Kidnapped Radar: Topological Radar Localisation using Rotationally-Invariant Metric Learning

Ştefan Săftescu, Matthew Gadd, Daniele De Martini, Dan Barnes, and Paul Newman
Oxford Robotics Institute, Dept. Engineering Science, University of Oxford, UK.
{stefan,mattgadd,daniele,dbarnes,pnewman}@robots.ox.ac.uk

Abstract—This paper presents a system for robust, large-scale topological localisation using Frequency-Modulated Continuous-Wave (FMCW) scanning radar. We learn a metric space for embedding polar radar scans using CNN and NetVLAD architectures traditionally applied to the visual domain. However, we tailor the feature extraction for more suitability to the polar nature of radar scan formation using cylindrical convolutions, anti-aliasing blurring, and azimuth-wise max-pooling; all in order to bolster the rotational invariance. The enforced metric space is then used to encode a reference trajectory, serving as a map, which is queried for nearest neighbours (NNs) for recognition of places at run-time. We demonstrate the performance of our topological localisation system over the course of many repeat forays using the largest radar-focused mobile autonomy dataset released to date, totalling $280 \text{ km}$ of urban driving, a small portion of which we also use to learn the weights of the modified architecture. As this work represents a novel application for FMCW radar, we analyse the utility of the proposed method via a comprehensive set of metrics which provide insight into the efficacy when used in a realistic system, showing improved performance over the root architecture even in the face of random rotational perturbation.

Index Terms—radar, localisation, place recognition, deep learning, metric learning

I. INTRODUCTION

In order for autonomous vehicles to travel safely at higher speeds or operate in wide-open spaces where there is a dearth of distinct features, a new level of robust sensing is required. FMCW radar satisfies these requirements, thriving in all environmental conditions (rain, snow, dust, fog, or direct sunlight), providing a $360^\circ$ view of the scene, and detecting targets at ranges of up to hundreds of metres with centimetre-scale precision.

Indeed, it has been shown that this class of radar can be effectively used for accurate motion estimation in various challenging environments using scan matching and data association of hand-crafted features [3, 4, 5]. Real-time deployment of this type of approach is possible by preprocessing the radar measurement stream and easing the data association burden [6]. The present state-of-the-art in Radar Odometry (RO) learns masks to apply to the radar data stream as well as an artefact- and distraction-free embedding in an end-to-end fashion [7].

With these modern capabilities, it is currently possible to apply FMCW radar to the construction of accurate map representations for use in an autonomy stack. Metric pose estimation in an unconstrained search over all places represented in the map is therefore feasible but would not scale well with the size of the environment. In the best case, when using heuristics for constraining the graph search, eventual drift in the ego-motion is likely to invalidate any reported poses.

This paper thus presents a system which reproduces and advances in the radar domain the current capabilities in
visual place recognition to produce topological localisation suggestions which we envision being used downstream for metric pose estimation. We believe that this represents the first occasion in which place recognition is performed for the FMCW class of radar. As our radar produces 360° scans we note that, unlike narrow field-of-view (FOV) cameras, the orientation of the sensor is irrelevant for place recognition: whether the vehicle is facing east or west on a street, the scan will be the same up to rotation. With this observation in mind, we design a Fully Convolutional Neural Network (FCNN) which is quasi-invariant to rotations of the input scans, and learn an embedding space which we can query for similarity between a reference trajectory and the live scan.

This paper proceeds by reviewing existing literature in Section V followed by a brief preliminary discussion of radar image formation in Section VI. Section VII gives an overview of our system and motivates the desired online operation, followed by a description in Section VIII of an offline learning stage which satisfies these design principles. Finally, Section VII presents necessary details for implementation as well as our experimental philosophy before Section VIII where we report our results.

II. RELATED WORK

Place recognition is a somewhat consolidated procedure in the camera sensor modality. A brief history of the community’s progress in this regard includes: probabilistic models around bag-of-words image representations [8], sequence-based approaches [9], and more recently by extracting features from the responses of Convolutional Neural Network (CNN) layers and subsequent use of these features for image comparison [10].

There has also been extensive investigation of Light Detection and Ranging (LiDAR)-based place recognition, often relying on geometrical features to overcome extreme appearance change, including systems based on: matching of 3D segments [11], semantic graph descriptor matching [12], learned discriminative global features [13], and combining the benefits of geometry and appearance by coupling the conventional geometric information from the LiDAR with its calibrated intensity return [14].

Besides its superior range and despite its lower spatial resolution, Millimetre-Wave (MMW) radar often overcomes the shortcomings of laser, monocular, or stereo vision because it can operate successfully in dust, fog, blizzards, and poorly lit scenes [15]. In [16] it is shown in the context of a Simultaneous Localisation and Mapping (SLAM) system that while producing slightly less accurate maps than LiDARs, radars are capable of capturing details such as corners and small walls.

Place recognition with Ultra Wide Band (UWB) radar is presented in [17] by matching received signals to a database of waveforms which represent signatures of places. Although the UWB class of radar is capable of very high update rates, it is shown in [18] that the FMCW class is superior in raw measurement quality, measured maximum range, and worst-case precision. As our system is eventually to be included in a larger framework which must yield precise pose estimation (c.f. Section I), the FMCW class is therefore our preferred sensor. Furthermore, evaluation in [17] was performed in indoor and forested environments, whereas our work is deployed in built environments representative of urban driving.

III. PRELIMINARIES

We use a FMCW scanning radar which rotates about its vertical axis while sensing the environment continuously through the transmission and reception of frequency-modulated radio waves. While rotating, the sensor inspects one angular portion (azimuth) of space at a time and receives a power signal that is a function of reflectivity, size, and orientation of objects at that specific azimuth and at a particular distance. The radar takes measurements along an azimuth at one of a number of discrete intervals and returns a list of power readings, or bins. We call one full rotation across all azimuths a scan, some examples of which are shown in Figures 1 and 3.

IV. SYSTEM OVERVIEW

Figure 2 depicts the overall method. As motivated in Section I we require a system which produces topological localisation suggestions, used downstream for metric pose estimation.

A. Design requirements

To satisfy our design outcomes, we do not require a mature SLAM system which models environment concurrently with estimating the state of the sensor [19]. Instead, we approach the place recognition problem as a nearest neighbour (NN) search in a multidimensional space, where each portion of the space describes a different place and points within a portion represent different views of a place.

We consider this approach well-posed as the invariance of radar measurements to changing environmental conditions (such as illumination, rain, fog, etc) implies that a map built from a single experience of the route will likely be of good utility over the course of several months or seasons, as only changes to the structure of the scene itself (e.g. building construction) will present significant variation in scene appearance.
to our network during training. Each scan shown is a range-versus-azimuth grayscale image. Batches are constructed such that there is no overlap in the radar sensing horizon between anchor scans (top row). Scans with returns painted with the same colour as an anchor (red, green, or blue) are marked as positive examples of topological localisations (columns).

Figure 3. A (contrast-enhanced) visualisation of a batch of radar scans input to our network during training. Each scan shown is a range-versus-azimuth grayscale image. Batches are constructed such that there is no overlap in the radar sensing horizon between anchor scans (top row). Scans with returns painted with the same colour as an anchor (red, green, or blue) are marked as positive examples of topological localisations (columns).

B. Offline learning

To achieve our requirements, good metric embeddings of the polar radar measurements are required which can be used to create a map of the environment in which the robot will operate by encoding places of interest offline.

Generation of these embeddings is delegated to an encoder network (c.f. Section V) which extracts information from the radar measurements and compresses them within the multidimensional space. The training procedure enforces that the network will learn this transformation.

Because of the geographical scale of the environment which must be encoded for representation (large urban centres), exploitation of common data structure techniques to discretise this space for fast lookups is essential to reduce the NN search complexity. In the results discussed in Section VII a k-dimensional tree [20] structure is used. This choice guarantees the exactness of the search result, thus not influencing our discussion of the accuracy of the learned representation. Other CPU- or GPU-based methods allow for faster, although approximate, searches [21][22].

C. Mapping and Localisation

At deployment time, inference on the network involves a feed-forward pass of a single radar scan, resulting in an embedding, i.e. a point in a multidimensional space. The metric nature of the learned space allows us to obtain a measure of similarity to the stored database and query it for topological localisation candidates. To this end, the discretised space discussed above is traversed for the closest places from the database of known locations within a certain embedding distance threshold. Alternatively, a fixed-size set, \(N\), of top scoring candidates are all considered.

V. LEARNING THE METRIC SPACE

To learn filters and cluster centres which help distinguish polar radar images for place recognition we use NetVLAD [23] with VGG-16 [24] as a front-end feature extractor. Specifically, we modify the implementation described in [25] to make the network invariant to the orientation of input radar scans. We refer to the original architecture as VGG-16/NetVLAD and our proposed architecture as OURS.

A. Feature extraction

With similar motivation to [26] we apply circular padding to the CNN feature spaces to reflect the fact that the polar representation of the assembled Fast Fourier Transform (FFT) returns has no true image boundary along the azimuth axis. This provides rotation equivalence throughout the network.

A common design in CNNs is to downsample feature maps every few layers to reduce computation cost and increase the input area a single network filter receives information from. As noted by [27], this breaks the translation equivalence CNNs are usually assumed to have and therefore also the rotational equivariance provided by circular padding. We apply the same solution from [27] in our network by replacing the usual max-pooling with stride 2 used for downsampling in VGG-16 with stride 1 max-pooling, followed by a stride 2 Gaussian blur. While this does not fully restore rotational equivariance, [27] show that it greatly reduces the aliasing introduced by downsampling.

Finally, to make the network rotationally invariant (up to the small aliasing that remains from downsampling), we perform max-pooling upon the last feature map along the azimuth axis. As the max function is commutative and associative, and the last feature map is rotationally equivariant, the result will be rotationally invariant.

B. Enforcing the metric space

To enforce the metric space, we perform online triplet mining and apply the triplet loss described in [28]. Loop closure labels are taken from a ground truth dataset, which will be discussed in Section VII. Batches, as illustrated in Figure 3, are constructed such that there is no overlap of the radar sensing horizon between a candidate radar scan and any anchor scan already sampled for the batch.

C. Training details and hyperparameters

Due to memory limitations on our graphical hardware, we crop the last 178 range bins and scale the width by a factor of 8 such that the original \(400\times3768\) polar radar scans are input to the network with resolution \(400\times450\) (c.f. Figure 3). As the azimuth axis remains unscaled, this does not affect rotational invariance.

When finetuning either the original architecture or our proposed modified architecture, we initialise internal weights with the publicly available checkpoint vdl16_pitts30k_conv5_3_vlad_preL2_intra_white, corresponding to the best performing model described in [25].

Our learning rate schedule applies a linear decay function initialised at \(1 \times 10^{-4}\) and settling to \(5 \times 10^{-6}\) at 5000

\[\text{github.com/uzh-rpg/netvlad_tf_open}\]
rotationally invariant design of our network architecture, as dataset. In particular, we did not randomly rotate the input
71° and a scan rotation rate of 0.04 s. We use two one-dimensional gaussian blur kernels with size 7 and standard deviation of 1.

No augmentation has been performed on the training dataset. In particular, we did not randomly rotate the input scans during training, in order to show the resilience of the rotationally invariant design of our network architecture, as assessed in Section VII.

VI. EXPERIMENTAL SETUP
This section details our experimental setup and philosophy.

A. Platform and sensor specifications
The experiments are performed using data collected from the Oxford RobotCar platform [31]. The vehicle, as described in the recently released Oxford Radar RobotCar Dataset [32], is fitted with a CTS350-X Navtech FMCW scanning radar without Doppler information, mounted on top of the platform with an axis of rotation perpendicular to the driving surface. This radar is characterised by an operating frequency of 76 GHz to 77 GHz, yielding up to 3768 range readings with a resolution of 4.38 cm (a total range of 165 m), each constituting one of the 400 azimuth readings with a resolution of 0.9° and a scan rotation rate of 4 Hz.

B. Ground truth location
For ground truth location, we manipulate the accompanying ground truth odometry described in [32], which is computed by a global optimisation using Global Positioning System (GPS), robust Visual Odometry (VO) [33], and visual loop closures from FAB-MAP [8].

As each ground truth odometry file does not begin at the same location in the urban environment, we manually selected (for each of the 32 trajectories) a moment during which the vehicle was stationary at a common point and manually aligned the ground traces. Furthermore, we align the ground traces manually by introducing a small rotational offset to account for differing attitudes at those instances.

The ground truth is preprocessed offline to capture the subsets of nodes that are at a maximum predefined distance, creating a graph-structured database that can easily be queried for triplets of nodes for training the system.

C. Dataset demarcation
Each approximately 9 km trajectory in the Oxford city centre was divided into three distinct portions: train, valid, and test. The maximum radar range was foreshortened due to the cluttered nature of the urban environment and we were thus able to specifically design the sets such that they did not overlap, padding the splits where necessary.

Figure 1 shows the GPS trace of the test split, which was specifically selected as the vehicle traverses a portion of the route in the opposite direction. The valid split selected was quite simple, consisting of two straight periods of driving separated by a right turn.

ori.ox.ac.uk/datasets/radar-robotcar-dataset

32019-01-10-11-46-21-radar-oxford-10k and 2019-01-10-14-50-05-radar-oxford-10k from ori.ox.ac.uk/datasets/radar-robotcar-dataset/datasets

Figure 4. Average PR curves and one standard-deviation bounds when using the learned representation to map an environment once and repeatedly localise consecutive trajectories against the static reference trajectory. The corresponding maximum $F_1$ scores are $0.70 \pm 0.04$ for ours (green) as compared to $0.68 \pm 0.04$ for VGG-16/NetVLAD (blue). The corresponding maximum $F_2$ scores are $0.71 \pm 0.04$ for ours as compared to $0.69 \pm 0.04$ for VGG-16/NetVLAD. The corresponding maximum $F_2$ scores are $0.77 \pm 0.04$ for ours as compared to $0.76 \pm 0.04$ for VGG-16/NetVLAD.

Figure 5. Average PR curves and one standard-deviation bounds when using the learned representation to map a difficult, unseen environment with backwards traversals, repeatedly localising consecutive trajectories in the test split. The corresponding maximum $F_1$ scores are $0.53 \pm 0.03$ for ours (green) as compared to $0.48 \pm 0.02$ for VGG-16/NetVLAD (blue). The corresponding maximum $F_2$ scores are $0.60 \pm 0.03$ for ours as compared to $0.57 \pm 0.02$ for VGG-16/NetVLAD. The corresponding maximum $F_2$ scores are $0.55 \pm 0.03$ for ours as compared to $0.46 \pm 0.02$ for VGG-16/NetVLAD.
We then apply the learned metric space to encode an entire trajectory from the dataset (c.f. Section VI-C), including data from all splits (train, valid, and test). This encoded trajectory is used as a static map along which all other trajectories in the dataset are localised against. We exclude the pair of trajectories which we use to train the network. Figure 4 shows average PR curves with one standard-deviation bounds. The corresponding AUC are 0.75 ± 0.06 for ours as compared to 0.72 ± 0.05 for VGG-16/NetVLAD. This experiment serves to indicate that our proposed modifications result in measurable performance improvements over the baseline system.

We then better illustrate the rotational invariance of our proposed architecture by showing in Figure 5 average PR curves when only data from the test split is used for mapping and subsequent localisation. This part of the environment was not seen by the network during training, and consists of a backwards traversal during which the vehicle is driving on the opposite side of the street (c.f. Figure 1). Here, the corresponding AUC is 0.52 ± 0.04 for ours as compared to 0.41 ± 0.03 for VGG-16/NetVLAD.

Next, we use the static map built with all splits to localise incoming query frames which have been randomly perturbed along the azimuth axis, to probe the resilience to rotational disturbance. Figure 6 shows degradation in the performance of VGG-16/NetVLAD while ours maintains the ability to recognise places. Admittedly, we did not expect VGG-16/NetVLAD to perform well under these conditions, as we have performed no data augmentation which would account for this perturbation. However, considering that ours was also trained on upright scans, this result vindicates the proposed architectural modifications.

Figure 6. PR curves representing localisation of all dataset trajectories against a static map when the input frames are randomly perturbed along the azimuth axis. In comparison to Figure 4, the performance of VGG-16/NetVLAD degrades; with maximum $F_1$, $F_{0.5}$, and $F_2$ scores of $0.23 \pm 0.02$, $0.25 \pm 0.02$, and $0.30 \pm 0.02$ respectively. In contrast, our system – having been designed to be quasi-rotationally invariant – maintains approximately the same performance level; with maximum $F_1$, $F_{0.5}$, and $F_2$ scores of $0.69 \pm 0.04$, $0.70 \pm 0.04$, and $0.77 \pm 0.04$ respectively.
% of localisation failures, \( d \), within threshold (m)

Figure 7. Variation in our system’s performance when randomly rotated input frames are used for inference, as the fractional deviation from the ideal performance level when inference is performed upon unaltered, upright frames. The dependent axis is a percentage, not a fraction. The range of absolute values for the ideal performance against which this variation is measured is, for “frames correctly localised”, (90.82\%, 97.59\%). This means that, even when only considering the 1-NN in embedding space, we are able to localise correctly 90.82\% of the time. On the scale of the journeys represented here (27 localised trajectories of about 9 km each), this corresponds to approximately 221 km of good localisations over a 243 km drive.

Figure 8 shows the performance of ours as a set of relative measures, where the upright condition is taken as the baseline signal, and the rotationally-perturbed condition is enumerated as fractional variation away from this ideal performance. The independent variable shown in Figure 7 is the number of database candidates which are closest in the embedding metric space which would have to be disambiguated by a downstream process (e.g. geometric verification in scan matching). This is different to the threshold sweep used to generate the PR curves (c.f. Figures 4 to 6), which corresponds to a ball in the multidimensional space. Each of these quantities is averaged over all 27 localisation trajectories against the same static map, as above. The ranges of change in corresponding absolute values from \( N = 1 \) to \( N = 50 \) are (90.82\%, 97.59\%) for frames correctly localised, (94.67\%, 78.78\%) for precision, and (0.93\%, 34.55\%) for recall, all for the upright condition. We observe again that our system is extremely robust to rotational disturbances, where each of these systems metrics are within 0.6\% of the ideal, upright condition.

Finally, Figure 8 shows the performance of our system as histograms of failure severity which are measured as the proportion of drop-outs in correct localisation results during which the vehicle moves a certain distance. We observe that over 90\% of the failures are limited to a driven distance of less than 3.75 m, even when only the closest embedding in the metric space is taken as the localisation result. As the number of candidates considered increases, this proportion tapers off as the worst failures are alleviated. The worst measured failure for \( N = 1 \) is 20.00 m, which decreases to 9.33 m for \( N = 50 \).

VIII. CONCLUSIONS AND FUTURE WORK

We have presented a system for radar-only place recognition using a metric feature space which is learned in a rotationally-invariant fashion. We described adjustments to off-the-shelf image recognition frameworks for better suitability to the native radar polar scan representation. We demonstrated the efficacy of our approach on the largest radar-focused autonomous driving dataset collected to date, showing improved localisation capability compared to the naive application to the radar domain of competitive vision-based approaches, especially in the face of severe rotational disturbance.

In the future we plan to integrate the system presented in this paper with our mapping and localisation pipeline which is built atop of the scan-matching algorithm of [36] and to deploy the combined system in a teach-and-repeat autonomy scenario using the platform we have presented in [35], the conception of which was in large part concerned with deploying our FMCW radar scanner.

ACKNOWLEDGMENTS

This project is supported by the Assuring Autonomy International Programme, a partnership between Lloyds Register Foundation and the University of York. We would also like to thank our partners at Navtech Radar. Șтеfan Săftescu is supported by UK’s Engineering and Physical Sciences Research Council (EPSRC) through the Centre for Doctoral Training (CDT) in Autonomous Intelligent Machines and Systems (AIMS) Programme Grant EP/L015897/1. Matthew Gadd is supported by Innovate UK under CAV2 – Stream 1 CRD (DRIVEN). Dan Barnes is supported by the UK EPSRC Doctoral Training Partnership. Daniele De Martini and Paul Newman are supported by the UK EPSRC programme grant EP/M019918/1.
REFERENCES

[1] OpenStreetMap contributors, “Planet dump retrieved from https://planet.osm.org,” [https://www.openstreetmap.org], 2017.

[2] M. Haklay and P. Weber, “OpenStreetMap: User-generated street maps,” IEEE Pervasive Computing, vol. 7, no. 4, pp. 12–18, 2008.

[3] H. Cen and P. Newman, “Precise ego-motion estimation with millimeter-wave radar under diverse and challenging conditions,” Proceedings of the 2018 IEEE International Conference on Robotics and Automation, 2018.

[4] S. Cen and F. Newman, “Radar-only ego-motion estimation in difficult settings via graph matching,” in Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), Montreal, Canada, 2019.

[5] R. Aldera, D. De Martini, M. Gadd, and P. Newman, “What Could Go Wrong? Retrospective Radar Odometry in Challenging Environments,” in IEEE Intelligent Transportation Systems (ITSC) Conference, Auckland, New Zealand, October 2019.

[6] R. Aldera, D. De Martini, M. Gadd, and P. Newman, “Fast radar motion estimation with a learnt focus of attention using weak supervision,” in Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), Montreal, Canada, 2019.

[7] D. Barnes, R. Weston, and I. Posner, “Masking by moving: Learning distraction-free radar odometry from pose information,” arXiv preprint arXiv:1909.03752, 2019. [Online]. Available: [https://arxiv.org/pdf/1909.03752]

[8] M. Cummins and P. Newman, “FAB-MAP: Probabilistic localization and mapping in the space of appearance,” The International Journal of Robotics Research, vol. 27, no. 6, pp. 647–665, 2008.

[9] M. J. Milford and G. F. Wyeth, “Sceqml: Visual route-based navigation for sunny summer days and stormy winter nights,” in 2012 IEEE International Conference on Robotics and Automation, IEEE, 2012, pp. 1643–1649.

[10] A. Khalij, S. Ehsan, M. Milford, and K. McDonald-Maier, “A holistic visual place recognition approach using lightweight cnns for severe viewpoint and appearance changes,” arXiv preprint arXiv:1811.03032, 2018.

[11] R. Dubé, D. Dugas, E. Stumm, J. Nieto, R. Siegwart, and C. Cadena, “Segmatch: Segment based place recognition in 3d point clouds,” in 2017 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2017, pp. 5266–5272.

[12] A. Gavel, C. Del Don, R. Siegwart, J. Nieto, and C. Cadena, “X-view: Graph-based semantic multi-view localization,” IEEE Robotics and Automation Letters, vol. 3, no. 3, pp. 1687–1694, 2018.

[13] M. Angelina Uy and G. 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, 2018, pp. 4470–4479.

[14] J. Guo, P. V. Borges, C. Park, and A. Gavel, “Local descriptor for robust place recognition using lidar intensity,” IEEE Robotics and Automation Letters, vol. 4, no. 2, pp. 1470–1477, 2019.

[15] G. Reina, D. Johnson, and I. Underwood, “Radar sensing for intelligent vehicles in urban environments,” Sensors, vol. 15, no. 6, pp. 14661–14678, 2015.

[16] M. Mielle and A. L. A. J. Magnusson, Martin, “A comparative analysis of radar and lidar sensing for localization and mapping,” in Proceedings of the European Conference on Mobile Robotics (ECMR), 2018.

[17] K. S. Takeuchi, A. Elfes, and J. Roberts, “Localization and place recognition using a ultra-wide band (uwb) radar,” in Workshop on Positioning, Navigation and Communications (WPNC). IEEE, 2016, pp. 1–6.

[18] C. Cadena, L. Carlone, H. Carrillo, Y. Latif, D. Scaramuzza, J. Neira, I. Reid, and J. J. Leonard, “Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age,” IEEE Transactions on robotics, vol. 32, no. 6, pp. 1309–1332, 2016.

[19] J. L. Bentley, “Multidimensional binary search trees used for associative searching,” Communications of the ACM, vol. 18, no. 9, pp. 509–517, 1975.

[20] Y. A. Malkov and D. A. Yashunin, “Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs,” CoRR, vol. abs/1603.09320, 2016.

[21] P. Wieschollek, O. Wang, A. Sorkine-Hornung, and H. P. A. Lensch, “Efficient large-scale approximate nearest neighbor search on the gpu,” in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2016.