A Different Point of View: Investigating Use of Neural Radiance Field View Synthesis for Viewpoint Invariance

Matthew Bradley — Vision and Image Processing Lab, University of Waterloo, ON, N2L 3G1, Canada

Georges Younes — Vision and Image Processing Lab, University of Waterloo, ON, N2L 3G1, Canada

John Zelek — Vision and Image Processing Lab, University of Waterloo, ON, N2L 3G1, Canada

Email: {m7bradle, georges.younes, jzelek}@uwaterloo.ca

Abstract

Place Recognition Systems provide the ability for vision systems to detect places they visit. Central to this is the use of image descriptors to summarize the visible scene in an image, for future comparison in the future. Current applications of VPR including self driving vehicles are pushing the limit of existing techniques, requiring long operational lives that cover large and diverse geographical areas. This introduces challenging new changes to scenes, like illumination due to the time of day/night, or seasonal variations. Neural-network-based image descriptors provide promising improvements in this area but a trade-off has emerged: training for robustness appearance changes or viewpoint changes between visits to a scene. One approach is in the use of synthetic views, which decouple the two problems by allowing the difference in viewpoint to be artificially corrected. Here we evaluate a promising method (Neural Radiance Fields, or NeRF) when trained on a series of images captured from a trajectory representative of real-world application in VPR. We compare the frames produced to ground truth frames with the same viewpoint to gauge the performance that can be expected and the potential issues of applying this technique. Overall we find promising examples where the quality of the synthetic frames may allow for use in VPR. We also suggest future work, improving on the quality and reliability of the views obtained.

1 Introduction

Critical to a variety of navigation tasks is the problem of visual place recognition or VPR. This is the ability for a vision system to detect when it revisits a place which has been previously seen. A common example is loop-closure in SLAM (Simultaneous Localization and Mapping) systems which are a staple of mobile robotics.

Core to the operating principles of VPR is the description of scenes which are visited, for storage and future comparison. These descriptions must be robust to all of the changes which a scene can undergo, while still being able to differentiate it from others which likewise may change. With self driving vehicles and other recent applications which require long operational lives over large geographical areas new challenges are introduced. The longer lifespan means that changes in illumination through the day or other environmental factors must be tolerated, and some applications even require robustness to seasonal appearance changes.

Techniques leveraging more powerful neural image descriptions have shown promise in overcoming these challenges, however a tension has emerged. Many of these descriptions are sensitive to the large-scale structures in an image which can improve performance when differentiating similar locations. However, this also makes them sensitive to the viewpoint from which the image was taken, which is not always guaranteed to be consistent. This is in contrast to approaches like BoVW and VLAD [1] which focus on image patches and local features. A compromise must be made between training for invariance to appearance changes but also to changes in viewpoint. [2]

One suggestion for addressing this problem has been to make use of synthetic views. This allows for the creation of artificial images with an alternate viewpoint for comparison, in turn decoupling the problems of viewpoint and appearance by solving them independently. If a consistent viewpoint is presented to the descriptor, focus can be placed on appearance changes. One demonstration of this is by [3] through the use of RGBD scan data to produce alternate views during verification of a match. However, [2] mention [4] as a possible method of offline view synthesis which requires no 3D data. Instead only regular images and poses are needed, which are already naturally generated in many mobile vision systems. This would greatly easy the deployment of view synthesis techniques to the wider world of VPR.

Here we describe the synthesis of alternate viewpoints using Neural Radiance Fields using real trajectory data from a SLAM system, operating on a common SLAM dataset. We train the model on an initial traversal and upon revisiting the same location generate several synthetic views with the same pose as newly captured frames. We compare these synthetic views with this ground truth to provide an early examination of the suitability of NeRF to generating synthetic frames in VPR. We highlight some of the difficulties encountered and provide some recommendations for how some may be overcome.

In the organization of this paper we first give a brief survey of various work to use synthetic views in easing the burden on visual place recognition systems. We then describe the conditions of the experiments undertaken and the factors which influenced their implementation. Finally we present and discuss the results of our preliminary investigation, as well as present promising avenues for improvement.

2 Background Review

2.1 Invariance of Various Description Methods

Core to VPR systems are the descriptor(s) used to describe scenes captured by the system for later matching. As discussed by [2], [5], and others, there are competing goals in the design of a VPR descriptor to be robust to both changes in appearance (appearance invariance, eg illumination or seasonal changes), and to the viewpoint of the camera (viewpoint invariance).

Many past descriptors including BoVW [6], VLAD [1], and Fisher vectors [7] process individual image features and patches, ignoring the image's spatial information. This lends them a comparatively high degree of viewpoint invariance. Recent approaches, and especially those leveraging neural networks like HybridNet and AMOSNet [8], have shown to be much more powerful descriptors and more robust to illumination and appearance changes. The robustness over techniques like BoVW makes them very attractive but is often accomplished by recording spatial information into the descriptor which approaches like BoVW ignore. This is because large scale elements of a scene are often reliable not to change over time, though they are virtually guaranteed to be captured differently from different viewpoints.

2.2 Use of Synthetic Views

Various methods in the past have been proposed to generate synthetic views for VPR and related tasks. [3] re-project dense RGBD data to generate alternate viewpoints during verification of a potential place recognition. [9] also produces synthetic views by generating a neural depth estimate and then applying a simple lateral shift to pixels based on their depth. [10] make use of 3D data and panoramas from Google Street View to increase the size of their matching database. [11] render point clouds into synthetic views for comparison with images, as no image data that corresponds to the point cloud exists. The most notable difference between these methods and NeRF is that NeRF can operate on simple images and their poses, without requiring a depth map, scan data, 3D models, or other extra 3D information. This greatly simplifies it's deployment.

2.3 Neural Radiance Fields for View Synthesis

One of the most widely cited methods for view synthesis in general synthetic view literature is Neural Radiance Fields or NeRF

[4]. It has spawned countless derivative works, for example Bundle-Adjusted Radiance Fields or BARF [12] which uses bundle adjustment to eliminate the need to relative pose in training images. [2] suggest Neural Radiance Fields as one method that may be assistive to the problem of viewpoint invariance in VPR. To our knowledge we provide the first experimental examination of it's suitability for this purpose.

3 Methods

3.1 Experimental Preparation

To train a NeRF model [4], a series of images showing the scene to be rendered are required, plus their relative poses. These are then used during training to develop a spatial representation of the scene from which synthetic views are rendered. With the goal of training NeRF on realistic data available to a live system, the ORB-SLAM3 [13] SLAM system was used to generate pose estimates for sequences from the TUM VI [14] dataset. The sequences selected are monocular, with inertial measurements available, which is a realistic assumption for many mobile devices and robotic platforms today. The specific NeRF implementation used was [15].

The collected poses were normalized by subtracting their average position, centering them on the origin. A scale factor was also provided to NeRF to rescale the largest displacement to a unit of 1.0. This normalization is a strongly encouraged technique for NeRF to preserve training of fine details. [4] Every N frames were taken from the set as consecutive frames are often similar an contain redundant information, and dramatically increase training time. Values of N of 2 and 4 were used for these experiments. The 512x512 images where also rescaled to one third their original size. Training was typically run for a maximum of 150,000 to 200,000 iterations as is common for NeRF models, though progress was monitored every few thousand iterations. The graphics card used was an RX 3090 and training times typically lasted 12+ hours.

In selecting segments of the TUM IV dataset's available sessions, the objective was to find portions where the same area (eg. a room) was visited more than once in the same session, preferably with good visibility of the area, from multiple viewing angles and beyond a simple straight-line path. More straight and linear segments are expected to perform well due to consistent appearance and are a target for future exploration.

3.2 Trials

The first traversal considered is in the same direction as the training sequence, entering and then exiting the room in the same direction. The training sequence is approximately frames 612 to 665 of the corridor3 session, corresponding to visitation of a room containing equipment. The test set is a second traversal soon after, following roughly the same path from frames 713 to 759 (entering and exiting via the same doors). A traversal in the opposite direction is available near the conclusion of the session from approximately frames 4789 to 4829. A NeRF model was trained on the training sequence above and then for the pose of each test frame a corresponding synthetic frame was generated.

An additional trial was also undertaken with training frames spaced more widely in time and encompassing more of the lead in and exit of the room. This was to provide more observability early in the sequence and more viewpoint difference between training images.

4 Preliminary Experimental Results

In Figure 1 is shown the trial along a similar trajectory to that of the training set (that with more closely sampled frames), passing through the same place in the same direction. The frames on top are the synthetic versions of the ground truth frames on the bottom. These frames are representative of the overall sequence from entry to exit. They are frames 721, 729, 737, and 745 in the corridor3 session of TUM IV. These are distributed through the course of the traversal segment taken, with frames generated near the middle having higher quality.

Also conducted was a training run with more spaced frames (every four frames instead of every two) as noted above. This followed the same path but the spaced out frames allowed for more of the environment to be captured, with hopefully more variation in viewpoint

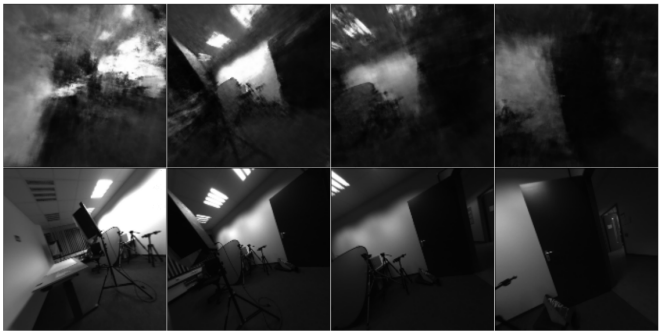

Fig. 1: From left to right: synthetic frames 721, 729, 737, and 745 with their ground truth images.

Fig. 2: An abruptly clear synthetic frame (690), likely coinciding with similar views at the end of the training set.

for the neural network to learn from. The quality of these frames was not noticeably different from the ones show from the prior run, except for the initial and concluding frames where it is believed the reduced sampling rate spread available training images more thinly and produced worse starting/ending synthetic frames which are occasionally unintelligible.

One unusual synthetic frame from the very beginning of the run stands out as being particularly clear and is presented in Figure 2. As discussed below, it is of the hallway outside the room and is believed to have coincided with a few training images depicting the hallway.

5 Discussion

5.1 Discussion of Results

In the initial trial covering the entrance and exit of the room, taking every second frame for training and testing, there are a few frames in the middle of the sequence that are the highest quality and in which large-scale features and many smaller features of the room are visible. (Figure 1, center) This is expected to be due to the contents of these frames which having been observed from as many viewpoints as possible as the camera turns and enters the room. Overall the presence of large spatial features in many of the synthetic images suggests that the synthetic images generated will be of use with those VPR descriptors most sensitive to the spatial distribution of large elements for the reliable identification of a scene.

The set of images with more spaced out sampling of images produced similar results, with perhaps negligible improvement. It is believed that to obtain more consistently good results, not only are more examples needed, but a wider variety of viewpoints would be best. One may prefer to train the model on regions which receive frequent traversals, preferably from multiple viewpoints. This provides a possible guide for where to allocate valuable training resources, and suggests some form of complementary nature with descriptors with poorer viewpoint invariance. The best trained NeRF models may occur where there is the largest variation in viewpoint and the worst descriptor performance.

Notably, one synthetic frame at the beginning of the trial with more widely spaced frames has abruptly higher clarity (Figure 2) compared to those immediately after it (which were similar to the leftmost image of Figure 1). The frame in question is believed to have a similar viewpoint to 1-2 images in the training set, looking down an adjacent hallway.

This does highlight one lacking aspect of gathering frames from continuous traversals of an environment. They tend to vary in viewpoint gradually as opposed to the carefully curated multiple views

of an environment NeRF is normally trained on. In addition to reinforcing the point above about preferring traversals with variation in viewpoint (or multiple traversals) during training, this also highlights a point of caution. In linear environments which tend to result in especially similar trajectories, hallways being a prime example, there may be a possibility of poor generalization and thus generation of different viewpoints.

In future work we aim to make use of multiple traversals of more environments and examine how the variety of training examples provided affects the model's ability to respond to changes in the requested synthetic viewpoints, gauging the requirements of generalization on trajectory frames. Also still ahead is the implementation and testing on VPR descriptors upon NeRF's synthetic images, performing tests to determine how performance benefit the best synthetic views can bring to the task of places recognition. This investigation lays the foundation of incorporation of NeRF virtual views into VPR systems, though reduction of NeRF's training overhead will also be required.

5.2 Additional Suggested Exploration

Training of NeRF models is very intensive and it is desirable to reduce this burden. One way in which the continuous nature of trajectories may be exploited is the gradual introduction of new scene geometry. It may be possible in future to gradually evolve NeRF models as the system travels and new views become available, saving model copies periodically for past regions. If true, this may help to amortize the cost of NeRF training.

One observation made regarding even the best synthetic frames and their ground truth is that there is frequently some small viewpoint difference between them despite the same pose. A possible explanation for this is drift in the SLAM system's estimate of pose or scale. A method worth considering in this case is the use of registration to better align the estimate of pose between the training images and the current reference frame when requesting a synthetic image. The easiest way to accomplish this may be to perform an additional registration step using observed SLAM mappoints or other features.

6 Conclusions

Visual place recognition is a critical requirement for a variety of navigational applications, but extensions to the conditions of it's deployment also pose new challenges. The need to compensate for changes in appearance has lead to a tension with robustness to changes in viewpoint. Synthetic views which can mimic alternate viewpoints present a possible solution by decoupling these problems and providing consistent viewpoints to appearance-invariant descriptors which underpin VPR's recognition. We find training NeRF synthetic view models on SLAM trajectories produces encouraging results in some situations. We recommend ensuring adequate variation in viewpoint is provided which may be difficult with linear trajectories. If possible, train from multiple trajectories to ensure generalization. We expect that with the suggestions made here and continual improvement of the costs of NeRF training, Neural Radiance Fields and similar view synthesis methods will become a valuable tool to future VPR methods.

Acknowledgments

I would like to thank the Canadian National Engineering Science and Research Council (NSERC) for supporting this research.

References

[1] H. Jégou, M. Douze, C. Schmid, and P. Pérez, "Aggregating local descriptors into a compact image representation," in *2010 IEEE computer society conference on computer vision and pattern recognition*. IEEE, 2010, pp. 3304–3311.

[2] S. Garg, T. Fischer, and M. Milford, "Where is your place, visual place recognition?" *arXiv preprint arXiv:2103.06443*, 2021.

[3] H. Taira, M. Okutomi, T. Sattler, M. Cimpoi, M. Pollefeys, J. Sivic, T. Pajdla, and A. Torii, "Inloc: Indoor visual localization with dense matching and view synthesis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7199–7209.

[4] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," in *ECCV*, 2020.

[5] X. Zhang, L. Wang, and Y. Su, "Visual place recognition: A survey from deep learning perspective," *Pattern Recognition*, vol. 113, p. 107760, 2021.

[6] J. Sivic and A. Zisserman, "Video google: A text retrieval approach to object matching in videos," in *Computer Vision, IEEE International Conference on*, vol. 3. IEEE Computer Society, 2003, pp. 1470–1470.

[7] F. Perronnin, Y. Liu, J. Sánchez, and H. Poirier, "Large-scale image retrieval with compressed fisher vectors," in *2010 IEEE computer society conference on computer vision and pattern recognition*. IEEE, 2010, pp. 3384–3391.

[8] Z. Chen, A. Jacobson, N. Sünderhauf, B. Upcroft, L. Liu, C. Shen, I. Reid, and M. Milford, "Deep learning features at scale for visual place recognition," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2017, pp. 3223–3230.

[9] M. Milford, C. Shen, S. Lowry, N. Suenderhauf, S. Shirazi, G. Lin, F. Liu, E. Pepperell, C. Lerma, B. Upcroft *et al.*, "Sequence searching with deep-learnt depth for condition-and viewpoint-invariant route-based place recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2015, pp. 18–25.

[10] A. Torii, R. Arandjelovic, J. Sivic, M. Okutomi, and T. Pajdla, "24/7 place recognition by view synthesis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1808–1817.

[11] D. Sibbing, T. Sattler, B. Leibe, and L. Kobbelt, "Sift-realistic rendering," in *2013 International Conference on 3D Vision-3DV 2013*. IEEE, 2013, pp. 56–63.

[12] C.-H. Lin, W.-C. Ma, A. Torralba, and S. Lucey, "Barf: Bundle-adjusting neural radiance fields," *arXiv preprint arXiv:2104.06405*, 2021.

[13] C. Campos, R. Elvira, J. J. Gomez, J. M. M. Montiel, and J. D. Tardós, "ORB-SLAM3: An accurate open-source library for visual, visual-inertial and multi-map SLAM," *arXiv preprint arXiv:2007.11898*, 2020.

[14] S. Klenk, J. Chui, N. Demmel, and D. Cremers, "Tum-vie: The tum stereo visual-inertial event dataset," in *International Conference on Intelligent Robots and Systems (IROS)*, 2021.

[15] B. Trabucco, "Nerf," https://github.com/brandontrabucco/nerf, 2021.