SeqNetVLAD vs PointNetVLAD: Image Sequence vs 3D Point Clouds for Day-Night Place Recognition

Sourav Garg   Michael Milford

QUT Centre for Robotics, Queensland University of Technology

{s.garg, michael.milford}@qut.edu.au

Abstract

Place Recognition is a crucial capability for mobile robot localization and navigation. Image-based or Visual Place Recognition (VPR) is a challenging problem as scene appearance and camera viewpoint can change significantly when places are revisited. Recent VPR methods based on "sequential representations" have shown promising results as compared to traditional sequence score aggregation or single image based techniques. In parallel to these endeavors, 3D point clouds based place recognition is also being explored following the advances in deep learning based point cloud processing. However, a key question remains: is an explicit 3D structure based place representation always superior to an implicit "spatial" representation based on sequence of RGB images which can inherently learn scene structure. In this extended abstract, we attempt to compare these two types of methods by considering a similar "metric span" to represent places. We compare a 3D point cloud based method (PointNetVLAD) with image sequence based methods (SeqNet and others) and showcase that image sequence based techniques approach, and can even surpass, the performance achieved by point cloud based methods for a given metric span. These performance variations can be attributed to differences in data richness of input sensors as well as data accumulation strategies for a mobile robot. While a perfect apple-to-apple comparison may not be feasible for these two different modalities, the presented comparison takes a step in the direction of answering deeper questions regarding spatial representations, relevant to several applications like Autonomous Driving and Augmented/Virtual Reality.

1. Introduction

Visual Place Recognition (VPR) is crucial for mobile robot localization and is typically challenging due to significant changes in scene appearance and camera viewpoint during subsequent visits of known places [14, 21]. Re-
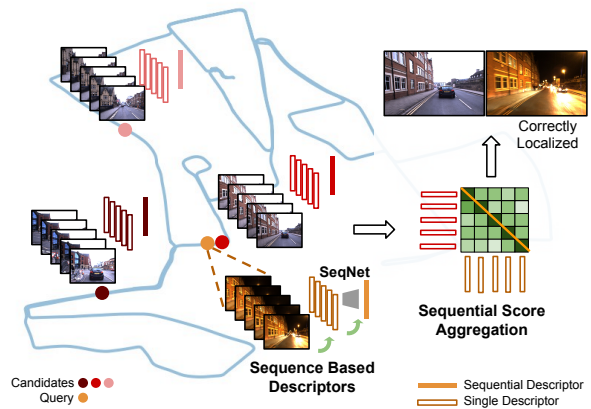

Figure 1. Sequence-based hierarchical visual place recognition. SeqNet learns short sequential descriptors that generate high performance initial match candidates and enables selective control sequence score aggregation using single image learnt descriptors.

searchers have explored a variety of methods to deal with this problem ranging from traditional hand-crafted techniques [7, 24] to modern deep learning-based solutions [2, 30]. Many of these systems aim to push the performance of single image based place recognition by learning better image representations as global descriptors [2, 6, 28, 29] or local descriptors [9, 11, 4] and matchers [31].

To further improve the accuracy of such techniques, researchers have also explored the use of sequential information inherent within the problem of mobile robot localization. However, most of these methods only focus on robustly aggregating *single image* match scores along a sequence [24, 25, 22, 34], where single image representations are agnostic to this post sequential processing. More recent methods have proposed *sequential descriptors* [15, 12, 26, 3, 5] that generate place representations considering sequential imagery *before* any sequence score aggregation. [16] proposed SeqNet and a hierarchical VPR pipeline where sequential descriptors are used to select match hypotheses for single image based sequence score aggregation, as shown in Figure 1.
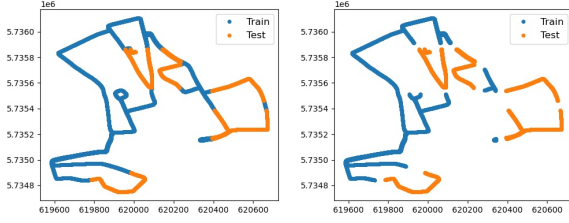
1

Figure 2. (Left) Train and test splits for the Oxford Robotcar dataset as used in [1] and (right) an updated train split instance ensuring no visual overlap between train and test imagery.

A parallel line of research for place recognition exists with regards to using 3D data in the form of point clouds, as done in PointNetVLAD [1], DH3D [10] and others [20, 18, 32, 33, 19]. Instead of using single images or image sequences, these methods rely on point cloud data, typically captured through a LiDAR sensor. Using 3D information in this fashion has significant advantages when considering extreme appearance variations, for example, matching data across day vs night, where single image-based solutions fail catastrophically. [1] demonstrates this behavior by comparing their method against NetVLAD [2]. However, it is not known how well image-based methods compare against a 3D point cloud based technique when a sequence of images is considered. In this extended abstract, we conduct additional experiments with SeqNet [16], showcasing that image sequences can potentially outperform 3D point cloud based methods given a similar metric span and localization radius. As a preliminary investigation's late-breaking result, we only consider a single dataset for this analysis: Oxford Robotcar [23], which was originally used by both SeqNet [16] and PointNetVLAD [1].

We refer the readers to the original works [1, 16] for detailed methodology description. Here, we primarily focus on the experimental settings and results.

2. Experimental Settings

2.1. Dataset

We use two traverses from the Oxford Robotcar dataset: one from day time (2015-03-17-11-08-44) and the other from night time (2014-12-16-18-44-24). Both these traverses were used in the original works, however, train and test splits differed. Hence, we use the splits defined by PointNetVLAD[1] to train and test SeqNet[2]. Note that the training and test splits are captured from geographically disparate locations and each split comprises its own reference (day) and query (night) database.

The LiDAR based 3D point cloud data in the Oxford dataset used for benchmarking PointNetVLAD [1]

---

[1]https://github.com/mikacuy/pointnetvlad
[2]https://github.com/oravus/seqNet

Table 1. Performance Comparison - Oxford (Day vs Night): Recall@K (1,5,20)

Method	Oxford Robotcar
Single Image Descriptors:	
NetVLAD [2]	0.54/0.74/0.89
NetVLAD-FT (S_1 [16])	0.62/0.83/0.94
NetVLAD-FT* (S_1 [16])	0.59/0.78/0.92
Point Clouds Descriptors:	
PointNetVLAD (Base) [1]	0.77/0.92/0.96
PointNetVLAD (Refine) [1]	0.76/0.90/0.94
Sequential Descriptors:	
SmoothNetVLAD [15]	0.66/0.75/0.87
DeltaNetVLAD [15]	0.41/0.64/0.84
SeqNetVLAD (S_5)	*0.87/0.94*/**0.99**
SeqNetVLAD* (S_5)	**0.85**/*0.91*/**0.98**
Sequential Score Aggregation:	
SeqMatch-NetVLAD [24]	0.67/0.78/0.89
SeqMatch-NetVLAD-FT [24]	0.84/0.92/0.98
SeqMatch-NetVLAD-FT* [24]	0.79/0.88/0.96
SeqNetVLAD-HVPR (S_5 to S_1)	**0.88/0.96/0.99**
SeqNetVLAD-HVPR* (S_5 to S_1)	*0.83*/**0.93/0.98**

strictly captures the local surroundings. On the other hand, forward-facing RGB images can comprise information projected from locations much farther away from the camera. Thus, the train-test split defined in [1] (see Figure 2 (left)) potentially leads to visual overlap between both the splits when using image data. Therefore, keeping the test split the same, we additionally created an updated instance of the train split avoiding such visual overlap between the two splits, as shown in Figure 2 (right). In Table 1, * marked results presented in gray font color correspond to the usage of the revised train split. For both the split settings, the best (bold) and the second best (italics) results are formatted independently in Table 1.

2.2. Metric Span

For PointNetVLAD, we use the authors' provided implementation for extracting point clouds and corresponding place representations, which have a metric span of 20 meters per place. For SeqNet and other sequence based methods, we use an image sequence of length 5 with a fixed frame separation of 2 meters between adjacent frames, leading to a metric span of 10 meters per place[3].

2.3. Descriptors Details

PointNetVLAD descriptors are of size 256 as the authors observed no notable performance gain with further

---

[3]Since SeqNet results with a metric span of 10 meters were found to be superior to PointNetVLAD, we did not conduct further experiments with a larger metric span for SeqNet.

doubling of descriptor dimensions [1]. For SeqNet, 4096-dimensional PCA'd NetVLAD descriptors were used as the underlying single image representations and output was a 4096-dimensional sequential descriptor, as per the original setting [16]. From here on, we refer to this SeqNet descriptor based on NetVLAD as *SeqNetVLAD*.

2.4. Evaluation

We use Recall@K ($K \in \{1, 5, 20\}$) as the evaluation metric as also used by both PointNetVLAD and SeqNet. Localization radius is set to be 25 meters as used by PointNetVLAD.

3. Results

Table 1 shows performance comparison between different types of approaches to place recognition: **1)** Single Image Descriptors including NetVLAD [2] and its nighttime fine-tuned version NetVLAD-FT (trained as S_1, as described in [16]); **2)** Point Cloud Descriptors including PointNetVLAD with both its *base* version trained only on the Oxford Robotcar dataset and *refine* version trained on multiple datasets; **3)** Sequential Descriptors including Smoothed and Delta Descriptor defined using NetVLAD, as described in [15] and SeqNetVLAD [16]; and **4)** Sequential Score Aggregation including SeqSLAM-based [24] sequence matching defined on single image descriptors using NetVLAD and NetVLAD-FT, referred to as SeqMatch-NetVLAD and SeqMatch-NetVLAD-FT respectively, and a hierarchical approach as per [16], referred to as SeqNetVLAD-HVPR, where SeqNetVLAD is used as a sequential descriptor to select top matching candidates for SeqMatch-NetVLAD-FT.

It can be observed from Table 1 that sequence-based methods like SmoothNetVLAD (0.66), SeqMatch-NetVLAD (0.67) improve performance on top of single image only techniques NetVLAD (0.54) and NetVLAD-FT (0.59/0.62) but do not approach performance of PointNetVLAD (0.77). However, SeqMatch-NetVLAD-FT (0.79/0.84), SeqNetVLAD (0.83/0.87) and SeqNetVLAD-HVPR (0.85/0.88) surpass PointNetVLAD's performance. This demonstrates that not only the sequential information is a strong cue for place recognition under challenging appearance conditions, trained sequential descriptors [16] might be learning the underlying 3D scene structure implicitly, leading to better performance than what was achievable through traditional sequence-based methods [24].

The experiments conducted in this preliminary investigation have their limitations as a perfect apple-to-apple comparison between RGB image sequence and LiDAR point clouds is not trivial. Both the sensor modalities have complementary characteristics. As compared to RGB cameras, active range sensors like LiDARs are not drastically affected by variations in environmental conditions such as time of day and seasonal cycles. However, the inherent in-formation richness of RGB image sensors, compared to LiDAR point clouds which are typically sparse, makes room for advanced image processing techniques, potentially leading to improved performance even under challenging environmental conditions. Furthermore, as a robot moves through an environment, the data accumulation strategy also plays a key role in determining the robustness of a place representation. For example, the strategy of feeding single images to sequence score aggregation methods [24, 25, 22] or sequential descriptor networks [16, 12, 5] can also be emulated for point cloud based techniques that currently *pre-process* individual point clouds to form a relatively larger one [1, 10] before learning any place representations.

4. Conclusion

With recent advances in deep learning, several novel methods have been developed for place (spatial) representations including both 3D point cloud based [1, 10] and those based on image sequences [16, 12, 15]. Both these modalities have their own inherent characteristics and there remain several questions unanswered in terms of what might constitute an ideal representation of the world perceived by a mobile robot [8, 17]. The analysis presented in this extended abstract takes an initial step towards answering such questions with preliminary investigations. Future work will investigate the scope of combining image sequences and 3D information for an even further improved spatial understanding as also explored recently in [13, 27].

References

[1] Mikaela Angelina Uy and Gim Hee Lee. Pointnetvlad: Deep point cloud based retrieval for large-scale place recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4470–4479, 2018. 2, 3

[2] Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Netvlad: Cnn architecture for weakly supervised place recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5297–5307, 2016. 1, 2, 3

[3] Roberto Arroyo, Pablo F Alcantarilla, Luis M Bergasa, and Eduardo Romera. Towards life-long visual localization using an efficient matching of binary sequences from images. In *2015 IEEE international conference on robotics and automation (ICRA)*, pages 6328–6335. IEEE, 2015. 1

[4] Bingyi Cao, Andre Araujo, and Jack Sim. Unifying deep local and global features for image search. In *Eur. Conf. Comput. Vis.*, pages 726–743, 2020. 1

[5] Marvin Chancán and Michael Milford. Deepseqslam: A trainable cnn+ rnn for joint global description

and sequence-based place recognition. *arXiv preprint arXiv:2011.08518*, 2020. 1, 3

[6] Zetao Chen, Adam Jacobson, Niko Sünderhauf, Ben Upcroft, Lingqiao Liu, Chunhua Shen, Ian Reid, and Michael Milford. Deep learning features at scale for visual place recognition. In *IEEE Int. Conf. Robot. Autom.*, pages 3223–3230, 2017. 1

[7] Mark Cummins and Paul Newman. Fab-map: Probabilistic localization and mapping in the space of appearance. *Int. J. Robot. Res.*, 27(6):647–665, 2008. 1

[8] Andrew J Davison. Futuremapping: The computational structure of spatial ai systems. *arXiv preprint arXiv:1803.11288*, 2018. 3

[9] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *IEEE Conf. Comput. Vis. Pattern Recog. Worksh.*, pages 224–236, 2018. 1

[10] Juan Du, Rui Wang, and Daniel Cremers. Dh3d: Deep hierarchical 3d descriptors for robust large-scale 6dof relocalization. In *European Conference on Computer Vision*, pages 744–762. Springer, 2020. 2, 3

[11] Mihai Dusmanu, Ignacio Rocco, Tomas Pajdla, Marc Pollefeys, Josef Sivic, Akihiko Torii, and Torsten Sattler. D2-net: A trainable CNN for joint description and detection of local features. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 8092–8101, 2019. 1

[12] Jose M Facil, Daniel Olid, Luis Montesano, and Javier Civera. Condition-invariant multi-view place recognition. *arXiv preprint arXiv:1902.09516*, 2019. 1, 3

[13] Sourav Garg, V Babu, Thanuja Dharmasiri, Stephen Hausler, Niko Suenderhauf, Swagat Kumar, Tom Drummond, and Michael Milford. Look no deeper: Recognizing places from opposing viewpoints under varying scene appearance using single-view depth estimation. In *IEEE Int. Conf. Robot. Autom.*, pages 4916–4923, 2019. 3

[14] Sourav Garg, Tobias Fischer, and Michael Milford. Where is your place, visual place recognition? In *IJCAI*, 2021. 1

[15] Sourav Garg, Ben Harwood, Gaurangi Anand, and Michael Milford. Delta descriptors: Change-based place representation for robust visual localization. *IEEE Robot. Autom. Lett.*, 5(4):5120–5127, 2020. 1, 2, 3

[16] Sourav Garg and Michael Milford. SeqNet: Learning descriptors for sequence-based hierarchical place recognition. *IEEE Robot. Autom. Lett.*, 2021. 1, 2, 3

[17] Sourav Garg, Niko Sünderhauf, Feras Dayoub, Douglas Morrison, Akansel Cosgun, Gustavo Carneiro, Qi Wu, Tat-Jun Chin, Ian Reid, Stephen Gould, et al. Semantics for robotic mapping, perception and interaction: A survey. *Foundations and Trends® in Robotics*, 8(1–2):1–224, 2020. 3

[18] Le Hui, Mingmei Cheng, Jin Xie, and Jian Yang. Efficient 3d point cloud feature learning for large-scale place recognition. *arXiv preprint arXiv:2101.02374*, 2021. 2

[19] Giseop Kim and Ayoung Kim. Scan context: Egocentric spatial descriptor for place recognition within 3d point cloud map. In *IEEE/RSJ Int. Conf. Intell. Robot. Syst.*, pages 4802–4809. IEEE, 2018. 2

[20] Zhe Liu, Shunbo Zhou, Chuanzhe Suo, Peng Yin, Wen Chen, Hesheng Wang, Haoang Li, and Yun-Hui Liu. Lpd-net: 3d point cloud learning for large-scale place recognition and environment analysis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2831–2840, 2019. 2

[21] Stephanie Lowry, Niko Sünderhauf, Paul Newman, John J Leonard, David Cox, Peter Corke, and Michael J Milford. Visual place recognition: A survey. *IEEE Transactions on Robotics*, 32(1):1–19, 2016. 1

[22] Simon Lynen, Michael Bosse, Paul Timothy Furgale, and Roland Siegwart. Placeless place-recognition. In *3DV*, pages 303–310, 2014. 1, 3

[23] Will Maddern, Geoffrey Pascoe, Chris Linegar, and Paul Newman. 1 year, 1000 km: The oxford robotcar dataset. *IJ Robotics Res.*, 36(1):3–15, 2017. 2

[24] Michael J Milford and Gordon F Wyeth. Seqslam: Visual route-based navigation for sunny summer days and stormy winter nights. In *Robotics and Automation (ICRA), 2012 IEEE International Conference on*, pages 1643–1649. IEEE, 2012. 1, 2, 3

[25] Tayyab Naseer, Luciano Spinello, Wolfram Burgard, and Cyrill Stachniss. Robust visual robot localization across seasons using network flows. In *Twenty-Eighth AAAI Conference on Artificial Intelligence*, 2014. 1, 3

[26] Peer Neubert, Stefan Schubert, and Peter Protzel. A neurologically inspired sequence processing model for mobile robot place recognition. *IEEE Robotics and Automation Letters*, 4(4):3200–3207, 2019. 1

[27] Amadeus Oertel, Titus Cieslewski, and Davide Scaramuzza. Augmenting visual place recognition with structural cues. *IEEE Robot. Autom. Lett.*, 5(4):5534–5541, 2020. 3

[28] Filip Radenović, Giorgos Tolias, and Ondřej Chum. Fine-tuning cnn image retrieval with no human annotation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(7):1655–1668, 2018. 1

[29] Jerome Revaud, Jon Almazán, Rafael S Rezende, and Cesar Roberto de Souza. Learning with average precision: Training image retrieval with a listwise loss. In *Int. Conf. Comput. Vis.*, pages 5107–5116, 2019. 1

[30] Paul-Edouard Sarlin, Frederic Debraine, Marcin Dymczyk, and Roland Siegwart. Leveraging deep visual descriptors for hierarchical efficient localization. In *Conference on Robot Learning*, pages 456–465, 2018. 1

[31] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4938–4947, 2020. 1

[32] Qi Sun, Hongyan Liu, Jun He, Zhaoxin Fan, and Xiaoyong Du. Dagc: Employing dual attention and graph convolution for point cloud based place recognition. In *Proceedings of the 2020 International Conference on Multimedia Retrieval*, pages 224–232, 2020. 2

[33] Kavisha Vidanapathirana, Peyman Moghadam, Ben Harwood, Muming Zhao, Sridha Sridharan, and Clinton Fookes. Locus: Lidar-based place recognition using spatiotemporal higher-order pooling. In *IEEE Int. Conf. Robot. Autom.*, 2021. 2

[34] Olga Vysotska, Tayyab Naseer, Luciano Spinello, Wolfram Burgard, and Cyrill Stachniss. Efficient and effective matching of image sequences under substantial appearance changes exploiting gps priors. In *IEEE Int. Conf. Robot. Autom.*, pages 2774–2779. IEEE, 2015. 1