 7. AUC PR-curves for Berlin Halenseestrasse dataset are presented here; Left: PR-curves of our proposed Region-VLAD and [7] employed on AlexNet with VLAD and BoW encodings. Right: Comparison with state-of-the-art VPR approaches employing VGG-16.

2) Berlin Kudamm: Due to urban environment, too many dynamic and confusing objects such as vehicles, trees and pedestrians with homogeneous scenes lead to perceptual aliasing coupled with severe viewpoint changes makes it a challenging dataset. Figure 8 (a) shows that our proposed Region-VLAD approach still manages to achieve better results. AlexNet combined with Cross-Region-BoW claims state-of-the-art results with 2.6*k* regional vocabulary. RMAC and SPP again underperformed. This is apparently because, VPR is different from other image retrieval and recognition systems where a single object majorly covers the whole image, thus, Sum-, Max-pool and RMAC which perform relatively well in those vision based tasks actually not performed in VPR under severe viewpoint changes.

In Figure 8 (b), due to the resembles among the places captured in sequence, Whole and SeqSLAM with their whole-image based approaches have shown better performances. With higher precision at start and as recall increases, Region-VLAD PR-curves are quite similar but covering more AUC than Whole, SeqSLAM, Cross-Pool and VGG-16 Cross-Region-BoW which mimics the usefulness of our novel CNN based regional approach merged with VLAD.

3) Berlin A100: This dataset exhibits moderate viewpoint and moderate lightning changes with evaluated PR-curves are displayed in Figure 9. It is quite evident that our Region-VLAD approach in (a) achieves similar results as state-of-the-art VGG-16 Cross-Region-BoW [7] (b). AlexNet combined with cross-regional approach

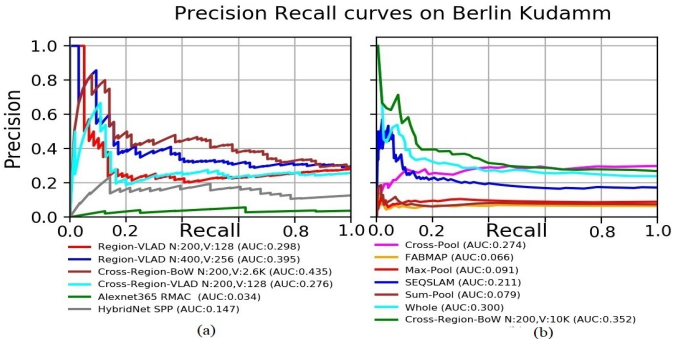

Fig. 8. AUC PR-curves for Berlin Kudamm dataset are presented here; Left: PR-curves of our proposed Region-VLAD and [7] employed on AlexNet with VLAD and BoW encodings. Right: Comparison with state-of-the-art VPR approaches employing VGG-16.

[7] achieves similar and better results for BoW and VLAD. SPP employed on HyridNet again found not to be convincing.

Against our approach, RMAC on AlexNet achieves comparable but better performance than FABMAP and other pooling techniques including Sum-, Max- and Cross-Pool. Condition and viewpoint changes are not much stronger, therefore, RMAC and other approaches have shown better performance on this dataset. A deep analysis on the datasets reveals varied time interval between the captured frames, thus, SEQSLAM underperformed on this dataset. Overall, our proposed Region-VLAD achieved second best performance after VGG-16 Cross-Region-BoW [7].

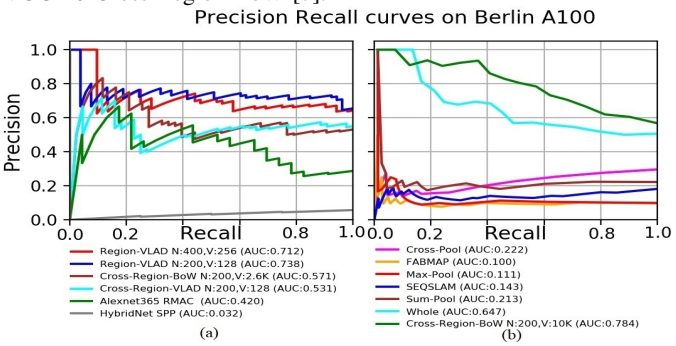

Fig. 9. AUC PR-curves for Berlin A100 dataset are presented here. Left: PR-curves of our proposed Region-VLAD and [7] employed on AlexNet with VLAD and BoW encodings. Right: Comparison with state-of-the-art VPR approaches employing VGG-16.

4) Synthesized Nordland: PR-curves in Figure 10 (a) show that our proposed approach works relatively well when comparing it from all other approaches excluding RMAC and SPP which achieve state-of-the-art performances. Employing object-centric VGG-16, other approaches including Max- and Sum-Pool have not work well on this dataset and similar whole-image based processing techniques i.e. SEQSLAM and Whole have also shown the similar PR-curves.

Fine-tuning with SPED induced condition invariance into the HybridNet, thus, SPP on HybridNet has shown better performance than scene-centric AlexNet. AlexNet with Cross-Region-BoW and

Cross-Region-VLAD outperformed VGG-16 based Cross-Region-BoW [7] pre-trained on ImageNet. This also indicates the importance of the CNN training.

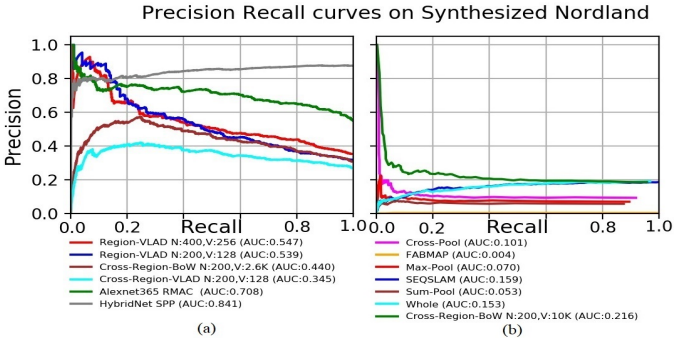

Fig. 10. AUC PR-curves for Synthesized Nordland dataset are presented here. Left: PR-curves of our proposed Region-VLAD and [7] employed on AlexNet with VLAD and BoW encodings. Right: Comparison with state-of-the-art VPR approaches employing VGG-16.

5) Garden Point: Both the *Garden Point* traverses exhibit stronger lightning variations with adequate temporal coherence between the frames. Figure 11 shows that our Region-VLAD approach manages to achieve similar and better performance than Cross-Region-BoW, Cross-Region-VLAD, Whole, RMAC and SPP. SEQSLAM takes advantage from the sequential information and shown state-of-the-art performance. Cross-Region-BoW and Cross-Region-VLAD integrated with AlexNet and VGG-16 exhibit similar performances but approaches including Sum-, Max-Pool and FABMAP relatively underperformed.

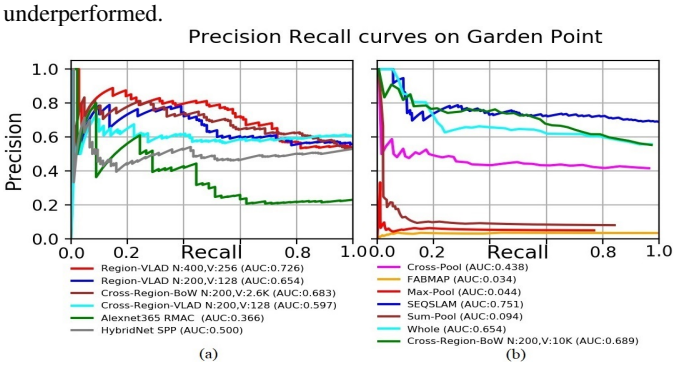

Fig. 11. AUC PR-curves for Garden Point dataset are shown here; Left: PR-curves of our proposed Region-VLAD and [7] employed on AlexNet with VLAD and BoW encodings. Right: Comparison with state-of-the-art VPR approaches.

E. Matching Score thresholding

By nature, PR curves do not consider the True Negative cases (correctly missed the non existing events/classes) [12]. So, to tackle such tricky situations, we employed T test traverse and R' reference traverse from all the datasets so that $T - T'$ queries can be treated as new places (see Table I). Figure 12 visualized the results of proposed Region-VLAD framework on modified datasets before and after match score thresholding. Matching scores in y-axis differentiate the TP, FN, FP and TN events[2] shown with different colored curves where length of the curves in x-axis denotes the number of images which the events contain. The threshold is an average of all the true negative scores calculated over all the modified datasets. Due to limited space, results are reported for two datasets only. Upon thresholding in Figure 12(b), *Berlin Halenseestrasse* missed $FN = 2$ correctly matched images and successfully not matched 10 queries out of $TN = 17$. The same behavior is observed for *Berlin A100*. It is evident that Region-VLAD not only boost up the AUC under PR-curves but also skilful in assigning low scores to TN queries (green curves). Knowing that under scenarios when some queries are new places, and for every query, it's quite challenging to successfully retrieve the correct match while reducing incorrect retrievals. Therefore, 100% performance is impractical to achieve.

[2]TP for True Positives, FN for False Negatives, FP for False Positives and TN for True Negatives

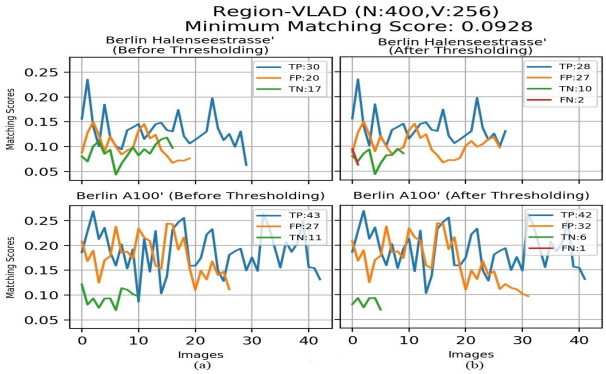

Fig. 12. Left column presents graphs for *Berlin Halenseestrasse* and *Berlin A100* before thresholding and right column graphs showcase the change in TP, FP, TN and FN upon thresholding. Our proposed Region-VLAD framework assign low score to the T-T' or TN queries.

F. Analysis

Figure 13 and Figure 14 illustrate some of the matched and mismatched scenarios. For the correct matches, taking advantage from CNN's scene-centric training, Region-VLAD identifies the common regions shown with different coloured boxes under simultaneous viewpoint and appearance changes. For the mismatched scenarios, the identified top novel regions with coloured boxes (trees, lamp posts) show the areas where the system confuses in and matches the scenes but wrongly recognizes the places. We have seen that Cross-Region-BoW [7] when integrated with AlexNet showed comparable performance but at high time computation cost. However, our Region-VLAD still outperformed Cross-Region-BoW [7] with smaller dictionary and low retrieval time. Also, cross-regional approach of [7] when combined with the VLAD shown inferior results which confirms the performance boost in our Region-VLAD encouraged with our novel regional approach. Datasets and results are placed at [44] and the author intends to open-source the code upon publication.

V. Conclusion

For Visual Place Recognition on resource-constrained mobile robots, achieving state-of-the-art performance/accuracy with lightweight CNN architectures is highly desirable but a challenging problem. This paper has taken a step in this direction and presented a holistic approach targeted for a CNN architecture comprising a small number of layers pre-trained on a scene-centric image database to reduce the memory and computational cost for resource-constrained mobile robots. The proposed framework detects novel CNN-based regional features and combines them with the VLAD encoding methodology adapted specifically for computation-efficient and environment Invariant-VPR problem. The Proposed method achieved state-of-the-art AUC-PR curves on severe viewpoint- and condition-variant benchmark place recognition datasets.

In future, it would be useful to analyse the performance of the proposed framework on other shallow/deep CNN models individually trained/fine-tuned on place recognition-centric datasets. Furthermore, instead of employing defined number of novel regions, it would be interesting to investigate the dynamic regional features selection at runtime and their performances on multiple regional vocabularies.

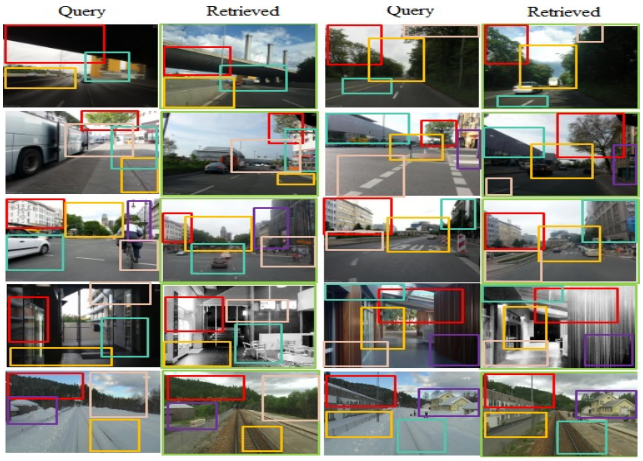

Fig. 13. Sample correctly retrieved matches using the proposed Region-VLAD methodology are presented here; our proposed system identifies common regions across the queries and the retrieved images under strong environment variations.

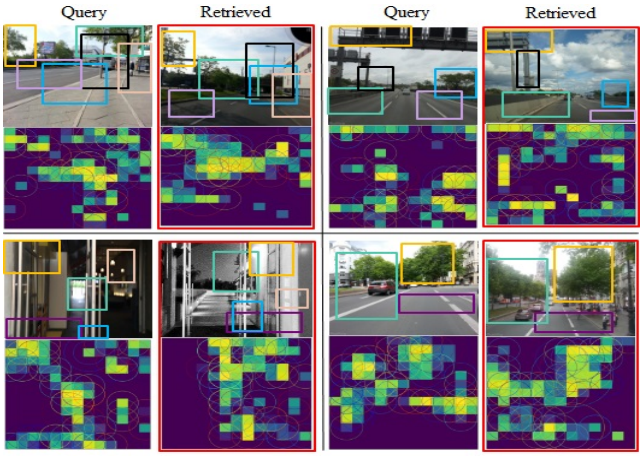

Fig. 14. Sample incorrectly retrieved matches using the proposed approach are presented here; both the queries and the recognized images are geographically different but exhibiting the similar scenes and conditions.

REFERENCES

[1] S. e. a. Lowry, "Visual place recognition: A survey," *IEEE T-RO*, vol. 32, no. 1, pp. 1–19, 2016.

[2] L. e. a. Zheng, "Sift meets cnn: A decade survey of instance retrieval," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 5, pp. 1224–1244, 2018.

[3] H. e. a. Bay, "Surf: Speeded up robust features," in *ECCV*. Springer, 2006, pp. 404–417.

[4] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *IJCV*, vol. 60, no. 2, pp. 91–110, 2004.

[5] Z. e. a. Chen, "Convolutional neural network-based place recognition," *arXiv preprint arXiv:1411.1509*, 2014.

[6] N. e. a. Sünderhauf, "Place recognition with convnet landmarks: Viewpoint-robust, condition-robust, training-free," *Proceedings of Robotics: Science and Systems XII*, 2015.

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

[8] ——, "Learning context flexible attention model for long-term visual place recognition," *IEEE RAL*, vol. 3, no. 4, pp. 4015–4022, 2018.

[9] K. e. a. Simonyan, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[10] A. e. a. Krizhevsky, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[11] J. e. a. Sivic, "Video google: A text retrieval approach to object matching in videos," in *null*. IEEE, 2003, p. 1470.

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

[13] M. e. a. Cummins, "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.

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

[15] L. e. a. Liu, "Cross-convolutional-layer pooling for image recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 11, pp. 2305–2313, 2017.

[16] A. e. a. Babenko, "Aggregating local deep features for image retrieval," in *ICCV*, 2015, pp. 1269–1277.

[17] G. e. a. Tolias, "Particular object retrieval with integral max-pooling of cnn activations," *arXiv preprint arXiv:1511.05879*, 2015.

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

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

[20] R. e. a. Arandjelovic, "All about vlad," in *CVPR*, 2013, pp. 1578–1585.

[21] N. e. a. Sünderhauf, "On the performance of convnet features for place recognition," in *IROS*. IEEE, 2015, pp. 4297–4304.

[22] P. e. a. Panphattarasap, "Visual place recognition using landmark distribution descriptors," in *ACCV*. Springer, 2016, pp. 487–502.

[23] P. e. a. Sermanet, "Overfeat: Integrated recognition, localization and detection using convolutional networks," *arXiv preprint arXiv:1312.6229*, 2013.

[24] Z. e. a. Chen, "Deep learning features at scale for visual place recognition," in *ICRA*. IEEE, 2017, pp. 3223–3230.

[25] J. Yue-Hei Ng, F. Yang, and L. S. Davis, "Exploiting local features from deep networks for image retrieval," in *CVPR workshop*, 2015, pp. 53–61.

[26] A. F. M. Agarap, "A neural network architecture combining gated recurrent unit (gru) and support vector machine (svm) for intrusion detection in network traffic data," in *Proceedings of the 2018 10th International Conference on Machine Learning and Computing*. ACM, 2018, pp. 26–30.

[27] J. Sánchez, F. Perronnin, T. Mensink, and J. Verbeek, "Image classification with the fisher vector: Theory and practice," *IJCV*, vol. 105, no. 3, pp. 222–245, 2013.

[28] H. e. a. Jin Kim, "Predicting good features for image geo-localization using per-bundle vlad," in *ICCV*, 2015, pp. 1170–1178.

[29] J. e. a. Matas, "Robust wide-baseline stereo from maximally stable extremal regions," *Image and vision computing*, vol. 22, no. 10, pp. 761–767, 2004.

[30] T. e. a. Sattler, "Are large-scale 3d models really necessary for accurate visual localization?" in *CVPR*, 2017, pp. 1637–1646.

[31] M. e. a. Jaderberg, "Spatial transformer networks," in *Advances in neural information processing systems*, 2015, pp. 2017–2025.

[32] M. e. a. Teichmann, "Detect-to-retrieve: Efficient regional aggregation for image search," *arXiv preprint arXiv:1812.01584*, 2018.

[33] A. S. e. a. Razavian, "Visual instance retrieval with deep convolutional networks," *ITE Transactions on Media Technology and Applications*, vol. 4, no. 3, pp. 251–258, 2016.

[34] G. e. a. Tolias, "Image search with selective match kernels: aggregation across single and multiple images," *IJCV*, vol. 116, no. 3, pp. 247–261, 2016.

[35] R. e. a. Tao, "Locality in generic instance search from one example," in *CVPR*, 2014, pp. 2091–2098.

[36] D. e. a. Yu, "Multi-level attention networks for visual question answering," in *CVPR*. IEEE, 2017, pp. 4187–4195.

[37] e. a. Jin Kim, "Learned contextual feature reweighting for image geo-localization," in *CVPR*, 2017, pp. 2136–2145.

[38] A. e. a. Torii, "24/7 place recognition by view synthesis," in *CVPR*, 2015, pp. 1808–1817.

[39] H. e. a. Jégou, "On the burstiness of visual elements," in *CVPR*. IEEE, 2009, pp. 1169–1176.

[40] T.-T. e. a. Do, "From selective deep convolutional features to compact binary representations for image retrieval," *arXiv preprint arXiv:1802.02899*, 2018.

[41] C. e. a. Cadena, "Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age," *IEEE TRO*, vol. 32, no. 6, pp. 1309–1332, 2016.

[42] "Cross-region-bow source code," https://github.com/scutzetao/IROS2017_OnlyLookOnce.

[43] "Fabmap source code," http://www.robots.ox.ac.uk/~mjc/Software.htm.

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