Microwave Engineering
Ulm University

universität ulm

A Ground Truth System for Radar Measurements of Humans

Nicolai Kern, Adrian Holzbock, Timo Grebner, Vasileios Belagiannis, Klaus Dietmayer, Christian Waldschmidt

DOI:

A Ground Truth System for Radar Measurements of Humans

Nicolai Kern[1], Adrian Holzbock[2], Timo Grebner[1], Vasileios Belagiannis[2],
Klaus Dietmayer[2], Christian Waldschmidt[1]

[1]Ulm University, Institute of Microwave Engineering, 89081 Ulm, Germany
[2]Ulm University, Institute of Measurement, Control, and Microtechnology, 89081 Ulm, Germany
E-Mail: nicolai.kern@uni-ulm.de

***Abstract*— Radar simulations of human targets can be deployed to reduce the measurement effort linked to dataset generation for tasks such as gesture or activity classification. However, simulations require realistic human motion data in order to capture the dynamics of the simulated activities. For this purpose, this paper proposes a stereo camera-based system that enables simultaneous recording of radar data and the corresponding human pose ground truth. By introducing a camera-radar calibration procedure, the 3D human poses and the radar system are synchronized both in time and space. Thus, the system enables the one-by-one re-simulation of the captured measurements for the investigation of simulation quality or sensor studies. The performance of the calibration procedure and the feasibility of direct re-simulations is shown with measurements of an exemplary gesture. In addition, the straightforward extension of the proposed approach to radar sensor networks is demonstrated.**

***Keywords*— motion capture, radar gesture recognition, radar simulation**

I. Introduction

Machine learning algorithms for radar-based gesture recognition have been successfully applied for tasks ranging from fine-grained finger gesture recognition [1] to large-scale gestures involving the whole human body [2]. While radars can enable the robust, privacy-sensitive detection of gestures, one bottleneck has been data scarcity due to the high effort entailed by extensive measurement campaigns. This is particularly restrictive in fields such as traffic gesture recognition, where datasets have to cover a wide range of scenarios and variations. One recently explored way to generate more diverse datasets with acceptable effort are radar simulations [2], [3].

Simulations of human gestures or activities require realistic human motion data in order to build simulation models that capture the motion dynamics. The motion data generation can either be purely simulative, such as modelling in computer graphics programs [4], or involve motion capture measurements. For the latter, measurements need to be conducted with a motion capture system. This creates additional effort, but it can also provide unique benefits when the motion capture is accompanied by simultaneous radar measurements. Generating the pose ground truth of real measurements of humans enables the re-simulation of actual measurements e.g. with modifications such as different sensors or modulation parameters, and it provides an elegant way of assessing simulation quality by one-to-one comparison

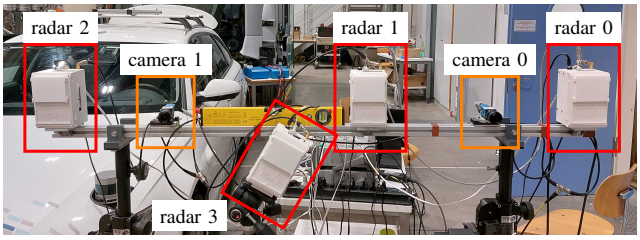


Fig. 1. Photography of the measurement system with the radar sensors and the cameras.

with the corresponding radar measurement. However, in order to fully exploit this potential, the recorded motion data has to be synchronized with the measured radar data both in time and space, such that pose data are not just available in the motion capture system's inherent coordinate system but also in the local coordinate systems (LCSs) of the radar sensors. Since this is hard to achieve with the popular Kinect sensor approach [5], [3], this paper proposes a stereo camera system for human pose ground truth generation. With our approach, 3D human poses are inferred for simultaneous short range radar measurements based on 2D skeletal keypoints. The stereo camera and the radar system are synchronized in space by a preceding camera-radar calibration whose accuracy is validated by data from a tacheometer. It is demonstrated that the system provides accurate ground truth data enabling one-by-one re-simulations of an exemplary gesture and that the approach can be easily extended to multiple radar sensors, thus facilitating simulations of radar sensor networks.

II. Measurement System

As shown in Fig. 1, the measurement system consists of a radar sensor network for recording of the human radar responses and is augmented by a stereo camera system for the simultaneous capture of the corresponding 3D pose ground truth. The radar sensor network consists of four chirp sequence (CS)-multiple-input multiple-output (MIMO) sensors. Three of the radar sensors are mounted on a rail similar to the setup used for radar gesture recognition in [6]. The fourth sensor is placed and oriented freely to demonstrate the calibration procedure's feasibility to account for arbitrary network configurations. All radar sensors are operated at 79 GHz with range and

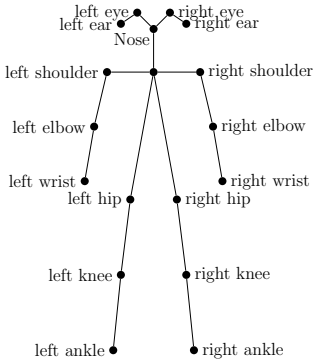


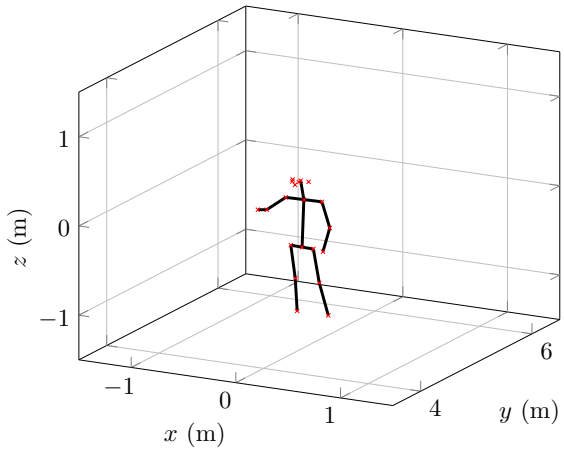
Fig. 2. 2D skeletal keypoints detected by Detectron2 in the COCO format.

velocity resolutions of $\Delta R = 4.5\,\mathrm{cm}$ and $\Delta v = 11\,\mathrm{cm\,s^{-1}}$, respectively. For beamforming, each radar sensor has 12 virtual Rx channels with up to eight elements for azimuth beamforming and two for elevation beamforming. All radars receive a common trigger and operate with slight time offsets, such that responses belonging to transmit signals of the other sensors are suppressed by baseband filtering, and only monostatic responses are processed. The radar sensor network is complemented by two RGB cameras with a resolution of 1280×1024 that are calibrated for stereo vision. Calibration measurements are conducted with a checkerboard of size $1.73\,\mathrm{m} \times 0.77\,\mathrm{m}$ and for ranges of $3\,\mathrm{m}$ to $6\,\mathrm{m}$. For accurate re-simulations, the 3D positions obtained by the stereo camera system should have higher precision than the radar sensors, whose accuracy is mainly limited due to the angle estimation with a relatively small number of virtual channels. Accurate stereo positions can be ensured by thorough stereo calibration and are proven by additional tacheometer measurements. For simultaneous measurements, images are recorded synchronously with the radar frames by common triggering, and the whole measurement system is operated at $30\,\mathrm{fps}$.

III. Generation of Human Pose Ground Truth Data with the Stereo Camera System

In principle, the stereo camera system allows the computation of the 3D position of any object in its field of view (FoV), thus e.g. enabling the generation of depth maps based on pixel correspondences. However, as we are only interested in the positions of the skeletal joints of the observed person, it is sufficient to detect and match just the relevant keypoints in both cameras' images. For that purpose, Detectron2 [7] is applied for person keypoint detection, which provides the pixel positions $\vec{p}_{i_\mathrm{c},i_\mathrm{kp}}$ for 17 skeletal keypoints i_kp defined in the COCO dataset format [8] and both cameras i_c. The skeletal model of the COCO dataset is illustrated in Fig. 2. The keypoints are computed for each measurement frame. An example is shown for both cameras in Figs. 3a and 3b. In order to remove faulty detections and to reduce the jitter over the course of the measurements, outliers are removed, and a smoothing spline is fitted to each sequence of keypoint pixel

(a) (b)



(c)

Fig. 3. Estimation of the 2D keypoints in the images of the left (a) and right (b) camera and computation of their 3D counterparts with the stereo calibration information (c).

positions. Based on the smoothed locations, the 3D positions of the skeletal keypoints are inferred by triangulation, i.e. finding the 3D points that minimize the overall reprojection error [9]

$$\vec{x}_{i_\mathrm{kp}} = \min_{\vec{x}} \sum_{i_\mathrm{c} \in 0,1} \left| \mathbf{C}_{i_\mathrm{c}} \vec{x}_{i_\mathrm{kp}} - \vec{p}_{i_\mathrm{c},i_\mathrm{kp}} \right|^2 , \tag{1}$$

with the 3D keypoint position $\vec{x}_{i_\mathrm{kp}}$. The matrix $\mathbf{C}_{i_\mathrm{c}}$ describes the reprojection from global coordinates to the image plane of camera i_c and is obtained by the stereo camera calibration process. The resulting 3D human pose for the exemplary frame is visualized in Fig. 3c. Since small but unfavourable errors in the pixel positions such as errors of different sign at the two cameras can lead to larger jitter after the triangulation step, the 3D keypoints are smoothed, too. Moreover, an additional skeletal center point is computed from the right and left hip, as shown in Fig. 3c.

IV. Camera-Radar Calibration Procedure

While the measurement ground truth provided by the stereo camera system is time-synchronized with the simultaneously collected radar data, it is still in the camera coordinate system (CCS). However, for proper re-simulations—especially for

radar sensor networks—the human pose information has to be transformed to the LCSs of the radar sensors. Therefore, a camera-radar calibration is proposed to obtain the required transformations $\mathbf{T}_{\mathrm{LCS}_{i_\mathrm{n}},\mathrm{CCS}}$ with the radar sensor node index $i_\mathrm{n} \in \{0, 1, 2, 3\}$. For calibration, a common calibration target detectable by both radar and camera is moved through the sensors' common FoV, and the transformations between the coordinate systems are obtained based on the calibration target trajectories. The joint target consists of a small checkerboard mounted on a corner reflector, with the checkerboard center aligned with the corner reflector's phase center. In addition, a prism is attached to the joint target to enable tracking by a high-precision tacheometer for assessment of the calibration quality.

A. Radar Signal Processing

Radar signal processing is carried out for each radar sensor independently: The 2D fast Fourier transform (FFT) is computed along the samples and chirps to obtain the range-Doppler maps. Zero-padding by a factor of two is applied for higher precision with respect to the target ranges. From the range-Doppler maps, targets are extracted by applying the ordered statistics constant false-alarm rate algorithm with subsequent peak search. From the resulting list of targets, the corner reflector serving as calibration target can easily be found due to its high radar cross section (RCS). For the corner reflector, the azimuth and elevation angles ϕ and θ, respectively, are computed by comparing the measured steering vectors with the theoretical ones for all incidence angles within the antennas' beamwidth [10]. Formally, with the incidence direction described by $\Phi = (\phi, \theta)$, first the cross correlation

$$\vec{v}\,(\Phi) = \frac{|\mathbf{C}^H(\Phi) \cdot \vec{a}|}{|\mathbf{C}\,(\Phi)\,||\vec{a}|} \tag{2}$$

is computed with the measured steering vector $\vec{a} \in \mathbb{C}^{12\times 1}$, the calibration matrix $\mathbf{C} \in \mathbb{C}^{N_\phi N_\theta \times 12}$, and its complex conjugate $\mathbf{C}^H$. The calibration matrix contains the $N_\phi N_\theta$ theoretical steering vectors for the N_ϕ tested azimuthal and N_θ tested elevation angles that are also corrected for the phase offsets between the receive channels. The estimated direction of arrival is then given by

$$\hat{\Phi} = \arg\max_{\Phi} \vec{v}\,(\Phi)\ . \tag{3}$$

From the range and direction of arrival the trajectories of the corner reflector are obtained. Since the elevation angle estimation is error-prone in cases with more than a single target in a range-Doppler cell e.g. due to clutter, detections with velocity of the corner reflector below $0.3\,\mathrm{m\,s^{-1}}$ are removed. In addition, frames experiencing a significant drop in target RCS are filtered out as they suffer from a misaligned corner reflector.

B. Stereo Camera Signal Processing

For computation of the trajectories of the stereo camera reference target, the small checkerboard mounted on the corner reflector is detected in the images from both cameras, and the 3D position of the checkerboard center is computed using 1 for the frames in which the target is in the FoV of both cameras.

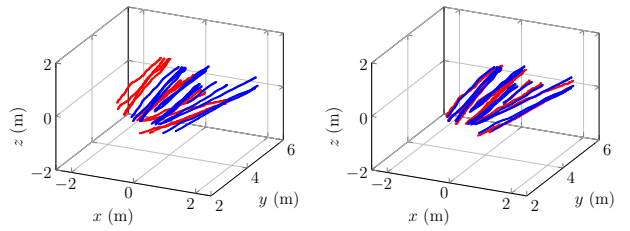


Fig. 4. Trajectories of the radar (blue) and stereo camera (red) calibration target before (a) and after (b) the minimization of the Euclidean distance.

Table 1. Mean trajectory deviation after applying the transformations.

Transformation Type	**Radar 0**	**Radar 1**	**Radar 2**	**Radar 3**
camera-radar	11.4 cm	9.5 cm	11.9 cm	9.6 cm
camera-tacheometer	3.3 cm	3.3 cm	3.4 cm	3.3 cm
radar-tacheometer	14.2 cm	11.4 cm	12.3 cm	11.4 cm

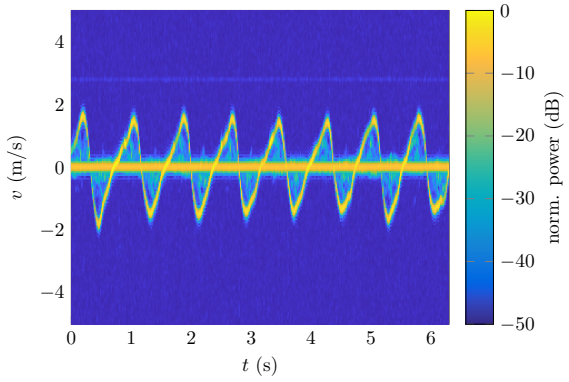
C. Transformations to the Radar Coordinate Systems

With the common calibration target detected in both the radar and stereo camera data, valid sequences are identified, in which detections are available for both sensor types. Since valid sequences differ between the radar sensors, each element of the radar sensor network has its own set of calibration data. Subsequently, the transformations $\mathbf{T}_{\mathrm{LCS}_{i_\mathrm{n}},\mathrm{CCS}}$ linking the CCS and the LCSs are computed by treating the detections of the calibration target in the radar and stereo camera data as two distinct point sets and applying the iterative closest point algorithm for point cloud registration [11]. The minimization of the Euclidean distance between the point sets results in the transformations and the poses of the radar sensors in the CCS. The calibration process for the freely positioned radar sensor 3 is illustrated in Fig. 4. Applying these transformations to the 3D keypoints resulting from the procedure in Sec. III, the human pose ground truth data is also synchronized in space and is directly ready for radar simulations without any necessity of taking into account the positions and orientations of the radar sensors.

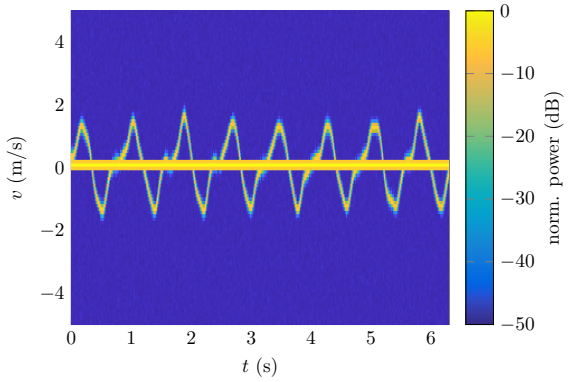
V. Experimental Results

A. Evaluation of the Sensor and Calibration Accuracy

In order to validate the camera-radar calibration, the average deviation between the trajectories of the common calibration target after the transformation is evaluated. As can be seen in Table 1, the calibration errors after transforming the stereo camera detections to the LCSs are comparable over all radar sensors, demonstrating the ability of the approach to generate spatially-synchronized ground truth data for arbitrary sensor orientations. In addition, the precision of



(a)



(b)

Fig. 5. Spectrograms of a measured "Stop" gesture (a) and its re-simulation based on the proposed ground truth system (b).

the 3D positions is assessed by computing the calibrations between the sensors and the highly-precise data provided by the tacheometer, with the deviations with respect to the tacheometer 3D positions being used as criterion. The results in Table 1 demonstrate the suitability of the stereo camera system for accurate ground truth generation, with the mean errors after camera-tacheometer calibration at around 3.3 cm being lower than the errors after the radar-tacheometer calibration. Moreover, the measurement of the 3D position of the corner reflector contributes most to the error in camera-radar calibration, mainly due to higher errors in elevation angle estimation.

B. Gesture Re-Simulation

In addition to the calibration accuracy, the feasibility of straightforward re-simulation is demonstrated by simulation an exemplary "Stop" gesture with a corner reflector held in the executing hand. The measured Doppler spectrogram for this gesture is illustrated in Fig. 5a, where the velocity profile of the corner reflector becomes directly visible. This gesture is re-simulated based on the simultaneously captured human pose ground truth data. After transforming the keypoints to the LCS of radar sensor 0, the right wrist is used as a single point target for simulation. As can be seen from the resulting spectrogram in Fig. 5b, both the zero-crossings as well as the velocity peaks are reflected in the simulation. Thus, the ground truth system leads to simulations highly similar to the original measurements, enabling novel opportunities for human target simulations.

VI. Conclusion

In this paper, a novel approach for the simultaneous recording of radar data and the corresponding ground truth data for measurements involving human motion is presented. The 3D human pose information is obtained by a stereo camera system that enables the triangulation of 2D skeletal keypoints. In addition, common triggering and camera-radar calibration allow the direct transformation of the motion data into the radar coordinate systems, which facilitates the re-simulation of measurements for arbitrarily positioned and oriented radar sensors. Finally, the accuracy of the stereo camera system, the camera-radar calibration, and the re-simulation of an exemplary gesture is demonstrated by measurement.

Acknowledgment

This work was supported by the Ministerium für Wissenschaft, Forschung und Kunst (MWK) Baden-Württemberg as part of the project INTUITIVER.

References

[1] S. Wang, J. Song, J. Lien, I. Poupyrev, and O. Hilliges, "Interacting with Soli: Exploring fine-grained dynamic gesture recognition in the radio-frequency spectrum," in *Proc. of the 29th Annu. Symp. User Interface Software and Technol.*, 2016, pp. 851–860.

[2] A. Ninos, J. Hasch, M. E. P. Alvarez, and T. Zwick, "Synthetic radar dataset generator for macro-gesture recognition," *IEEE Access*, vol. 9, pp. 76 576–76 584, 2021.

[3] M. S. Seyfioglu, B. Erol, S. Z. Gurbuz, and M. G. Amin, "Diversified radar micro-Doppler simulations as training data for deep residual neural networks," in *IEEE Radar Conf.*, 2018, pp. 0612–0617.

[4] K. Ishak, N. Appenrodt, J. Dickmann, and C. Waldschmidt, "Advanced radar micro-Doppler simulation environment for human motion applications," in *IEEE Radar Conf.*, 2019, pp. 1–6.

[5] B. Erol and S. Z. Gurbuz, "A kinect-based human micro-doppler simulator," *IEEE Aerospace and Electronic Systems Magazine*, vol. 30, no. 5, pp. 6–17, 2015.

[6] N. Kern, M. Steiner, R. Lorenzin, and C. Waldschmidt, "Robust Doppler-based gesture recognition with incoherent automotive radar sensor networks," *IEEE Sens. Lett.*, vol. 4, no. 11, pp. 1–4, 2020.

[7] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, and R. Girshick, "Detectron2," https://github.com/facebookresearch/detectron2, 2019.

[8] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Eur. Conf. Comput. Vis.* Springer, 2014, pp. 740–755.

[9] M. Zhao, Y. Tian, H. Zhao, M. A. Alsheikh, T. Li, R. Hristov, Z. Kabelac, D. Katabi, and A. Torralba, "RF-based 3D skeletons," in *Proc. Conf. ACM Special Interest Group Data Commun.*, 2018, p. 267–281. [Online]. Available: https://doi.org/10.1145/3230543.3230579

[10] C. Vasanelli, F. Roos, A. Dürr, J. Schlichenmaier, P. Hügler, B. Meinecke, M. Steiner, and C. Waldschmidt, "Calibration and direction-of-arrival estimation of millimeter-wave radars: A practical introduction," *IEEE Antennas Propag. Mag.*, vol. 62, no. 6, pp. 34–45, 2020.

[11] P. J. Besl and N. D. McKay, "Method for registration of 3-D shapes," in *Sensor fusion IV: control paradigms and data structures*, vol. 1611. International Society for Optics and Photonics, 1992, pp. 586–606.