HODET: Hybrid Object DEtection and Tracking using mmWave Radar and Visual Sensors

Joseph St. Cyr[a], **Joshua Vanderpool**[b], **Yu Chen**[a,*], **Xiaohua Li**[a]

[a]Dept. of Electrical & Computer Engineering, Binghamton University, Binghamton, NY 13902
[b]The Raymond Corporation, Greene, NY 13778

Abstract. Image sensors have been explored heavily in automotive applications for collision avoidance and varying levels of autonomy. It requires a degree of brightness, therefore, the use of an image sensor in nighttime operation or dark conditions can be problematic along with challenging weather such as fog. Radar sensors have been employed to help cover the various environmental challenges with visible spectrum cameras. Edge computing technology has the potential to address a number of issues such as real-time processing requirements, off-loading of processing from congested servers, and size, weight, power, and cost (SWaP-C) constraints. This paper proposes a novel Hybrid Object DEtection and Tracking (HODET) using mmWave Radar and Visual Sensors at the edge. The HODET is a computing application of low SWaP-C electronics performing object detection, tracking and identification algorithms with the simultaneous use of image and radar sensors. While the machine vision camera alone could estimate the distance of an object, the radar sensor will provide an accurate distance and vector of movement. This additional data accuracy can be leveraged to further discriminate a detected object to protect against spoofing attacks. A real-world smart community public safety monitoring scenario is selected to verify the effectiveness of HODET, which detects, tracks objects of interests and identify suspicious activities. The experimental results demonstrate the feasibility of the approach.

Keywords: Hybrid Detection and Tracking, mmWave Radar, Visual Sensor, Convolutional Neural Network..

*Corresponding Author: Yu Chen, ychen@binghamton.edu

1 Introduction

The unprecedented pace of urbanization poses many opportunities and challenges.[1] The recent concept of Smart Cities has attracted the attention of the urban planners and researchers to enhance the security and well-being of the residents.[2,3] One of the most essential smart community services is the intelligent resident surveillance.[4,5] It enables a broad spectrum of promising applications, including access control in areas of interest, human identity or behavior recognition, detection of anomalous behaviors, interactive surveillance using multiple cameras and crowd flux statistics and congestion analysis and so on.[6]

There is a considerable amount of interesting research being conducted in the field of real-time object identification and tracking.[7,8] In order to identify and track an object in real-time a few requirements need to be defined. One obvious requirement is that there must be some sort of interesting object (or objects) that is worth identifying and tracking. Another is that a system must be defined that will track and identify such an object.[9,10] A typical component of one of these systems is an image sensor. The image sensor allows the system to "see" the object. Once the object can be seen, the system needs to be able to process the image it is "seeing" in order to identify, track and possibly make other decisions. A common term to describe this technology is machine vision.

Image sensors have been explored heavily in automotive applications for collision avoidance and varying levels of autonomy.[11] Complementary metal-oxide-semiconductor (CMOS) imagers

1

convert photons to a proportional voltage that is read by the sensor to digitise the scene. This process requires a degree of brightness therefore use of an image sensor in nighttime operation or dark conditions can be problematic along with challenging weather such as fog.[12] Radar sensors have also been employed to help cover the various environmental challenges with visible spectrum cameras.[13] The frequency range of the radar sensor is an important characteristic when selecting an appropriate sensor for an application. For instance, Earths atmosphere contains a considerable amount of water vapor which leads to a high level of frequency absorption in the 60 GHz range. Typically, to overcome this limitation a higher frequency, such as 77 GHz, is utilized.[14] An additional benefit of using a higher frequency is that it provides a finer range resolution.

An adjacent and popular topic of research is edge computing.[15,16] This technology has the potential to address a number of issues such as real-time processing requirements, off-loading of processing from congested servers, and size, weight, power and cost (SWaP-C) constraints[17–19] to name a few. Much of the edge computing phenomenon is being driven by the ever-growing Internet-of-Things (IoT).[20] It is expected that the IoT trend will continue to dominate the future of the Internet and our data processing paradigms for many years to come.[21]

An application of the previously mentioned image and radar sensor data is object detection and classification (e.g. provide detection of objects and count the number objects classified as a human that occupy, enter, or exit the field of view of the sensors). There are established methods and algorithms to facilitate the decision making a system must compute to detect and classify objects.[22] A few examples of the open source software libraries available are OpenCV,[23] TensorFlow,[24] Saliency,[25] and YOLO.[26] These libraries make use of several techniques to detect and classify an object (e.g point and edge detection, machine learning, neural networks). An effective method to quickly detect and classify an object is to deploy machine learning with the use of a Convolution Neural Network (CNN).[27]

This paper proposes a novel hybrid object detection and tracking (HODET) using two types of sensors, mmWave radar and visual sensor. The HODET is an edge computing application of low SWaP-C electronics performing object detection, tracking and identification algorithms with the simultaneous use of image and radar sensors. While the machine vision camera alone could estimate the distance of an object, the radar sensor will provide an accurate distance and vector of movement. This additional data accuracy can be leveraged to further discriminate a detected object to protect against spoofing attacks.

The rest of the paper is organized as follows: Section 2 provides a review of related work. Section 3 presents our proposal of the hybrid detection and tracking using radar and vision sensors and the technical approach. Experimental results are presented in Section 4. Finally, Section 5 wraps up this paper with the conclusions.

2 Motivation and Related Work

This section briefly discusses some of the related work that has been completed in closely related applications that inspired our proposed HODET system. Basically, the HODET is an automotive application that draws upon a collection of object detection, tracking and identification algorithms. It leverages the knowledge and insights under the umbrella of edge computing.

A sensor fusion based pedestrian collision detection system was proposed, which used a monocular CMOS camera as the image sensor along with a millimeter-wave radar sensor.[28] Data from the

radar sensor was used with a Kalman filter to provide detection and tracking of objects. The image sensor was used to detect whether or not a pedestrian crosswalk was present. Combining these two data streams allowed them to calculate the probability of a pedestrian collision and trigger a warning system based on a threshold.

Another reported solution with pedestrian detection as its primary goal uses similar methods and highlights their use of a model-based detection algorithm.[29] Meanwhile, in a research vehicles and and bicycles are added as primary objects to be detected in addition to pedestrian.[30] This solution uses a charge-coupled device (CCD) camera with a millimeter-wave radar. They pointed out that, "*millimeter-wave radar offers advantages of higher reliability in bad weather conditions*" in contrast to longer-range radars such as Light Detection and Ranging (LIDAR).

The motivation of using multiple different types of sensors (i.e. image and radar) was summarized as "[A] *camera provides high spatial resolution but low accuracy in estimation of the distance to an object. The high spatial resolution of the camera can support the low directional resolution of the radar, and the high distance resolution of the radar can support the low accuracy in distance estimation of the camera.*"[30] A deep convolutional neural network is proposed as a method of sensor fusion and object classification for autonomous vehicles.[31]

In contrast to the previously mentioned solutions, a system was proposed that use a red/green/blue (RGB) color camera data and two different types of radar sensors, millimeter-wave and LIDAR.[31] The data from these sensors are fused and fed into an AlexNet[32] to classify pedestrians, bicycles, cars and trucks. After a supervised training, the deep CNN method provided an efficient and accurate classification of their four primary objects of interest. Meanwhile, a color and thermal camera was used for image sensing in conjunction with millimeter-wave and LIDAR radar sensors to deploy a self-driving vehicle.[33] The thermal image sensor is used to "*perceive objects in challenging driving conditions, such as at night and in fog.*" They propose detecting and classifying objects using motion and observation models of their primary objects of interest (i.e. pedestrians, bicyclists, vehicles).

Another solution based on image and radar sensor fusion focused on pedestrian safety, which used sensor fusion at two levels (low and high) of their architecture.[34] A symmetrical deep convolutional neural network is used to detect changes in heterogeneous images taken at different times and dates by optical and radar sensors for the purpose of urban growth tracking, land use monitoring, and disaster evaluation.[35] Although it is not directly related to an automotive application like the other papers mentioned above, it is still applicable to our work. An interesting aspect of their work is the employment of unsupervised CNN training.

It would be remissive of us to not highlight one of the critical tasks we will need to address as part of our research: sensor calibration. Each of the sensors in our system will produce data in its own local coordinate system. To use the sensor data effectively we will need to calibrate the image and radar sensor data to exist in a common coordinate system. This is necessary to accurately detect, track and classify objects in the field of view (FOV). Sensor calibration is mentioned in all of the papers previously discussed[28–31, 33, 35] except one,[34] which does not explicitly mention calibration however it is assumed. In general, a transformation matrix is used to translate data points from one coordinate system to another, although some slightly different methods are presented for obtaining the transformation matrix. Primarily, the transformation matrix was obtained through some method of data point collection and correlation.
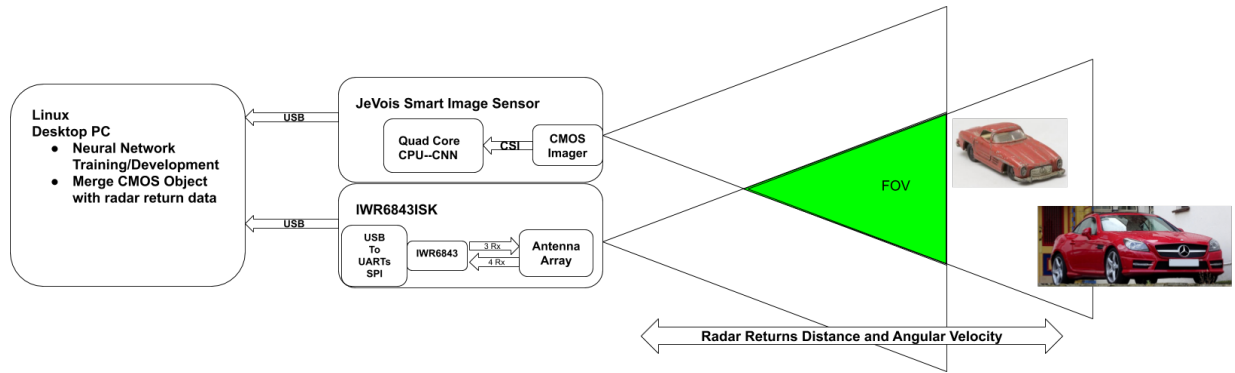
Fig 1 A System Level Illustration of HODET.

3 HODET: Hybrid Object DEtection and Tracking

3.1 System Architecture

Figure 1 shows the system architecture of the proposed HODET scheme, which is a system constructed of a Linux PC to collect and merge data streams from the CMOS *smart* image sensor and mmWave radar returns. The JeVois Smart Machine Vision Camera[36, 37] allows the use of the quad-core CPU to run one level of algorithms to identify objects. The radar sensor provides returns to help discriminate objects identified by the CMOS imager. The Radar could also be used in cases where it is not conducive for image sensors such as night time or in a climate weather. This system is built to explore the possibilities and would/could be merged into a more powerful SoC to run a centralized neural network at the edge. The diagram in Fig. 1 also illustrates the preliminary configuration where the JeVois and radar data will be collected, correlated, and processed by a development PC running the CNN algorithm.

Our technical approach will be incremental, building upon a baseline implementation. The JeVois Smart Machine Vision Camera has been chosen as our SWaP-C friendly image sensor and processor. JeVois is an open source machine vision camera developed by Professor Laurent Itti and his team at the University of Southern California (USC).[38] This camera has a substantial amount of user demo modules which can be used to familiarize ourselves with the product and establish a baseline implementation. There is also an established support community[36] which will be useful for debugging help. JeVois also has beta development tools and an extensive code base stored in GitHub[39] which can be leveraged for rapid development.

Once the baseline is established the radar sensor will be integrated into the solution. Texas Instruments TIDEP-01000 millimeter-wave radar sensor[40] has been selected. The radar sensor will provide accurate range, velocity, and angle information of the objects thus allowing further discrimination; Image sensors alone could be fooled with pictures or 3D models of objects. The JeVois will provide results via the USB webcam output and on the UART (serial port). This data will be merged with the radar sensor point cloud or object field given in three dimensions (i.e. X,Y,Z-axis coordinates). The radars field of view is 120° horizontal and 30° vertical. As previously mentioned the views will not be collinear nor matched in the FOV. Some transformations will be needed and accounted for using a calibration methodology.
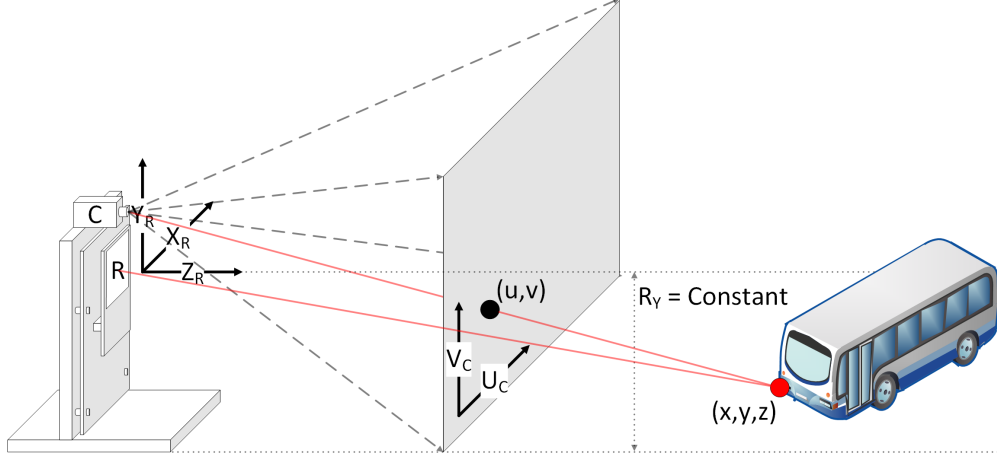
Fig 2 Evaluation Rig Calibration Diagram.

3.2 Calibration Method

From the research gathered we will leverage some of the techniques that were used to transform the two different coordinate systems to a unified field. This will be a cornerstone to the use of this data as an input into a CNN. Figure 2 illustrates the need for calibration. From the diagram you can see that the camera data exists in the U_C, V_C coordinate system while the radar data exists in the X_R, Y_R, Z_R coordinate system. There is a constant RY which is the constant distance the radar sensor is above the ground. The two data sets need to be merged together to accurately use the data for detection, tracking and identification. To do this we will follow a similar method as to[30] where an object or radar reflector is swept through the radar beam and camera view to collect and correlate coordinate points. Once the calibration data points are collected a transformation matrix, as seen in Eq. 1, can be used to convert to/from each coordinate system. Elements u and v are the camera coordinates while x, y, and z are radar coordinates. Elements P_{11} through P_{34} are the transformation matrix parameters that were collected as part of the calibration data collection. The data must also be synchronized in a deterministic manner to properly correlate the data sets.

$$\begin{pmatrix} u \\ v \\ 1 \end{pmatrix} = \begin{pmatrix} P_{11} & P_{12} & P_{13} & P_{14} \\ P_{21} & P_{22} & P_{23} & P_{24} \\ P_{31} & P_{32} & P_{33} & P_{34} \end{pmatrix} \begin{pmatrix} x \\ y \\ z \\ 1 \end{pmatrix} \tag{1}$$

4 Experimental Results

In order to validate the correctness and effectiveness of the HODET scheme, two sets of experimental studies have been conducted. On the one hand, a proof-of-concept prototype was built using a JeVois camera and a TIDEP-01000 mmWave radar sensor, which has been tested in a real-world scenario; on the other hand, the processing algorithm has been further tested using the Oxford Radar RobotCar Dataset.[41]

Fig 3 Evaluation Rig.

4.1 Study on the Proof-of-Concept Prototype

The initial test plan as follows:

- Configure JeVois Image Sensor

- Use a pre-learned CNN for object detection

- Configure the mmWave Sensor

- Mechanically fix the two sensors to prepare for calibration

- Merge data sets from objects gathered from the JeVois then cascade this in conjunction with the Radar data for another CNN.

The development hardware was placed on a 3D printed platform to enable the stable and consistent operation of the aperture of the CMOS imager and the radar sensor array. This is shown in Figure 3.

The initial test of the hardware setup provides a range in the radar field of view along with the JeVois sensor running a lightweight CNN. It is noted that the radar had a frame rate of 10 frames per second (FPS) while the inferred objects was about 2 FPS. This is not ideal and requires some optimizations that may or may not be possible with the base hardware in the provided development kits. Figure 4 shows the initial test output from the evaluation systems. Notice on the left hand side of the diagram the radar data spike denoting a detected object which can be correlated with the pedestrian on the lower right side of the diagram. Also note that the pedestrian and vehicle are detected by the JeVois camera denoted by the green boxes.

The performance of the hardware will be critical to meet our goal of a low power edge compute engine to track objects of interest with a frame rate as to not miss objects altogether. The image sensors specification are as follows:[36]
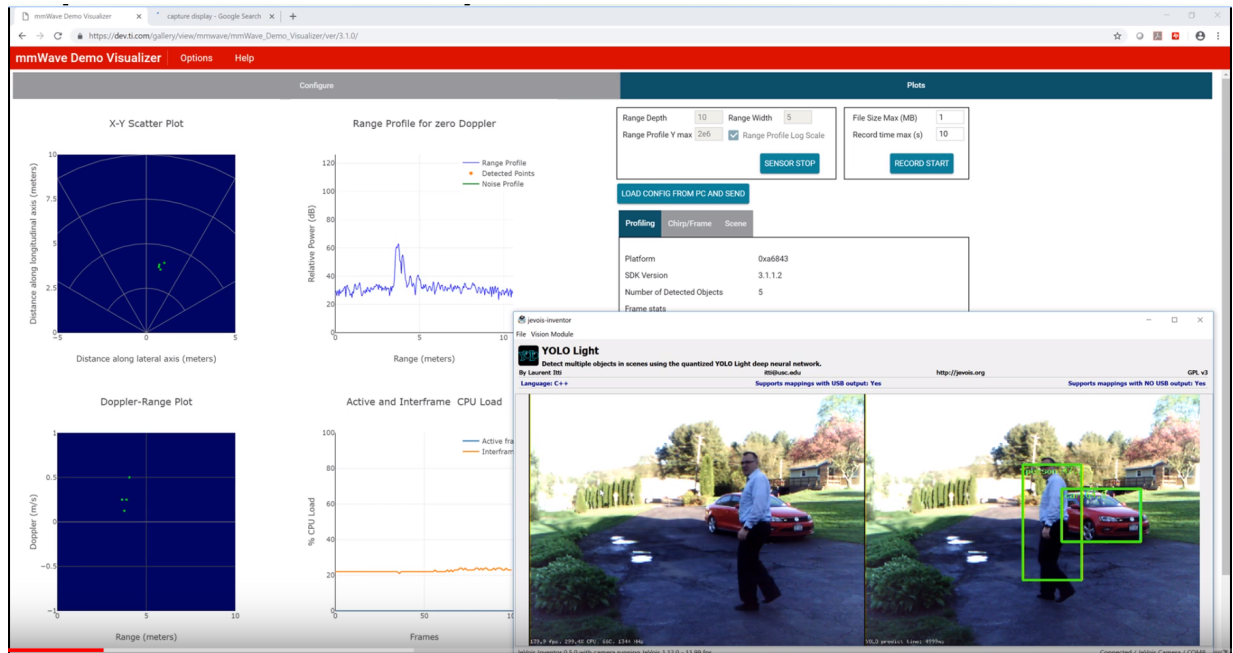
- 1.3MP camera capable of video capture at

6

Fig 4 Initial Test Output from the Evaluation Systems.

- **–** SXGA (1280 x 1024): up to 15 FPS (frames/second)
- **–** VGA (640 x 480): up to 30 FPS
- **–** CIF (352 x 288): up to 60 FPS
- **–** QVGA (320 x 240): up to 60 FPS
- **–** QCIF (176 x 144): up to 120 FPS
- **–** QQVGA (160 x 120): up to 60 FPS
- **–** QQCIF (88 x 72): up to 120 FPS
- **–** Rolling shutter, F2.8, 65° horizontal field of view

- Quad-core ARM Cortex A7 processor, default clock 1.35GHz. Supports hard floating-point operations (VFPv4) and NEON SIMD multimedia instructions.

- Dual-core MALI-400 GPU (graphics processing unit), supports OpenGL-ES 2.0

- 256MB DDR3-1600 SDRAM

The above specs allow some exploration and provide a low power platform for some image sensing applications however with initial testing the camera appears to limit the frame rate to 15 FPS. This does not allow much time for the CNN which is implemented in software rather than making use of the GPU or other specialized hardware found in some new Integrated Circuits (ICs) being produced for such applications in automotive or industrial environments. Nvidia is a leading developer of such ICs that provide plenty of GPU chains to allow the parallelization of the

	Jetson Nano™
GPU	NVIDIA Maxwell™ architecture with 128 NVIDIA CUDA® cores
CPU	Quad-core ARM® Cortex®-A57 MPCore processor
Memory	4 GB 64-bit LPDDR4
Storage	16 GB eMMC 5.1
Video Encode	4K @ 30 (H.264/H.265)
Video Decode	4K @ 60 (H.264/H.265)
Connectivity	Wi-Fi requires external chip
Camera	12 lanes (3x4 or 4x2) MIPI CSI-2, DPHY 1.1 (1.5 Gbps)
Size	69.6 mm x 45 mm
Mechanical	260-pin edge connector

Fig 5 Jetson Nano.

networks. Figure 5 shows comparison hardware which will be compared to the sensor we have chosen to begin development with.

The mmWave Radar sensor appears it will meet our needs when an antenna geometry is chosen to provide the ideal FOV to match the image sensor. This initial evaluation unit provides roughly 10 meters of unambiguous range however when a custom antenna is designed the range could reach upwards of 50 to 70 meters. The IWR6843 evaluation unit provides 3 transmit (TX) and 4 receive
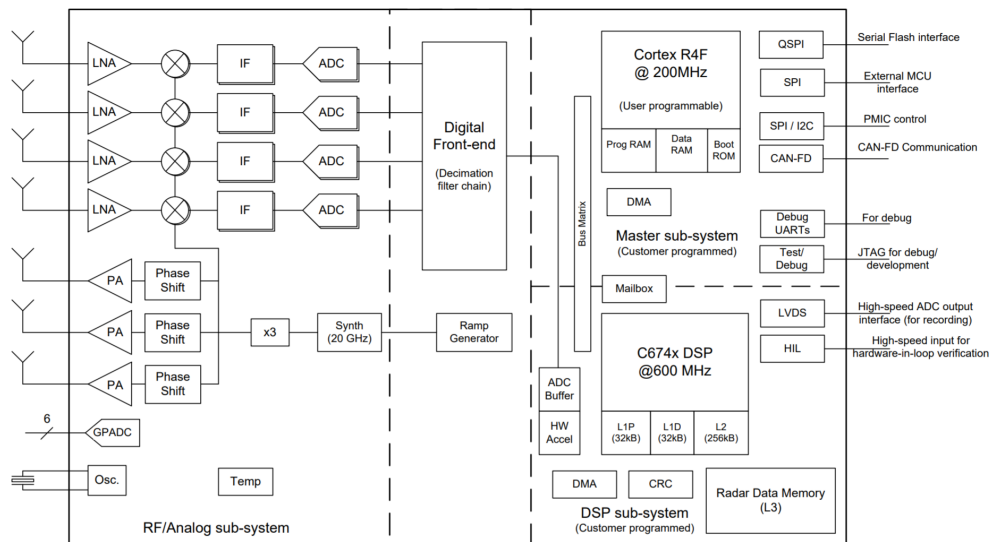

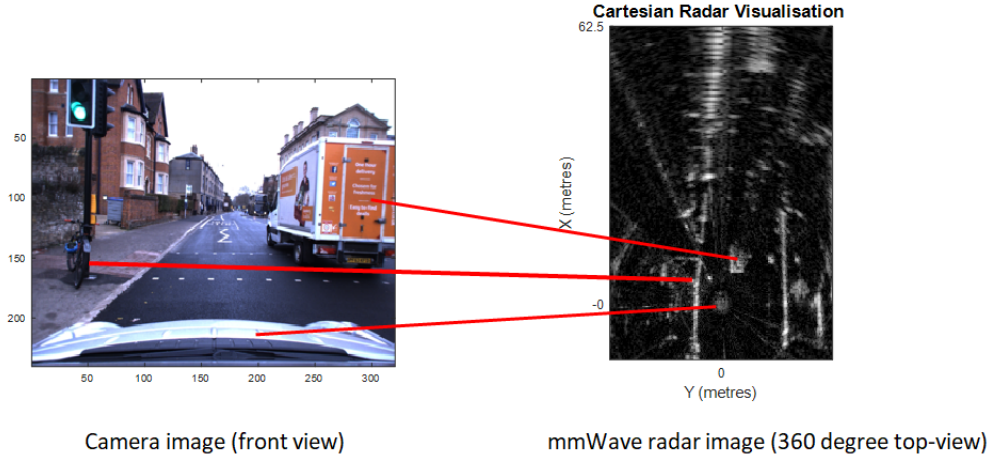**Fig 6** mmWave Sensor Block Diagram.

8

Fig 7 Sample images (a camera image and a mmWave radar image) of the Oxford Radar Robotcar dataset. Red lines indicate the same objects in these two images: truck, light pole, and the car (self).

(RX) channels to provide 108° azimuth and 44° of elevation. The fourth RX channel can be used to perform beam-forming techniques. Figure 6 provides the block diagram of the mmWave sensor.

4.2 Test using the Radar Robotcar Dataset

Figure 7 shows a sample set of images (a camera image and a mmWave radar image) of the Oxford Radar Robtcar Dataset. The dataset consists of several tens hours of video and mmWave radar images, which can be used to train our proposed object detection network. Based on the trained network, we can use transfer learning to adapt the network into our own data. This provides us with a quick way of proof of concept, and can also greatly reduce our workload of acquiring training data. To furnish this, the following tasks will be conducted.

1. We need to set up a calibration function to match the corresponding point positions between camera images and radar images. The original dataset does not include such a calibration example.

2. We need to do a heavy hand labeling work to label typical objects in camera images and radar images for training purpose. The original dataset unfortunately does not have labeling information.

3. We need to train CNN classifiers and YOLO object detection networks from scratch because none of them had been trained directly over mmWave radar images. Retrain of the object detection YOLO network will be a nontrivial task.

After training the object detection networks with the Oxford Radar Robotcar dataset, we will then apply transfer learning to fine tune these networks over our own data acquired by our HODET prototype. This is necessary because the mmWave radar images in the Oxford dataset are somewhat different from the our dataset. Specifically, the Oxford dataset gives 360° top-view mmWave images because it has a 360° revolving mmWave sensor. But it provides only azimuth and range information. In contrast, our mmWave sensor can provide zaimuth-elevation-range information but has a limited azimuth FoV with a fixed sensor.

9

5 Conclusions

There is an increasing demand for effective, efficient, and reliable surveillance solutions to maintain situational awareness (SAW) in many mission-critical delay-sensitive tasks, such as battlefield monitoring, smart public safety, disaster monitoring and recovery, etc. While the optical video surveillance system is the most popular approach, it is insufficient.

In this paper, we propose a novel Hybrid Object DEtection and Tracking (HODET) using mmWave Radar and Visual Sensors at the edge. Through the initial demo applications provided for the JeVois camera it appears a low resolution image will need to be used to reach acceptable frame rates, this can be considered a cost of speed. Initial use of the YoloLight network has an inference time of 500ms. The JeVois Single Board Computer does not have a direct input that could support the mmWave so this would be an improvement in a system moving forward to allow a CNN to run on one device rather than more ICs, more space and more power. Considering the small amount of data obtained using our own prototype, a larger scale data set from the Oxford Radar RobotCar Dataset has been identified, which allows us to exam the effectiveness and correctness of the core algorithms of our HODET scheme.

This work has just begun and appears it could benefit from studying some alternate hardware systems to optimize the solution. Ultimately, a desirable hybrid sensor that meets the demands of an edge computing device can provide accurate object tracking through adverse weather with the ability to delineate objects, which could try to fool (i.e. spoof) such a camera system. There are a lot of open questions to be addressed, including the performance, security, privacy, and robustness of such a powerful but complicated system. The authors hope this preliminary study will inspire more active discussions in the community.

References

1 N. Chen, Y. Chen, Y. You, *et al.*, "Dynamic urban surveillance video stream processing using fog computing," in *Multimedia Big Data (BigMM), 2016 IEEE Second International Conference on*, 105–112, IEEE (2016).

2 N. Chen and Y. Chen, "Smart city surveillance at the network edge in the era of iot: opportunities and challenges," in *Smart Cities*, 153–176, Springer (2018).

3 N. Chen, Y. Chen, X. Ye, *et al.*, "Smart city surveillance in fog computing," in *Advances in mobile cloud computing and big data in the 5G era*, 203–226, Springer (2017).

4 S. Y. Nikouei, Y. Chen, S. Song, *et al.*, "Smart surveillance as an edge network service: From harr-cascade, svm to a lightweight cnn," in *2018 IEEE 4th International Conference on Collaboration and Internet Computing (CIC)*, 256–265, IEEE (2018).

5 R. Wu, B. Liu, Y. Chen, *et al.*, "A container-based elastic cloud architecture for pseudo real-time exploitation of wide area motion imagery (wami) stream," *Journal of Signal Processing Systems* **88**(2), 219–231 (2017).

6 W. Hu, T. Tan, L. Wang, *et al.*, "A survey on visual surveillance of object motion and behaviors," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* **34**(3), 334–352 (2004).

7 R. Xu, S. Y. Nikouei, Y. Chen, *et al.*, "Blendmas: A blockchain-enabled decentralized microservices architecture for smart public safety," in *2019 IEEE International Conference on Blockchain (Blockchain)*, 564–571, IEEE (2019).

8 R. Xu, S. Y. Nikouei, Y. Chen, *et al.*, "Real-time human objects tracking for smart surveillance at the edge," in *2018 IEEE International Conference on Communications (ICC)*, 1–6, IEEE (2018).

9 S. Y. Nikouei, Y. Chen, S. Song, *et al.*, "Kerman: A hybrid lightweight tracking algorithm to enable smart surveillance as an edge service," in *2019 16th IEEE Annual Consumer Communications & Networking Conference (CCNC)*, 1–6, IEEE (2019).

10 R. Wu, Y. Chen, E. Blasch, *et al.*, "A container-based elastic cloud architecture for real-time full-motion video (fmv) target tracking," in *2014 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*, 1–8, IEEE (2014).

11 B. Hosticka, W. Brockherde, A. Bussmann, *et al.*, "Cmos imaging for automotive applications," *IEEE Transactions on Electron Devices* **50**(1), 173–183 (2003).

12 M. Bigas, E. Cabruja, J. Forest, *et al.*, "Review of cmos image sensors," *Microelectronics journal* **37**(5), 433–451 (2006).

13 U. Gessner, M. Machwitz, T. Esch, *et al.*, "Multi-sensor mapping of west african land cover using modis, asar and tandem-x/terrasar-x data," *Remote Sensing of Environment* **164**, 282–297 (2015).

14 M. Marcus and B. Pattan, "Millimeter wave propagation: spectrum management implications," *IEEE Microwave Magazine* **6**(2), 54–62 (2005).

15 E. Ahmed and M. H. Rehmani, "Mobile edge computing: opportunities, solutions, and challenges," *Future Generation Computer Systems* **70**(5), 59–63 (2017).

16 W. Shi, J. Cao, Q. Zhang, *et al.*, "Edge computing: Vision and challenges," *IEEE Internet of Things Journal* **3**(5), 637–646 (2016).

17 R. C. Allen, W. B. Blanton, E. Schramm, *et al.*, "Strategies for reducing swap-c and complexity in dve sensor systems," in *Degraded Environments: Sensing, Processing, and Display 2017*, **10197**, 101970M, International Society for Optics and Photonics (2017).

18 S. Y. Nikouei, R. Xu, Y. Chen, *et al.*, "Decentralized smart surveillance through microservices platform," in *Sensors and Systems for Space Applications XII*, **11017**, 110170K, International Society for Optics and Photonics (2019).

19 D. Nagothu, R. Xu, S. Y. Nikouei, *et al.*, "A microservice-enabled architecture for smart surveillance using blockchain technology," in *2018 IEEE International Smart Cities Conference (ISC2)*, 1–4, IEEE (2018).

20 M. Satyanarayanan, "The emergence of edge computing," *Computer* **50**(1), 30–39 (2017).

21 M. Wollschlaeger, T. Sauter, and J. Jasperneite, "The future of industrial communication: Automation networks in the era of the internet of things and industry 4.0," *IEEE industrial electronics magazine* **11**(1), 17–27 (2017).

22 S. Y. Nikouei, Y. Chen, A. Aved, *et al.*, "I-safe: Instant suspicious activity identification at the edge using fuzzy decision making," in *Proceedings of the 4th ACM/IEEE Symposium on Edge Computing*, 101–112 (2019).

23 G. Bradski and A. Kaehler, *Learning OpenCV: Computer vision with the OpenCV library*, " O'Reilly Media, Inc." (2008).

24 M. Abadi, P. Barham, J. Chen, *et al.*, "Tensorflow: A system for large-scale machine learning," in *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*, 265–283 (2016).

25 R. Zhao, W. Ouyang, H. Li, *et al.*, "Saliency detection by multi-context deep learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1265–1274 (2015).

26 J. Redmon, S. Divvala, R. Girshick, *et al.*, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 779–788 (2016).

27 S. Y. Nikouei, Y. Chen, S. Song, *et al.*, "Real-time human detection as an edge service enabled by a lightweight cnn," in *Edge Computing, the IEEE International Conference on*, (2018).

28 S. Suzuki, P. Raksincharoensak, I. Shimizu, *et al.*, "Sensor fusion-based pedestrian collision warning system with crosswalk detection," in *2010 IEEE Intelligent Vehicles Symposium*, 355–360, IEEE (2010).

29 S. Milch and M. Behrens, "Pedestrian detection with radar and computer vision," (2001).

30 S. Sugimoto, H. Tateda, H. Takahashi, *et al.*, "Obstacle detection using millimeter-wave radar and its visualization on image sequence," in *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, **3**, 342–345, IEEE (2004).

31 H. Gao, B. Cheng, J. Wang, *et al.*, "Object classification using cnn-based fusion of vision and lidar in autonomous vehicle environment," *IEEE Transactions on Industrial Informatics* **14**(9), 4224–4231 (2018).

32 M. Z. Alom, T. M. Taha, C. Yakopcic, *et al.*, "The history began from alexnet: A comprehensive survey on deep learning approaches," *arXiv preprint arXiv:1803.01164* (2018).

33 H. Cho, Y.-W. Seo, B. V. Kumar, *et al.*, "A multi-sensor fusion system for moving object detection and tracking in urban driving environments," in *2014 IEEE International Conference on Robotics and Automation (ICRA)*, 1836–1843, IEEE (2014).

34 M. Tons, R. Doerfler, M.-M. Meinecke, *et al.*, "Radar sensors and sensor platform used for pedestrian protection in the ec-funded project save-u," in *IEEE Intelligent Vehicles Symposium, 2004*, 813–818, IEEE (2004).

35 J. Liu, M. Gong, K. Qin, *et al.*, "A deep convolutional coupling network for change detection based on heterogeneous optical and radar images," *IEEE transactions on neural networks and learning systems* **29**(3), 545–559 (2016).

36 JeVois, "Jevois smart machine vision camera," *http://www.jevois.org/* (2020).

37 JeVois, "Jevois smart machine vision," *https://www.jevoisinc.com/* (2020).

38 JeVois, "Jevois smart embedded machine vision toolkit," *http://www.jevois.org/doc/* (2020).

39 JeVois, "Jevois," *https://github.com/jevois* (2020).

40 T. Instruments, "People counting and tracking reference design using mmwave radar sensor," *http://www.ti.com/tool/TIDEP-01000* (2018).

41 D. Barnes, M. Gadd, P. Murcutt, *et al.*, "The oxford radar robotcar dataset: A radar extension to the oxford robotcar dataset," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, (Paris) (2020).