People Counting Solution Using an FMCW Radar with Knowledge Distillation From Camera Data

Conference Paper · November 2021

DOI: 10.1109/SENSORS47087.2021.9639798

5 authors, including:

Michael Stephan
Friedrich-Alexander-University of Erlangen-Nürnberg
8 PUBLICATIONS **35** CITATIONS
SEE PROFILE

Souvik Hazra
Infineon Technologies
14 PUBLICATIONS **153** CITATIONS
SEE PROFILE

Avik Santra
Infineon Technologies
71 PUBLICATIONS **484** CITATIONS
SEE PROFILE

People Counting Solution Using an FMCW Radar with Knowledge Distillation From Camera Data

Michael Stephan[12*], Souvik Hazra[1*], Avik Santra[1], Robert Weigel[2], Georg Fischer[2]
[1]*Infineon Technologies AG*, Neubiberg, Germany
[2]Friedrich-Alexander-University Erlangen-Nuremberg, Erlangen, Germany
E-mail: {souvik.hazra, avik.santra}@infineon.com
{michael.stephan, georg.fischer, robert.weigel}@fau.de
*equal contribution

***Abstract*—Radar systems enable remote sensing of multiple persons within their field of view. In this paper, we propose a novel architecture to perform people counting using a 60 GHz Frequency Modulated Continuous Wave radar trained on supervised radar data and knowledge distillation performed using synchronized camera data. In the evaluation phase, only the radar encoder with Range - Doppler Images (RDI) as input is used and tested on a dataset consisting of scenarios recorded in a different setup than the training recordings with up to 6 persons present. In this paper we focus on showing the benefit of using the cross-modal camera information compared to the same unimodal model. In spite of the low-cost radar sensor, the proposed architecture achieves an accuracy of 71% compared to 58% for the test data from a different sensor with a different orientation and aspect angle, and an accuracy of 89% compared to 74% for test data from the same radar sensor when training without knowledge distillation.**

***Index Terms*—Cross Learning, People Counting, FMCW Radar, Deep Learning.**

I. Introduction

Automatic People counting solutions have several industrial and consumer applications. Automatically counting the number of people in a room or a floor in buildings, office spaces, and smart homes can help regulate the power consumption of lighting, heating, and air conditioning (HVAC) systems [1], saving energy and also offering a low carbon footprint solution. In public places, with limits imposed due to government regulations during a pandemic, or system constraints such as in an elevator, camera-based people counting offers high reliability in counting people in both less dense and densely crowed environments [2] [3] [4] [5] [6] [7]. However, due to the privacy intruding aspect and a dependency on illuminating conditions, cameras are not a preferred modality in indoor deployment. Radars offer a promising choice for people counting solutions, and there are several such systems proposed in literature [8] [9] [10] [11] [12] [13] [14]. However, there are numerous challenges for a robust and accurate people counting solution using radars due to large variability in radar data for a particular people count, measurement noise, along with artifacts due to occlusions, multi-path reflections, and ghost targets.

Such challenges and limitations can be addressed using cross-modal learning from a paired superior network or sensor. The superiority of the cross-modal solution could be due to higher resolution sensor data, a deeper complex network, a larger training dataset, or better feature images for a specific task. Several papers in the literature propose various cross-modal learning frameworks for respective tasks to improve the performance of a target neural network. In [15] a simple form of distillation is proposed whereby knowledge from a source model is transferred to the target model by training it on a transfer set while training via soft target distribution using a temperature variable. In [16], a general cross-learning technique is proposed wherein learned representations from a large labeled RGB image dataset are used as supervised information for unlabeled but paired depth and optical flow images. The framework demonstrates significant performance improvements for both depth and optical flow cross-modal supervision transfers. In [17], authors propose a teacher-student setting for cross-modal learning, whereby supervised learning from camera DCNN is used to train a WiFi-based DCNN so that eventually, the WiFi-based DCNN can learn tasks that are seemingly intractable for WiFi DCNNs, such as human pose estimation during inference. A correlation-based similarity metric is proposed in [18], wherein supervised information is passed from one to another modality for gesture sensing. In [19], a multi-modal cross-learning framework for people counting, using frequency modulated continuous wave (FMCW) radars, is proposed, wherein additional supervised information is passed from camera heatmaps to train a radar DCNN, which demonstrates superior performance compared to unimodal learning. In [20], the authors propose cross-modal learning whereby a network trained for action recognition on RGB videos was adapted to recognize actions from 3D pose data using cross-entropy loss knowledge distillation along with mutual supervised training on a small dataset. In [21], authors have proposed and deployed cross-modal learning using a teacher-student methodology for tasks such as emotion recognition from human speech data, where obtaining labeled data is a difficult challenge. The work proposes distilling knowledge from the visual domain to the speech domain through cross-modal distillation and demonstrates good learned representation learning. This paper proposes a people counting solution based on a novel multi-modal cross-learning framework using FMCW radar. The training process involves learning from supervised radar data while enforcing the pre-final layer of the network to learn similar embedding semantics as that of another DCNN network trained with triplet loss on corresponding camera data. The significant difference in the performance of the proposed network over a DCNN trained on

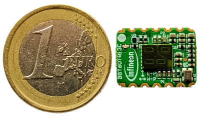


Fig. 1. *Infineon*'s *BGT60TR13C* 60-GHz radar sensor.

supervised radar data only is demonstrated in indoor scenarios with up to 6 people present.

II. FMCW Radar System Design

The proposed solution uses Infineon's BGT60TR13C FMCW radar chipset displayed in Fig. 1. It operates in the most common mode of having a sequence of frequency chirps with short ramp-times and delays between chirps and a sequence of chirps for saving power and data pre-processing respectively. The Analog to Digital Converter (ADC) digitizes the data in 12 bit which is then sent to the PC from the evaluation board via USB for further processing. The chipset configuration and derived parameters are presented in Tab 1.

TABLE I
Operating Parameters.

Parameters, Symbol	Value
Ramp start frequency, f_{min}	60.5 GHz
Ramp stop frequency , f_{max}	61.5 GHz
Number of samples per chirp, NTS	128
Sampling frequency, fs	2 MHz
Chirp time, T_c	64 μs
Number of chirps, PN	64
Number of Tx antennas, N_{Tx}	1
Number of Rx antennas, N_{Rx}	3

III. Data Preparation Processing

A. Radar Data Processing

A 2D matrix of shape PN x NTS is formed by acquiring an intermediate frequency signal from a chirp with NTS samples across consecutive PN chirps for each Rx Channel. Next, the Range Doppler Image (RDI) is computed for each channel by taking a 1D Fast Fourier Transform (FFT) along the fast time for all PN chirps to obtain range information followed by another 1D FFT along the slow time for all the range bins to obtain doppler information. Before performing the respective 1D FFTs, a mean subtraction is performed along both fast time and slow time. Further, a moving target indicator (2D MTI) is employed over the micro and macro RDIs to remove reflections caused by static objects present in the scenario. In a people-sensing setup, a person can exhibit macro motions (eg. walking, running , etc.) and micro motions such as breathing or very small body movements while being static. Micro motions are captured by computing an additional RDI from a virtual frame formed by stacking the summations of all the chirps in a physical frame for 32 consecutive physical frames [1]. A moving window with a stride of 1 and a size of 32 is employed over the physical frames to generate the virtual frame.

B. Camera Embedding Generation

Each corresponding RGB camera image is fed to a pre-trained OpenPose network. The OpenPose network is a real-time, multi-person detection, and pose estimation system that involves a multi-stage Convolutional Neural Network (CNN). In the network, the input image is passed through a baseline model comprising of the first 10 layers of the VGG-16 network. The generated feature map is then fed to another network that predicts part affinity fields (PAFs) and the output of it is further used by an additional network to refine the confidence map prediction. The predicted PAFs and the confidence map are fed to a greedy matching algorithm for final processing. The PAFs provide a connection between different limbs that belong to the same person. This refinement is critical to map correct limbs to its body. For our use case, we use 13 confidence maps which represent the possibility of a particular limb being present in a given pixel. The mean confidence maps and corresponding people count labels are then fed to a triplet network which has a 32D embedding layer with linear activation as the final layer is trained to generate clusters for each person count class. A 70-30 train-test split was performed and an accuracy of 98.75% was measured by using a k-Nearest Neighbor algorithm for classifying the embedding outputs. The embeddings for all of the confidence maps were generated for further use.

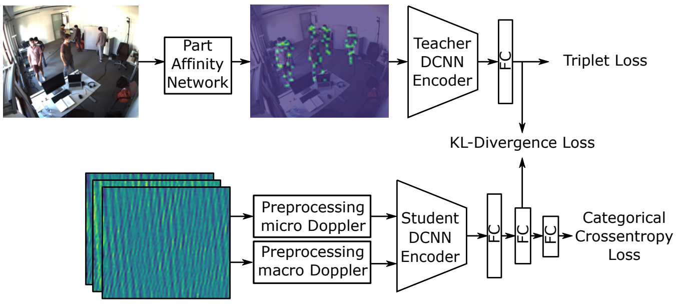


Fig. 2. Proposed Architecture

IV. Proposed Solution - Learning

A. Encoder

The input to the encoder consists of RDIs with 12 channels representing the real and complex part of the RDIs separately, generated from 3 channels for macro and micro motions, respectively. The encoder architecture comprises of Residual Blocks (ResBlock), which enforce the network to learn identity mapping. A residual block consists of a 'skip connection' that adds input activation to output activation of the last layer in the block. The input dimension may be different from that of the output due to the use of convolutional layers as intermediate layers. In order to match the dimensions, a 1x1 convolution layer is employed to project the input. Our ResBlock contains two 3x3 convolution layers, each followed by batch normalization and a relu activation function. The input is passed through a 1x1 convolution layer before adding it to the output. The encoder architecture consists of 5 such ResBlocks followed by three fully connected layers with 128, 32, and 7 hidden units. The number of hidden units in the last layer is the same as the number of total classes, and the second last fully connected layer is the same as the embedding dimension of the camera data, with both layers having softmax activation. The proposed architecture is depicted in Fig 2.

B. *Loss Functions*

In order to perform knowledge distillation from the corresponding camera embeddings to the encoder E, the Kullback-Leibler divergence L_{KL} between the embeddings generated from the encoder, i.e., the output of the second last fully connected layer, E^{emb} and the softmax output of the camera embeddings C^{emb} is minimized, which helps the encoder learn the camera embedding space semantics. The L_{KL} can be written as:

$$L_{\text{KL}}(E^{\text{emb}} \parallel C^{\text{emb}}) = \sum_{x \in \mathcal{X}} E^{\text{emb}} \log\left(\frac{E^{\text{emb}}}{C^{\text{emb}}}\right). \tag{1}$$

The Categorical Cross Entropy (CCE) is used on the output of the last layer of the encoder E^{cls} for the purpose of classification and can be written as:

$$L_{\text{CCE}} = -\sum_{i}^{C} t_i log(E_i^{\text{cls}}), \tag{2}$$

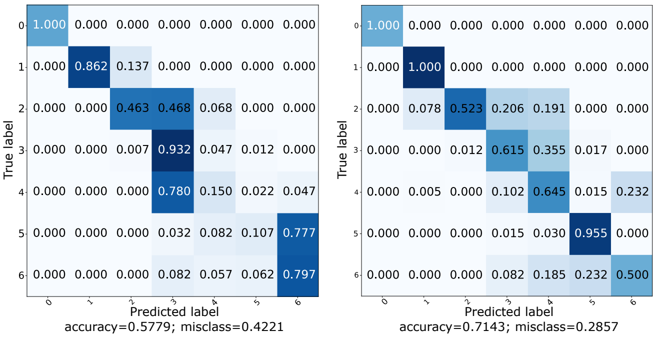
where t_i is the ith element in the ground truth (one-hot encoded class label), C is the number of classes, and E_i^{emb} is the ith element in E^{emb}. The loss terms L_{CCE}, and L_{KL} are jointly optimized.

C. *People Count Classification*

In inference mode, the raw radar ADC data goes through the same pre-processing as mentioned in Section II (A). The output is fed to the trained encoder, and the index of the maximum value of the element in the encoder output E^{cls} is predicted as the class. Note that the camera embeddings are only used during the training process and are not involved during inference.

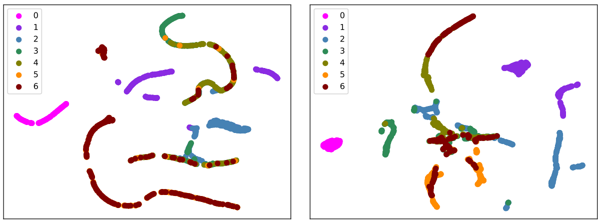
V. Experimental Results

Our training set consists of preprocessed camera and radar data, recorded for zero to six people in an office environment. For the seven classes, we have around 5000 radar frames with the corresponding camera pictures each for people moving around and people standing still. In total, the training set consists of around 65000 radar frames and camera embeddings. For the test set, we use 2600 radar frames with 200 frames recorded for each class, moving and standing. The data for the test set was recorded with another radar sensor at a different position to make sure that the model generalizes to different radar scenes, positions, and viewing angles. No camera pictures were taken for the test recordings, as those are only required during training. The results in terms of accuracy on the test recordings for all the classes are shown in Fig 3(a) for the model without knowledge distillation, and in 3(b) for the same model but with knowledge distillation. The base encoder model achieves an accuracy of 58% on the test set from a different radar sensor with a different orientation and aspect angle and 74% for the test data from the same radar sensor, while the model with knowledge distillation gets 71% accuracy on the test set from a different radar sensor and 89% accuracy for the test data from the same radar sensor. Both models work better from zero to three targets. Especially for zero and one target, both models achieve nearly perfect accuracy. This is expected, as the RDIs for zero targets are quite distinct, and there are fewer multipath and occlusion effects affecting the RDI in the one target case. For more than three targets, both models struggle, but the model without knowledge distillation does noticeably worse, as it completely fails for four and five targets, while the other one still does reasonably well. Fig. 4(a) and Fig. 4(b) show the embedding spaces of the test examples for the respective trained model mapped to two dimensions with Uniform Manifold Approximation and Projection (UMAP). Due to the KL-Divergence loss term, the clusters in Fig. 4(b) form tighter and better-separated clusters than those in Fig. 4(a). In Fig. 4(a), it is also visible that the embeddings belonging to the three to six people classes are likely hard to separate.



(a) (b)

Fig. 3. Confusion matrix for people count classification on the test set (a) without and (b) with knowledge distillation for data from a different radar sensor with a different position and orientation



(a) (b)

Fig. 4. Embedding space of the second to last layer visualized in 2D after applying UMAP (a) without and (b) with knowledge distillation for data from a different radar sensor with a different position and orientation

VI. Conclusion

This paper proposes a people counting solution for indoor scenarios using a low-cost radar sensor in conjunction with preprocessing and a neural network. The main idea we propose is to use knowledge distillation from a superior sensor, here a camera, to learn the embedding space semantics by adding a KL divergence loss term between those two embeddings. Testing the models trained with and without this additional loss term on radar data from a different sensor shows that adding that term heavily improves the test accuracy and generalization.

References

[1] A. Santra, R. V. Ulaganathan, and T. Finke, "Short-range millimetric-wave radar system for occupancy sensing application," *IEEE sensors letters*, vol. 2, no. 3, pp. 1–4, 2018.

[2] C. Wang, H. Zhang, L. Yang, S. Liu, and X. Cao, "Deep people counting in extremely dense crowds," in *Proceedings of the 23rd ACM international conference on Multimedia*. ACM, 2015, pp. 1299–1302.

[3] L. Boominathan, S. S. Kruthiventi, and R. V. Babu, "Crowdnet: A deep convolutional network for dense crowd counting," in *Proceedings of the 24th ACM international conference on Multimedia*. ACM, 2016, pp. 640–644.

[4] Z. Shi, L. Zhang, Y. Liu, X. Cao, Y. Ye, M.-M. Cheng, and G. Zheng, "Crowd counting with deep negative correlation learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 5382–5390.

[5] Z. Zou, X. Su, X. Qu, and P. Zhou, "Da-net: Learning the fine-grained density distribution with deformation aggregation network," *IEEE Access*, vol. 6, pp. 60 745–60 756, 2018.

[6] Y. Li, X. Zhang, and D. Chen, "Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1091–1100.

[7] C. Gao, P. Li, Y. Zhang, J. Liu, and L. Wang, "People counting based on head detection combining adaboost and cnn in crowded surveillance environment," *Neurocomputing*, vol. 208, pp. 108–116, 2016.

[8] C. Will, P. Vaishnav, A. Chakraborty, and A. Santra, "Human target detection, tracking, and classification using 24-ghz fmcw radar," *IEEE Sensors Journal*, vol. 19, no. 17, pp. 7283–7299, Sep. 2019.

[9] J. W. Choi, D. H. Yim, and S. H. Cho, "People counting based on an ir-uwb radar sensor," *IEEE Sensors Journal*, vol. 17, no. 17, pp. 5717–5727, 2017.

[10] J. He and A. Arora, "A regression-based radar-mote system for people counting," in *2014 IEEE International Conference on Pervasive Computing and Communications (PerCom)*. IEEE, 2014, pp. 95–102.

[11] X. Yang, W. Yin, and L. Zhang, "People counting based on cnn using ir-uwb radar," in *2017 IEEE/CIC International Conference on Communications in China (ICCC)*. IEEE, 2017, pp. 1–5.

[12] X. Yang, W. Yin, L. Li, and L. Zhang, "Dense people counting using ir-uwb radar with a hybrid feature extraction method," *IEEE Geoscience and Remote Sensing Letters*, vol. 16, no. 1, pp. 30–34, 2018.

[13] M. Stephan and A. Santra, "Radar-based human target detection using deep residual u-net for smart home applications," in *Proceedings of the 18th IEEE international conference on machine learning applications (ICMLA)*. IEEE, 2019.

[14] A. Santra and S. Hazra, *Deep learning applications of short-range radars*. Artech House, 2020.

[15] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," in *NIPS Deep Learning and Representation Learning Workshop*, 2015. [Online]. Available: http://arxiv.org/abs/1503.02531

[16] S. Gupta, J. Hoffman, and J. Malik, "Cross modal distillation for supervision transfer," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2827–2836.

[17] M. Zhao, T. Li, M. Abu Alsheikh, Y. Tian, H. Zhao, A. Torralba, and D. Katabi, "Through-wall human pose estimation using radio signals," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7356–7365.

[18] M. Abavisani, H. R. V. Joze, and V. M. Patel, "Improving the performance of unimodal dynamic hand-gesture recognition with multimodal training," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1165–1174.

[19] C. Y. Aydogdu, S. Hazra, A. Santra, and R. Weigel, "Multi-modal cross learning for improved people counting using short-range fmcw radar," in *2020 IEEE International Radar Conference (RADAR)*, 2020, pp. 250–255.

[20] F. M. Thoker and J. Gall, "Cross-modal knowledge distillation for action recognition," in *2019 IEEE International Conference on Image Processing (ICIP)*, 2019, pp. 6–10.

[21] S. Albanie, A. Nagrani, A. Vedaldi, and A. Zisserman, "Emotion recognition in speech using cross-modal transfer in the wild," in *ACM Multimedia*, 2018.