2019 International Conference on Robotics and Automation (ICRA)
Palais des congres de Montreal, Montreal, Canada, May 20-24, 2019

Fast Radar Motion Estimation with a Learnt Focus of Attention using Weak Supervision

Roberto Aldera, Daniele De Martini*, Matthew Gadd*, Paul Newman
Oxford Robotics Institute, Dept. Engineering Science, University of Oxford, UK.
{roberto,daniele,mattgadd,pnewman}@robots.ox.ac.uk

***Abstract*— This paper is about fast motion estimation with scanning radar. We use weak supervision to train a focus of attention policy which actively down-samples the measurement stream before data association steps are undertaken. At training, we avoid laborious manual labelling by exploiting short-term sensor coherence from multiple poses in the presence of an external ego-motion estimator (for example, wheel odometry). In this way, we generate copious annotated measurements which can be used for training a learning algorithm in a weakly-supervised fashion. We demonstrate the validity of the approach in the context of a Radar Odometry (RO) task, pre-filtering raw data with a popular image segmentation network trained as presented. We evaluate our system against 26 km of data collected in Central Oxford and show consistent motion estimation with greatly reduced radar processing times (by a factor of 2.36).**

***Index Terms*— radar, sensing, ego-motion estimation, field robotics**

I. Introduction

Radar is a sensor that continues to receive relatively little attention as a source of pose information for field robots in unstructured environments – typically being relegated to collision avoidance systems in domains such as driving assist systems. Frequency-Modulated Continuous-Wave (FMCW) scanning radars, in contrast to more widely used cameras and lidar sensors, operate well under variable weather and lighting conditions. Due to the long wavelength of radio waves and processing techniques, radar sensors receive multiple returns from a single azimuthal transmission and can operate out to many hundreds of metres. In many ways this sounds like the ideal sensor. However, radar measurements are complex. The beam is not narrow and tightly focused, returns are affected by various noise sources, and the interaction of the electromagnetic wave in the environment is far more complex than that of time-of-flight (TOF) lasers [1]. Consequently, comprehension of radar scans for precise ego-motion estimation requires dealing with complex measurement patterns which are not intuitive. In the end, as it often does, this boils down to a data association problem: "what detail in the last frame is relevant to the detail seen in this frame?" This paper is about making that task simpler.

Figure 1 shows an example of the Cartesian projection of a complete raw scan as collected by a FMCW scanning radar, where the peaks of the indicated reflections cannot be completely resolved into objects at those locations – even by a human expert. Indeed, the energy reflected back to the source sensor is a complex function of both geometry and material properties. A small nail in a wall at 400 m range is an ideal natural feature, but may appear as "bright" and as large as a car or be considered a nuisance glint – only briefly visible.

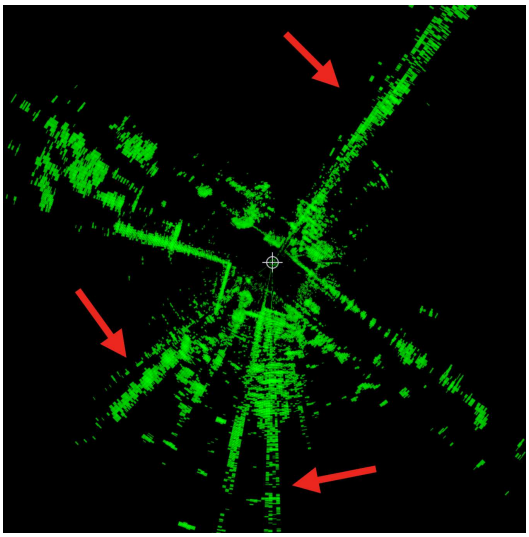


Fig. 1: An unfiltered radar scan: returns which are not spatially coherent from multiple viewpoints have been highlighted by red arrows.

This paper presents in Section IV a simple but effective method for down-sampling the measurements in radar scans to ease the computational complexity of data association between two scans. As our goal is ego-motion estimation or localisation, we leverage the straightforward point that ideally we would only operate with returns from artefacts in the scene that are visible from multiple views. Accordingly, we set ourselves the task of building a filter that only passes through measurements which, in the context of the entire scene, appear to be visible across wide baselines. We approach this task by building a classifier which marks individual reflections as either suitable for the data association stage or not.

The training of such a classifier may be challenging. Certainly manual labelling is arduous, but even if that were ignored, it is not a trivial task to decide what detail in one scan is visible in another given the remarkable way radar data is liable to change under small pose increments (totally unlike first return lidars or cameras). We create annotations

* The authors contributed equally to this work.

in a weakly-supervised fashion – simply accumulating radar returns in consecutive scans and looking for spatial coincidence. To do this, of course, one needs an external ego-motion source at training time, for which we use Visual Odometry (VO) [2].

Section V demonstrates the feasibility of the procedure in training a segmentation network based on U-Net [3]. We present and analyse in Section VI the impact of this approach upon the performance of a state-of-the-art Radar Odometry (RO) algorithm [4], where we achieve consistent motion estimation with radar processing times reduced from $168.12\,\mathrm{ms}$ to $71.10\,\mathrm{ms}$ as averaged over three representative trials in an urban environment.

II. Related works

Many techniques have been developed to extract landmarks from radar data, mostly based on probabilistic distributions and integrated directly into landmark extraction algorithms, e.g. in Constant False Alarm Rate (CFAR) processor and derivates [5, 6, 7] or in [4]. Although Deep Learning (DL) techniques have been proven to be effective in many applications for noise filtering [8, 9, 10], there is little or no work that exploits them in directly filtering range measurements [11], although many can be found on end-to-end classification, mostly concerning Automatic Target Recognition (ATR) problems [12, 13, 14].

Indeed, the drawback of such techniques is the need for an exhaustive set of annotated training examples in learning how to perform a specific task. The data annotation procedure can be a costly process in terms of both time and expertise since, in many situations, ground truth annotation can not only take much time, but it may require a specific knowledge of the system that only an expert can provide. Thus, it is desirable to build Machine Learning (ML) frameworks that can work with weak supervision. In [15] three categories of Weakly-Supervised Learning (WSL) are described – the proposed method falling into the *inaccurate supervision*, i.e. where information in the ground truth data presented to the model may suffer from errors.

Similar approaches can be found in other domains, where accurate human annotation has been replaced by less accurate, but automatic, labelling procedures. Most closely related to our method is the work of Barnes et al. [16] which uses a learned model to mask sensor observations for the purposes of improving a vision-based odometry system. As in this work, the authors use automatic annotation from a localisation system to generate ground truth pixel-wise mask labels for training a CNN by analysing geometric consistency. Additionally, the authors look at geometric consistency over multiple traversals of the same route, whereas we look only at consistency over a single traversal. However, we are the first to apply this method to radar domain data.

In a similar fashion [17] use lidar records for training a DL model to perform a road marking segmentation on images. They exploit material reflectance to detect road markings with a push-broom laser scanner and build the ground truth images by projecting the results onto the video stream from the camera as the vehicle moves using odometry information.

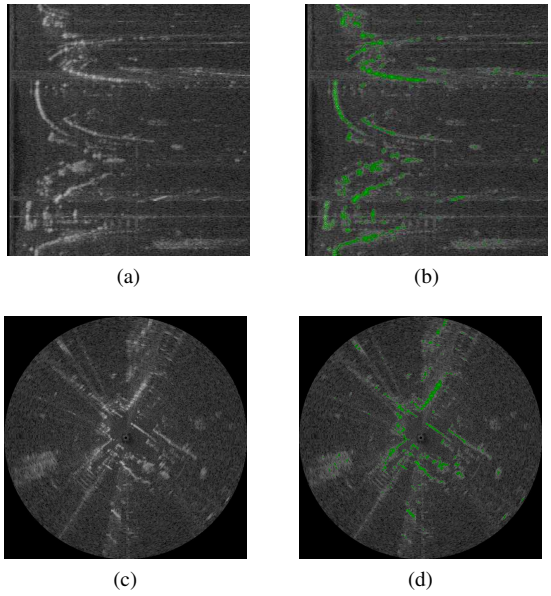


Fig. 2: Examples of scans assembled from successive azimuth readings from a FMCW scanning radar before and after the proposed annotation procedure. (a) and (c) show the raw data in polar (range vs azimuth) and Cartesian space respectively, whereas the corresponding annotated scans (green labels indicative of class **KEEP**) are shown in (b) and (d) respectively. For clarity, these examples have been cropped to display 384 of the full 2000 bins available that are used to train the network discussed in Section V.

Finally, in [18], in the same spirit as us, the authors apply a WSL approach to train a network for object detection on dynamic grid maps. They exploit temporal and spatial relationships to extract moving objects and their shapes using a lidar sensor. Indeed, once an object is observed, it continues being observed until it exits the field of view, updating shape and trajectory information. Then, this information is propagated backwards in time to refine the annotation for more consistency in the labels.

III. Preliminaries

The sensor we employ is a FMCW scanning radar which rotates about its vertical axis while continuously sensing the environment through the transmission and reception of frequency-modulated radio waves. While rotating, the sensor inspects one angular portion (*azimuth*, α) 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, ρ. The radar takes measurements along an azimuth at one of M discrete intervals and returns N power readings which we refer to as *bins*. One full rotation across all M azimuths is called a *scan*, $\mathcal{S}$. Furthermore, let $\mathbf{s}(k) \in \mathbb{R}^{N\times 1}$ be the power-range readings at time step k, where $t(k) = t_k$ is the time value at k and $\alpha(k) \in A$ is

the azimuth associated with the measurement. The element $s_i(k) \in \mathbf{s}(k)$ is the power return at the i-th range bin, with $i \in \{1, \ldots, N\}$. The measurement range is given by $\rho_i(k) = \beta(i - 0.5)$, where β is the range resolution of the radar.

IV. Methodology

In RO tasks such as [4], data association involves correlating the distance between pairs of landmarks identified in subsequent frames. The aim of this work is to develop a front end that simplifies this data association burden by focusing its attention on a stream of raw azimuth messages that are a filtered subset of reflections likely to be visible across wide baselines.

To this end, we automatically annotate each portion of the sensor's polar space representation as belonging to the class **KEEP** or not. Here **KEEP** is defined as a power bin that contains an object which has been observed to be wide baseline visible. This means that noise effects, dynamic objects, and reflections from objects that are highly specular or not visible from multiple views constitute the **REJECT** class.

The details of the proposed pipeline will be described in more detail from Section IV-A to Section IV-C, and are summarised here as:

(A) collection of a whole-scan measurement and conversion into Cartesian space;
(B) comparison of the spatially accumulated measurements and annotation of the power bins by thresholding;
(C) updating the internal representation of the environment and conversion of the current scan back to polar space.

A. Data Accumulation in Cartesian Space

The labelling process is carried out upon scans $\mathcal{S}(\bar{k}, \bar{k} + M)$. To this end, M azimuth measurements $\mathbf{s}(k), k \in (\bar{k}, \bar{k} + M]$ are collected and processed to represent the Cartesian environment prior to annotation. From now on, we will abbreviate $\mathcal{S}(\bar{k}, \bar{k} + M)$ as $\mathcal{S}_{\bar{k}}$, where $\bar{k}$ represents the time step at the last azimuth reading of the previously assembled scan, $\mathcal{S}_{\bar{k}-M}$.

Each measurement bin is processed separately to project it into Cartesian space. Assuming that the vehicle is moving on a flat ground plane, the Cartesian representation of the i-th bin within an azimuth reading is given by

$$\mathbf{x}_i(k) = \left[\rho_i(k)\cos\alpha(k), \rho_i(k)\sin\alpha(k), 0, 1\right]^T \quad (1)$$

specified using a projective space, $\mathbb{P}^3$, and the associated homogeneous co-ordinate system.

The rigid-body transformation $T(\bar{k}, k) \in \mathbb{SE}(3)$ between the pose of the robot at time steps coinciding with the start of the window and the current azimuth reading[1] ($\bar{k}$ and k respectively) can be used to project this sensor-centric Cartesian point $\mathbf{x}_i(k)$ into a common frame of reference as $\bar{\mathbf{x}}_i(k) = T(\bar{k}, k) \cdot \mathbf{x}_i(k)$ where the transformation

$$T(\bar{k}, k) = \begin{bmatrix} \mathbf{R}(\bar{k}, k) & \mathbf{t}(\bar{k}, k) \\ \mathbf{0} & 1 \end{bmatrix}$$

is obtained by exploiting any reliable external odometry measurement, in our case Visual Odometry (VO) [2]. In general, however, let $\bar{h}$ and h be the closest time steps of the external odometry source with respect to the radar measurement samples such that $t_{\bar{h}} \leq t_{\bar{k}}$ and $t_h \geq t_k$. Then, the transformation $T(\bar{k}, k)$ can be calculated as

$$T(\bar{k}, k) = T(\bar{k}, \bar{h}) \cdot T(\bar{h}, h) \cdot T(h, k) \quad (2)$$

and

$$T(\bar{h}, h) = \prod_{l=\bar{h}}^{h-1} T(l, l+1) \quad (3)$$

The terms $T(\bar{k}, \bar{h})$ and $T(h, k)$ are obtained by interpolation in $[\bar{h}, \bar{h} + 1]$ and $[h - 1, h]$, respectively: the rotational component is obtained by Spherical Linear Interpolation (SLERP) on a spherical surface traced by a unit quaternion, as described in more detail in [19], and the translational component is obtained by a constant velocity interpolation. Since the time step $\bar{k}$ defines the last time step of the previous scan, the term $T(\bar{h}, \bar{k})$ is available from previous computations.

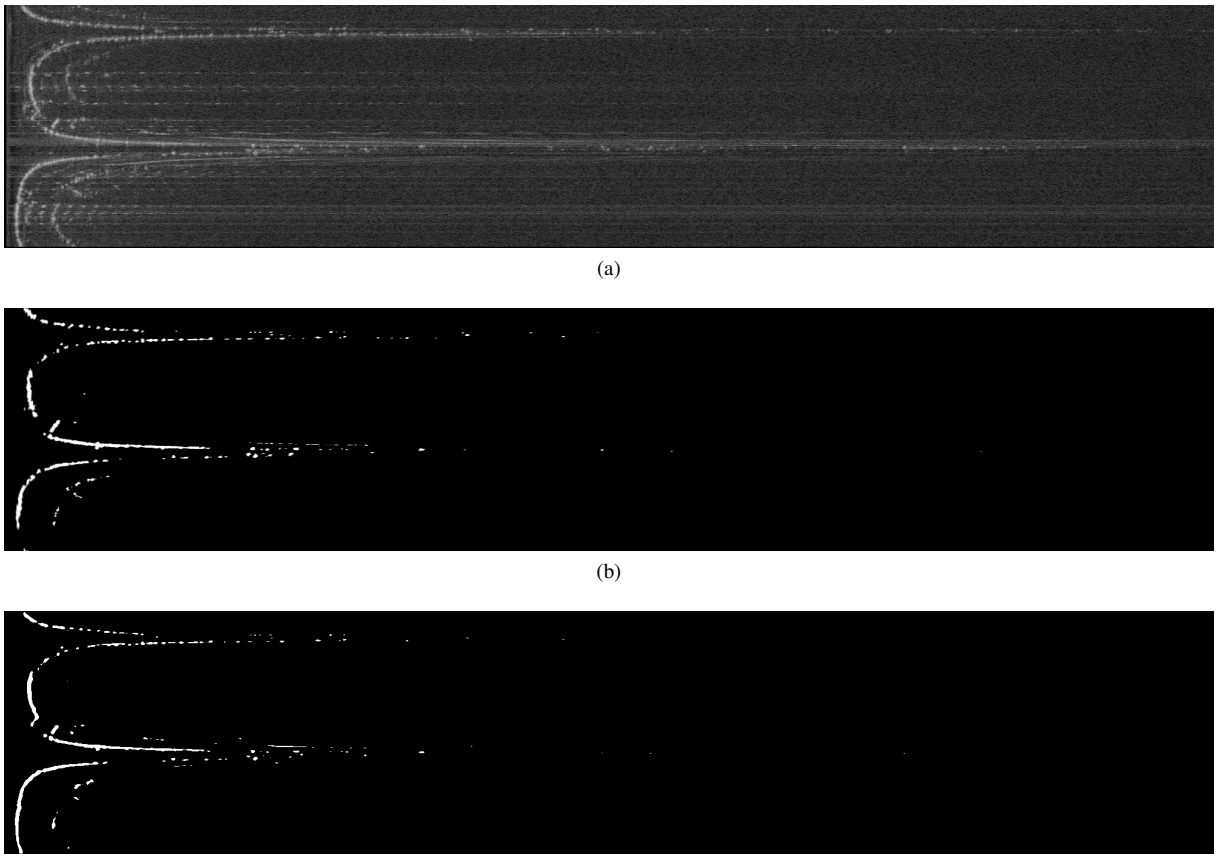
Once the position of each bin is projected into Cartesian space with a common reference frame, it is possible to build the Cartesian representation of the current scan, $\mathcal{S}_{\bar{k}}$. The chosen representation for the Cartesian space is a fixed size 2-dimensional grid, $\mathcal{G}_{\bar{k}}$, of size $2N \times 2N$, the values of which are derived from the power readings from incoming and projected bins. The two representations, Cartesian and polar, do not match perfectly, since the sparsity of the information in polar space depends on the distance from the radar, requiring the interpolation of each cell power value $\mathcal{G}_{\bar{k}}(u, v)$, with $u, v \in [-N, N]$, among the four closest bin measurements. For this task we employed a weighted polar to Cartesian data conversion as described in [1]. Figures 2a and 2c show the polar and Cartesian representations of a whole scan respectively.

B. Labelling Bins in Cartesian Space

Let $\mathcal{W}(k) = \{\mathcal{G}_{k-l*M} : l \in \{1, \ldots, w\}\}$ be a vector of Cartesian grids each of size $2N \times 2N$. $\mathcal{W}$ contains, at the time step k, the previous w scans represented as described above. Moreover, the relative transformations between them are recorded in $\mathcal{T}(k)$ by chaining transformations computed by eq. (2).

At this point, we can construct a histogram $\mathcal{H}_k$ of size $2N \times 2N$ by projecting each grid in $\mathcal{W}(k)$ into a common reference attached to the current grid, $\mathcal{G}_k$, and by counting the number of overlapping cells which have a power value greater than some label threshold, τ. This count is assigned to $\mathcal{H}_k(u, v)$, with $u, v \in [-N, N]$.

[1] In practice, we find a single pose for an entire scan (and not for every azimuth) to yield comparable performance in our weakly-supervised approach at a much improved processing speed for label generation.



(a)

(b)

(c)

Fig. 3: Visual comparison of a complete radar scan in polar coordinates (range vs azimuth): (a) shows the original data, while (b) and (c) show classification masks from the labelling procedure and the U-Net network respectively.

If the number of occurrences in a single histogram cell exceeds a given threshold, σ, the corresponding cell in $\mathcal{G}_k$ is annotated as **KEEP**, otherwise it is assigned to the **REJECT** class. Figure 2d shows the results of the labelling task performed on the measurements. In addition to reflections and noise, dynamic objects have not been labelled as **KEEP**s.

C. Updating the Environment's Polar Space Representation

Once the histogram calculation is completed, it is possible to update the internal representation of the environment; discarding the oldest scan grid from $\mathcal{W}$ and transformation from $\mathcal{T}$ and replacing them as more azimuth readings become available. Since annotated bins in polar space are required, the last step is to transform the annotated grid, $\mathcal{G}_k$, back into polar space. To do so, we exploit the weights derived by the projection into Cartesian space and annotate as **KEEP** each bin that most contributed to the power value in each corresponding **KEEP** grid cell. Figure 2b shows the results of the classification task on the polar measurements.

V. Experimental Setup

We discuss in this section an integration of our pre-filtering technique within a RO system as described in [4], leveraging U-Net [3]. The raw scans (indicated in the results as **ro**), scans with generated masks applied directly (labelled as **gt**), scans filtered by the network output (**unet**), and outputs from VO (**vo**) [2] are all compared with each other.

A. Hardware

The tests are performed using data collected from the *Oxford RobotCar* platform [20]. We employ 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. It is characterised by an operating frequency of $76\,\mathrm{GHz}$ to $77\,\mathrm{GHz}$, yielding up to 2000 range readings with a resolution of $0.25\,\mathrm{m}$, each constituting one of the 400 azimuth readings with a resolution of 2° and a scan rotation rate of $4\,\mathrm{Hz}$.

Images (for VO) were gathered by a Point Grey Bumblebee XB3 camera, mounted on the front of the platform facing towards the direction of motion. The camera is characterised by $1280\times960\times3$ resolution, $16\,\mathrm{Hz}$ FPS, $1/3''$ Sony ICX445 CCD, global shutter, $3.8\,\mathrm{mm}$ lens, 66° HFoV, $12/24\,\mathrm{cm}$ baseline.

The training and the evaluation have been run on a Dell PowerEdge machine with 3.33 GHz Intel Xeon processors, 192 GB RAM, 2666 MT/s DDR4.

B. Odometry

Our implementation of VO uses FAST [21] corners combined with BRIEF [22] descriptors, RANSAC [23] for outlier

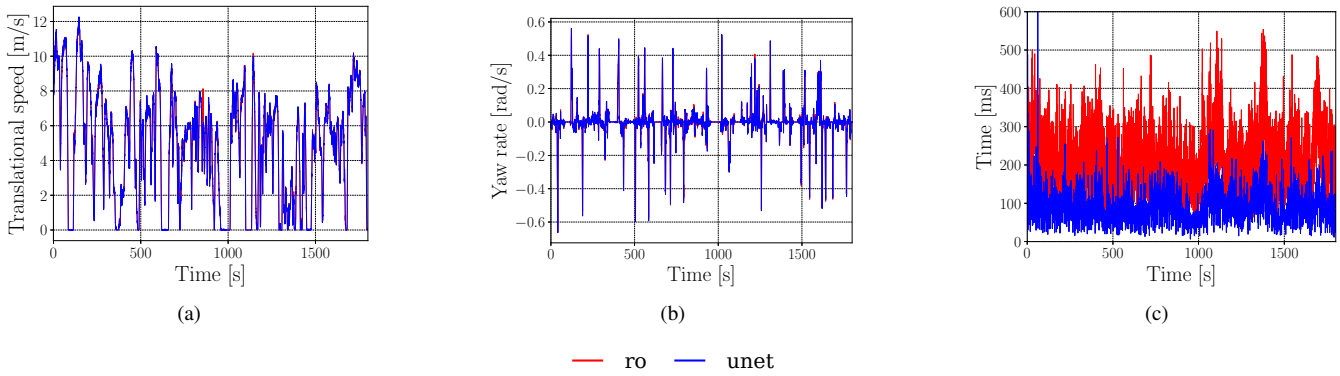


Fig. 4: Profiles of translational (a) and rotational (b) velocities with computational time for odometry estimation (c), for a $9\,\mathrm{km}$ portion of a $10\,\mathrm{km}$ loop, corresponding to the Cloudy condition summarised in Table I.

rejection, and nonlinear least-squares refinement. In order to verify the compatibility of the labelling method with existing landmark extraction algorithms, we exploit the RO method as described in [4], without any changes to the algorithm.

C. U-Net Training

The network is trained by providing it with whole-scan images of raw data in polar space and the corresponding annotated image obtained from the proposed method, where each pixel row corresponds to a single azimuth measurement and each pixel column corresponds to a bin, the value of which is set to `255` if the bin belongs to the **KEEP** class or `0` otherwise. As shown in Figure 3, each single-channel grayscale image is of resolution $M \times N$, where $M = 400$ and $N = 2000$ are the cardinalities of the azimuth set in a single scan and of the bin set in a single azimuth reading respectively, as described in Section IV.

The loss function we exploited is a sum of Dice coefficient and cross-entropy loss and we apply early stopping at 5 epochs, obtaining a validation Dice coefficient of $64.32\,\%$ and a validation loss of 0.37.

The dataset employed consists of three trials, collected in different weather conditions, one of which is used for training the network. This trial consists of a single loop under sunny weather conditions (favourable to VO). It has been divided by a proportion of $90\,\%$-$10\,\%$ for training and validation sets respectively, shuffled in order to avoid learning environment-specific geometric features (such as street or building layout) present within training data and absent from validation, or vice versa. The remaining trials (Cloudy, Rainy, and Night) are used for the evaluation of the proposed method, as described below.

Augmentation on the dataset has been performed as random noise on the input image to mimic speckle noise in radar data and random flips on the horizontal axes. While horizontal flips can be interpreted as a reverse in perspective, vertical flips are not physically meaningful due to the polar nature of the measurements.

	Translational speed $[\mathrm{mm\,s^{-1}}]$		
	ro	gt	unet
Cloudy	103.86	110.75	110.87
Rainy	97.06	100.67	99.89
Night	198.23	208.51	206.79
	Rotational speed $[\mathrm{mrad\,s^{-1}}]$		
	ro	gt	unet
Cloudy	5.54	5.64	5.72
Rainy	4.94	5.25	5.36
Night	6.88	7.26	7.14
	Estimation time $[\mathrm{ms}]$		
	ro	gt	unet
Cloudy	232.17	107.92	99.05
Rainy	174.11	84.17	74.97
Night	98.08	36.02	39.27

TABLE I: Summary statistics for several sorties exhibiting distinct environmental conditions. The vehicle drove $8834.44\,\mathrm{m}$, $8379.14\,\mathrm{m}$, and $8846.90\,\mathrm{m}$ during the Cloudy, Rainy, and Night outings, respectively. The median error quantities are presented in millimetres and milliradians per second to expose the significant digits. Estimation time is summarised by a RMS value.

D. U-Net Inference

At run-time we produce filtering masks at an average of $13.52\,\mathrm{Hz}$ using a single GTX 1080 Ti GPU, which is more rapid than the baseline RO performance ($5.95\,\mathrm{Hz}$) and our implementation which produces ground truth labels ($3.78\,\mathrm{Hz}$) and as such it does not need to be considered when discussing timing results. The masks generated by **unet** and **gt** are applied directly onto the landmark sets to **KEEP** bins that are relevant between frames and **REJECT** those that are not.

VI. Results

This section presents metrics for the performance of our approach when deployed as described above, summarised in Table I.

Figure 4 shows a comparison of odometry estimation as compared to the baseline **ro** discussed above. Here, the vehicle has traversed approximately $9\,\mathrm{km}$ of a $10\,\mathrm{km}$ urban

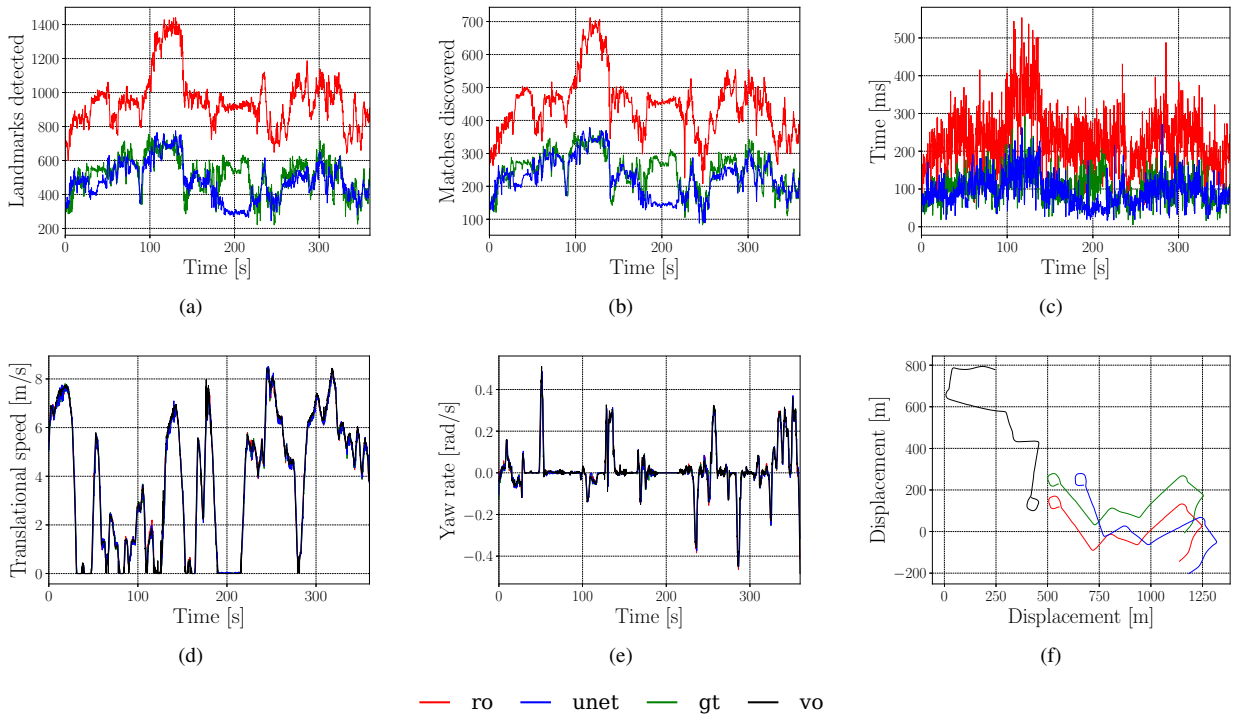


Fig. 5: Number of landmarks detected (a), number of matches between landmarks discovered (b), computational time for odometry estimation (c), translational velocity (d), rotational velocity (e), and integrated pose (f) on a short portion of a 10 km loop, corresponding to the Cloudy condition summarised in Table I. In (f) the absolute pose with respect to the start of the foray is shown, where it is clear that the drift experienced by the camera system is distinct from any of the radar systems. Nevertheless, the shape of the recovered poses, and thus the inherent egomotion, is consistent. In (a) and (b) it should be noted that by default the motion estimation algorithm attempts to match only half of the detected landmarks between frames.

route in mixed traffic and consequently exhibits a range of movement types (including regular stopping and starting at traffic control points as well as accelerating and decelerating along straight sections of road and around corners). As can be seen in Figure 4, the translational and rotational velocities recovered by using scans pre-filtered by our approach, **unet**, is consistent with the baseline **ro**. The median estimation time of 99.05 ms for this trial, as read from Table I and shown for the full trial in Figure 4, is a marked improvement over the baseline of 232.17 ms and approximates that of the ground truth annotation (107.92 ms).

In Figure 5 we consider a briefer portion of this same trial in more detail. Figure 5a includes for comparison the landmarks detected by applying the ground truth mask as discussed in Section V above (**gt**). Additionally, Figure 5d and Figure 5e include for comparison the translational and angular velocities recovered by VO.

Table I presents some summary statistics over three forays with distinct atmospheric conditions that do not correspond to the (Sunny) condition of the training data. The performance achieved by the baseline **ro**, **unet**, and **gt** is presented in terms of the median translational and angular velocity errors using **vo** as the benchmark signal. We observe in Figure 5f the integrated pose of **unet** – our fully integrated system – tracks visually that of the baseline **ro** faithfully.

VII. Conclusions

This paper demonstrates the feasibility of using a weakly-supervised learning framework to pre-filter radar measurements for use in radar-only navigation. By retaining only the information that is wide-baseline visible, our approach is able to reduce processing times (by a factor of 2.36) while maintaining the accuracy of the recovered motion. We evaluate our system over 26 km of urban driving and show that it rivals baseline RO despite having access to significantly fewer features. This contribution is more than a neat efficiency improvement – by focusing attention on relevant landmarks, the task of motion estimation can now be run in real-time on a robot using radar alone.

Acknowledgements

Roberto Aldera and Daniele De Martini are supported by the UK Engineering and Physical Sciences Research Council (EPSRC) programme grant EP/M019918/1. Matthew Gadd is supported by Innovate UK under CAV2 – Stream 1 CRD (DRIVEN). Paul Newman is supported by EPSRC Leadership Fellowship Grant EP/J012017/1. The authors would also like to thank Paul Murcutt for his software engineering support.

REFERENCES

[1] E. J. M. Adams, J. Mullane and B. Vo, *Robot Navigation and Mapping with Radar*. Artech House, 2012.

[2] D. Nistér, O. Naroditsky, and J. Bergen, "Visual odometry," in *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, vol. 1. Ieee, 2004, pp. I–I.

[3] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds. Cham: Springer International Publishing, 2015, pp. 234–241.

[4] S. Cen and P. Newman, "Precise ego-motion estimation with millimeter-wave radar under diverse and challenging conditions," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, May 2018.

[5] M. A. Khalighi and M. H. Bastani, "Adaptive CFAR processor for nonhomogeneous environments," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 36, no. 3, pp. 889–897, July 2000.

[6] P. P. Gandhi and S. A. Kassam, "Optimality of the cell averaging CFAR detector," *IEEE Transactions on Information Theory*, vol. 40, no. 4, pp. 1226–1228, July 1994.

[7] B. Magaz, A. Belouchrani *et al.*, "Automatic threshold selection in OS-CFAR radar detection using information theoretic criteria," *Progress In Electromagnetics Research*, vol. 30, pp. 157–175, 2011.

[8] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Beyond a gaussian denoiser: Residual learning of deep CNN for image denoising," *IEEE Transactions on Image Processing*, vol. 26, no. 7, pp. 3142–3155, 2017.

[9] J.-M. Valin, "A hybrid dsp/deep learning approach to real-time full-band speech enhancement," *arXiv preprint arXiv:1709.08243*, 2017.

[10] Y. Wang, L. Tu, J. Guo, and Z. Wang, "Residual learning based RF signal denoising," in *2018 IEEE International Conference on Applied System Invention (ICASI)*. IEEE, 2018, pp. 15–18.

[11] E. Mason, B. Yonel, and B. Yazici, "Deep learning for radar," in *2017 IEEE Radar Conference (RadarConf)*, May 2017, pp. 1703–1708.

[12] S. Chen and H. Wang, "SAR target recognition based on deep learning," in *Data Science and Advanced Analytics (DSAA), 2014 International Conference on*. IEEE, 2014, pp. 541–547.

[13] B. Xue and N. Tong, "DIOD: Fast and Efficient Weakly Semi-Supervised Deep Complex ISAR Object Detection," *IEEE Transactions on Cybernetics*, no. 99, pp. 1–13, 2018.

[14] H. Wang, S. Chen, F. Xu, and Y.-Q. Jin, "Application of deep-learning algorithms to MSTAR data," in *Geoscience and Remote Sensing Symposium (IGARSS), 2015 IEEE International*. IEEE, 2015, pp. 3743–3745.

[15] Z.-H. Zhou, "A brief introduction to weakly supervised learning," *National Science Review*, vol. 5, no. 1, pp. 44–53, 2017.

[16] D. Barnes, W. Maddern, G. Pascoe, and I. Posner, "Driven to distraction: Self-supervised distractor learning for robust monocular visual odometry in urban environments," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 1894–1900.

[17] T. Bruls, W. Maddern, A. Morye, and P. Newman, "Mark Yourself: Road Marking Segmentation via Weakly-Supervised Annotations from Multimodal Data," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, May 2018.

[18] S. Hoermann, P. Henzler, M. Bach, and K. Dietmayer, "Object Detection on Dynamic Occupancy Grid Maps Using Deep Learning and Automatic Label Generation," *arXiv preprint arXiv:1802.02202*, 2018.

[19] K. Shoemake, "Animating rotation with quaternion curves," in *ACM SIGGRAPH computer graphics*, vol. 19, no. 3. ACM, 1985, pp. 245–254.

[20] W. Maddern, G. Pascoe, C. Linegar, and P. Newman, "1 Year, 1000km: The Oxford RobotCar Dataset," *The International Journal of Robotics Research (IJRR)*, vol. 36, no. 1, pp. 3–15, 2017.

[21] E. Rosten, G. Reitmayr, and T. Drummond, "Real-time video annotations for augmented reality," in *International Symposium on Visual Computing*. Springer, 2005, pp. 294–302.

[22] M. Calonder, V. Lepetit, M. Ozuysal, T. Trzcinski, C. Strecha, and P. Fua, "BRIEF: Computing a local binary descriptor very fast," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 7, pp. 1281–1298, 2012.

[23] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.