Michela Raimondi¹, Gianluca Ciattaglia¹, Antonio Nocera¹, Linda Senigagliesi¹, Susanna Spinsante¹, and Ennio Gambi¹

¹Affiliation not available

February 14, 2024
mmDetect: YOLO-based Processing of mm-Wave Radar Data for Detecting Moving People

Michela Raimondi, Gianluca Ciattaglia, Member, IEEE, Antonio Nocera, Student Member, IEEE, Linda Senigagliesi, Member, IEEE, Susanna Spinsante, Senior Member, IEEE, and Ennio Gambi, Senior Member, IEEE

Abstract—The application of millimeter-Wave (mmWave) Radar sensors for people monitoring raised a lot of interest in the context of Active Assisted Living (AAL), especially since the processing of Radar signals can provide interesting information about the observed subjects. Correct recognition of the ongoing behavior, however, cannot disregard from detecting where the subject is positioned. Detection approaches, based on Constant False Alarm Rate (CFAR) algorithms, sometimes fail to correctly identify the presence of targets within the observed scenario, especially in complex environments such as indoors. This paper proposes the use of a mmWave Multiple Input Multiple Output (MIMO) Radar in combination with a You Only Look Once (YOLO) neural network-based algorithm for the detection of moving people in indoor environments by processing all the data cube information at the same time. Results are validated through experimental tests which involve subjects walking in linear or random mode, different Radar configurations, and different indoor environments. By exploiting at the same time information such as the angle, Doppler, and range distance of the target, the proposed approach proves to be very effective in the examined scenarios. Experimental results will be discussed in this work to demonstrate the effectiveness of the proposed method.

Index Terms—Classification, FMCW Radar, people detection, target recognition, YOLO

I. INTRODUCTION

The ability to identify and recognize moving people in a sequence of images (optical ones or derived from other sensors) is crucial to successfully tracking, and recognising them or their activities, and in general, understanding human behavior in indoor applications. Indoor surveillance is important not only for security reasons but also for health monitoring purposes. Collecting information about a moving target within a room, in fact, allows detecting possible falls [1]–[4], or monitoring activities of daily living with the purpose of assessing a person’s level of autonomy in performing these actions independently.

The problem of indoor surveillance can be approached in different ways and with different types of sensors. The most common solutions make use of RGB cameras. These methods, however, exhibit some problems and limitations; for example, they are affected by poor illumination of the detection area and may also raise some privacy concerns [5]–[7], while adverse indoor conditions like smoke presence can affect their detection capability [8]–[10].

The problem of privacy and poor illumination can be overcome by using RGB-D sensors, which provide a depth image less prone to release confidential information. The privacy issue is addressed but at the expense of a limitation on the maximum detectable distance. Sensors able to preserve privacy and reach longer detection distances are Lidars. Their usage is becoming common in vision applications of autonomous systems as they can provide depth images with very high resolution [11], [12], but they present some critical issues. In fact, Lidars rely on light emission, and in indoor environments, they cannot be used in the presence of smoke, such as in case of fire. They are also very expensive with respect to other types of sensors.

Radar systems are a powerful alternative to the previously mentioned sensors. They can perform detection at greater distances than other sensors, reaching, in the case of automotive Radar, distances up to 30 m in short-range configuration, or greater than 200 m for the long-range one [13]–[15]. Compared with Lidars and RGB cameras, Radar sensors have several advantages, including greater achievable range, and lower cost than Lidars or high-resolution RGB cameras. Radars can work in the presence of smoke or generally in difficult optical vision situations. Thanks to these characteristics,
Radar systems are excellent candidates for indoor target detection and applications such as activity recognition and Active Assisted Living (AAL) [16]–[19].

A. State-of-the-art

Modern commercial automotive Radar sensors widely exploit the frequency range from 76 GHz to 81 GHz, mainly because the automotive market uses the W-Band for Radar systems [20]–[22]. Thanks to advances in this field, there are many development boards in the market designed for automotive, which can be used also for indoor applications [20], [23]. The detection and identification of targets is a relevant research field in Radar applications. The classical approach to perform these operations, both indoor and outdoor, makes use of Constant False Alarm Rate (CFAR) thresholds on Radar processed signals [24]. However, their performance degrades rapidly in non-homogeneous environments, which represent the most common situation. In most cases, the CFAR algorithms are difficult to manage to achieve a correct solution to complex identification tasks [25]–[27]. The CFAR algorithms have been modified and improved over time proposing different solutions, such as Cell Averaging (CA) CFAR or Ordered Sort (OS) CFAR [28]. Other algorithms have also been developed recently, such as Comp-CFAR [29], or CFAR based on the so called “zlog” estimator [30]. All current CFAR algorithms perform target detection through a reference window applied on Radar maps and process the data contained therein. The reference window is usually adopted to estimate the average interference power representing the range-Doppler side. This is used to obtain the detection threshold, which should be set high enough to limit the false alarm rate to an acceptable small percentage, but using reference windows reduces the efficiency of target detection. The main purpose of the CFAR is to define a threshold and all values above will be considered as a target. Many CFAR-based techniques have been proposed with the purpose of not only revealing a target but also classifying it [31]. Unfortunately, these methods are very sensitive to the configuration parameters of the CFAR which must be calibrated and subjected to careful tuning. A more general approach should work without calibration of the algorithm for each specific environment, and this is one of the aims of the methodology proposed in this work.

With the progressive advancement in the performance of Machine Learning (ML) algorithms, data collected by the Radar can be processed as images and used to improve traditional Radar techniques. Experimental results demonstrate that the ML approach exhibits high robustness with respect to CFAR thresholds in noisy environments [32], [33]. Initially, the most common ML techniques used for image classification were based on Convolutional Neural Networks (CNNs) [32]. Today, real-time methods for target recognition such as You Only Look Once (YOLO), a network called the “single pass network”, are preferred in conditions where the latency of the algorithm is relevant. YOLO reduces processing time compared to other ML techniques: the network was created for fast object recognition in images or videos, like in [34], and was later applied to Radar signals. In [35], a Multiple Input Multiple Output (MIMO) Radar able to exploit a bi-dimensional array is used to obtain an image similar to one obtained from an RGB sensor. Detection and classification of the target, which can be a person, a fence, or a road sign, are performed by applying a YOLO neural network using the combination of the sensors. The results show that the use of Radar helps the YOLO network in dark conditions or when the lens of the camera is dirty, but the problem of privacy remains open. The obtained results also demonstrate that the Radar systems without the joint usage of the camera can reach 84% of accuracy in classification. Authors in [36] illustrate how YOLO achieves good results on range-Doppler maps using vehicles as targets. These maps are obtained from the processing of the Radar signals and make it possible to measure the target’s velocity and range distance from the sensor. Range-Doppler maps are also used in [37], where a YOLO neural network is applied for the classification of three different targets (pedestrians, vehicles, and bikes). In [38], [39], authors apply YOLO to range-angle Radar images and demonstrate the possibility of applying deep learning algorithms to high-resolution Radar sensor data, particularly in the range-angle domain. Furthermore, they show that the use of YOLO instead of CNN improves classification performance. The research presented in [40] exploits semantic segmentation methods applied to range-Doppler maps with good results. In this case, the U-net network, that was proposed in [41], is applied using an appropriate loss function in order to replace the CFAR algorithm and achieve accurate clustering, while in this paper clustering techniques are not intended to be used, but detection is performed only on the basis of the YOLO network output.

Range-azimuth maps are notoriously difficult to analyze because of clutter, especially in indoor conditions [42]. These maps are subject to reflection problems and multipath, especially if the target of interest has a low Radar Cross Section (RCS). A method named Deep Image Prior (DIP) is proposed for denoising the range-azimuth map in [42]. This method is based on ML, but it is used to help the application of CFAR thresholds. In [43], the authors propose a method that exploits not only the range-azimuth map but also the range-Doppler map to improve the performance. Range and Doppler information is also considered in [44] to simplify detection on range-azimuth maps. In this article, Jiang et al. test the functioning of a CNN, by relating it to conventional methods. However, the data in this experiment are simulated and not tested in real environments.

B. Main work contribution

In this work, we propose an approach that makes use of a mmWave W-Band MIMO Radar, together with a ML-based detection technique, for indoor target recognition and detection. By taking advantage of all the information from Radar signal processing (i.e., angle, Doppler, and range), it is possible to detect a moving person in an indoor scenario. The proposed method involves the use of two YOLO networks trained on different datasets: the former uses range-Doppler maps and the latter is applied to Doppler-azimuth maps. In
this work the used version of the YOLO is the “v3”, and in the following the indication of YOLO will refer to this version of the network.

The information obtained from the dual use of the networks is combined to detect a moving person on the range-angle map. The main novelty compared to the current state of the art is the use of the three axes of the Radar data cube. Existing works in the literature, in fact, are mainly based on the application of ML methods on range-Doppler maps or Range-azimuth maps. The idea behind the present paper is to use all the information contained in the Radar data cube, as well as the Doppler-angle map. This is possible only if the Radar used can reach a good angular resolution, given by its MIMO capabilities. In this case, the Doppler-angle map can be used to improve the performances of a ML technique applied to Radar data. The inclusion of the Doppler-azimuth map in the processing allows better detection of the target in the range-azimuth map. In addition, unlike CFAR, the proposed approach can be applied directly to Radar images avoiding the application of thresholds and clustering algorithms. The proposed method needs a training step but the network does not need to be re-trained if the Radar is moved in a different environment. This represents a great advantage over traditional methods.

The rest of the paper is organized as follows. Section II describes the Frequency Modulated Continuous Wave (FMCW) Radar used and its basic operating principles. Section III introduces the main concepts concerning YOLO and the processing of the proposed method is explained. Section IV describes the experimental tests performed and the results derived from the application of the proposed algorithm. Conclusions are drawn in Section V.

II. RADAR SYSTEM AND SIGNAL PRE-PROCESSING

A. FMCW Radar

The sensor used is a Texas Instruments mmWave FMCW Radar, model name TIDEP - 01012, equipped with twelve transmitters and sixteen receivers [45]. A basic block scheme of an FMCW quasi-monostatic Radar is depicted in Fig. 1.

Fig. 1. Radar block scheme: the transmitted signal is generated by the Chirp synthesizer; the reflected back signal is collected by the receiver antenna and the mixer performs the mixing between them to obtain the IF signal.

The chirp signal has a starting frequency $f_{\text{start}}$ of 77 GHz and a stop frequency $f_{\text{stop}}$, which depends on the Radar configuration. The maximum usable Bandwidth is 4 GHz and the value depends on the slope of the chirp and the time $t_{\text{chirp}}$. The time that the transmitted chirp takes to go from the initial frequency $f_{\text{start}}$ to the final frequency $f_{\text{stop}}$ is called the chirp time and is indicated with $t_{\text{chirp}}$. The difference between the transmitted chirp and the received chirp is indicated as $\Delta_t$. Only in the time window called $t_{\text{overlap}}$ the IF signal will be sampled by the Analog-to-Digital Converter (ADC). The velocity and distance of the target can be obtained by processing the samples of IF signals.

As mentioned above, the Radar used is equipped with MIMO technology, and to estimate the Angle of Arrival (AoA) indicated with $\theta$, it is necessary to use at least two receiver antennas. This limitation is met as the Radar used has a virtual array consisting of 86 elements. Each physical receiver antenna has a dedicated ADC that samples the related IF signal. This can be called “spatial sampling” and the angular information can be obtained from the processing of these samples. The signal transmitted by the nine transmitting antennas along the azimuth axis is reflected back by the target with an angle $\theta$, with respect to the Radar receiver virtual array boresight. For each virtual array element, there is a phase difference in the received signal, named $\Delta_\phi$. For example, for the $n$-th element of the array $R_{\theta_n}$ and the subsequent $R_{\theta_{n+1}}$, the phase difference between the received signals is $\Delta_\phi$. It is possible to calculate the AoA as

$$\theta = \arcsin \left( \frac{\Delta_\phi \lambda}{2\pi d} \right),$$  \hspace{1cm} (1)

where $\Delta_\phi$ is the aforementioned phase difference, $\lambda$ the
wavelength of the carrier frequency and \(d\) the distance between the virtual receiver antennas [46], [47].

With the considered MIMO FMCW Radar, it is possible to calculate the position of the target with respect to the Radar position and also their relative velocity. The limits of the obtainable measurements depend on the device configuration. The configuration must be customized for the testing area where the acquisitions are conducted.

**B. Radar signal pre-processing**

The Radar system used in this work can be configured with a specific software provided by Texas Instruments, namely mmWave studio used to set the configuration parameters [48]. The main parameters to be set for this work are:

- \(t_{idle}\): it is the time between chirp transmissions. Is used to restore the internal ramp generator from one transmission to another;
- \(t_{ramp}\): chirp time duration. This parameter affects the used Radar Bandwidth;
- \(f_{sampling}\): it is the sampling frequency of the beat signal;
- \(f_{start}\): it is the starting frequency of the chirp;
- \(n_{ADC}\): number of samples in each chirp;
- \(n_{chirp}\): number of chirp in each frame;
- \(t_{overlap}\): time over which the beat signal is sampled. Depends on the number of samples \(n_{ADC}\) and the sampling time.

The operating mode of the device groups the transmissions of the chirps by frame. Each frame is composed of a certain number of transmissions that can be set with the parameter \(n_{chirp}\). Each transmitted chirp will produce an IF signal sampled with a configurable number of samples indicated with \(n_{ADC}\). Considering only one couple transmitter-receiver and one frame, the so-called Fast-Time/Slow-Time matrix can be obtained by placing side-by-side the sample vectors of each chirp. This process is depicted in Fig. 3.

![Fig. 3. Radar IF signal samples organization: Fast-Time/Slow-Time map.](image)

The Fast-Time axis contains the samples of one chirp transmission and is composed of \(n_{ADC}\) samples, while the Slow-Time axis contains samples of different chirps and is composed of \(n_{chirp}\) elements. Extending the consideration to multiple couples transmitter-receiver, it is possible to obtain a cube that is called Radar data cube, whose representation is reported in the left part of Fig. 4. The “Spatial Sampling” is the sampling along different receivers. All processing to extract the information about the target is based on this data organization. A data cube is obtained from each transmitted frame, so organizing the data in this form is an easy way to represent and manage the Radar IF signals’ samples. To obtain the distance of the target from the Radar, its velocity and AoA, the most simple way is to compute a Fast Fourier Transform (FFT) along the different axes. For the purposes of this work, the computation is performed bi-dimensionally, resulting in the so-called detection maps. These maps are:

- range-Doppler map, for the computation along the Fast-Time and the Slow-Time;
- range-azimuth map, for the computation along the Fast-Time and the Spatial Sampling;
- Doppler-azimuth map, for the computation along the Slow-Time and the Spatial Sampling.

A schematic representation of how the maps are obtained from the Radar data cube is given in Fig. 4.

![Fig. 4. Extraction of maps from the Radar data cube: range-azimuth (green), range-Doppler (yellow), Doppler-azimuth (white).](image)

The conversion of the axis from FFT bins can be done with the classical Radar FMCW equations but with particular attention to the Doppler/Velocity axis. The transmission operating mode of the Radar is Time Division Multiplexing (TDM), so this means that all the transmitters must transmit their chirp signal before transmitting another one. For example, in case where the transmitters are indicated as Tx1 to Tx\(n_{T_x}\), with \(n_{T_x}\) the number of transmitters; Tx1 start the transmission, then the next transmitters until Tx\(n_{T_x}\), then Tx1 restart another transmission. By defining the Pulse Repetition Interval (PRI) \(T\) from one transmitter point of view as

\[
T = n_{T_x}(t_{idle} + t_{ramp}),
\]

where \(n_{T_x}\) is the number of used transmitters, the maximum detectable target velocity can be computed as

\[
v_{max} = \frac{\lambda}{4T} = \frac{\lambda}{4n_{T_x}(t_{idle} + t_{ramp})},
\]

where \(\lambda\) is the center wavelength of the transmitted signal and \(n_{T_x}\) the number of active transmitters in the Radar.
configuration [49]. Since the value \( n_F \) affects the maximum detectable target velocity, this must be taken into consideration in the choice of the configuration parameters.

III. OBJECT DETECTION USING YOLO

The proposed approach aims to identify a person’s position within a room and the YOLO network can be exploited to achieve this goal. YOLO is a detection algorithm that makes predictions of images in a single run. It employs CNNs to detect multiple objects within a single image. This unique approach not only predicts the object’s class but also identifies the position of the object in the image. The initial segment of the YOLO network, dedicated to feature extraction, utilizes a CNN, primarily designed for image classification. In the first stage, the image, regardless of the sensor by which it is obtained, is divided into several grids where each grid cell of side \( B \) will detect objects that appear within it, with a certain value of confidence. Fig. 5 shows how YOLO works. YOLO takes an input image and outputs a vector containing information about the position and class of the target to be identified. In particular, the first four elements of the vector determine the position given by the bounding box, and contain information about:

- vertex of the element at the top right of the box \((x, y)\);
- width of the element \((w)\);
- height of the element \((h)\).

The remaining part of the output vector contains the probabilities that the target belongs to a certain class \((p_0, \ldots, p_n)\), being \( n \) the total number of classes.

**Fig. 5.** Example application of the YOLO on the range-azimuth map.

Confidence is determined using the Intersection over Union (IoU) method, which is an evaluation metric used to measure the accuracy of an object detector on a particular dataset [50]. To apply IoU, it is necessary to know the ground-truth bounding boxes (i.e., the bounding boxes manually labelled in the test set which specify where the object is in the image) and the bounding boxes provided by the used model.

The IoU is given by the ratio between the overlap area and the union area; the overlap area is the area between the predicted bounding box and the ground-truth bounding box, while the union area corresponds to the area enclosed by both the predicted bounding box and the ground-truth bounding box. If the IoU is larger than 75\% the prediction is considered good. When a new input arrives, the YOLO network estimates the position and class of the object for that input. Several bounding boxes may be generated for a single ground truth, and in order to choose which of these is the most important one, non-maximum suppression (NMS) is used, i.e., only the one with the highest value is extracted.

The YOLO network is trained on the basis of a dataset whose labels we already know or, a dataset to which we apply labels. The labels contain information about the class and the bounding box. Typical parameters used in a YOLO training model include:

- Learning Rate: it adjusts how quickly the model updates its weights according to the gradient calculated during training;
- Mini Batch Size: fixed number of training examples that is less than the actual dataset;
- Penalty Threshold: detections that overlap by less than this value are penalized;
- Warm-up period: it represents the period during which the desired learning rate to be achieved;
- Augmentation: it allows the model to be trained on different versions of the available data to avoid overfitting.

A. Description of the proposed method

Starting from the Radar data cube, range-Doppler, azimuth-Doppler and range-azimuth maps are obtained. Information about range and angle is obtained from the first two maps, respectively. The first stage of the proposed method is based on the realisation of two datasets, one containing range-Doppler maps and the other containing azimuth-Doppler maps; two YOLO networks are trained separately on these two datasets. For the sake of brevity, we will henceforth refer to these networks as YOLO-Doppler and YOLO-azimuth. A third dataset containing unprocessed range-azimuth maps is also collected, on which an additional YOLO network is trained, which will be used only to evaluate the results. To distinguish it from the others, it is referred to as YOLO-range.

The application of YOLO-Doppler to the range-Doppler maps gives as output the bounding box and then the value of the top left vertex, width \((w)\) and height \((h)\). The \( x_1 \) value in Fig. 6(a) represents the y-coordinate of the vertex, while the \( x_2 \) value is given by \( x_1 + w \). The portion of the image between \( x_1 \) and \( x_2 \) contains the target and gives information about the distance at which the target is located. Anything outside the area bounded by \( x_1 \) and \( x_2 \) is not taken into account. The same process is then applied to azimuth-Doppler maps using the YOLO-azimuth network, as shown in Fig. 6(b). In this case, the x-coordinate of the vertex represents value \( \theta_1 \) and \( \theta_2 \) is given by \( \theta_1 + \omega \). Portion of the map contained between \( \theta_1 \) and \( \theta_2 \) gives us information concerning a portion of the angle covered by the target.

Finally, the presence of the target is determined using the third side of the Doppler Radar cube, namely the range-azimuth map. Having acquired distance and angle information through the aforementioned procedures, we can consider the indirect detection of the target, as shown in Fig. 6(c). In order to cut out the portion of no interest in the range-azimuth map, all values of \( x \) outside the range \([x_1, x_2]\) and all values of \( \theta \) outside the range \([\theta_1, \theta_2]\) are eliminated. This allows us to obtain a map that emphasizes the precise area where the target is situated, completely removing any portion that might have been erroneously identified as the target.
An example of what happens in practice is depicted in Fig. 7. The YOLO-Doppler and YOLO-azimuth networks are fed with the range-Doppler and the Doppler-azimuth maps, respectively, to obtain the bounding boxes. To emphasize the results, we report the range-azimuth map in Fig. 7(a), which has not been processed, while Fig. 7(d) shows the post-processed image obtained by applying the proposed algorithm to maps shown in Figs. 7(c) and 7(b).

To generalise the proposed approach, acquisitions are processed without removing the background. Also, a good visualisation can be obtained using a specific colour map made available by Matlab, called Colorcube. This contains many regularly spaced colours in RGB colourspace to provide more steps of grey, red, green, and blue. The fact that the chosen colour map is divided into slots each containing different colour intensities helps to visualise the targets in a more evident way. The steps of the scale and their limits are set automatically by the colorbar function on the basis of the provided images. The YOLO network is based on a CNN network, so having uniform images helps to increase performance, which is why this representation was adopted.

An example of maps represented with “Colorcube” is shown in Figs. 7(c) and 7(b), where a range-Doppler image and an azimuth-Doppler map are reported, respectively.

B. YOLO parameters

For the tests, we consider a YOLOv3 [34] and a SqueezeNet as backbone CNN. SqueezeNet is a pre-trained model on ImageNet [51], to which layer freezing and transfer learning are applied, and represents one of the most lightweight solutions...
in terms of complexity and the number of parameters. Freezing a layer means that its weights cannot be changed further. The main advantage of transfer learning is that it mitigates the problem of insufficient training data limiting the number of parameters that can be updated during the training process. The architecture of SqueezeNet has been adapted to make it compatible with the requirements of YOLO. SqueezeNet consists of 8 convolutional blocks and presents 68 layers for feature extraction; the extracted features are the input for the YOLO detection layer used for object detection. When the freezing layer is applied, all the layers of the SqueezeNet are frozen and only two CNN-based “heads” are trained.

The SqueezeNet network is one of the networks with fewer parameters in the literature [52]. In fact, the number of learnable parameters is 1.2 million. When the YOLO network is added, the number of parameters increases to 6.4 million. By applying the freezing layer technique, the number of parameters does not decrease much (it drops to 5.8 million), because the most of them is given by the YOLO head, and the final training network size is 20 mega bytes. The parameters chosen for the network are reported in Table I.

<table>
<thead>
<tr>
<th>TABLE I</th>
<th>YOLOv3 PARAMETERS.</th>
</tr>
</thead>
<tbody>
<tr>
<td>N_Epochs</td>
<td>70</td>
</tr>
<tr>
<td>LearningRate</td>
<td>0.001</td>
</tr>
<tr>
<td>MiniBatchSize</td>
<td>8</td>
</tr>
<tr>
<td>Penalty threshold</td>
<td>0.5</td>
</tr>
<tr>
<td>Warm-up Period</td>
<td>500</td>
</tr>
<tr>
<td>Input size</td>
<td>227 x 227 x 3 pixel</td>
</tr>
<tr>
<td>Augmentation 1</td>
<td>Color jitter</td>
</tr>
<tr>
<td>Augmentation 2</td>
<td>Random horizontal flip</td>
</tr>
<tr>
<td>Augmentation 3</td>
<td>Random scaling by 10%</td>
</tr>
</tbody>
</table>

In order to avoid overfitting problems, the dataset is augmented before training. The augmentation process includes the use of several methods described in [53], in this case, the following were used: random horizontal flipping, random x/y scaling, and finally the application of jitter color. Jitter color is a technique that allows varying the brightness, contrast, hue, and saturation of the images.

IV. EXPERIMENTAL RESULTS

A. Dataset realization

The experimental tests were conducted at the Department of Information Engineering (DII) of Università Politecnica delle Marche where the targets are people moving within a room. The data collection is divided into two stages, the former one is used to collect the data to train the algorithm and the latter to test it. The chosen first Radar configuration parameters are shown in Table II and will be used for the training and testing phases; in the test phase, some parameters will be changed to demonstrate the independence of the proposed technique from the adopted configuration.

The Radar system is placed on a stand, connected to the control computer and to the power supply. The room in which the experiments are conducted is 8500 ± 2 mm long and 3 ± 2 mm high. The measurements are made with a DTAPE laser distance meter (model DT50).

<table>
<thead>
<tr>
<th>TABLE II</th>
<th>RADAR PARAMETERS FOR THE TRAINING SETUP.</th>
</tr>
</thead>
<tbody>
<tr>
<td>n_ADC</td>
<td>500</td>
</tr>
<tr>
<td>n_chips</td>
<td>128</td>
</tr>
<tr>
<td>f_Bandwidth</td>
<td>3.8 [GHz]</td>
</tr>
<tr>
<td>f_sampling</td>
<td>10000 [kHz]</td>
</tr>
<tr>
<td>n_frame</td>
<td>78</td>
</tr>
<tr>
<td>f_start</td>
<td>77 [kHz]</td>
</tr>
<tr>
<td>t_idle</td>
<td>20 [μs]</td>
</tr>
<tr>
<td>t_train</td>
<td>52 [μs]</td>
</tr>
<tr>
<td>Range max</td>
<td>10.2 [m]</td>
</tr>
</tbody>
</table>

The training setup involves two different subjects who perform two tests; the walking modes asked to perform during each acquisition are:
- Linear mode, as shown in Fig. 8 indicated as (1);
- Random mode, as shown in Fig. 8 indicated as (2).

The subjects repeat each activity three times so a total number of six acquisitions will be obtained.

As three different types of maps can be obtained from the Radar data cube, each acquisition provides 78 range-Doppler images, 78 images related to the azimuth-Doppler side, and 78 images related to the range-azimuth side, where each set of images can be considered as a separate dataset. At this point, for each type of map, a total number of 312 images is obtained. To add generality to the dataset, the acquisitions are made with the Radar at two different heights: the former is at 950 ± 2 mm above the ground, while the latter is at 1200 ± 2 mm. At this point, the datasets are composed of 624 images.

Images derived from range-Doppler maps and images derived from azimuth-Doppler maps are used to train YOLO-Doppler and YOLO-azimuth networks, respectively. The datasets are divided into two parts, 80% for training, and 20% for validating the networks. Images derived from range-azimuth maps will be only used at the end of the processing as a benchmark.

Following the training and validation of the network, the actual detection is performed on four different tests. Several settings and setups are considered, involving different Radar positions, different subjects and different ways of walking, as described in the following:
- Test01: test obtained from an acquisition in which a person walks in a linear mode. The Radar is raised up...
to 1800 ± 2 mm and tilted by 15°. Radar configuration and room remain the same as in the training setup (refer to Table II);

- Test02: the room remains unchanged, while the position and the height of the Radar vary from the previous test. In addition, there is a change in the Radar configuration: the one used in this test is described in Table III. With this configuration, two different acquisitions are performed, denoted in the following as “Test02 a)” and “Test02 b)”;
- Test03: same setting as Test02 and three different acquisitions are performed, denoted as “Test03 a)” and “Test03 b)” and “Test03 c)”;
- Test04: changing the environment, this acquisition is made in a hallway and in presence of obstacles like metallic shelves. This test aims to verify the robustness of the algorithm, by installing the Radar system in scenarios not included in the network training process. In Fig. 9 the acquisition set-up for Test04 is depicted.

The labels a), b) and c) indicate a different subject involved in the test. Tests 02, 03 and 04 have a different height of the Radar board with respect to Test01. This parameter is not measured in these tests to show that this feature has no impact on the performance evaluation.

![Fig. 9. Setup used during the acquisitions of Test04.](image)

### TABLE III

| Radar configuration parameters used for the tests: Test02, Test03 and Test04. |
|---------------------------------|---------------------------------|
| \( n_{ADC} \)                  | 500                             |
| \( n_{chirp} \)                 | 128                             |
| \( f_{bandwidth} \)            | 3.8 GHz                         |
| \( f_{sampling} \)             | 20000 kHz                       |
| \( n_{frame} \)                | 78                              |
| \( f_{start} \)                | 77 GHz                          |
| \( t_{idle} \)                 | 5 \( \mu \)s                     |
| \( t_{chirp} \)                | 52 \( \mu \)s                     |
| Range max                       | 20.4 [m]                        |

### B. Results and discussion

Since we are not dealing with an actual classification but with the detection of a target, some of the common evaluation metrics (such as classification matrices) are not suitable for measuring the performance of the proposed method. The accuracy of the network is therefore evaluated by considering precision, recall, and average precision on the validation set. In fact, these metrics are the most widely used to evaluate object detection algorithms [54]–[56]. Precision indicates the model’s ability to avoid false positive predictions. Recall measures the ability of the model to avoid false negative predictions. It indicates the percentage of correctly identified positive samples out of the total number of positive samples in the dataset.

Fig. 10 shows the trend of precision and recall related to three networks trained on three different datasets, i.e., the range-Doppler dataset, the azimuth-Doppler dataset and the range-azimuth dataset. From the figure it can be seen that the results of the network trained on range-azimuth images are not good, in fact, high precision values indicate that there are no false positives, while low recall values indicate the presence of many false negatives. On the other hand, high levels of accuracy and precision are achieved by the other two networks, meaning that they are able to provide accurate predictions and with a good ability to correctly identify positive objects. These results suggest that the model has a good ability to generalize and is effective in the specific application for which it was trained. Using the different YOLO networks on both sides for indirect detection on the range-azimuth map allows to decrease the number of false negatives compared with the case where the network was applied directly on the range-azimuth maps.

![Fig. 10. YOLO performance related to validation test: range-azimuth (blue), range-Doppler (red), azimuth-Doppler (yellow).](image)

To evaluate the performance of the algorithm in the four tests considered, average accuracy, intended as the number of true detections over the total number of maps (78), cannot be used because the bounding-boxes on range-azimuth are obtained indirectly using YOLO-Doppler and YOLO-azimuth networks on the other maps. For this reason, the detection performance is evaluated by counting the number of frames in which the bounding-box correctly detects the target presence. To count the frames in which detection occurred correctly a handmade label is made on the test sets. After that, a comparison with the result obtained from the network can be made. To count a frame in which correct detection occurred, it is sufficient to check whether there was a match between the output of the mmDetect approach and the handmade label. Considering that the total number of frames for each acquisition is 78, Table IV shows the percentage of frames in which the bounding-box correctly reveals the target in the range-Doppler map and, the percentage of frames in which detection occurred in the azimuth-Doppler side in the four different tests under consideration. The number in parentheses indicates the number of maps with correct detection. The last
The results show that it is very difficult to apply YOLO directly on range-azimuth maps, due to the numerous false negatives that are obtained. To improve the performance, all three sides of the Radar data cube are exploited, which are range-Doppler side, Doppler-azimuth side, and range-azimuth side. The proposed approach can work also without the application of background removal techniques. Due to the possibility of exploiting the Doppler effect of the target, the proposed method can perform detection of the target with better performances than the traditional methods based on CFAR thresholds. Unlike CFAR, which requires a careful choice of parameters to determine the threshold, our YOLO-based approach is able also to perform classification due to its convolutional neural network architecture. This feature is not used in this work but offers a future possibility of improvement of what is here proposed. Considering experimental tests which included different Radar configurations, different subjects, and indoor environments, the feasibility of the proposed approach showed that it is able to detect the target on average in 92.49% of the cases with also very good generalization capabilities.

**V. CONCLUSIONS**

In this paper, an approach for target detection based on the joint use of FMCW Radar and YOLOv3 neural network is proposed. The results show that it is very difficult to apply YOLO directly on range-azimuth maps, due to the numerous false negatives that are obtained. To improve the performance, all three sides of the Radar data cube are exploited, which are range-Doppler side, Doppler-azimuth side, and range-azimuth side. The proposed approach can work also without the application of background removal techniques. Due to the possibility of exploiting the Doppler effect of the target, the proposed method can perform detection of the target with better performances than the traditional methods based on CFAR thresholds. Unlike CFAR, which requires a careful choice of parameters to determine the threshold, our YOLO-based approach is able also to perform classification due to its convolutional neural network architecture. This feature is not used in this work but offers a future possibility of improvement of what is here proposed. Considering experimental tests which included different Radar configurations, different subjects, and indoor environments, the feasibility of the proposed approach showed that it is able to detect the target on average in 92.49% of the cases with also very good generalization capabilities.

**REFERENCES**


Fig. 11. Application of classical pipeline for target detection from radar acquisition: (a) Results of CA-CFAR application on range-azimuth map with a false alarm probability of $10^{-5}$, a training band size of $[4, 4]$ pixels and a guard band size of $[2, 2]$ pixels. (b) Results of DBSCAN application with a neighbourhood search radius of 4 and a minimum number of neighbours of 3. (c) Result obtained with the proposed approach.


[21] E. T. S. Institute, Electromagnetic compatibility and Radio spectrum Matters (ERM); System Reference document (SRdoc); Technical characteristics of Radio equipment to be used in the 76 GHz to 77 GHz band; Short-Range Radar to be fitted on fixed transport infrastructure, ETSI TR 103 148 V1.1.1 ed. European Telecommunications Standards Institute, 2014.


