2021 22nd IEEE International Conference on Industrial Technology (ICIT) | 978-1-7281-5730-6/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICIT46573.2021.9453703

ToF/Radar early feature-based fusion system for human detection and tracking

Feryel Zoghlami
Automation, Maintenance and Factory Integration
Infineon Technologies Dresden GmbH & Co. KG
Dresden, Germany
Feriel.Zoghlami@infineon.com

Okan Kamil Sen
Technische Universität München
München, Germany
okankamil.sen@tum.de

Harald Heinrich
Automation, Maintenance and Factory Integration
Infineon Technologies Dresden GmbH & Co. KG
Dresden, Germany
Harald.Heinrich@infineon.com

Germar Schneider
Automation, Maintenance and Factory Integration
Infineon Technologies Dresden GmbH & Co. KG
Dresden, Germany
Germar.Schneider@infineon.com

Emec Ercelik
Technische Universität München
München, Germany
emec.ercelik@tum.de

Alois Knoll
Technische Universität München
München, Germany
knoll@in.tum.de

Thomas Villmann
University of Applied Sciences Mittweida, Department of MPI
Mittweida, Germany
villmann@hs-mittweida.de

***Abstract*—Industry 4.0 has become a general keyword over the last years. It is based on the inclusion of automation by increasing connectivity in various tasks during the production process. This fact did not exclude the human's effort whose presence remains important, especially the interaction between humans and robots will be a key element in the future manufacturing. In automated production lines, we find both humans and robots operating side-by-side in hybrid workplaces. The major focus for this workplaces today and in the future is to establish a safe work environment. However, what if safety meets "collaborative efficiency"? The system presented in this paper relies on the fusion of data coming from a Time of Flight (ToF) sensor and a 60 GHz radar sensor. The data are analyzed and evaluated using deep learning (DL) algorithms. The purpose is to detect humans and track their movements in the observed area. The resulted perception system can be installed somewhere in a room or on a moving system. A first demonstrator has been developed, tested and evaluated. An additional graphical interface was developed to show in real time the capability of the data fusion system. The system can detect up to 5 persons in a selected area with 98% confidentiality. The so-described system is able as well to estimate each person DoM and the person's instantaneous speed and position. Based on the output of our developed system, it is possible to define industrial use cases as well as many other different applications in different fields.**

***Index Terms*—sensor fusion, human/robotic collaboration, deep learning, industry 4.0, machine learning, automated fabrication, radar sensor, time of flight camera**

I. Introduction

Over the last years, the industry image has been changed from a purely manual work achieved by human workers to automated tasks ensured mainly by machines and robotic systems operating at 24/7. Today, the main tasks of workers in the production are changed from operation to control tasks. However, when a failure occurs, the human should be present near to the failure source and must immediately solve the problem. In a hybrid workplace, machines should observe the human intentions and stop immediately to avoid accidents. However, a continuous human interaction can lead to excessive stops in the production line due to the false classification of human intentions by the machines. Researches are focusing today on overcoming the passive usage of the sensors implemented everywhere in the fabrication area. Today these sensors are used only to inform the workers about the machine state and to establish a safe shared workplace in order to minimize the risk of further accidents. It is important to improve the collaboration between humans and machines and make both parts communicate with each other. This starts by a smart exploitation of the information we receive from implemented sensors. In this context, Internet of Things (IoT) [1] has received a big interest but it remains an approach that focuses more on establishing a harmonic communication between the different sensors. As a result, human is partially included in

978-1-7281-5730-6/21/$31.00 ©2021 IEEE

this communication. Sensor fusion [2] is more focusing on the perception of a dynamic environment including humans. In fact, for settling an advanced collaborative workplace, it is worth to use two or more sensors in one compact system to observe the surrounding area and to collect continuously relevant information used to guide machines and robots in various situations.

The power of the sensor data fusion consists in ensuring a spatial and temporal coverage extension with improvement of the global system resolution. The reference to both redundant and new data format helps in increasing the confidence in measurements and reducing ambiguities by being able to classify objects and behaviors in a complex environment. We develop in this paper a method combining a ToF and a Radar sensor data. ToF gives the information about the shape and the exact position of the target in the 3D world and the radar confirms this position and gives additional information about the velocity of each moving target. Since we receive from both sensors thousands of data, which cannot be all helpful for our development, we refer to machine learning (ML) methods and to deep learning (DL) in order to filter received data and select relevant features useful for our development.

II. Related Work

During the last few years, human-robot interaction (HRI) and human-robot collaboration (HRC) have gained a lot of interest among researchers and industry. This collaboration aims mainly at exploiting the different but complementary skills of both the human workers and the programmable robots in order to achieve a common goal. Establishing a safe shared workplace in this case is important and challenging at the same time. When previously robotic systems were isolated in closed operating rooms in order to ensure the safety of the human workers, today, with the appearance of new technologies, human workers and robots can work in the same room and the area protection is ensured with the help of different sensor modularities implemented either on operating robots, at a stationary positions inside the room or both. Lidar scanners based-technology, with conformance and certification against the International Organization for Standardization (ISO) 10218-1/2 [3] and ISO/TS 15066 [4], remains the most exploited mean for settling a safe manufactory as illustrated in [5]–[8]. These scanners work with 2D scan and with an opening view of 270 degrees. Therefore, these kind of sensors need to be implemented on robots in different directions in order to ensure a total coverage of the surrounding area. There are thousands of cobots operating and transferring products between different stations in a fully automated factory like semiconductor [9] and automotive [10] industries. With this state of the art technology, we need today more than 6 expensive scanner systems to protect the working area covered by one robotic system operating in only one station. In 2020, Mariane D. in [11] estimates that the number of industrial robots will reach over 3 million, which is almost double of the current number. Therefore, with the current solution there would be a need of heaps of such scanners. Such a huge number needs from the company a high cost investment. In order to overcome these losses, studies are currently oriented to look for alternative solutions for better interaction between humans and robots. The major goal is to ameliorate productivity by reducing the number of successive stops with a minimum of collision risks. The robot's trajectory re-planning is one of the proposed solution and there are various methods adopted in this purpose. Emam Fathy Mohamed introduces in his paper [12] the utility of creating a potential field in the robot workspace with repulsive or attractive pressures on the surface of the obstacle and the target respectively, which helps in the robot's collision-free path planning. However, this method could have an indetermination when both repulsive and attractive forces are equal or similar. The second studies axis interested in ensuring a complementary and safe collaborative human-robot work is based on vision monitoring. Companies like MetraLabs, LG Business were for example among the first companies, which use 3D Kinect camera or ToF systems for obstacle detection and avoidance in the robotics field. The company Pilz has developed a camera system, which they call a SafetyEYE [15]. This system enables the configuration of multiple safety zones for the human-robot workspace. Various industries are today exploiting this system to monitor the violations of these predefined zones and forces of the robot, (e.g. to decrease the speed or stop when movement is detected in a certain zone). Flacco et al. presents in his papers [15], [16] a fast method to calculate the distance between a number of points and moving obstacles using a depth camera. The vision based monitoring approach explained above uses only one 3D vision sensor, which cannot be efficient in case of complicated collaborative tasks in very dynamic environments. Therefore, recent studies go in the direction of combining different sensors in the same system to improve reliability regardless the external requirements of an industrial workspace. studies in this field have a focus on finding robust solutions in mainly the automotive area for both Advanced Driver Assistance System (ADAS) like in the paper of Ziguo Zhong [18] and gesture recognition applications like in [19] where authors explain a data fusion-based system using both a radar sensor and an RGBD camera. These outdoor applications require a wide detecting range and a big sustainability against the environment conditions to ensure the maximum of safety for human lives. These environment restrictions are of more tolerance in the industries, where machines run in lower speeds and don't need a high resolution to explore the environment. Therefore, the development of the sensor platforms in this case reveals more flexible and gives the chance to develop various new technologies that prove reliability and high accuracy in recent studies for different applications.

III. Methodology

A. Hardware Choice

Our sensor fusion based system is composed of a CamBoard pico Monstar ToF camera [31], which is produced by the company PMD technologies AG with a resolution 352x287 pixels and an opening field of view of 100°Hx85°V. This depth camera deliver a point cloud of the scene in its field of view as well as an aligned 2D amplitude image of the same scene. We chose a frame rate of 35fps for our application. The ToF camera is installed in a compact system together with a 60GHz Frequency-Modulated-Continuous-Wave (FMCW) radar sensor with 2 transmitters and 4 receivers (2RX/4TX) [32]. We configure our radar to work with a bandwidth of 1GHz between 59.5GHz and 60.5GHz and a sampling frequency of 200000Hz. With this configuration, the radar sends 24 chirps per frame and 64 samples per chirp. The used radar operates in a maximum range of 4.5m. Both sensors presented in the fig. 1 are based on chips produced in the company Infineon Technologies AG.

(a)

(b)

Fig. 1: (a) 60 GHz radar sensor module (Project Soli website,2019) [32]. (b) CamBoard pico Monstar [31]

B. Data Preprocessing

Before starting the fusion approach, it is important to analyze the data coming from both sensors separately and modify their format depending on our fusion goals. While using depth and amplitude matrices from the ToF camera, we have used the reflected signals of the detected targets for each radar antenna. Radar's signal processing starts with creating Intermediate Frequency (IF) signal by mixing the reflected and transmitted signal. It is a sinusoidal signal and includes phase (Φ) and frequency (f) components as can be seen in the following equation.

$$IF = sin^{-1}(2\pi ft + \Phi) \tag{1}$$

It is later used in radar signal processing chain for creating useful information in fusion. In this section, we focus on the processing of the radar raw data and its transformation in relevant information exploited for our development. The data flow from both sensors is illustrated in the fig. 2

1) Range Doppler Map (RDM): It includes both the range and the velocity information of the targets around peaks. The following steps show how it is generated.

a) Range FFT: Fast Fourier Transform (FFT) is the most common technique used in digital signal processing. In FMCW radars, FFT of IF will give us information about the range of the target. IF signal frequencies of each object in front of radar will be proportional to its range. Therefore, applying this algorithm will create peaks in several locations for each target on FFT signal, which is called range FFT.

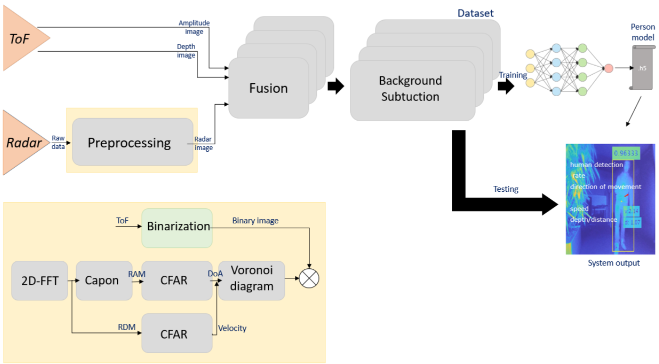
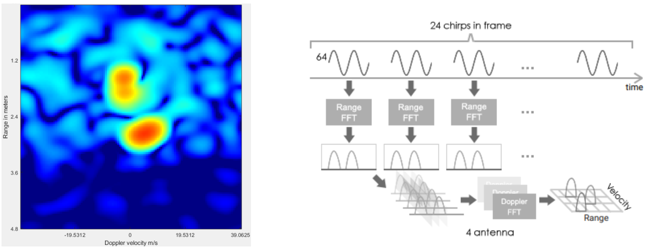


Fig. 2: Sensors' data flow chart: from raw data until human detection and tracking

b) Range Doppler/ 2D-FFT: It is important to have the frequency component of the IF signal in order to calculate the range, however it is not enough to find out small displacements. In this case, phase of the IF signal can be used to distinguish closely located targets. It is also possible to extract velocity while utilizing period of the radar TX signal from the following phase difference formula.

$$\Delta\Phi = \frac{4\pi v T_c}{\lambda} \tag{2}$$

If there are multiple objects in the same range with different velocities, the peak value in range-FFT includes phase components of each target. Similar to range FFT, doppler FFT can be applied to resolve these objects. The result is called range-Doppler Map (RDM) presented in the Fig. 3. We construct one RDM for each of the four antennas.



(a)

(b)

Fig. 3: (a) RDM with 2 detected targets, (b) RDM construction from IF radar signal

2) Direction of Arrival (DoA) estimation:

a) Range Angle Map (RAM): The radar sensor provided by the company Infineon Technologies AG is a sensor with 2x2 URA array shape, which contains four antennas. To create the range angle map we use only two either horizontal or vertical antennas' signals. Due to the low angle resolution of the sensor due to the limited number of antennas, an accurate DoA of the detected targets is difficult to estimate. Therefore, we adopt an additional algorithm in order to improve our estimation.

b) Minimum Variance Distortionless Response (MVDR) beamformer: MVDR [20] or also known as Capon algorithm is used for DoA estimation. Capon beamformer is an enhanced version of the Bartlett method [20]. For targets that are close to each other, Capon provides more precise peaks than Bartlett algorithm. It maximizes the sensitivity in one direction only. An illustration of both algorithms' performances is presented in Fig.4

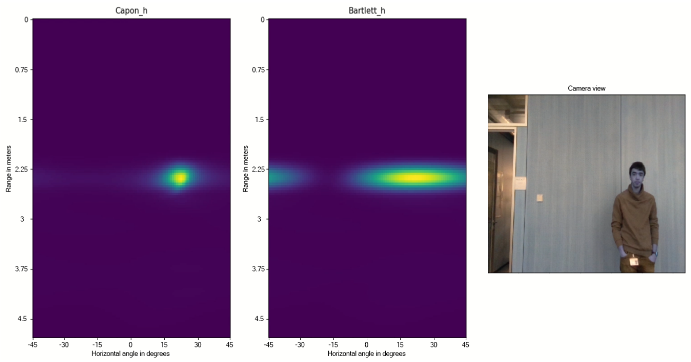


Fig. 4: Comparison of the capon and the Bartlett beam formers for range angle construction

Based on multiple test, we choose to apply Capon algorithm in order to have more precise detection of the moving target in the radar sensor view and therefore better angle estimation.

3) 2D CA-CFAR vs. 2D OS-CFAR: CFAR stands for Constant False Alarm Rate is an adaptive thresholding algorithm which is used in radar signal processing to detect peaks (targets in our case) while neglecting background noise. During our developments, we consider the computation time as one of our main constraint. Therefore, we choose to work with the Cell Averaging (CA) CFAR algorithm, which is robust enough in homogeneous environment [24], shows good performance in pure noise situations and it is faster than the Ordered Statistics (OS) CFAR algorithm [25].

4) DBSCAN: DBSCAN [26] can identify noisy clusters of arbitrary shape and size. For each received frame we construct four RDMs (4 RX) and one RAM and then we apply for each the 2D CA-CFAR algorithm and finally the DBSCAN clustering algorithm. In the first phase, we detect regions of peaks designating the detected moving targets. In the second phase, we cluster the peaks into one peak. For each CFAR region, we define initially random clusters' centers inside the region, afterwards we draw around each initial point a boundary with radius equal to a chosen epsilon value. We go through all density reachable points from each starting point and we define a cluster each time we find a minimum number of data points inside the drawn boundary. After matching the peaks between RDMs and with the RAM based on the range values, we extract the corresponding velocity values from the RDM and the corresponding angle of arrival values from the RAM.

5) 1D CA-CFAR: Since we ignore the exact number and locations of our targets in the radar map, we run the 2D CFAR algorithm, which uses sliding windows, over the whole RDM to determine the range and velocity of each target. The whole algorithm spends 1.8s for each of the 4 antennas regardless the number of the moving targets. In order to optimize the computation time spent in this phase, we consider only the first antenna (we believe that detected targets are seen at the same time in the same place with the same velocity by all the antennas) and we refer to the 1D CFAR algorithm. Therefore, from the RDM we create a new 1D signal, which presents the mean over velocities for each range slot. We apply to this signal 1D CFAR to find peaks and note their corresponding ranges. Fig. 5 illustrates the method explained above.

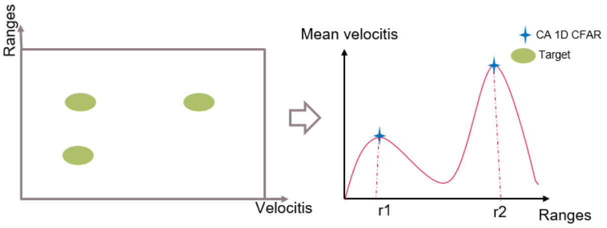


Fig. 5: Targets' detection using only 1D CFAR. (a) 2D RDM with detected targets. (b) 1D signal extracted from RDM with CFAR peaks

6) Cartesian coordinates calculation: We construct two RAMs the horizontal RAM, which uses the two horizontal antennas and the vertical RAM, which uses the two vertical antennas. We apply 1D CFAR to signals corresponding to the ranges computed in the previous section. From the horizontal RAM, we determine azimuth angle and from the vertical RAM, we determine the elevation angle. In the end, we calculate the 3D cartesian coordinates of each detected target by applying the following function:

$$\begin{aligned} &\phi : IR^4 \to IR^3 \\ &(r_1, \theta_1, r2, \theta_2) \to (x, y, z) = \quad (r1sin\theta_1, r2sin\theta_2, \\ &\qquad\qquad\qquad\qquad \frac{r1sin\theta_1 + r2sin\theta_2}{2}) \end{aligned} \tag{3}$$

Where (r_1, θ_1) and (r_2, θ_2) are the parameters extracted for each target from respectively the horizontal RAM and the vertical RAM.

The whole process from RDMs and RAMs construction until cartesian coordinates calculation takes only 0.004s. As a result, the system is around 450 times faster compared to the system using the old approach.

C. Sensor calibration

In order to get data from both sensors on the same coordinate system, we start by calibrating data on the x, y and z axes. We install both sensors side by side and we track the center of a spherical ball moving in all directions in front [21]. We use a sphere rather than any other form in order to guarantee a unique reflection from the same target especially on the radar map.

D. Fusion Approach

We adapt in this paper an early fusion feature based approach. From the ToF camera, we consider both depth and amplitude images. From the radar, we consider the detected targets with their 3D positions and velocities. The fusion approach is illustrated by the fig.6.

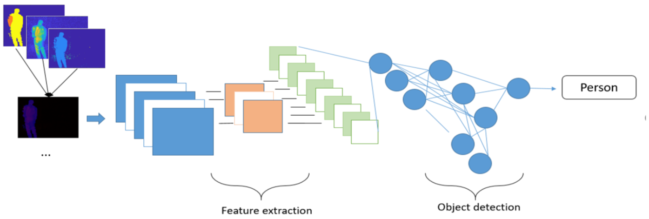


Fig. 6: human detector training architecture

The following sections explain in details the adopted fusion procedure.

1) Dataset creation and annotation: Since there is not a dataset for indoor applications available for both ToF and radar, we create our own dataset. We install our system with both sensors and run it to get synchronized data inside the cleanroom at the wafer fabrication site of Infineon Technologies Dresden. Data sampling is sequential and run on a raspberry pi3 board with sampling frequency of 1 sample/s. Our focus is to detect persons. Therefore, we collect samples from humans with different shapes and postures in different places inside the cleanroom. We collect as well samples from different persons outside the cleanroom to enhance our detection performance. In the end we collect 1500; One sample is an xml file that contains synchronized raw data from both sensors. Before moving to the training phase, we parse the generated xml files and create out of the raw data images fused and then used for training the person detection model.

2) Radar Image: From the raw data in the radar map, we construct a so called radar image, which has the same shape information from the ToF images but with additional information about the moving object's velocity. The construction of the radar image follows the steps summarized in the fig.7.

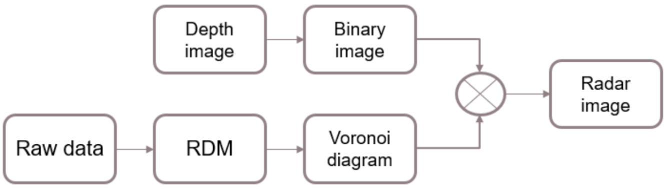
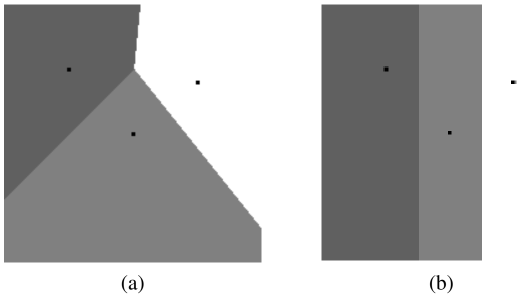


Fig. 7: Radar image construction flow

From radar, we detect random points from the detected objects in front. For each target point, we calculate as explained in III-B6 the 3D position and the radial velocity. Based on this values we construct for each frame a 200x200 voronoi diagram, which relies on the segmentation of the frame into regions in the pixel scale and affect to each of this region an intensity value proportional to the velocity of the target inside this region. The construction of one voronoi diagram takes around 1s. Due to the real time constraints of our application, we look for accelerating the built of such diagram by using only the x values from the detected targets for the construction of the voronoi diagram's regions. As a result, we obtain a voronoi diagram with rectangular areas. Both old and modified voronoi diagram are depicted in the Fig. 8 .


(a)

(b)

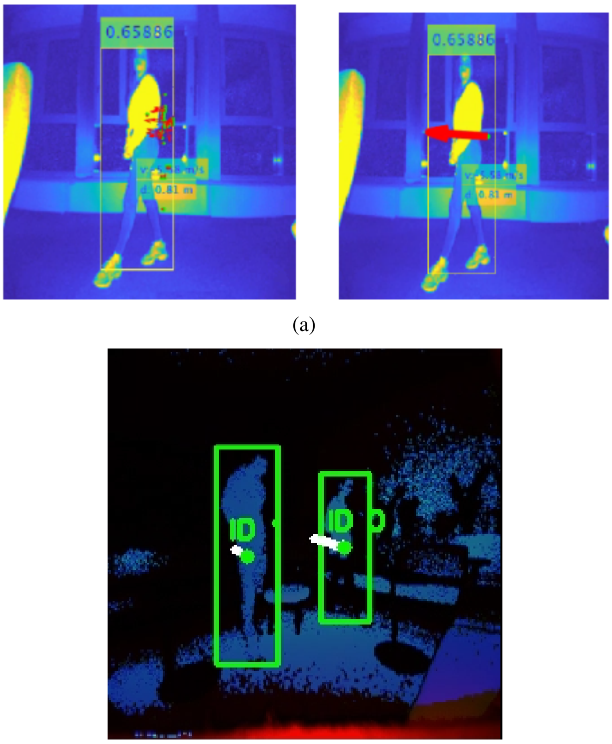
Fig. 8: Voronoi diagram of three detected targets (depicted by black points): (a) old method, (b) modified diagram

With this new approach, the diagram creation is 100 times faster than before. It is important to affect each velocity from the radar to its corresponding object. Therefore, we start by resizing the voronoi diagram to the ToF size. Afterwards, we use a 2D binary image that we construct from the ToF depth image to mask the resized voronoi diagram. As a result, we get a 2D radar image with both shape and velocity information of the objects present in the aligned radar and ToF view.

3) 3-channel inputs: From ToF we receive depth images, each presents the z values from the point cloud in each pixel. We consider as well an aligned amplitude image, which is a 2D matrix where pixels present the strength of the reflected signal from the active illumination unit. From radar we construct the so called radar image. Our created dataset is composed of 1500 3-channel samples (depth, amplitude, radar). We structure Our dataset as follows: 1200 samples for training, 150 samples for validation and 150 samples for testing. The dataset result is used to train a model for human detection.

4) Direction of movement estimation: In order to understand more the human behavior, we track his movements by estimating his direction of movement (DoM). For this purpose, we test and evaluate two different approaches.

a) Lukas Kanade optical flow: We base our calculation on the amplitude image and search for the optical flow regarding each pixel. In fact, we apply the Lukas Kanade algorithm, which is a sparse optical flow algorithm that allows the tracking of features in consecutive images. Changes in the brightness are used to define the movement of each feature. During the testing phase, we consider the pixels inside the bounding box around the detected person from the amplitude channel. We start by searching for the relevant Harris features' inside the region of interest in the first frame. Afterwards, we track these features brightness change in the next frame. Finally, the average displacement on both x and y direction is calculated and we consider this value as the general DoM of the detected person.



(a)

(b)

Fig. 9: Direction of movement estimation based on optical flow relevant Harris feature tracking on the left average direction magnitude on the right in (a), and on centroid tracking method (c)

b) Bounding boxes tracking: The algorithm consists in tracking the center of each bounding box resulting from the detection phase. For each new target means each new bounding box around the detected human, we attribute a unique ID. Each time a new target appears in the system view, a new ID will be assigned to it. From one frame to the next frame, one target keeps the same ID. Fig. 9 illustrate the output from both methods tested in real time.

IV. Experimental Results and Discussion

A. System outputs

Our developed system aims to detect humans running around and to track their movements. In this paper, we limit our work to visualize the system outputs in real time. Fig. 10 illustrates one frame where we detect one person in the area. The detection rate is displayed on top of the bounding box and in the middle the depth value (d) in meters and the velocity value (v) in meters per second are shown. The DoM of the person is displayed with the red arrow in the center of the human body.

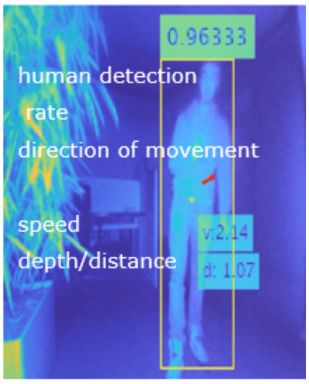
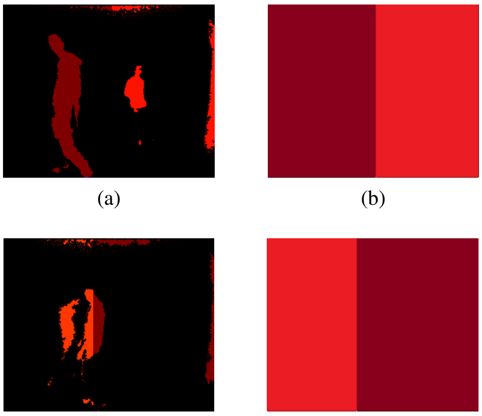


Fig. 10: Human detection and tracking fusion system outputs

B. System evaluation

In this section, we focus on the evaluation of the radar image constructed in the section III-D2 as well as the comparison of the performances of the 3 tested Convolutional Neural Networks (CNN)-based neural networks used for human detection. Finally, we evaluate the outputs from our whole fusion system.

a) Radar image: The radar image is a combination of visual features coming from the ToF camera aligned with targets detected by the radar sensor. The result is a masked voronoi diagram fusing both shape and velocity information. Examples from two voronoi diagrams and their corresponding so-called radar images are presented in fig.11.



(a)

(b)

(c)

(d)

Fig. 11: examples of voronoi diagram (on the second colomn) and resulted radar images (on the first colomn)

In the second raw of the graph, we have a misinterpretation of velocity pixels; the person (on the right) shows on a part of his body velocity value that belongs to the other person (on the left). This failure happens mainly in the scenario when persons are close to each other due to the low angle resolution of the radar sensor.

b) Direction of movement evaluation: In this section, we compare both approaches explained in the section III-D4 for estimating the DoM of the detected persons. Therefore, we calculate two metrics, which are the average angle error (AAE) and the inference time. The AAE is calculated as follows:

$$AAE = arccos\left(\frac{u_c u_e + v_c v_e + 1}{\sqrt{(u_c^2 + v_c^2 + 1)(u_e^2 + v_e^2 + 1)}}\right) \quad (4)$$

where (u_e, v_e) and (u_c, v_c) are the estimated and the reference displacement values (ground truth values) respectively on the x and y axis in the 2D image.

Our test consists in testing different sequences of successive images with standard movements (moving left, moving right, moving forward, moving backward).We compare the predicted directions from the two approaches with the ground-truth references. The results of both of the techniques are presented for a single target in TABLE.I. For the Lucas Kanade tests, we have used different number of features selected from images. Neighboring size of 30x30, 15x15, 5x5 and 1x1 pixels are used to get features in these tests. The aim of these neighboring size is to reduce the effect of the features that are inside the bounding box but not on the human body.

TABLE I: Direciton of movement techniques evaluation

	AAE in radians	**Inference time in seconds**
Centroid tracking	**0.146**	0.007
Lukas Kanade 30x30	1.18	**0.003**
Lukas Kanade 15x15	1.24	0.005
Lukas Kanade 5x5	1.41	0.032
Lukas Kanade 1x1	1.48	0.819

Only the points inside the detected bounding boxes are considered in all tests. Inside one bounding box, beside the human target, there are also some pixels, which belong to the background. These pixels confuse the DoM estimation algorithm. This explain why the 30x30 neighboring window performs better than the other Lucas Kanade window's sizes since there are only a few points which doesn't belong to the target. However, the Lukas Kanade algorithm still doesn't perform well in the case of non-rigid corps tracking, like tracking a person, since there are some parts of the body moving in totally different directions simultaneously. This disadvantage is overcame with the centroid based tracking algorithm since only the centroid (one pixel) of the bounding box is tracked.

c) Human detection: In this paper we compare three neural networks used for human detectors training, which are: faster RCNN [27], YOLOv3 [28] and Retinanet [27]. For consistent comparison, we use the same dataset. The Nets are compared regarding the inference time and the mean average precision for both 50% and 75% Intersection over Union (IoU) between ground truth and the predicted objects. In order to accelerate the training convergence, we use pretrained weights on the COCO dataset and fine tune them on our own dataset. All chosen Networks are CNN-based networks. For YOLOv3 we use Darknet53 as a features extraction network and Resnet50 for both Faster RCNN and Retinanet.

Our dataset is composed of training samples divided into 80% for training 10% for validation and 10% for testing (for evaluation). Each of the neural networks cited above is trained three times in order to enhance the robustness of the trained human models. The performance of the trained models is summarized in the TABLE.II.

TABLE II: Trained neural network performances

IoU	**Model**	**Precision**	**Recall**	**AP**	**FPS**
50% IoU	FRCNN resnet50	**0.99**	0.98	**0.98**	13.7
	YOLOv3 darknet53	**0.99**	**0.99**	**0.98**	**29.2**
	Retinanet resnet50	0.98	0.97	0.95	14.4
75% IoU	FRCNN resnet50	0.88	**0.90**	**0.87**	13.7
	YOLOv3 darknet53	**0.89**	0.89	0.81	**29.2**
	Retinanet resnet50	0.85	0.84	0.81	14.5

The precision metric, which describes the prediction performance of the trained model and the recall metric, which define the ability of how ground truth values are found when testing the trained model, are calculated as follows:

$$Precision = \frac{TP}{TP + FP} \quad (5)$$

$$Recall = \frac{TP}{TP + FN} \quad (6)$$

where TP is the True Positive detection rate (there is a person and it is detected), FP is the False Positive rate (there is no person but there is detection), and FN is the False Negative rate (there is a person and it is not detected). Starting from the results presented in the TABLE. II, the faster RCNN network, which gain a lot of interest in the research field over the last years has the highest mean average precision value compared to the other networks. This is the advantage of the two-stage network with Region Proposal Network (RPN) [27], however as a trade-off, faster RCNN is slower than the other networks. On the other hand, YOLOv3 has a small inference time but with slightly less precision. For a real-time transmission of the fusion system output the human model trained on YOLOLv3 based Darknet 53 is considered for our system development. In this case, the system has the highest frame rate and a relatively high accuracy.

d) Comparison to the state-of-the-art solution: In this section we compare our system performance to the state-of-the-art solution (safety scanners) regarding functionalities and testing run-time. TABLE III summarizes the results from the two solutions comparison.

While our proposed solution outperform the state-of-the-art solution in terms of human detection and tracking capabilities, it presents a lower frame rate. This can be explained by the voluminous network (YOLOv3 Darknet53-based network) we were using for human detection. We could run our system with up to 3.5fps on Nvidia Geforce GTX 1060 GPU. This frame rate should be optimized in the future to reach at least 5fps in order to respect the real time constraint for a real industrial application.

TABLE III: Comparison between state-of-the-art and proposed solutions

performance metrics	safety laser scanner	Our fusion system
Obstacle detection	+	+
Human detection	-	+
Human localisation	-	+
Velocity estimation	-	+
DoM estimation	-	+
Testing frame rate	-	+

Conclusion

We present in this paper a 3D fusion system based on the combination of data coming from a 3D ToF camera and a 60 GHz radar sensor. The data is fused on the feature level and the system is trained to detect persons walking around with a confidence up to 98%. Besides detection and positioning of humans in the 3D workplace, the system tracks them and gives information about each person's speed and DoM. The system works with a frame rate of 3.5fps, which should be improved by optimizing the considered algorithms in order to reach the real time requirements for uploading the system in a real use case. Besides, the system presents velocity estimation errors in case of multi-targets detection. In fact, the low angle resolution of the used radar results in the misinterpretation of each person velocity especially in crowded areas where people are very close to each other. Therefore, in a next step we plan to try different fusion approaches (e.g. late fusion) to improve the system robustness or to think about more sensors integration depending on the application restrictions.

Acknowledgment

This project has received funding from the ECSEL Joint Undertaking (JU) under grant agreement No 826452. The JU receives support from the European Union's Horizon 2020 research and innovation program and Sweden, Austria, Spain, Poland, Germany, Italy, Czech Republic, Netherlands, Belgium, Latvia, Romania, France, Hungary, Portugal, Finland, Turkey, Norway, Switzerland.

References

[1] Smart Sensor Technology for the IoT. Techbriefs Media Group. https://www.techbriefs.com/component/content/article/tb/features/articles/33212

[2] Sensor Fusion. https://www.sciencedirect.com/topics/engineering/sensor-fusion

[3] ISO 10218-1/2:2011 (2011) Robots and Robotic Devices Safety Requirements for Industrial Robots Part 1: Robots/Part 2: Robot Systems and Integration.

[4] ISO/TS 15066:2016 (2016) Robots and Robotic Devices Collaborative Robots.

[5] Y. Kakigi, K. Inoue, M. Hijikata, K. Ozaki, "Development of Flexible Cowl Covered Mobile Robot in Consideration with Safety and Design Property", Journal of Robotics and Mechatronics, 2017.

[6] A. Jack J, S. Mason, C. Richard, B. Saad "Collision Avoidance System for Unmanned Aerial Vehicles Using LiDAR and Optical Flow",2018.

[7] Á. M. Guerrero-Higueras , C. Álvarez-Aparicio , M. Carmen Calvo, "Tracking People in a Mobile Robot From 2D LIDAR Scans Using Full Convolutional Neural Networks for Security in Cluttered Environments", 2018.

[8] L. Jun ; T. Keng Peng ; Ch. Lawrence , "A perception system for robot arms to convey objects to in-car passengers",Asia Pacific Signal and Information Processing Association Annual Summit and Conference(APSIPA ASC), 2017.

[9] H. Heinrich, G.Schneider, F. Heinlein, S. Keil, "Pursuing the Increase of Factory Automation in 200mm Frontend Manufacturing to Manage the Changes Imposed by the Transition from High-Volume Low-Mix to High-Mix Low-Volume Production", 8th ASMC Conference Boston

[10] G. Michalos, S. Makris, J. Spiliotopoulo, "ROBO-PARTNER: Seamless Human-Robot Cooperation for Intelligent, Flexible and Safe Operations in the Assembly Factories of the Future", Conference on Assembly Technologies and Systems, 2014.

[11] M. Davids "A Look into Fully Automated Futuristic Factories", 2019.

[12] E. F. Mohamed, Kh. El-Metwally, A.R. Hanafy "An improved Tangent Bug method integrated with artificial potential field for multi-robot path planning" Innovations in Intelligent Systems and Applications (INISTA), 2011.

[13] M. Khansari-Zadeh, Seyed, A. Billard, "A dynamical system approach to realtime obstacle avoidance", Auton. Robots 32 (4), 2012.

[14] R. Meziane, Martin J.-D. Otis Hassan Ezzaidi, "Human-robot collaboration while sharing production activities in dynamic environment: SPADER system", 2017.

[15] F. Flacco, T. Kröger, A. De Luca, O. A. Khatib, "depth space approach to human-robot collision avoidance", IEEE International Conference on Robotics and Automation, 2012.

[16] F. Flacco, T. Kröger, A. De Luca, O. A. Khatib , "Depth Space Approach for Evaluating Distance to Objects: with Application to Human-Robot Collision Avoidance" Journal of Intelligent and Robotic Systems: Theory and Applications, 2014.

[17] L. Wang, B. Schmidt, A. Y. C. Nee, "Vision-guided active collision avoidance for human-robot Collaborations", Manufacturing Letters, 2013.

[18] Z. Zhong, S. leyLiu, M. Mathew and A. Dubey, "Camera Radar Fusion for Increased Reliability in ADAS Applications", 2018.

[19] P. Molchanov, Sh. Gupta, K. Kim, and K. Pulli "Multi-sensor System for Driver's Hand-Gesture Recognition" NVIDIA Research, Santa Clara, California, USA.

[20] S. Bhuiya, F. Islam, and M. Matin. "Analysis of direction of arrival techniques using uniform linear array." In: International Journal of ComputerTheory and Engineering 4.6 (2012), pp. 931–934.

[21] Molchanov, P., Gupta, S., Kim, K., & Pulli, K. (2015, May). Short-range FMCW monopulse radar for hand-gesture sensing. In 2015 IEEE Radar Conference (RadarCon) (pp. 1491-1496). IEEE.

[22] Khodja, M., Belouchrani, A., & Abed-Meraim, K. (2012). Performance analysis for time-frequency MUSIC algorithm in presence of both additive noise and array calibration errors. EURASIP Journal on Advances in Signal Processing, 2012(1), 94.

[23] Ahmad, M., & Zhang, X. (2016). Performance of MUSIC Algorithm for DOA Estimation. ICAYS Engineering Mechanics and Interdisciplinary.

[24] Hong, S. W., & Han, D. S. (2014). Performance analysis of an environmental adaptive CFAR detector. Mathematical Problems in Engineering, 2014.

[25] Kronauge, M., & Rohling, H. (2013). Fast two-dimensional CFAR procedure. IEEE Transactions on Aerospace and Electronic Systems, 49(3), 1817-1823.

[26] Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996, August). A density-based algorithm for discovering clusters in large spatial databases with noise. In Kdd (Vol. 96, No. 34, pp. 226-231).

[27] Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. In Advances in neural information processing systems (pp. 91-99).

[28] Redmon, J., & Farhadi, A. (2018). YOLOv3: An incremental improvement. arXiv preprint arXiv:1804.02767.

[29] Lin, T. Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. In Proceedings of the IEEE international conference on computer vision (pp. 2980-2988).

[30] Kiong, T. S., Salem, S. B., Paw, J. K. S., Sankar, K. P., & Darzi, S. (2014). Minimum variance distortionless response beamformer with enhanced nulling level control via dynamic mutated.

[31] monstar: picofamily https://pmdtec.com/picofamily/monstar

[32] 60GHz - Infineon Technologies AG https://www.infineon.com/cms/en/product/promopages/60GHz/