Loop closure with a low power millimeter wave radar sensor using an autoencoder

Pieter Meiresone[1], David Van Hamme[1], and Wilfried Philips[1]

[1]Ghent University, IPI-imec
{pieter.meiresone, david.vanhamme, wilfried.philips}@ugent.be

Abstract

In this paper, we will consider place recognition, more commonly know as loop closure, with a low resolution single-chip millimeter wave (mmWave) radar in indoor environments. It is an essential part in simultaneous localization and mapping (SLAM) systems to avoid drift. By using a novel method to create descriptors or latent codes with an autoencoder in combination with exploiting the temporal similarity between our latent codes, we are able to successfully extract loop closures with a radar-only system without requiring ground truth. Our proposed method is validated in an industrial IoT lab on an Unmanned Aerial Vehicle (UAV) and on a cargo bike in a parking building.

1 Introduction

Loop closure can be divided in two parts, firstly a sensor reading has to be translated in to a descriptor or latent code which can be used to compare frames. Similar descriptors need to be extracted and verified as a genuine loop closure. In [1], an image is translated into a bag of words which is matched with other frames. Variations exist where multiple images are used to create one descriptor[2], or by comparing the descriptor of one image to a sequence of images[3, 4].

Recently, neural networks have been used to provide descriptors. An excellent overview is given by [5]. These methods can use fully connected (FC) layers as descriptors, feature maps from intermediate CNN layers and/or pooling layers to provide a more compact descriptor. The current state-of-the-art [6] uses a combination of convolutional layers and generalized max/average pooling layers.

When descriptors fail to distinguish between visually similar but not exactly the same scenes, a Visual Similarity Matrix (VSM) [7, 8] is constructed between all image pairs in which long sequences of high similarity are extracted as loop closures by using, for example, the Smith-Waterman [9] algorithm.

All the methods above mainly work on place recognition based on camera images. The release of the Oxford radar dataset [10] launched more loop closure research with radar sensors. In [11], a pose estimation neural network is proposed and re-used for loop closure based on predicted keypoint descriptors. Based on these, the cosine similarity is calculated between all pairs of radar images. All pairs above a certain threshold are extracted as loop closures. In [12], a modified NetVLAD architecture is proposed for the radar domain. A comparison with the standard NetVLAD architecture for the visual domain is made and the benefits of specific processing for radar data are made clear. The descriptor contains 4096 dimensions. Thanks to the highly detailed radar images, the descriptors contain sufficient detail to unambiguously compare different locations. For a low power FMCW radar sensor, we are interested in descriptors that are significantly smaller due to the limited range and azimuth resolution.

There are also a lot of similarities between loop closure on Lidar point clouds and radar images. In [13] an overview is given, methods can be divided in 2 categories: segmentation and descriptor (both global and local) based methods. While being similar sensors, these methods require a high resolution.

To the best of our knowledge, no research has been performed yet on using a low power mmWave radar only-system for loop closure. In [14], joint radar and event based camera (DVS) latent codes are used for loop closure detection on an Unmanned Aerial Vehicle (UAV). The DVS camera and a spiking radar image is fed to multiple Spiking Neural Networks (SNN). The output spikes from the different SNN are used as latent codes for loop closure. While the authors use also a low power radar, no results are reported for a radar-only system. Furthermore, the loop closure detection method does not take into account perspective changes when there is a loop closure.

In this paper, our main contribution is to use an autoencoder to obtain a good descriptor from which loop closures can be extracted by using the temporal similarity between the descriptors. We also propose a method to determine the exact point of loop closure by taking into account perspective changes.

Proceedings of the 5th Northern Lights Deep Learning Conference (NLDL), PMLR 233, 2024.

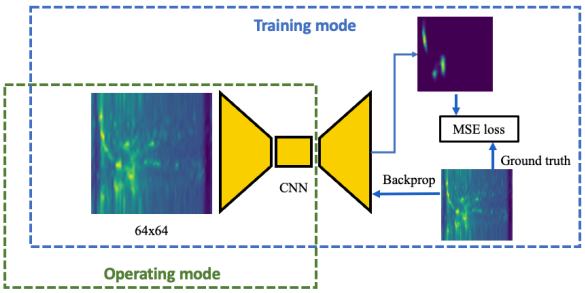
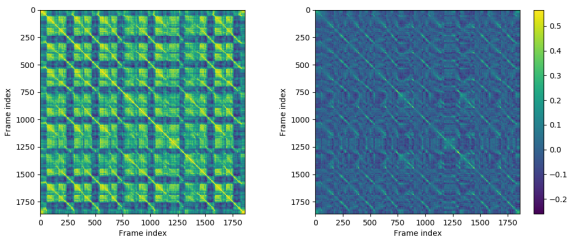


Figure 1. Training and operating mode of the neural network. In operating mode, the decoder part is skipped and the output at the bottleneck is used for loop closure detection.



(a) S_0 **(b)** S_2

Figure 2. Similarity matrix and the rank-reduced version. Loop closures can be seen in the long sequences of high similarity around the sub-diagonals

2 Proposed method

We will perform loop closure in an indoor environment with a low resolution radar sensor on a UAV and a cargo bike. Firstly, range-azimuth radar images are converted to latent codes by an autoencoder. By looking at the similarity and temporal information between latent codes, we will detect loop closures. Finally, since a loop closure never happens at exactly the same point we propose a method to calculate the perspective change or pose transformation. Recording and processing of the radar data happens in the local coordinate system of the UAV. In practice, roll and pitch angles are small. By approximating them by 0, processing is simplified from a 6 to 3 degrees of freedom problem.

2.1 Creating latent codes using autoencoders

An autoencoder network consists of an encoding part and a decoding part. The network is trained to have an output identical to the input, but it is forced to learn patterns in the input data due to the bottleneck. Mathematically, an autoencoder can be expressed as follows. Let $f(\mathbf{x})$ be the encoder, and $g(\mathbf{f}(\mathbf{x}))$ the decoder. The goal of the neural network is then to minimize a loss function $L(\mathbf{x}, g(\mathbf{f}(\mathbf{x})))$ where L is a loss function penalizing $\mathbf{x}$ being dissimilar from $g(\mathbf{f}(\mathbf{x}))$, in our case the mean squared error (MSE). The output of the neural network, after training, is a denoised range-azimuth image which shows better the features of the environment. The neural network tries to capture all the relevant information in the bottleneck from which the input image can be the most accurately approximated. The representation at this point is thus the most suited to be used as latent code. An overview of the autoencoder is shown in Fig. 1.

The similarity between latent codes $\mathbf{t}_i$ and $\mathbf{t}_j$ for respectively radar frame i and j is calculated using the cosine similarity δ as follows:

$$\delta_{i,j} = \delta(\mathbf{t}_i, \mathbf{t}_j) = \frac{\mathbf{t}_i \cdot \mathbf{t}_j}{\|\mathbf{t}_i\| \|\mathbf{t}_j\|}$$

The cosine similarity is calculated for every possible frame pair resulting in a similarity matrix S with $S_{ij} = \delta_{i,j}$. A sequence of n radar frames results in a matrix of size $n \times n$.

This matrix is shown in Fig. 2(a). The checkerboard pattern is caused by similar areas in our warehouse between the different aisles. An aisle looks very similar for the radar, independently of its position in the aisle. This leads to high similarity between the latent codes.

2.2 Ambiguity reduction with a singular value decomposition

As illustrated by the authors in [7], it is possible to reduce the ambiguity in the similarity matrix by looking at the singular value decomposition (SVD) and removing the biggest eigenvalues which are responsible for the ambiguities. Their method is summarized below since it is an important part of our algorithm. Since S is a symmetric real $n \times n$ matrix, there is an orthogonal matrix Q and a diagonal matrix Λ such that $S = Q\Lambda Q^T$. The columns of Q are eigenvectors and the diagonal elements of Λ are eigenvalues.

When removing an eigenvalue, we reduce the rank of our similarity matrix. We refer to our rank reduced matrix as S_r.

$$S_r = Q\Lambda_r Q^T \tag{1}$$

$$= Q \begin{bmatrix} 0 & \dots & \dots & 0 \\ \vdots & \lambda_r & & \vdots \\ \vdots & & \ddots & \vdots \\ 0 & \dots & \dots & \lambda_n \end{bmatrix} Q^T \tag{2}$$

For S_2, shown in Fig. 2(b), many ambiguities have been removed and loop closures can be easier identified.

2.3 Sequence detection using temporal information

From a rank reduced similarity matrix, we want to calculate frame sequences $A = [a_1, ..., a_i, ..., a_n]$ and $B = [b_1, ..., b_i, ..., b_n]$ where a_i, b_j are frame indexes. The tuple (a_i, b_i) is a corresponding frame pair within a corresponding sequence of frames noted as the tuple (A, B). In [7], a modified version of the Smith-Waterman algorithm is used. It is a dynamic programming method by using a scoring matrix H to keep track of the matching scores for all possible sequences. We modified the algorithm in [7] to extract multiple corresponding subsequences. Each element $H_{i,j}$ can be considered as the cumulative score of a subsequence (A, B). When aligning sequences, there are 3 possible moves through the matrix S: diagonal, horizontal and vertical. Horizontal and vertical receive a penalty δ, since we assume the vehicle always has a non-zero velocity, however due to velocity differences they have to be allowed. The matrix H is initialized to 0 and then constructed recursively as follows:

$$H_{i,j} = \begin{cases} 0 & \text{if } S_{i,j} < \alpha \\ H_{i-1,j-1} + S_{i,j} & \text{if } H_{i-1,j-1} = H^{max}_{i,j} \\ H_{i,j-1} + S_{i,j} - \delta & \text{if } H_{i,j-1} = H^{max}_{i,j} \\ H_{i-1,j} + S_{i,j} - \delta & \text{if } H_{i-1,j} = H^{max}_{i,j} \end{cases} \tag{3}$$

$$H^{max}_{i,j} = \max(H_{i-1,j-1}, H_{i,j-1}, H_{i-1,j}) \tag{4}$$

The end of a possible sequence is signaled by $S_{i,j} < \alpha$, where α is empirically chosen. A smaller value of α favors more weak associations and leads to longer extracted subsequences, however if the value of α is chosen too small the probability of extracting false positive loop closures is increased.

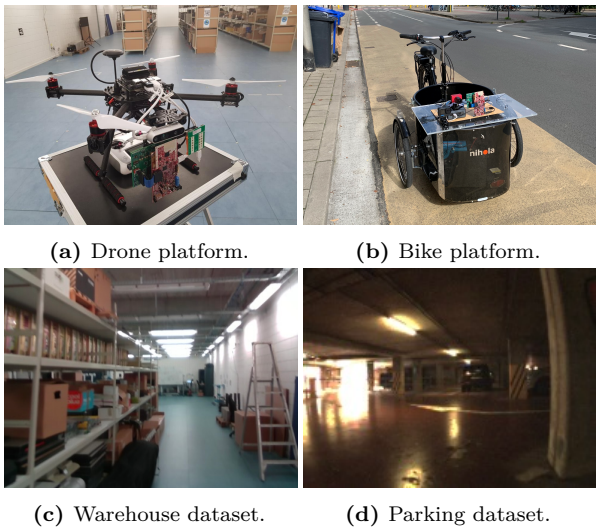
The maxima of H represent the endpoint of sequences with the best cumulative scores. To extract different maxima, a threshold value has to be determined. The criterion used for selection is $H_{ij,peak} > \beta * \max(H)$ where $\beta < 1$ and empirically chosen. For each maximum, we backtrack in matrix H according to the reverse of the steps in equation 3. The row indices of the obtained sequence in H form A and the column indices form B.

2.4 Loop closure extraction

Once a sequence has been obtained, the point of loop closure has to be selected. Furthermore, the pose transformation in $SE(2)$ between the two radar scans has to be calculated. We use the method in [15] to obtain a proposed transformation and motion score.

The extension in this paper determines the exact frame pair and verifies that it is not ambiguous with other frame pairs. If the proposed pair cannot be distinguished from other pairs by their motion score, we cannot confidently conclude this is the correct loop closure. In consequence, it is discarded to reduce the amount of false positives. The procedure goes as follows:

- Given two subsequences $A = [a_1, a_2, ...]$ and $B = [b_1, b_2, ...]$ resulting from section 2.3. The center element a_l from A is taken fixed.
- For each b_j in $[b_{l-25}, ..., b_l, b_{l+25}]$, execute the procedure summarized in [15]. A transformation is obtained for each frame pair a_l and b_j. For each pair (a_l, b_r) the mean squared error (MSE) is calculated between the overlapping regions in the radar scans. The pair with the lowest score is taken as a potential candidate for loop closure.
- To avoid false positive loop closures, the MSE is calculated in a window left and right from the potential candidate (a_l, b_r). With $L = \frac{1}{m}\sum_{i=1}^{m} MSE(a_l, b_{r-i})$ and $R = \frac{1}{m}\sum_{i=1}^{m} MSE(a_l, b_{r+i})$, it is required that $\alpha MSE(a_l, b_r) < L$ and $\alpha MSE(a_l, b_r) < R$. α is chosen empirically and in our experiments set to 1.1.



(a) Drone platform. **(b)** Bike platform.

(c) Warehouse dataset. **(d)** Parking dataset.

Figure 3. Capturing platform and dataset overview for the two datasets.

3 Experiments and implementation details

We use a TI IWR1443 mmWave radar (detailed in [15]) on a drone platform and a cargo bike. We evaluated our method on two datasets (see Fig. 3). The first dataset is recorded in a warehouse and consists of loops through the different aisles in a random order as shown in Fig. 4. The aisles are difficult to disambiguate and make for a challenging dataset. The dataset consists of 5 drone flights totaling around

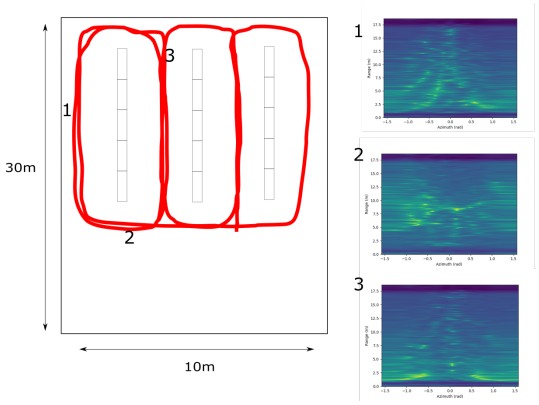


Figure 4. Layout of the warehouse with a potential trajectory included in red. On the right, the radar range-azimuth response is displayed in various positions in the warehouse.

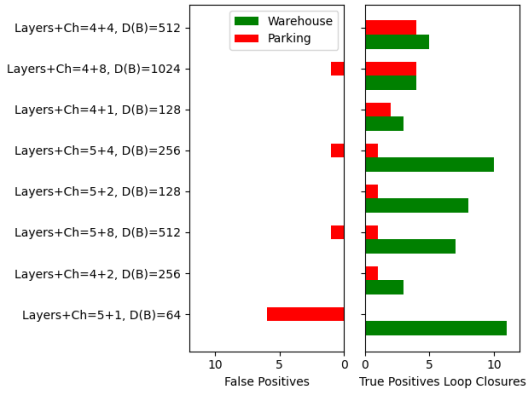


Figure 5. Results on different tested configurations. Higher is better. Both the amount of false positive and true positive loop closures as indicated as negative and positive bars. The size of the bottleneck is indicated as $D(B)$.

35 minutes of flight time (9000 frames). One flight is used as test set. The second dataset is recorded in a parking building consisting of two floors. Training data is recorded on both floors clockwise, test data is recorded counterclockwise on the first floor. This dataset consists of around 13000 radar frames.

We vary the number of layers and channels in the autoencoder to find the optimal trade-off between size of the bottleneck and performance. The amount of channels is doubled with every layer. The input size of the image is 64×64. Experiments with smaller input sizes, failed to capture meaningful features.

Results are shown in Fig 5, our method results in a high quality number of loop closures. Each loop closure is manually verified since we have no ground-truth information in our warehouse and classified as a true or false positive. It shows that architectures where the dimension of the bottleneck is smaller than 512 fail to extract a good descriptor for loop closure. We observe that if a neural network is extremely good in one dataset, it performs poorly in one dataset. Likely this is a result of the bottleneck being too small and forces the network to focus on one dataset.

To adapt to new environments retraining will be likely necessary. However, since usually the operation environment is known beforehand this is not necessarily a problem.

4 Conclusion and future work

In this paper, we propose to use an autoencoder to obtain qualitative latent codes from a low power radar sensor that can be used for loop closure in indoor environments without requiring ground truth for training. While it is hard to obtain hand-crafted features on low power radar sensors, the autoencoder provide us such a descriptor automatically.

After describing how we use the temporal similarity between latent codes, we also propose a method to determine the pose transformation in a loop closure. Our proposed algorithm is most suited for indoor environments and mobile platforms that have a restriction on sensor capability. We show that it is possible to extract loop closures, but the radar's lack of detail and susceptibility to clutter make it not ideal for this purpose. Our future research directions include a more in-depth study of the latent space and additional radar pre-processing steps to further improve the quality of the descriptor. Furthermore, our tested environments have been static. To consider dynamic environments, some form of doppler-filtering will be necessary to focus on the static scene. We will also investigate fusing other sensors to improve the robustness of our algorithm.

Acknowledgments

This research received funding from the Flemish Government under the "Onderzoeksprogramma Artificiele Intelligentie (AI) Vlaanderen" program.

References

[1] A. Angeli, D. Filliat, S. Doncieux, and J.-A. Meyer. "Fast and Incremental Method for Loop-Closure Detection Using Bags of Visual Words". In: *IEEE Transactions on Robotics* 24.5 (Oct. 2008). Conference Name: IEEE Transactions on Robotics, pp. 1027–1037. ISSN: 1941-0468. DOI: 10.1109/TRO.2008.2004514.

[2] R. Arroyo, P. F. Alcantarilla, L. M. Bergasa, and E. 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)*. 2015 IEEE International Conference on Robotics and Automation (ICRA). ISSN: 1050-4729. May 2015, pp. 6328–6335. DOI: 10.1109/ICRA.2015.7140088.

[3] M. J. Milford and G. F. Wyeth. "SeqSLAM: Visual route-based navigation for sunny summer days and stormy winter nights". In: *2012 IEEE International Conference on Robotics and Automation*. 2012 IEEE International Conference on Robotics and Automation. ISSN: 1050-4729. May 2012, pp. 1643–1649. DOI: 10.1109/ICRA.2012.6224623.

[4] E. Johns and G.-Z. Yang. "Feature Co-occurrence Maps: Appearance-based localisation throughout the day". In: *2013 IEEE International Conference on Robotics and Automation*. 2013 IEEE International Conference on Robotics and Automation. ISSN: 1050-4729. May 2013, pp. 3212–3218. DOI: 10.1109/ICRA.2013.6631024.

[5] C. Masone and B. Caputo. "A Survey on Deep Visual Place Recognition". In: *IEEE Access* 9 (2021). Conference Name: IEEE Access, pp. 19516–19547. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2021.3054937.

[6] F. Radenović, G. Tolias, and O. Chum. "Fine-Tuning CNN Image Retrieval with No Human Annotation". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41.7 (July 2019). Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 1655–1668. ISSN: 1939-3539. DOI: 10.1109/TPAMI.2018.2846566.

[7] K. L. Ho and P. Newman. "Detecting Loop Closure with Scene Sequences". In: *International Journal of Computer Vision* 74.3 (Sept. 1, 2007), pp. 261–286. ISSN: 1573-1405. DOI: 10.1007/s11263-006-0020-1. URL: https://doi.org/10.1007/s11263-006-0020-1 (visited on 03/21/2023).

[8] M. Klopschitz, C. Zach, A. Irschara, and D. Schmalstieg. "Generalized detection and merging of loop closures for video sequences". In: (Jan. 1, 2008).

[9] T. F. Smith, M. S. Waterman, et al. "Identification of common molecular subsequences". In: *Journal of molecular biology* 147.1 (1981). Publisher: Elsevier Science, pp. 195–197.

[10] D. Barnes, M. Gadd, P. Murcutt, P. Newman, and I. Posner. "The Oxford Radar RobotCar Dataset: A Radar Extension to the Oxford RobotCar Dataset". In: *2020 IEEE International Conference on Robotics and Automation (ICRA)*. 2020 IEEE International Conference on Robotics and Automation (ICRA). ISSN: 2577-087X. May 2020, pp. 6433–6438. DOI: 10.1109/ICRA40945.2020.9196884.

[11] D. Barnes and I. Posner. "Under the Radar: Learning to Predict Robust Keypoints for Odometry Estimation and Metric Localisation in Radar". In: *2020 IEEE International Conference on Robotics and Automation (ICRA)*. 2020 IEEE International Conference on Robotics and Automation (ICRA). ISSN: 2577-087X. May 2020, pp. 9484–9490. DOI: 10.1109/ICRA40945.2020.9196835.

[12] S. Saftescu, M. Gadd, D. De Martini, D. Barnes, and P. Newman. "Kidnapped Radar: Topological Radar Localisation using Rotationally-Invariant Metric Learning". In: *2020 IEEE International Conference on Robotics and Automation (ICRA)*. 2020 IEEE International Conference on Robotics and Automation (ICRA). ISSN: 2577-087X. May 2020, pp. 4358–4364. DOI: 10.1109/ICRA40945.2020.9196682.

[13] S. Arshad and G.-W. Kim. "Role of Deep Learning in Loop Closure Detection for Visual and Lidar SLAM: A Survey". In: *Sensors* 21.4 (Jan. 2021). Number: 4 Publisher: Multidisciplinary Digital Publishing Institute, p. 1243. ISSN: 1424-8220. DOI: 10.3390/s21041243. URL: https://www.mdpi.com/1424-8220/21/4/1243 (visited on 05/16/2023).

[14] A. Safa, T. Verbelen, I. Ocket, A. Bourdoux, H. Sahli, F. Catthoor, and G. Gielen. "Fusing Event-based Camera and Radar for SLAM Using Spiking Neural Networks with Continual STDP Learning". In: *2023 IEEE International Conference on Robotics and Automation (ICRA)*. 2023 IEEE International Conference on Robotics and Automation (ICRA). May 2023, pp. 2782–2788. DOI: 10.1109/ICRA48891.2023.10160681.

[15] P. Meiresone, D. Van Hamme, W. Philips, and T. Verbelen. "Ego-motion estimation with a lowpower millimeterwave radar on a UAV". In: *International Conference on Radar Systems (RADAR 2022)*. International Conference on Radar Systems (RADAR 2022). Vol. 2022. Oct. 2022, pp. 371–376. DOI: 10.1049/icp.2022.2346.