Non-cooperative Personnel Tracking with Cross Modality Learning in 5G-enabled Warehouse Application

Yang Zhao
Software and Analytics Lab
GE Research
Schenectady, NY, USA

Gangliang Zhao
Optimization Lab
GE Research
Schenectady, NY, USA

Prabhu Janakaraj
Embedded Computing Lab
GE Research
Schenectady, NY, USA

Lynn Derose
Software and Analytics Lab
GE Research
Schenectady, NY, USA

Austars Schnore
Controls and Optimization
GE Research
Schenectady, NY, USA

Hasan SM
Embedded Computing Lab
GE Research
Schenectady, NY, USA

***Abstract*—Asset and personnel visibility is crucial for improving workflow efficiency and reducing waste in smart facility, e.g., warehouse applications. 5G networks and technologies provide the high bandwidth and low latency necessary for communicating and fusing multi-modality sensor data, such as high-definition video, time series with high temporal resolution. In this work, we propose to use cross modality learning to develop a self-learning system for locating and tracking indoor personnel with video and WiFi channel state information (CSI) data. We use video data and our computer vision system to provide location annotation automatically, and train a feedforward neural network model for WiFi CSI data in our localization algorithm. Our experimental results show that our localization system is capable of locating a person with submeter accuracy in real-time without laborious manual data annotation.**

***Index Terms*—Ambient Intelligence, Context Awareness, Cyber-physical Systems, Machine Learning, Internet of Things**

I. INTRODUCTION

5G networks provide the speed and capability that enable a multitude of technologies in digital transformation of industrial asset and operations applications. For an industrial warehouse application, asset and personnel visibility is the key to warehouse inventory management. 5G networks and technologies provide the high bandwidth and low latency needed to acquire and analyze real-time multi-modality sensor data for improving workflow efficiency and reducing waste and cost. In this work, we propose to leverage the 5G testbed at GE Research to develop a personnel tracking system with multi-modal sensors and cross modality learning for our warehouse application [1].

The location of the warehouse personnel is one of the most important context information for staff safety, security, and asset and operations management purposes. 5G networks and technologies provide the bandwidth and capability to collect data from multi-modal sensors to locate and track personnel in real-time. Various techniques have been proposed and developed to achieve accurate and robust indoor people localization solutions. For example, computer vision (CV)-based people tracking system has become one of the most accurate localization systems in good light condition, due to the recent development of deep learning techniques. However, a CV-based system suffers from the occlusion problem and is sensitive to light condition of an environment. On the other hand, a radio frequency (RF)-based system does not suffer from the occlusion issue, but the performance is significantly affected by the multi-path effect of an indoor environment [2]. For our smart warehouse applications, we have investigated both CV-based localization system and various RF-based systems, which include Bluetooth, ultra wide-band (UWB), mmWave Radar and WiFi channel state information (CSI) systems. In this paper, we show our research effort and progress in developing a hybrid CV and WiFi CSI-based personnel localization system.

Since CV and RF sensing modalities are complimentary in human sensing, recent studies have shown significant performance improvement in object detection, human identification and localization by fusing these two sensing modalities [3]–[5]. While these state-of-the-art systems focus on different sensing aspects, e.g., object detection in [3], localization in [4], and identification in [5], they have used one common approach: cross modality learning. That is, they use one sensing modality in its favorite work condition, e.g., CV system in good light condition, to train a deep neural networks model for the other modality. The major difference between this work and the state-of-the-art vision-WiFi localization work in [4] is that we do not rely on the people to carry any WiFi devices, instead, we use our CV system to train our WiFi CSI system to locate people, who do not cooperate with the localization system by carrying any devices. That is, we apply the cross modality learning approach to passive non-cooperative people

Corresponding author: zhao.yang@ieee.org

978-1-6654-4308-1/21/$31.00 ©2021 IEEE
DOI 10.1109/5GWF52925.2021.00041

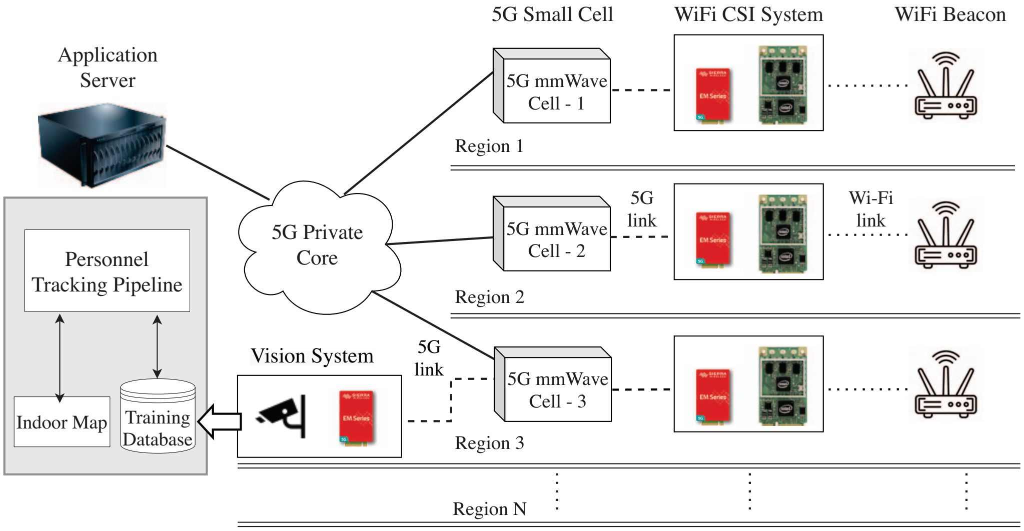

Fig. 1: 5G-enabled personnel tracking system architecture.

localization and tracking [6]–[8].

To achieve WiFi CSI-based device-free localization with cross modality learning, we first integrate and synchronize our CV tracking system with a WiFi CSI system. Then, we train a multi-layer perceptron (MLP) neural network model for the CSI-based system, in which the data annotation, i.e., location ground truth labeling can be automatically provided by the CV system in good light condition. When the light condition is poor or for regions without cameras, the tracking system will switch to the CSI-only online mode using the trained MLP model. The overall personnel tracking system architecture with 5G connectivity is shown in Figure 1. Note that recent studies propose data fusion method to achieve centimeter level accuracy for joint communication and device-based localization using the same 5G infrastructure [9]. For device-free wireless sensing, recent study in [10] proposes to use 5G devices as the sensing modality for human occupancy and activity detection. In this work, we focus on device-free localization, and we aim to leverage existing WiFi network infrastructure as sensing purpose, while using 5G testbed for communication purpose.

Finally, we perform real-world experiments to evaluate our method and system in a typical office environment. We investigate the performance of MLP neural network with different hyperparameters. The experimental results show that our localization system is capable of tracking a person in real-time with less than 0.5m root mean squared error (RMSE) for over 40 tests. The rest of this paper is organized as follows. Systems and methods are discussed in Section II. Experiments and results are presented in Section III. Section IV concludes the paper.

II. Systems and Methods

In this section, we describe our 5G testbed, vision and RF hardware. We also discuss how we use cross modality learning to develop a vision-WiFi-based personnel localization system.

A. Systems

First, we describe the 5G testbed system, the device-free people localization system that we have developed, and the commodity systems that we have deployed and used for the ambient intelligence, e.g., smart warehouse applications. The in-house localization systems include a computer vision-based system, and two WiFi CSI-based systems. The commodity systems include a mmWave Radar system, a UWB system, and a motion capturing system to provide 3D positioning with millimeter accuracy.

1) 5G System for Smart Warehouse: Modern warehouse is highly dependent on timely collaboration among multiple sensing devices for efficient operations. One of the key enabler of smart warehouse is the availability of data from multi-modal sensor for planning and execution. In this work, we propose a multi-modal sensing solution using commodity WiFi and vision sensors for tracking personnel location within a warehouse. The CSI data from WiFi sensor and the video data from camera sensor should be communicated in real-time to achieve continuous tracking. 5G mmWave based communication backbone is a perfect candidate solution that enables us to transfer data in near real-time using wireless medium and in a fraction of cost avoiding expensive wired infrastructure. As shown in the system architecture in Figure 1, we envision our sensors are equipped with 5G mmWave communication modems. We also consider 5G small cell based deployment scenario due to mmWave's limited coverage area. The small cell base stations are then connected to the 5G private core, where we can host an application server, which performs personnel tracking that we discuss in details next. As 5G communication devices become more widely deployed, we believe 5G technologies and devices can be used for both human sensing and data communication purposes. We leave this as future research topic.

2) WiFi CSI-based tracking system: As the IEEE 802.11n standard and WiFi MIMO devices are widely used nowadays, WiFi channel state information (CSI)-based localization becomes a promising solution using low-cost commodity hardware. We have investigated two commodity WiFi hardware devices for developing our WiFi CSI-based people tracking systems: Intel 5300 network interface card (NIC) with open-source IEEE 802.11n toolkit [11], and regular 802.11n WiFi devices with Nexmon CSI monitor [12].

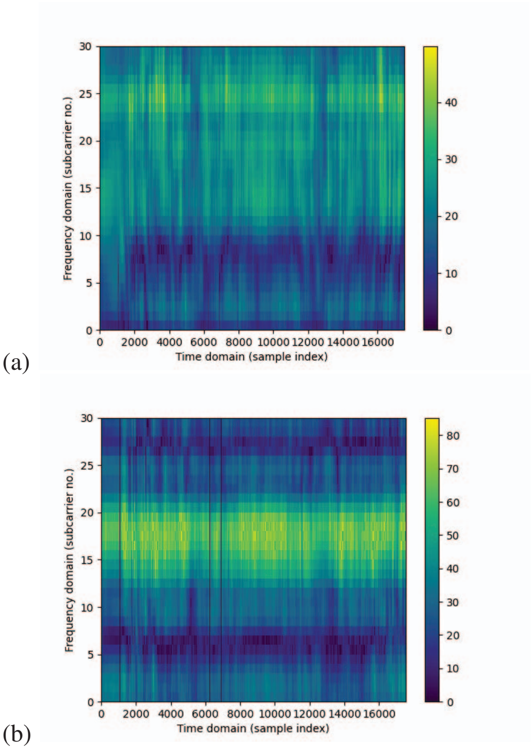

Fig. 2: Amplitude data of CSI measurements from two MIMO antenna pairs (a) Tx1-Rx1, (b) Tx3-Rx3.

For the Intel 5300 system, we collect CSI measurements between a WiFi beacon, i.e., WiFi router with three antennas, and an Intel 5300 NIC also with three antennas. The Intel 5300 NIC reports CSI from 3x3 multiple-input multiple-output (MIMO) antennas on 30 OFDM subcarriers with a total bandwidth of 20 MHz [11]. The amplitude of the CSI data from two MIMO pairs recorded in the same experiment are shown in Figure 2, from which we can clearly see the antenna diversity of the CSI data. For the Nexmon CSI system, we use a Raspberry Pi board with the Nexmon toolkit as the monitor to collect the CSI data between a WiFi access point and its connected IEEE 802.11n devices. We can obtain CSI data from 56 OFDM subcarrier, on which we investigate the indoor multi-path effect as in SectionIII-B. However, we cannot control its packet data rate as the Intel 5300 system. Thus, we choose to use the Intel CSI system to test the idea of cross modality learning in personnel localization. However, the general approach is applicable to the Nexmon system, which we leave as future work.

3) Vision-based tracking system: Due to the recent development of computer vision and deep learning techniques, the vision-based people tracking system has become one of the leading systems in the multiple object tracking scenario. We use the Intel RealSense RGB-D camera as our system hardware, and we develop a people tracking software pipeline following the tracking-by-detection paradigm [13].

Our tracking pipeline includes three major components: object detection, position estimation, and multi-target tracking. The person detection task is achieved by applying the state-of-the-art YOLOv5 object detection algorithm to the RGB images captured by the camera. YOLOv5 is the latest you-only-look-once (YOLO) object detection architecture and model [14], which is capable of generating object detection results, i.e., the bounding box of the person in real-time, as shown in Figure 3. Once the bounding boxes of the detected people are generated, the perspective-n-point algorithm is used to project the 2D coordinates of the bounding boxes to the 3D world coordinate, using the intrinsic and extrinsic parameters of a calibrated camera. Finally, for the multiple targets tracking scenario, the DeepSort algorithm [13] incorporates both the object motion and appearance information for solving the object assignment problem, which achieves better performance than the classical Hungarian algorithm. Note that for the single person tracking scenario, the tracking algorithm can smooth the trajectory, and the previous study shows that our CV system can accurately track the 2D location of a person in good light condition in real-time [15].

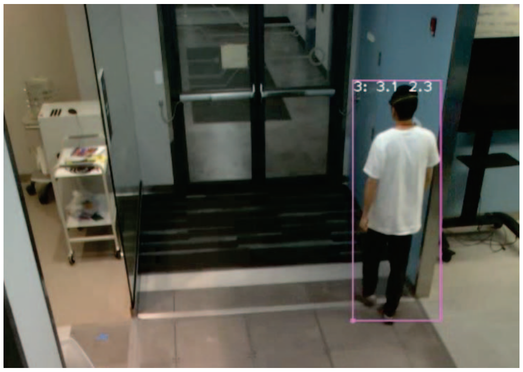

Fig. 3: Vision-based tracking result.

4) COTS tracking systems: In addition to the WiFi CSI and computer vision-based people tracking systems, we have deployed and evaluated other tracking systems with commercial-off-the-shelf (COTS) hardware as part of our testbed. The COTS tracking systems include the mmWave Radar system with IWR6843 sensor from TI, the ultra wide-band (UWB) system from Pozyx, and the motion capturing system from Vicon. We are also deploying and evaluating the Bluetooth tracking system from LocatorX with our 5G testbed system.

The cross modality learning approach can be applied to combinations of the in-house tracking systems and COTS systems as well. We leave it as future work as discussed in Section III-C.

B. *Cross Modality Learning*

Now we discuss how we apply cross modality learning to bootstrap two complimentary sensing systems, computer vision (CV) tracking system and WiFi CSI system.

First, we integrate and synchronize the CV and CSI systems. We use the CV system for people detection, and only record camera image data and WiFi CSI data when at least a person is detected. Once human presence is detected, we use time stamps to synchronize data collected on two systems. In good light condition, we use the location estimates from the CV system as the training data for the CSI system. When the light is turned off, we stop using the CV data for ground truth labeling purpose.

Once we have the CV and CSI data synchronized, we use the label from the CV system to train a feedforward MLP neural network model for the CSI system. Specifically, we build a MLP neural network with ReLU activation function, Adam optimizer, L2 regularization, and a constant learning rate. We have tested using different numbers of hidden layers, and also tested the effects of different maximum iterations used in the Adam optimizer, as shown in Section III-B. Note that we only use the amplitudes of the CSI measurements on all subcarriers and MIMO pairs as the inputs of the CSI localization system, since the CSI data show that the phase information of CSI is sensitive to even subtle human motion.

After the MLP model is trained with initial label from the CV system, the WiFi CSI system works even in poor light condition, or when the light is turned off. Once the light condition change is detected, the system can switch between online testing phase and offline training phase. As the offline training can be performed whenever the ambient light condition is sufficient, the CSI system can be calibrated automatically and continuously as the environment multi-path effect changes. Thus, the joint CV-CSI system become a continuous and self-learning system, which is more robust to light condition change and multi-path effect change.

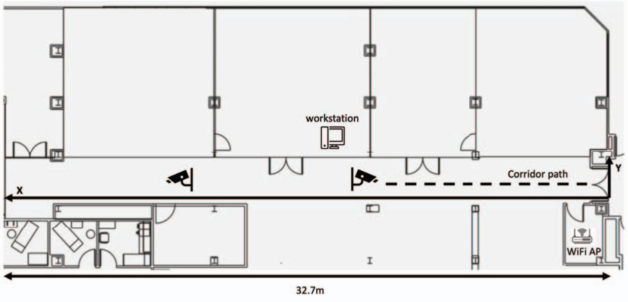

Fig. 4: Layout of experiments.

III. Experiments and Results

A. *Experiments*

We deploy our CV and CSI systems and perform experiments to track a person in the corridor of our lab space, as shown in the experiment layout in Figure 4. A WiFi beacon and a workstation with Intel 5300 NIC are deployed at two diagonal corners, and two RealSense cameras are deployed on the ceiling to monitor the pedestrian flow. During our experiments, we ask two volunteers to walk along predefined paths in the corridor space. Each volunteer performs experiments individually for 20 times, and we collect a dataset with 40 trials from our human subjects. From Figure 4 we can see that there are walls between our WiFi devices, and it is a multi-path rich environment as we verify on the CSI data next.

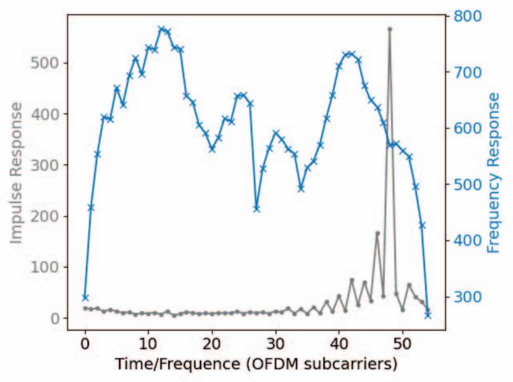

Fig. 5: Impulse and frequency responses showing multi-path effect from CSI data on 56 sub-carriers.

B. *Results*

As mentioned in Section I, the multi-path effect is one of the major challenges for RF-based indoor localization systems. Before we apply neural network models on the dataset, we first investigate the multi-path effect on the CSI data from the OFDM subcarriers of the IEEE 802.11n channel [11]. Here we use the CSI data collected from the Nexmon system. Since the frequency response of the wireless channel can be represented by the CSI values on the OFDM subcarriers [2], the time-domain impulse response can be computed by performing inverse Fourier transform on the CSI data. As shown in Figure 5, there are multiple peaks in the computed impulse response, which represent received radio signals from both line-of-sight (LOS) path and non-LOS path.

TABLE I: RMSE vs. test sample size.

Test size	10%	20%	30%	40%	50%
Max iteration=100	0.50	0.52	1.56	0.82	0.65
Max iteration=400	0.47	0.49	1.53	0.59	0.88
Max iteration=800	0.48	0.52	0.87	0.62	0.99

Now we apply the MLP model on the CSI data that we have collected to evaluate the system performance. We have

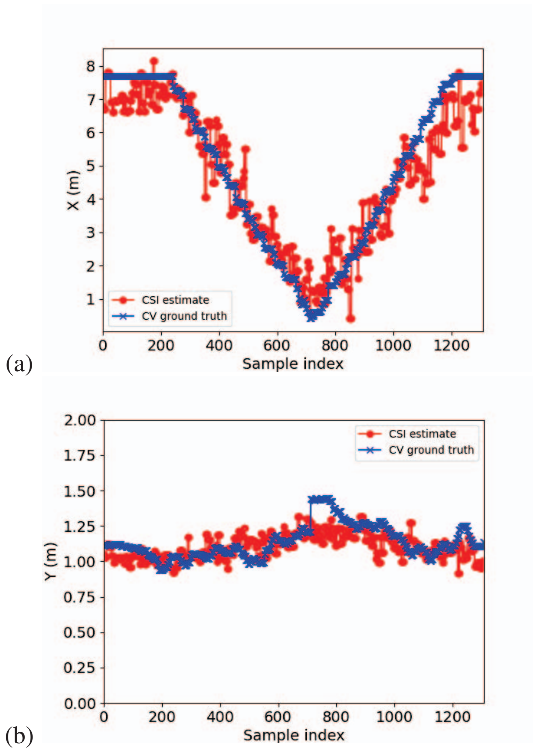

Fig. 6: Localization results of the 40th test trial: (a) X coordinate estimates, (b) Y coordinate estimates.

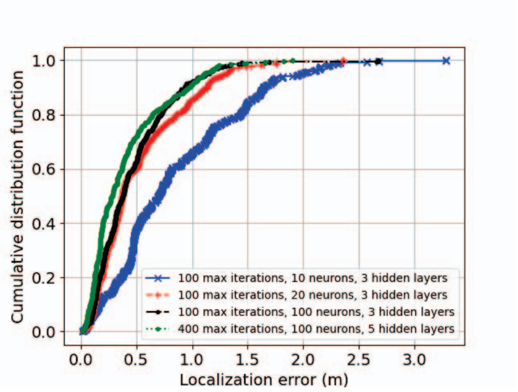

Fig. 7: CDF of localization errors from different hyperparameters.

performed leave-one-out (LOO) evaluation and test-train split cross validation. In our LOO evaluation, for each of the 40 trials, we use data from 39 trials to train the MLP model, and apply the model on that test case. Then, we calculate the root mean squared error (RMSE) and the cumulative distribution function (CDF) of the position error, for each of the 40 trials. For example, for the 40th trial, the estimated XY coordinates are shown in Figure 6 together with the location ground truth provided by our computer vision system. In addition, the CDF of the localization error is shown in Figure 7, with different hyperparameters, e.g., number of neurons and number of hidden layers in the neural network, and maximum iteration in the Adam optimizer. We see that the localization performance improves as the numbers of neurons and hidden layers increase. The median error is 0.7 m when we only use ten neurons, three hidden layers in the MLP neural network, and 100 maximum iterations in the optimizer; while the median error is less than 0.5 m when we have 20 neurons in the MLP network. A MLP with 100 neurons, 5 hidden layers and 400 maximum iterations achieves the best accuracy performance as shown in Figure 7. We also find that there is diminishing return in localization accuracy when the maximum iteration number is above 400.

We also have performed test-train split cross validation. For different test sample sizes from 10% to 50%, the root mean squared errors (RMSE) are listed in TableI. We see that as the test sample increases, the RMSE tends to increase, since there are fewer data samples used in training the MLP model. Finally, we also calculate the RMSEs for all 40 trials, and the average RMSE is 0.4 m. That is, our CSI system achieves submeter accuracy at this 32.7 m by 2.4 m corridor environment.

C. *Future Work*

We have deployed and evaluated our tracking system at only one indoor environment. We plan to deploy the system to other locations, perform experiments, collect more dataset and compare performance with other state-of-the-art device-free localization systems in the future. As mentioned earlier, we also plan to deploy and compare the Intel 5300 CSI system with the Nexmon CSI system, which use different system architectures. In addition, we focus on single person case in this work, but the cross modality learning and MLP neural networks model also apply to the multiple people tracking scenario. Finally, the cross modality learning approach is applicable to other sensing modality combination. For example, mmWave radar provides finer temporal resolution and can be integrated with CV system for localization and other human sensing applications as well.

IV. Conclusion

We have developed a multi-modal personnel localization and tracking system using vision and WiFi CSI data for our 5G smart warehouse application. We have applied the cross modal learning approach to device-free localization to bootstrap the individual vision and WiFi systems. We have performed real-world experiments in a typical office environment, and our experimental results show that we can achieve submeter tracking accuracy without labor-intensive manual annotation in neural network model training.

Acknowledgment

This work is supported by the DARPA Re-ID program under contract number N6600118C4006. We thank Jasmine Ratchford and Jim Simpson from Cynnovative for their discussions. We also thank our anonymous reviewers for their constructive comments and suggestions.

References

[1] GE News, https://www.ge.com/research/newsroom/ge-research-and-us-department-defense-partner-5g-initiative-create-smart-warehouses, 2021.

[2] D. Tse and P. Viswanath, *Fundamentals of wireless communication*. Cambridge university press, 2005.

[3] X. Shuai, Y. Shen, Y. Tang, S. Shi, L. Ji, and G. Xing, “millieye: A lightweight mmwave radar and camera fusion system for robust object detection,” in *Proceedings of the International Conference on Internet-of-Things Design and Implementation*, 2021, pp. 145–157.

[4] S. Fang, T. Islam, S. Munir, and S. Nirjon, “Eyefi: Fast human identification through vision and wifi-based trajectory matching,” in *2020 16th International Conference on Distributed Computing in Sensor Systems (DCOSS)*. IEEE, 2020, pp. 59–68.

[5] L. Fan, T. Li, R. Fang, R. Hristov, Y. Yuan, and D. Katabi, “Learning longterm representations for person re-identification using radio signals,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10 699–10 709.

[6] Y. Zhao, M.-C. Chang, and P. Tu, “Deep intelligent network for device-free people tracking: Wip abstract,” in *Proceedings of the 10th ACM/IEEE International Conference on Cyber-Physical Systems*, 2019, pp. 302–303.

[7] N. Patwari and J. Wilson, “Rf sensor networks for device-free localization: Measurements, models, and algorithms,” *Proceedings of the IEEE*, vol. 98, no. 11, pp. 1961–1973, 2010.

[8] R. M. Buehrer, C. R. Anderson, R. K. Martin, N. Patwari, and M. G. Rabbat, “Introduction to the special issue on non-cooperative localization networks,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, no. 1, pp. 2–4, 2013.

[9] O. Kanhere and T. S. Rappaport, “Position location for futuristic cellular communications: 5g and beyond,” *IEEE Communications Magazine*, vol. 59, no. 1, pp. 70–75, 2021.

[10] A. M. Ashleibta, A. Taha, M. A. Khan, W. Taylor, A. Tahir, A. Zoha, Q. H. Abbasi, and M. A. Imran, “5g-enabled contactless multi-user presence and activity detection for independent assisted living,” *Scientific Reports*, vol. 11, no. 1, pp. 1–15, 2021.

[11] D. Halperin, W. Hu, A. Sheth, and D. Wetherall, “Tool release: Gathering 802.11 n traces with channel state information,” *ACM SIGCOMM Computer Communication Review*, vol. 41, no. 1, pp. 53–53, 2011.

[12] F. Gringoli, M. Schulz, J. Link, and M. Hollick, “Free your csi: A channel state information extraction platform for modern wi-fi chipsets,” in *Proceedings of the 13th International Workshop on Wireless Network Testbeds, Experimental Evaluation & Characterization*, 2019, pp. 21–28.

[13] N. Wojke, A. Bewley, and D. Paulus, “Simple online and realtime tracking with a deep association metric,” in *2017 IEEE international conference on image processing (ICIP)*. IEEE, 2017, pp. 3645–3649.

[14] J. Redmon and A. Farhadi, “Yolov3: An incremental improvement,” *arXiv preprint arXiv:1804.02767*, 2018.

[15] J. Chen, M. Chang, T. Tian, T. Yu, and P. Tu, “Bridging computer vision and social science: A multi-camera vision system for social interaction training analysis,” in *2015 IEEE International Conference on Image Processing (ICIP)*, Sep. 2015, pp. 823–826.