Only look once, mining distinctive landmarks from ConvNet for visual place recognition

Conference Paper

Author(s):
Chen, Zetao; Maffra, Fabiola; Sa, Inkyu; Chli, Margarita

Publication date:
2017

Permanent link:
https://doi.org/10.3929/ethz-b-000174255

Originally published in:
https://doi.org/10.1109/IROS.2017.8202131

Funding acknowledgement:
644128 - Collaborative Aerial Robotic Workers (SBFI)

Only Look Once, Mining Distinctive Landmarks from ConvNet for Visual Place Recognition

Zetao Chen[1], Fabiola Maffra[1], Inkyu Sa[2] and Margarita Chli[1]
[1]Vision for Robotics Lab [2]Autonomous Systems Lab ETH Zurich, Switzerland

***Abstract* – Recently, image representations derived from Convolutional Neural Networks (CNNs) have been demonstrated to achieve impressive performance on a wide variety of tasks, including place recognition. In this paper, we take a step deeper into the internal structure of CNNs and propose novel CNN-based image features for place recognition by identifying salient regions and creating their regional representations directly from the convolutional layer activations. A range of experiments is conducted on challenging datasets with varied conditions and viewpoints. These reveal superior precision-recall characteristics and robustness against both viewpoint and appearance variations for the proposed approach over the state of the art. By analyzing the feature encoding process of our approach, we provide insights into what makes an image presentation robust against external variations.**

I. INTRODUCTION

Visual place recognition can be interpreted as an image retrieval task, which consists of determining a match between the current scene and previously visited locations. Motivated by the success of deep learning in computer vision, the focus of place recognition research has recently moved from utilizing traditional handcrafted features [1], such as SIFT [2] or SURF [3], to more generic deep learning-based features extracted from Convolutional Neural Networks (CNNs).

A fundamental question in utilizing deep learning for place recognition is how to generate an image representation from a pre-trained CNN. Generally, current approaches to this question fall into two broad categories that either (a) directly feed the whole image into a pre-trained CNN and extract its activations as the image representation [4-6] or (b) apply the pre-trained CNN to the regions of the input image and aggregate activations from each of these regions to create a final image representation [7, 8]. Usually, approaches in category (a), directly flatten activations from a single CNN layer, either a convolutional or a fully connected layer, to create a global image representation. Such global representations are computed from the entire image and are therefore, not robust against effects, such as partial occlusion or severe viewpoint variations. Features arising in category (b) are more viewpoint-invariant, as such approaches usually involve combining an external landmark proposal technique with CNN-based features to match image patches over extreme appearance and viewpoint variations [7, 8]. However, these approaches rely on external landmark detectors and require applying the pre-trained CNN to each landmark proposal resulting in high computational cost. In other computer visions tasks, such as image retrieval or recognition, there have been a range of investigations into how to efficiently encode the convolutional layer activations of a pre-trained CNN [9-11]. There methods are proposed for tasks that are different in nature from place recognition. In this paper, we propose a place recognition-specific feature encoding method and demonstrate its superior performance over existing methods.

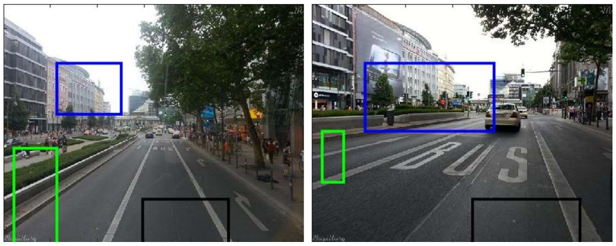

Figure 1. Corresponding image regions across two views of the same scene are identified using our novel CNN-based image features, visualized here with the same color. A thorough evaluation on benchmark datasets reveals better performance of this method under significant viewpoint and condition variations against the state of the art

We look at CNNs from a third, different perspective and propose a novel feature encoding method on CNN activations to tackle both viewpoint and appearance variations. Instead of relying on external landmark proposal techniques, the proposed method identifies salient regions by directly mining distinctive patterns based on activations of the convolutional layers. In particular, we utilize one convolutional layer for local feature extraction and another convolutional layer at a higher level, which usually embeds richer semantic information to discover meaningful image regions, from which local features can be extracted. Each image is then represented by a set of distinctive image regions (i.e. rectangular patches) and cross matching of these regions permits the comparison of two images. Figure 1 illustrates

[1] This research was supported by the Swiss National Science Foundation (SNSF, Agreement no. PP00P2 157585), EC's Horizon 2020 Programme under grant agreement n. 644128 (AEROWORKS), the European Union's Horizon 2020 research and innovation Programme under grant agreement No 644227 and the Swiss State Secretariat for Education, Research and Innovation (SERI) under contract number 15.0029

such an example. We evaluate our method against other state-of-the-art place recognition algorithms and feature encoding approaches on several benchmark datasets that exhibit both appearance and viewpoint variations. In particular, this paper makes the following two main contributions:

1) A novel, CNN-based feature encoding method to create image representations enabling the description of several different image regions without the need to feed multiple inputs to the CNNs;
2) A region-based visual place recognition system that can tackle variations both in viewpoint and conditions, simultaneously;

II. Related Work

In this section, we give a brief overview of previous work utilizing CNNs for place recognition and methods that have been developed to encode CNNs-based features.

A. *Visual Place Recognition with Convolutional Neural Networks*

The first step in visual place recognition is to extract image information that is salient in defining a particular place. The aim is not only to compress information captured along the camera's trajectory, but also to suppress of non-useful image regions (i.e. regions that do not aid the distinctive representation of a particular place). Traditional approaches either operate directly on raw pixels [12] or utilize a fixed set of handcrafted features [1] (i.e. manually defined, as opposed to learning-based representations). Recently, the success of deep learning in computer vision has triggered a range of investigations of its applicability to visual place recognition resulting to impressive first findings, such as the demonstration of the effectiveness of utilizing the intermedia layer activations of a CNN as feature vectors [4-6] in repeatedly recognising a place. The approaches proposed in [13, 14] train CNN architectures for the specific place recognition tasks at hand, demonstrating improved performance under strong condition-variations. In contrast, the approach proposed in this paper can be deployed on any pre-trained CNN architecture to create more compact and robust representations.

All aforementioned approaches extract global representations from an entire image, rendering them unsuitable in cases, where partial occlusions of the scene structure or severe viewpoint variations are expected. Instead, representations that break down an image into smaller regions, such as [15] can be more robust against scale and viewpoint variations. The approaches in [7, 8] combine an external landmark detector with CNN-based features to match regions over extreme viewpoint- and condition-variations. Our approach follows the direction of region-based representation, but does not require any external landmark proposal technique. The proposed approach directly identifies salient regions from the convolutional layer activations and only needs to run the network once for each image, significantly reducing the computation complexity.

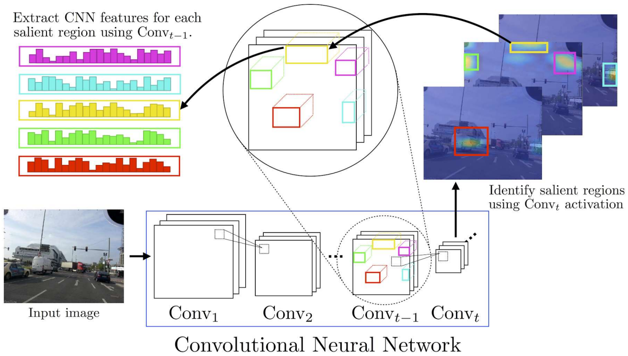
Extract CNN features for each salient region using Conv_{t-1}.
Identify salient regions using Conv_t activation
Input image
Conv_1
Conv_2
Conv_{t-1}
Conv_t
Convolutional Neural Network

Figure 2 A schematic illustration of the proposed feature encoding method. The input image is used to identify salient regions using late Convolutional activation, here visualized in a heat map (blue: not salient, orange: salient). CNN features are then extracted for each salient region.

B. *Existing Feature Encoding Approaches*

In many computer vision tasks, such as image retrieval or object detection, there have been extensive studies on deriving more powerful representations from the convolutional layer activations of a CNN. For example, the work in [10] demonstrates that sum-pooled convolutional feature works well in image retrieval, while [9] explores the use of spatial max-pooling to extract image representation from CNN layer activations for object retrieval. In [16], the Fisher vector is used to pool the extracted convolutional features for text recognition and segmentation. The works in [17, 18] are the most relevant to this work, employing pooling over convolutional activations for image recognition /classification. However, these approaches use the whole feature map to pool the features, which is demonstrated to achieve sub-optimal results for place recognition in our analysis in Section V.A.

III. Method

In this section, the key components of the proposed system are described in detail. We first describe the extraction of local features directly from convolutional layers, before we illustrate how to utilize a higher convolutional layer as guidance to pool extracted local features and create multiple region descriptors to represent each image. Analyzing the weighting scheme according to the importance of each individual region, we finally present how to calculate the similarity between two image representations. The schematic illustration of the proposed system is shown in Figure 2.

A. *Extracting Local Descriptors from Convolutional Activations*

This first stage aims at deriving local representations for a certain image region directly from the convolutional layer activations. Given a pre-trained CNN, for this step, we only consider its convolutional layers and discard all its fully connected layers. The major advantage of the convolutional

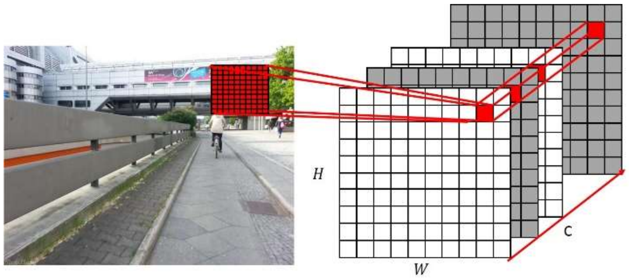

Figure 3. The structure of the convolutional layer activations. There are **C** feature maps and each feature map is a H × W matrix.

layer activations over fully connected ones is that the former usually embeds rich spatial information, from which local descriptors can be constructed. Given an image I , its activations at a certain convolutional layer can be arranged as a tensor of size H × W × C, where H and W denote the height and width of each feature map, respectively and C is the number of feature maps. As illustrated in Figure 3, activations at a certain spatial location across all feature maps can be concatenated into a C -dimensional local descriptor to represent a local image region. The size of this region is equal to the receptive field of the filter. In this way, the convolutional layer activations can be considered as a 2D array of a C-dimensional local descriptor, with each feature describing a certain local region, essentially very similar to a traditional, handcrafted local descriptor. Formally, we represent the convolutional layer activations $\mathrm{X} \in R^{\mathrm{H \times W \times C}}$ as:

$$\mathrm{X} = \{x_i \in R^C | i \in \{1, \dots, H \times W\}\}. \quad (1)$$

B. Encoding Regions from Local Descriptors

Features extracted during the previous step can only describe a region of the size of its filter's receptive field. In place recognition, however, each location is represented by a set of distinctive landmarks, which can take any shape and size. In order to represent a visual pattern at different sizes and in arbitrary shapes, a straightforward method would be to aggregate all local descriptors falling into that region to create a pooled feature vector. In other computer vision tasks, such as image retrieval or recognition, various feature pooling strategies have been proposed to aggregate the local descriptors, such as max-pooling [9], sum-pooling [10] or Fisher vectors [11]. Inspired by the lack of robustness of these approaches in dealing with viewpoint variance ubiquitous in place recognition tasks, the following section presents an alternative feature aggregating method that can significantly improve the features' robustness against viewpoint variance.

C. Mining Distinctive Patterns from a Late Convolutional Layer

In order to discover landmarks useful for place recognition, here we directly mine distinctive patterns from a late convolutional layer. Generally, a feature map generated by a convolutional filter can be interpreted with the detection scores obtained by applying the convolution filter on the input image. Locations in this map with high activation values indicate that there exist visual patterns around them that the filter is searching for. It is observed that feature maps at a late convolutional layer are generally very sparse and selective to visual patterns corresponding to some semantically meaningful regions, such as a shape or an object part [19]. As a result, when a place is visited from different viewpoints, some of these visual 'signatures' will be preserved and can be detected in subsequent images of the same region by applying the same convolutional filter.

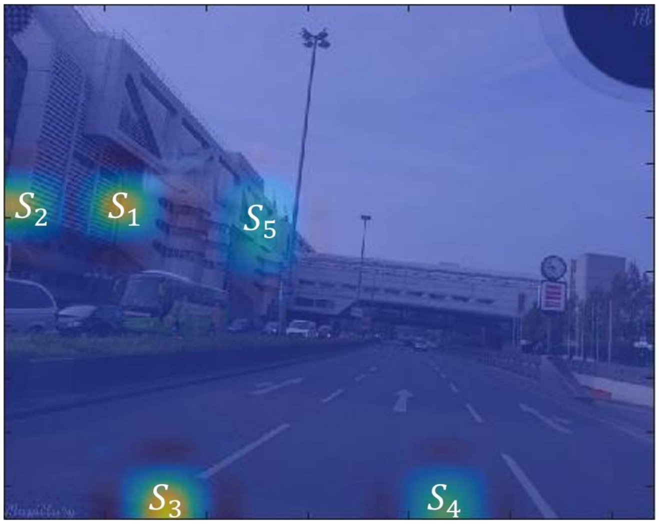

Figure 4. Illustration of clusters $S_{1\sim5}$, from the Convolutional layer activations in an example image.

Based on these observations, we propose a landmark discovery approach by searching for the strongest spatially localized regions at late convolutional feature maps. In particular, for each feature map of a particular convolutional layer from one image, we first group all non-zero and spatially proximal 8-connected activations into an individual cluster $S_j, \forall \mathrm{j} \in [1, \dots \mathrm{T}]$, where T denotes the number of clusters. Figure 4 illustrates the clusters obtained on an example image. For each activation cluster S_j, we calculate its energy by averaging over all activations that fall into the cluster:

$$E_j = \frac{1}{|S_j|} \sum_i a_j^i \,, \forall a_j^i \in S_j. \quad (2)$$

where a_j^i denotes the i-th activation in cluster S_j. We search for the M clusters with the highest energies and identify them as the basic Regions-Of-Interest (ROIs):

$$O_t, \ \forall\, t \in [1, .., M]. \quad (3)$$

Each of these ROIs represents a salient region. For each of these ROI, we aggregate all the local descriptors $x_i \in R^C$ falling into the respective ROI and generate a pooled feature vector to represent it. In total, this strategy creates M different pooled feature vectors P_t, to represent one image:

$$P_t = \sum_{i \in O_t} x_i a_t^i \,, t = 1, \ldots, M. \qquad (4)$$

where a_t^i denotes the i-th activation value in region/cluster O_t. It is worth noting that we do not concatenate these pooled vectors together to form a single image representation.

D. Assignment of Weight to Each Region

Each of the M pooled vectors describes one image region that is considered salient by the pre-trained CNN. Inspired by the bag-of-words approach [20], we develop a strategy to calculate the inverse-document-frequency of each of these salient regions. To this end, we built a separate training dataset of K ($K = 5000$) images and constructed M pooled vectors P_t from each of these images, resulting to a total of $K \times$ M pooled vectors. A vocabulary was then built by clustering these features into N different words, assigning a weight W_c to each visual word c as:

$$W_c = log_{10}(K/n_c), c = 1, \ldots, N. \qquad (5)$$

where K is the total number of training images and n_c is the number of images containing the visual word c. This weight is assigned to all feature vectors P_t that belong to the visual word c.

E. Image Matching

To determine the similarity between two images A and B, we perform cross matching between all region vectors P^A and P^B that were extracted from both images. The similarity between region i from A and region j from B can be calculated as:

$$s_{i,j} = \frac{{P_i^A}^T P_j^B}{\|P_i^A\| \|P_j^B\|} \text{ i} = 1, \ldots, \text{M; j} = 1, \ldots, \text{M}. \qquad (6)$$

Crosschecking is applied here to accept only mutual matches. As a result, the overall similarity between two images A and B can be calculated as:

$$Q_{A,B} = \frac{1}{M} \sum_{i,j} s_{i,j} \times W_i \times W_j. \qquad (7)$$

where W_i and W_j denote the weight of the word that features P_i^A and P_j^B belong to, respectively. The search for the best matching reference image A for the query image B goes through all reference images from the database and picks the one with the highest similarity score:

$$Y(B) = arg \max_A Q_{A,B} \qquad (8)$$

IV. Experimental Setup

This section describes the testbed used and the acquirement of ground truth used to conduct the evaluation of the proposed approach against the state of the art.

A. Datasets

We evaluated our proposed system on five benchmark place recognition datasets. These datasets capture different types of environments as well as exhibiting variations in both viewpoints and conditions. Details are summarized in Table 1. Each dataset consists of two traverses along the same route with the first traverse used for reference and the second one used for testing. Some example images are illustrated in Figure 5.

The *Gardens Point* dataset was collected at the Queensland University of Technology campus with the first traverse recorded at the daytime along the left side of the walkways and the second traverse taken at night from the right side of the walkways. It has been evaluated in a number of previous studies [4, 5, 7]. The *Synthesized Nordland* dataset was recorded from a camera mounted on a train. The first traverse was recorded in spring and the second in winter (see [21] for a more detailed introduction). The *Berlin A100*, *Berlin Halenseestrasse* and the *Berlin Kudamm* datasets were all downloaded from a crowdsourced photo-mapping platform called Mapillary[2]. It was first introduced in [7] as place recognition datasets. Each of the three datasets consists of two different sequences mapping the same route but uploaded by different users, exhibiting severe viewpoint and moderate appearance variations.

Table 1 DATASET DESCRIPTIONS

Dataset	No. of frames	Environment	Viewpoint variation	Condition variation
Garden Point	400	campus	strong	strong
Synthesized Nordland	970	train journey	moderate	strong
Berlin A100	166	urban	strong	moderate
Berlin Halenseestrasse	225	Urban +suburban	strong	moderate
Berlin kudamm	424	urban	strong	moderate

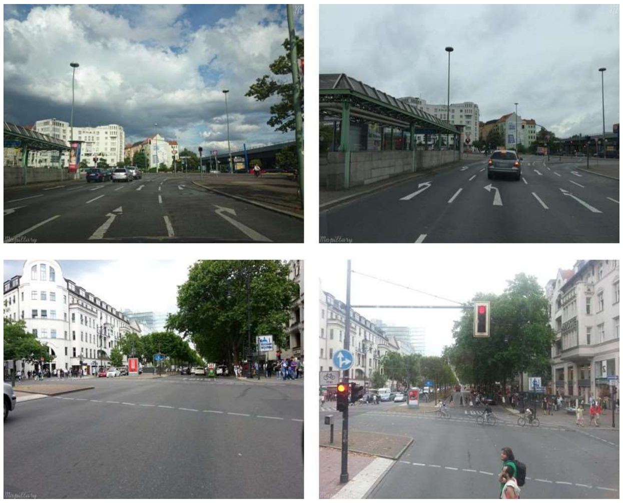

Figure 5. Examples of the *Berlin A100* (top row) and *Berlin Kudamm* (bottom row) datasets [7]. A strong viewpoint change can be observed in both examples.

[2] http://www.mapillary.com

B. Ground Truth

For the datasets *Garden Point*, *Berlin A100*, *Berlin Halenseestrasse* and *Berlin kudamm*, ground truth was obtained by manually parsing the frames and building frame-level correspondence. For the *Synthesized Nordland* dataset, we used the frame-level correspondence provided with the original dataset.

C. Implementation Details

We employed the VGG16 network [22] as the pre-trained CNN to evaluated our proposed approach. However, other pre-trained networks, such as ResNet [23], GoogleNet [24] or AlexNet [25], can also be employed. We utilized the second last convolutional layer to extract local descriptors and the last convolutional layer to discover salient regions. For all other baseline methods evaluated in our experiments, we also extracted the second last convolutional layer activation as image representation. For each image, we pick 200 ($M = 200$) regions with highest average activations and we build a vocabulary of 10000 words ($N = 10000$). All images are first resized to 224×224 before they are fed to CNN for feature extraction. The parameters are set once and used across all experiments.

V. RESULTS

The proposed approach was evaluated against other state-of-the-art feature encoding and place recognition methods recording the Area Under the Curve (AUC) [26] computed on precision-recall curves. We also visualize the mutual matches established by our approach to provide insights about the superiority of our method. Finally, the runtime performance of our approach is analyzed.

A. Precision-recall Characteristics

The AUC is recorded on all five testing datasets for our proposed place recognition system against the most relevant state-of-the-art approaches for place recognition and feature encoding. The higher the AUC, the better the performance is. In particular, we compare against the feature-based method FAB-MAP [27] and the sequence-based SeqSLAM [12]. Since our approach is focused on deriving a more efficient visual representation from a pre-trained CNN, we also compare with other existing feature encoding methods; namely, the whole image representation used in [4, 5], sum-pooling [10], max-pooling [28] and cross-layer pooling [17], which uses a similar idea of pooling over convolutional activations.

Figure 6 presents the AUC generated by these methods on the *Berlin Halenseestrass* dataset. It is evident that the proposed approach achieves significantly better performance than all other methods. Although FAB-MAP, as a feature-based place recognition system tackles viewpoint variation better than approaches using whole-image representations, such as SeqSLAM, it still underperforms on this challenging dataset. Cross-layer pooling achieves the closest performance to our method, indicating the potential benefits of utilizing late convolutional layer activations as the pooling guidance.

Figure 7 presents results on the *Berlin A100* dataset, with the proposed approach still outperforming the rest and the whole-image representation used in [4, 5] achieving the second best performance. It is interesting that applying SeqSLAM on the whole-image descriptors deteriorates the performance. A closer look at the dataset reveals that this is probably because there are varied intervals between consecutive frames and the temporal coherence required by SeqSLAM is violated.

Figure 8 and Figure 9 illustrate the AUC results on the *Berlin Kudamm* and *Synthesized Nordland* dataset. On both datasets, the proposed approach achieves the best performance. Once again, using a whole-image representation, despite its apparent simplicity, still delivers inferior performance. The other encoding methods, such as sum-pooling or max-pooling, which have been demonstrated to achieve impressive performance in image retrieval and recognition, do not perform well. This is probably due to the fact that place recognition is different in nature from other vision tasks, such as image recognition or retrieval, where there is often a single object occupying the biggest part of the image. In place recognition, there is no such structural constraint and therefore, our proposed approach of decomposing a location into multiple region elements may be the main reason behind its superior performance.

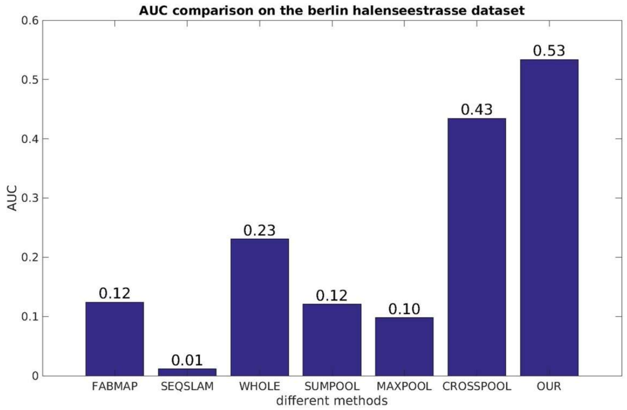

Figure 6. AUC levels on the *Berlin Halenseestrasse* dataset

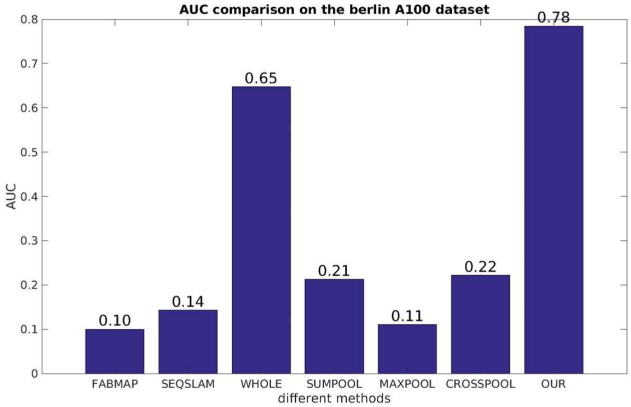

Figure 7. AUC levels on the *Berlin A100* dataset

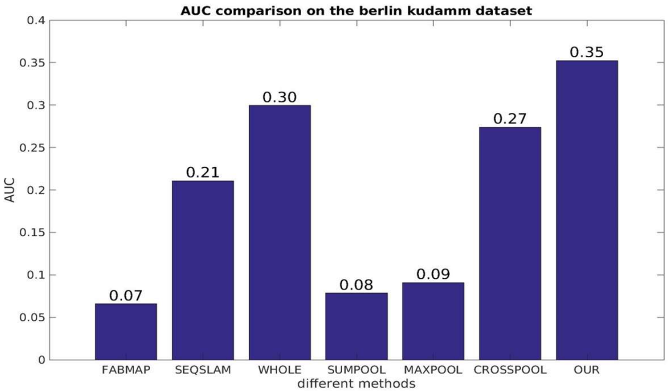

Figure 8. AUC levels on the *Berlin Kudamm* dataset

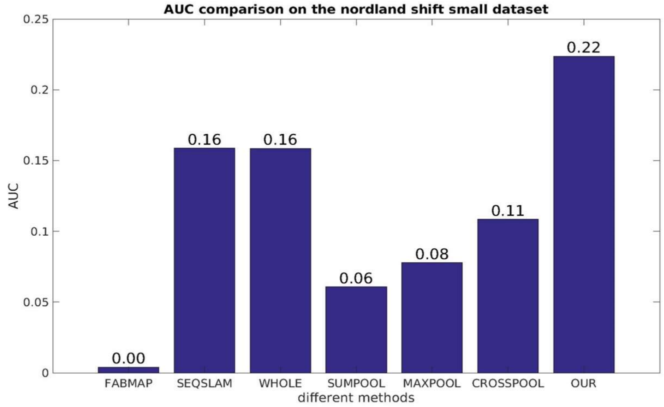

Figure 9. AUC levels on the *Synthesized Nordland* datast

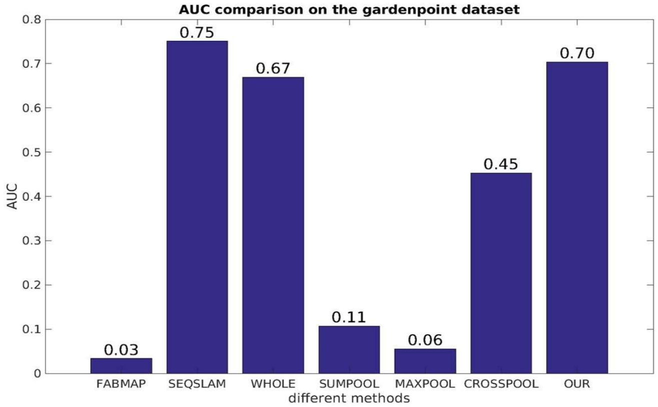

Figure 10. AUC levels on the *Garden Point* dataset

Figure 10 illustrates that SeqSLAM achieves similar performance with our approach in the *Garden Point* dataset. It is observed that in this dataset, there is strong temporal coherence and therefore, SeqSLAM can leverage the sequential information to improve its performance. However, it is worth noting that our approach only requires a single image instead of a sequence of images required in SeqSLAM. Other methods, such as sum-pooling or max-pooling, still do not perform well on this dataset, indicating the challenge of place retrieval in the presence of strong condition variations. The cross-pooling method again achieves close performance to our approach.

B. Qualitative Analysis of the Region-Matching

In this section, we visualize the place recognition results and analyze the difference between our approaches and other feature encoding methods qualitatively. Figure 11 illustrates three such examples, where our method can correctly match the query images (shown in the top row of Figure 11) against the database, while other methods fail. We also visualize the top three region matches identified by our approach (i.e. the black, blue and green rectangles). Observing the results in the first column for example, our approach successfully identifies the white box on the left as a distinctive landmark for that place. Other methods return images, where the global semantics are similar to the query image (such as the trees and roads), but are most often confused with different places. Similar phenomenon can be observed in the second and third column. These examples illustrate the potential benefits of using region-based representation, when there are strong viewpoint variations.

In Figure 12, we also visualize examples where our approach fails and others, such as cross-layer pooling, successfully recognize the place. In the first column, our approach identifies regions around a car as a mutual match, resulting in wrong recognition. It is worth noting the pre-trained CNN we utilized were trained on object datasets with many car images and this is probably why it focuses so much on car-like shapes. This indicates the influence of pre-trained CNN on the performance of our approach.

In Figure 13, we visualize more correctly matching results from our algorithm to provide more insights into our approach. As illustrated in the figure, our approach can match places across various degrees of viewpoint and condition changes.

C. Runtime Consideration

Deep learning approaches are computationally intensive and therefore, an evaluation of their runtime performance is particularly important in order to realize their employments for robotics applications. We run experiments on 1000 images and record the average runtime of the proposed method. For a single image, one forward pass through the VGG16 network costs approximately $59.4ms$ using Caffe on an NVIDIA Titan X Pascal GPU and encoding the CNN features using our proposed approach on the Matlab platform takes about 0.349s. Matching between two images using the Matlab implementation takes approximately $7ms$.

VI. Conclusion

Inspired by the success of region-based image representations for place recognition and the recent boom in deep learning techniques, in this paper, we present a novel feature encoding method to build image representations making use of CNN's convolutional layer activations. We utilize one convolutional layer for local feature extraction and another, late convolutional layer to identify salient regions, from which local features can be extracted. The derived image presentation encodes several distinctive image regions

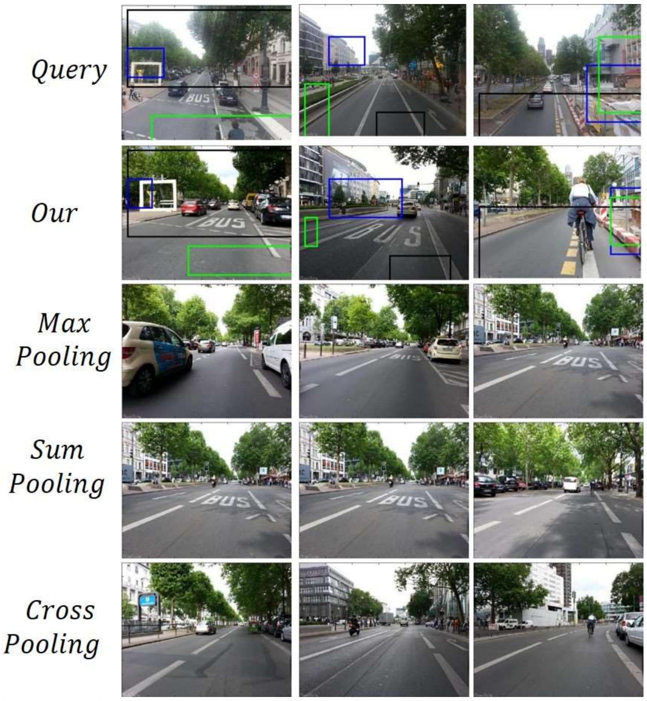

Figure 11. Visualization of the top scoring image resulting from each method in each column, when querying the database with the corresponding image from the top row. The first row describes three query images and the second to the last row respectively illustrates the images returned from our approach, max-pooling [28], sum-pooling [10] and cross-pooling [17].

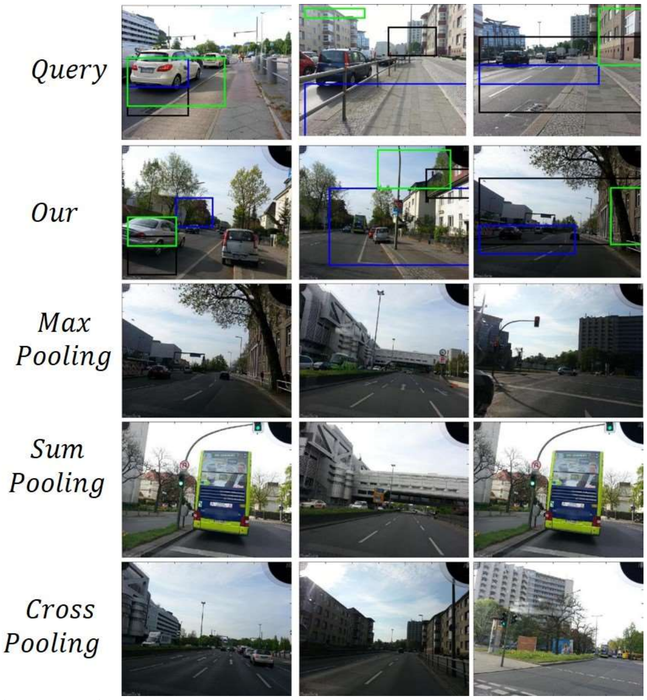

Figure 12. The images are arranged the same way as in Figure 11. However, in these three examples, we approach returns incorrect match and the cross-pooling method successfully match the place.

that can be used for cross matching in a later retrieval stage. Comparisons to state of the art techniques on extensive benchmarking datasets, demonstrate superior performance of the proposed method in place recognition tasks with strong viewpoint and condition variations.

The pre-trained CNN used in this paper is trained on object recognition dataset, so future directions involve investigations on whether encoding features from a CNN that is particularly trained for place recognition can further improve the performance. Moreover, we will study the integration of temporal information in a bid to improve the place recognition performance under more severe conditions, by propagating place recognition hypotheses over time.

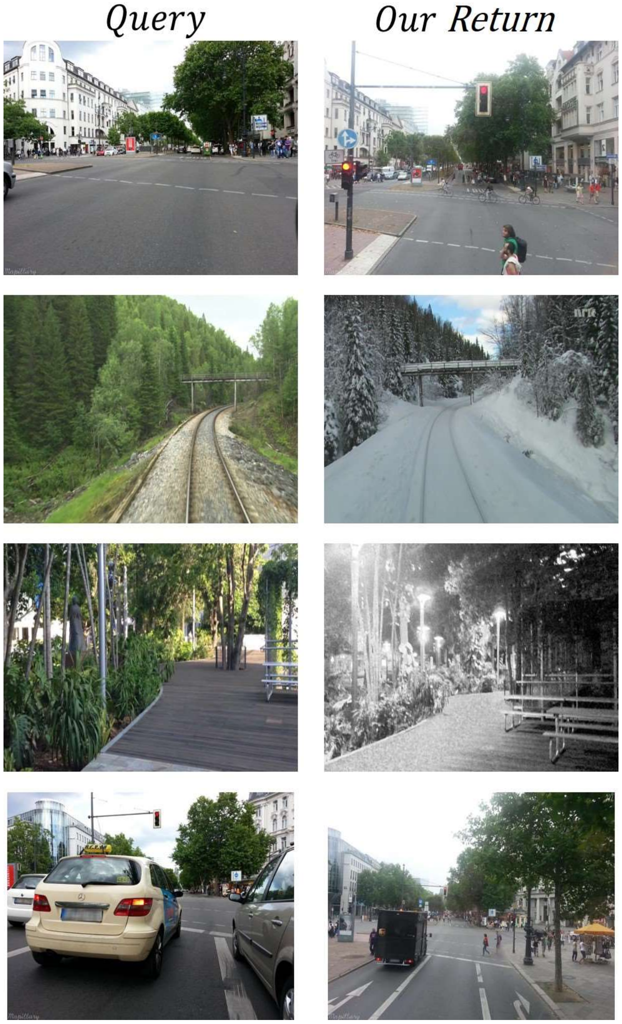

Figure 13 Correctly matching examples from our approach, where the left column indicate query images and the right are our returns.

Acknowledgement: We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan X Pascal GPU used for this research.

REFERENCES

[1] M. Cummins and P. Newman, "FAB-MAP: Probabilistic Localization and Mapping in the Space of Appearance," *International Journal of Robotics Research,* vol. 27, pp. 647-665, 2008.

[2] D. G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *International Journal of Computer Vision,* vol. 60, pp. 91-110, 2004.

[3] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (SURF)," *Computer vision and image understanding,* vol. 110, pp. 346-359, 2008.

[4] Z. Chen, L. Obadiah, A. Jacobson, and M. Milford, "Convolutional Neural Network based Place Recognition," presented at the Australiaian Conference on Robotics and Automation, Melbourne, Australia, 2014.

[5] N. Sünderhauf, F. Dayoub, S. Shirazi, B. Upcroft, and M. Milford, "On the Performance of ConvNet Features for Place Recognition," presented at the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Hamburg, Germany, 2015.

[6] R. Arroyo, P. F. Alcantarilla, L. M. Bergasa, and E. Romera, "Fusion and binarization of CNN features for robust topological localization across seasons," in *Intelligent Robots and Systems (IROS), 2016 IEEE/RSJ International Conference on*, 2016, pp. 4656-4663.

[7] N. Sunderhauf, S. Shirazi, A. Jacobson, F. Dayoub, E. Pepperell, B. Upcroft*, et al.*, "Place recognition with ConvNet landmarks: Viewpoint-robust, condition-robust, training-free," *Proceedings of Robotics: Science and Systems XII,* 2015.

[8] T. Kanji, "Self-localization from images with small overlap," in *Intelligent Robots and Systems (IROS), 2016 IEEE/RSJ International Conference on*, 2016, pp. 4497-4504.

[9] G. Tolias, R. Sicre, and H. Jégou, "Particular object retrieval with integral max-pooling of CNN activations," *arXiv preprint arXiv:1511.05879,* 2015.

[10] A. Babenko and V. Lempitsky, "Aggregating local deep features for image retrieval," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1269-1277.

[11] F. Perronnin, J. Sánchez, and T. Mensink, "Improving the fisher kernel for large-scale image classification," in *European conference on computer vision*, 2010, pp. 143-156.

[12] M. Milford and G. Wyeth, "SeqSLAM: Visual Route-Based Navigation for Sunny Summer Days and Stormy Winter Nights," in *IEEE International Conference on Robotics and Automation*, St Paul, United States, 2012.

[13] Z. Chen, A. Jacobson, N. Sunderhauf, B. Upcroft, L. Liu, C. Shen*, et al.*, "Deep Learning Features at Scale for Visual Place Recognition," *arXiv preprint arXiv:1701.05105,* 2017.

[14] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "NetVLAD: CNN architecture for weakly supervised place recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 5297-5307.

[15] C. McManus, B. Upcroft, and P. Newmann, "Scene signatures: Localised and point-less features for localisation," 2014.

[16] M. Cimpoi, S. Maji, and A. Vedaldi, "Deep filter banks for texture recognition and segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3828-3836.

[17] L. Liu, C. Shen, and A. van den Hengel, "The treasure beneath convolutional layers: Cross-convolutional-layer pooling for image classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 4749-4757.

[18] L. Liu, C. Shen, and A. van den Hengel, "Cross-convolutional-layer Pooling for Image Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 2016.

[19] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Computer vision–ECCV 2014*, ed: Springer, 2014, pp. 818-833.

[20] J. Sivic and A. Zisserman, "Video Google: a text retrieval approach to object matching in videos," in *Proceedings of the Ninth IEEE International Conference on Computer Vision, 2003*, 2003, pp. 1470-1477 vol.2.

[21] P. Neubert, N. Sünderhauf, and P. Protzel, "Superpixel-based appearance change prediction for long-term navigation across seasons," *Robotics and Autonomous Systems,* vol. 69, pp. 15-27, 2015.

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

[23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770-778.

[24] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov*, et al.*, "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1-9.

[25] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," presented at the Advances in neural information processing systems, 2012.

[26] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve," *Radiology,* vol. 143, pp. 29-36, 1982.

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

[28] H. Azizpour, A. Sharif Razavian, J. Sullivan, A. Maki, and S. Carlsson, "From generic to specific deep representations for visual recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2015, pp. 36-45.