19th International Conference on Information Fusion
Heidelberg, Germany - July 5-8, 2016

Fusion of Stereo Camera and MIMO-FMCW Radar for Pedestrian Tracking in Indoor Environments

Roman Streubel and Bin Yang
Institute of Signal Processing and System Theory
University of Stuttgart, Germany
Email: {roman.streubel, bin.yang}@iss.uni-stuttgart.de

***Abstract*—In this paper, we present a fusion of stereo camera and radar targets to significantly improve the tracking of pedestrians indoors especially suited for surveillance and security applications. Indoor environments pose bigger challenges for pedestrian tracking compared to outdoor environments. This makes pedestrian tracking with only a camera or radar difficult. In this work, we demonstrate that in some preliminary experiments our fusion system increases the multiple object tracking accuracy (MOTA) from -16.7% and 38.0% for camera or radar only tracking to 90.9% using a fusion of both. Furthermore, the proposed fusion system achieves a multiple object tracking precision ($MOTP_{3D}$) of 90.3% compared to 83.4% and 81.8% obtained via camera and radar only tracking, respectively.**

I. Introduction

Camera and radar fusion for object tracking has been intensively studied for decades [1], [2]. It is used in a wide range of applications such as topographic map creation, driver assistance and autonomous driving [3], [4]. In automotive applications, for example, the objects to be detected and tracked are mostly vehicles. In recent years, pedestrian detection and tracking on streets for an increased driving safety in urban environments has also attracted much attention. In comparison to vehicles, the radar detection and tracking of pedestrians suffers from additional difficulties: a low radar cross section (RCS), unpredictable reflection centers depending on clothes and diverse velocities of different body parts (e.g. swinging arms) of the same object (pedestrian) [5]–[9].

In this paper, we focus on a fusion application which has rarely been addressed in the literature, pedestrian detection and tracking in indoor environments for surveillance and security systems. These systems work 24 hours a day and 7 days a week and shall provide an automatic, highly sensitive and reliable object detection and tracking. In comparison to pedestrian tracking on streets, the indoor environment poses additional challenges. The occlusion of a pedestrian by various items and installations in a room or by other pedestrians in a crowded situation is one such challenge. This makes both camera and radar detection and tracking difficult. Other challenges include the rich radar reflections on ceilings, floors, walls and other items as well as installations in a room with an even higher RCS than the pedestrians.

The aim of this paper is to study this situation and to compare the camera and radar only tracking with that of sensor fusion. For this purpose, we use an experimental system consisting of a stereo camera and a multiple input multiple output (MIMO) frequency modulated continuous wave (FMCW) radar.

II. System Overview

Fig. 1 shows the overall pedestrian detection and tracking system which is assumed to be installed on a fixed base. It consists of a stereo camera providing rectified stereo images and a MIMO-FMCW radar delivering radar baseband signals. They are processed by an image processing chain (Section III) and a radar signal processing chain (Section IV) separately. Both chains return so-called targets which are then combined and tracked in a heterogeneous sensor fusion unit, which finally return objects and their trajectories over time. The sensor fusion is addressed in Section V, with quantitative performance evaluation results reported in Section VI.

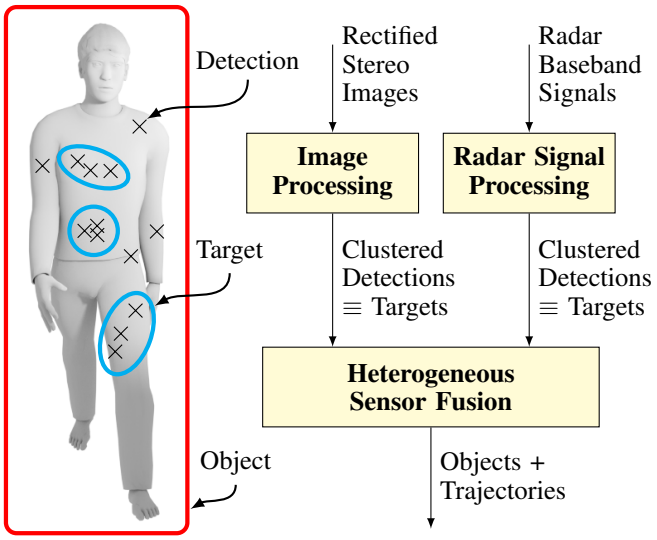

Fig. 1. Overview of the fusion system for indoor detection and tracking of pedestrians.

Both the camera and radar can make detections from pedestrians, upper bodies and other moving foreground items in the case of camera and radar reflections in the case of radar. The basic assumption here is that a pedestrian is by definition moving, at least in a long timescale in comparison to the stationary background. Based on the spatial coordinates, the detections can be clustered separately in the camera and radar signal processing chains. Each cluster is called a target. This clustering is necessary because at a close range, a pedestrian is not a point target and the pedestrian size is often larger than the range resolution of the radar. The clustering combines radar detections of the same pedestrian from different range bins. A second advantage of the clustering is to provide both a mean and covariance of the position of the targets before a Kalman filter based fusion and tracking.

A big difference to vehicle or aircraft tracking is that the radar Doppler (radial velocity) of the detections is less useful for clustering of pedestrian detections because different body parts (e.g. moving arms and legs) can show totally different

978-0-9964527-4-8

velocities [5]–[9]. In this paper, we first decided to completely ignore the radar velocity estimates in both clustering and tracking. The motivation is to investigate whether a simpler radar (e.g. pulse radar without Doppler estimate) is suitable for such an indoor pedestrian tracking system. In a future study, we will compare the tracking performance with and without the Doppler estimate.

After the clustering of detections, the camera and radar share a common spatial domain for fusion of targets and tracking. This happens in a heterogeneous way because the image and radar targets are temporally asynchronous. This simplifies the realization of the pedestrian tracking system by using off-the-shelf sensors.

The coordinate system used in this paper consists of x in horizontal dimension, y in height and z in depth.

III. Image Processing

Fig. 2 provides an overview about the image processing chain starting with the rectified stereo camera images $\boldsymbol{\mathcal{I}}_{\mathrm{L}}$ and $\boldsymbol{\mathcal{I}}_{\mathrm{R}}$ as input. They are captured with a BumbleBee2 stereo camera where the left camera center and right camera center are horizontally aligned with a distance of 12 cm. We assume a simple pinhole camera model. The camera is calibrated in both the intrinsic and extrinsic parameters.

In the following, various image processing steps are briefly described.

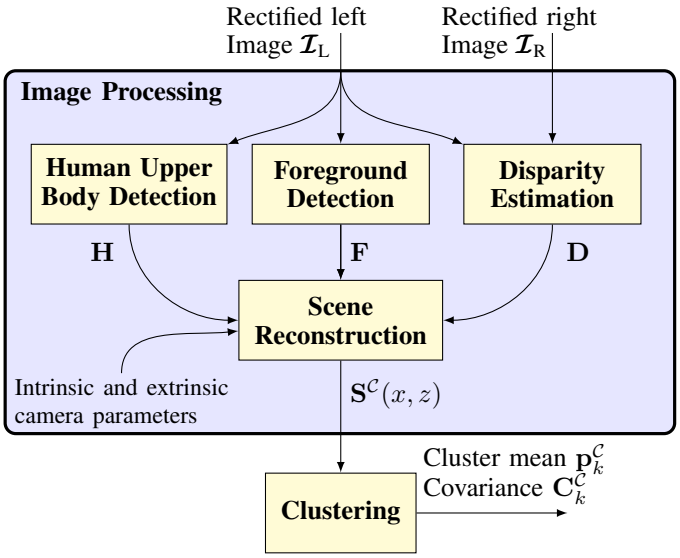

Fig. 2. Overview of the image processing chain.

A. Human Upper Body Detection

In indoor environments with many installations and items, in particular in a crowded situation with many pedestrians, occlusion is one of the major difficulties, for both the camera and the radar. In automotive pedestrian tracking applications, the conventional approach is to detect the pedestrian silhouette. This is no longer possible in indoor environments. The occlusion statistics in [10] show that the lower body is the most occluded body part. This implies that the detection of the upper body is more robust against occlusion than a human silhouette detection. We detect human upper bodies in the left image $\boldsymbol{\mathcal{I}}_{\mathrm{L}}$ by using a method based on the Viola-Jones algorithm [11]. It is a supervised learning algorithm and uses histogram of oriented gradients (HOG) features obtained from gray images. Front, back, left and right views of real and simulated pedestrians are used as training samples to train a classifier for each view, respectively. The human upper body detection is important when a pedestrian is standing still and there is no movement in the image because the parallel detection of moving foreground objects (next subsection) will fail in this case. The upper body detection provides bounding boxes including x_i and y_i position as well as width and height. The bounding box properties for all detected human upper bodies are stacked into the human upper body bounding box matrix $\mathbf{H}$ for further processing.

Fig. 3(a) shows 2 detected (green) upper bodies in one image in a scene where one pedestrian is walking away from the camera. One detection is from the back view classifier and is thus correct, while the second one is a false detection from the front view classifier, because the front and back views of an upper body look quite similar. We tolerate this situation in our work because a miss detection is far more serious than two similar detections which will be easily merged at a later stage.

B. Foreground Detection

The upper body detection is mostly successful, but fails in certain situations such as when the upper body itself is occluded, a pedestrian enters or leaves the field-of-view (FoV) or when the pedestrian is too far away from the camera to provide enough detail for a successful upper body detection. For this reason, we apply a parallel detection of moving foreground objects to the left image $\boldsymbol{\mathcal{I}}_{\mathrm{L}}$.

The foreground detection relies on a short-term and a long-term model which are both essentially background subtraction algorithms [12]. Those two models have different adaptation speeds. One with a low adaptation speed and the other with a high adaptation speed. The final foreground detection matrix $\mathbf{F}$ is found as the intersection of the foreground areas in both models.

Fig. 3(b) shows 3 detected moving foreground areas as the pedestrian is walking away from the camera: the left arm, the right arm, and the left foot. Note that the right foot is not moving and hence is not detected.

C. Disparity Estimation

A third parallel step is the disparity estimation using the left and right image to obtain depth information about the different detections. The disparity describes the horizontal pixel shift between both rectified images. With increasing distance of objects from the camera the disparity decreases. We use the block matching technique for this task, which computes the disparity by comparing the sum of absolute difference (SAD) of each block of pixels in the images [13].

Fig. 3(b) illustrates the estimated disparity image matrix $\mathbf{D}$ for the scene in Fig. 3(a). It is a grayscale depth map for the left image where brighter pixels correspond to closer points. Note that due to non-overlapping FoVs of the left and right camera, the disparity image matrix $\mathbf{D}$ has a blind area on the left side as highlighted which does not provide any depth information.

D. Scene Reconstruction

With knowledge about the intrinsic camera calibration parameters of the left and right camera, the extrinsic calibration parameters between both of them and the disparity image

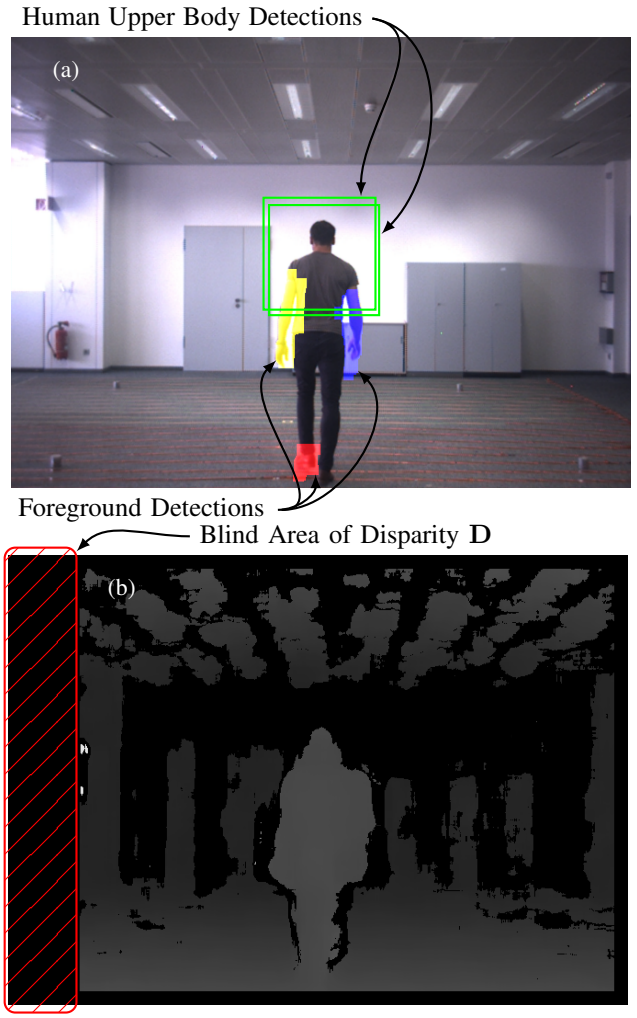

Fig. 3. (a) Results of upper body and moving foreground object detection from the left camera image. (b) Disparity map estimated from the left and right image.

matrix $\mathbf{D}$, we are now able to compute the 3D scene described by a point cloud matrix $\mathbf{S}(x, y, z)$. In order to avoid detections due to light reflections, we remove all points from $\mathbf{S}(x, y, z)$ with a height $y > 2.5$ m or $y < 0.5$ m representing the ceiling and floor. Since we assume upright walking and standing pedestrians, for each point in all bounding boxes of detected upper bodies in $\mathbf{H}$ and in all foreground detections in $\mathbf{F}$, we extract the 3D point position and ignore the height information y to form the 2D point cloud matrix $\mathbf{S}^{\mathcal{C}}(x, z)$. This means, each point cloud matrix $\mathbf{S}^{\mathcal{C}}$ contains the 2D coordinates (x, z) of relevant points from the detected upper bodies and foreground objects for each left image.

E. Clustering

As illustrated in Fig. 1, we use clustering to combine detections which are spatially close to each other. Each cluster is called a target and is characterized by a 2D mean position vector $\mathbf{p}_k^{\mathcal{C}} = \left[x_k^{\mathcal{C}}, z_k^{\mathcal{C}}\right]^{\mathrm{T}}$ and a 2×2 covariance matrix $\mathbf{C}_k^{\mathcal{C}}$, $1 \leq k \leq K^{\mathcal{C}}$. $K^{\mathcal{C}}$ is the number of found clusters for each camera image. We use the density-based spatial clustering of applications with noise (DBSCAN) method from [14] for clustering. It is a density-based clustering with a neighborhood radius parameter ε and a minimum-number-of-neighbors parameter $c_{\min}$ to create a cluster. We use $\varepsilon = 0.2$ and $c_{\min} = 100$.

Fig. 6(a) shows the clustering result for the scenario in Fig. 3(a). In this case, DBSCAN finds two targets in the image. The first target with the mean lateral position $x \approx 0.1$ m and the mean depth $z \approx 5.3$ m is a true target (pedestrian). The second target $(\mathbf{p}_2^{\mathcal{C}}, \mathbf{C}_2^{\mathcal{C}})$ is a ghost target and appears due to the large bounding box of upper body detection which also contains many pixels from the wall beyond 10 m, see Fig. 3(a).

IV. Radar Signal Processing

Fig. 4 shows an overview about the radar signal processing chain starting with the baseband signal $\mathbf{\Phi}$ captured by a MIMO-FMCW radar.

The experimental radar system operates at the carrier frequency of 76.5 GHz with a bandwidth $B = 900$ MHz. It uses a FMCW modulation. Each measurement cycle consists of $N_p = 32$ identical FMCW chirps with a chirp duration of $T_c = 16\,\mu$s. Each received chirp is sampled, resulting in a number of N_s samples per chirp. In addition, the radar has a MIMO structure consisting of M virtual antennas. They form a uniform linear array (ULA) along the x-axis for azimuth estimation (no elevation estimation). Hence the range resolution of this radar system is $\Delta R = c/2B \approx 16.7$ cm.

In the following, all radar signal processing steps are briefly described.

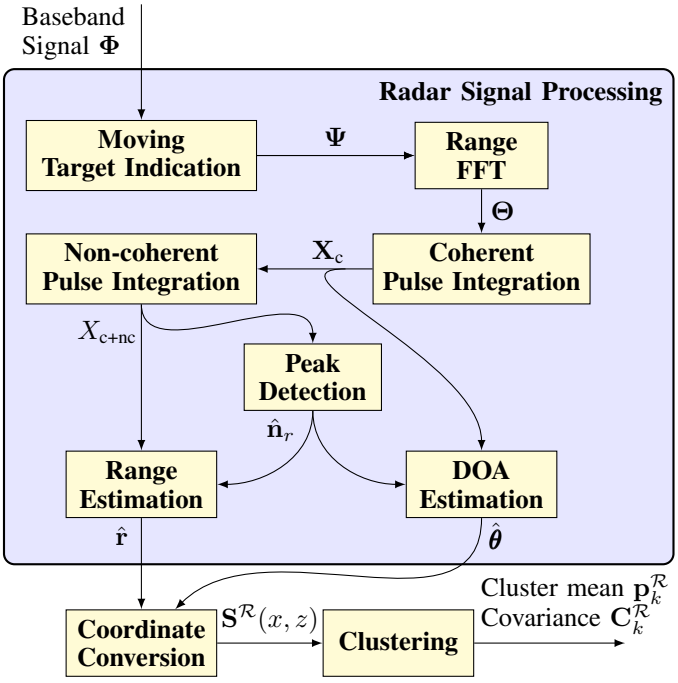

Fig. 4. Overview of the radar signal processing chain.

A. Radar Signal Sampling

For each radar measurement cycle, we obtain a 3D data cube $\mathbf{\Phi} \in \mathbb{C}^{N_p \times N_s \times M}$ containing the complex-valued baseband signal. The first dimension contains the different chirps $n_p = 0, \ldots, N_p - 1$ (slow time) for velocity estimation, the second dimension contains the different samples per chirp (fast time) $n_s = 0, \ldots, N_s - 1$ for range estimation, and the third dimension is composed of the different virtual antennas $m = 1, \ldots, M$ for azimuth estimation . We assume

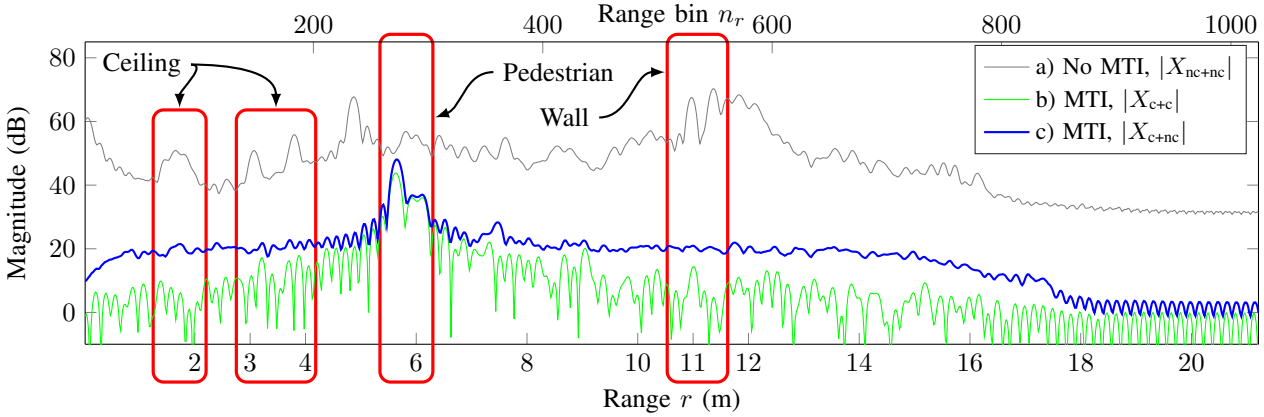

Fig. 5. Radar pulse integration for the scenario in Fig. 3(a).

the radar objects are stationary inside each measurement cycle. This is a reasonable assumption for pedestrian tracking. For the direction of arrival (DOA) estimation, a far-field model is used.

One important note regarding velocity estimation. The used radar is capable of range-Doppler processing. On the other side, it is well known from literature [9], [8], [15], [6], [5] that walking pedestrians are characterized by the micro-Doppler phenomena. Due to the diverse movements of different body parts, their velocities are, in general, different and even the velocity of the same body part is highly time-varying. This makes the velocity less useful for clustering of detections and tracking of objects than the position. In this paper, we ignore the velocity estimate in the radar signal processing and even in later tracking in order to study whether a simpler pulse radar (only range and DOA estimation) is sufficient for pedestrian tracking.

B. Moving Target Indication

Another difficulty of indoor pedestrian tracking are the rich radar reflections on ceilings, floors, walls, other installations and items in rooms. In order to avoid static reflections, we first apply a moving target indication (MTI) to the slow time (Doppler) dimension (n_p) of the data cube $\mathbf{\Phi}$. We extend the 2-pulse finite impulse response (FIR) highpass filter in [16] to a linearphase FIR bandpass with the impulse response $(1, 0, -1)$ in order to suppress both static reflections and those with large Doppler shifts (e.g. arms and legs). The result of MTI processing is the 3D data cube $\mathbf{\Psi}$.

C. Range FFT

For range estimation and further processing, an 1D fast Fourier transformation (FFT) is applied to the fast time dimension (n_s) of $\mathbf{\Psi}$, leading to the result $\mathbf{\Theta} \in \mathbb{C}^{N_p \times N_r \times M}$. Now the second dimension contains N_r of range bins $n_r = 0, \ldots, N_r - 1$.

D. Pulse Integration

The pulse integration aims to calculate the range profile $X(n_r)$ as a function of the range bin n_r by summing up the 3D date cube $\mathbf{\Theta}$ after MTI and range FFT over the chirp and antenna dimension in order to improve the signal-to-noise ratio. We consider three pulse integration techniques: non-coherent-plus-non-coherent:

$$X_{\text{nc+nc}}(n_r) = \sum_{m=1}^{M} \sum_{n_p=0}^{N_p-1} |\mathbf{\Theta}(n_p,\, n_r,\, m)| \in \mathbb{R}^{N_r}, \tag{1}$$

coherent-plus-coherent:

$$X_{\text{c+c}}(n_r) = \sum_{m=1}^{M} \mathbf{X}_{\text{c}}(n_r, m) \in \mathbb{C}^{N_r}, \tag{2}$$

coherent-plus-non-coherent:

$$X_{\text{c+nc}}(n_r) = \sum_{m=1}^{M} |\mathbf{X}_{\text{c}}(n_r, m)| \in \mathbb{R}^{N_r}, \tag{3}$$

with

$$\mathbf{X}_{\text{c}}(n_r, m) = \sum_{n_p=0}^{N_p-1} \mathbf{\Theta}(n_p,\, n_r,\, m) \in \mathbb{C}^{N_r \times M}. \tag{4}$$

In the non-coherent-plus-non-coherent case, both integration steps over chirp (n_p) and antenna (m) are non-coherent. In the coherent-plus-coherent case, the first integration step over chirp (n_p) and the second integration step over antenna (m) is coherent as well. In the coherent-plus-non-coherent case, the first integration step over chirp (n_p) is coherent while the second integration step over antenna (m) is non-coherent.

Fig. 5 compares the range profiles of three different cases:

a) no MTI filter, use $X_{\text{nc+nc}}(n_r)$
b) with MTI filter, use $X_{\text{c+c}}(n_r)$
c) with MTI filter, use $X_{\text{c+nc}}(n_r)$

clearly, without the MTI filter, the static radar reflections from ceilings, floors or walls dominate the reflections from the pedestrian. The use of the MTI filter successfully suppresses these static radar reflections and facilitates the detection of pedestrians. A comparison between b) and c) in Fig. 5 shows that coherent-plus-non-coherent technique in c) results in a more smooth noise floor in the range profile. $X_{\text{c-nc}}(n_r)$, making a peak detection simpler than in b).

E. Peak Detection

For the peak detection, we use an ordered statistic constant false alarm rate (OS-CFAR) detector

$$\hat{\mathbf{n}}_r = \text{OSCFAR}\left(|X_{\text{c-nc}}|,\, p_f,\, N_t,\, N_g,\, o\right) \in \mathbb{N}^{K^{\mathcal{R}}}$$

for a given false alarm rate p_f, the number of training cells N_t, the number of guard cells N_g, and the rank of order statistic o with $N_t/2 < o < 3N_t/4$ [17]. The vector $\hat{\mathbf{n}}_r$ contains the range bin indices of the $K^{\mathcal{R}}$ detected peaks.

F. Range Estimation

From the range bin vector $\hat{\mathbf{n}}_r$ containing the $K^{\mathcal{R}}$ highest peaks, the range of the radar detections is estimated by

$$\hat{\mathbf{r}} = \left(\frac{c}{2}\,\frac{T_c}{2\,N_r\,B\,T_s}\right)(\,\hat{\mathbf{n}}_r + \boldsymbol{\Delta}_{n_r}\,) \in \mathbb{R}^{K^{\mathcal{R}}} \tag{5}$$

via a quadratic interpolation of the peak location

$$\boldsymbol{\Delta}_{n_r} = \frac{1}{2}\,\frac{(\boldsymbol{\beta}_{-1} - \boldsymbol{\beta}_{+1})}{(\boldsymbol{\beta}_{-1} - 2\boldsymbol{\beta} + \boldsymbol{\beta}_{+1})} \in \mathbb{R}^{K^{\mathcal{R}}}. \tag{6}$$

$\boldsymbol{\beta} = |X_{\mathrm{nc}}(\hat{\mathbf{n}}_r)|$, $\boldsymbol{\beta}_{-1} = |X_{\mathrm{nc}}(\hat{\mathbf{n}}_r - 1)|$ and $\boldsymbol{\beta}_{+1} = |X_{\mathrm{nc}}(\hat{\mathbf{n}}_r + 1)|$ denote the magnitude of X_{nc} in the peak range bins and their left and right neighbor bins, respectively.

G. DOA Estimation

We assume a single target in each range bin. For the azimuth estimation

$$\hat{\theta}(i) = \arg\max_{\theta} \boldsymbol{\zeta}(\theta,\, i) \qquad \forall\, i \in \hat{\mathbf{n}}_r, \tag{7}$$

we maximize the Bartlett beamformer

$$\boldsymbol{\zeta}(\theta,\, i) = \left|\frac{\mathbf{a}(\theta)^{\mathrm{H}}\,\mathbf{R}(i)\,\mathbf{a}(\theta)}{\mathbf{a}(\theta)^{\mathrm{H}}\,\mathbf{a}(\theta)}\right| \tag{8}$$

over the azimuth angle θ. $\mathbf{a}(\theta)$ is the steering vector of the virtual ULA.

$$\mathbf{R}(i) = \mathbf{Y}(i)\,\mathbf{Y}(i)^{\mathrm{H}} \in \mathbb{C}^{M\times M} \tag{9}$$

is the sample correlation matrix of the antenna array where $\mathbf{Y}(i) = \left[\mathbf{X}_{\mathrm{c}}(i-1,:)^{\mathrm{T}},\, \mathbf{X}_{\mathrm{c}}(i,:)^{\mathrm{T}},\, \mathbf{X}_{\mathrm{c}}(i+1,:)^{\mathrm{T}}\right] \in \mathbb{C}^{M\times 3}$ consists of 3 snapshots of the array around the peak location $i \in \hat{\mathbf{n}}_r$. $\mathbf{X}_{\mathrm{c}}(i-1,:)$, $\mathbf{X}_{\mathrm{c}}(i,:)$ and $\mathbf{X}_{\mathrm{c}}(i+1,:)$ are the corresponding row vectors from $\mathbf{X}_{\mathrm{c}}$ after the coherent pulse integration.

For further processing, all DOA estimates $\hat{\theta}(i)$ are stacked into a single column vector $\hat{\boldsymbol{\theta}} \in \mathbb{R}^{K^{\mathcal{R}}}$.

H. Coordinate Conversion

The range estimates $\hat{\mathbf{r}}$ and the azimuth estimates $\hat{\boldsymbol{\theta}}$ of radar detections are converted to the Cartesian coordinates $(\mathbf{x}_k^{\mathcal{R}}, \mathbf{z}_k^{\mathcal{R}})$.

I. Clustering

As in Section III-E, we apply DBSCAN clustering to $(\mathbf{x}_k^{\mathcal{R}}, \mathbf{z}_k^{\mathcal{R}})$ in order to find clusters of radar detections. Each cluster is a radar target, see Fig. 1. We use the radar detections from the last 5 radar measurement cycles for the clustering to guarantee a high number of radar detections for a reliable estimate of the covariance matrix for each cluster. We use the neighborhood radius $\varepsilon = 0.2$ m and the minimum number of $c_{\mathrm{min}} = 3$ detections for DBSCAN. The results of this clustering are the mean $\mathbf{p}_k^{\mathcal{R}}$ and covariance $\mathbf{C}_k^{\mathcal{R}}$ of radar targets in the (x, z) plane.

Fig. 6(b) shows the result for the scenario in Fig. 3(a). In this case, DBSCAN finds two targets. Both targets originate from the same pedestrian and are due to the large spatial extent of the object in body height. As expected, we do not have any targets from the static radar reflections at the end of the radar signal processing chain.

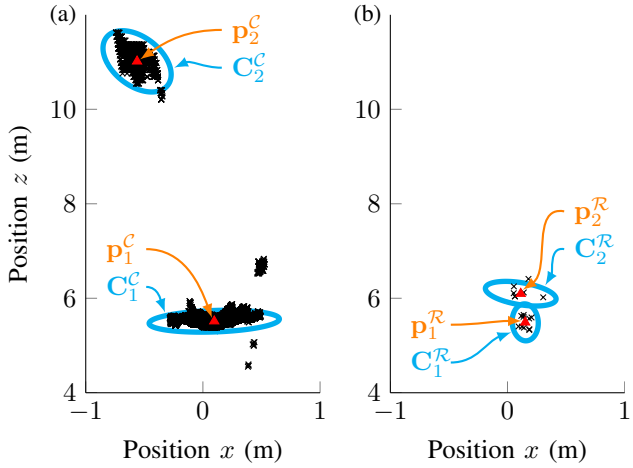

Fig. 6. Detected targets for the scenario in Fig. 3(a). (a) From image. (b) From radar signal.

V. Heterogeneous Sensor Fusion

Fig. 7 gives an overview about the Kalman filter based fusion tracking with camera $(\mathbf{p}_k^{\mathcal{C}}, \mathbf{C}_k^{\mathcal{C}})$ and radar targets $(\mathbf{p}_k^{\mathcal{R}}, \mathbf{C}_k^{\mathcal{R}})$ as input. The targets, separately detected in the image and radar signal, are now further processed using a heterogeneous sensor fusion in order to obtain objects (pedestrians). In the following, we define the camera domain $\mathcal{C} = (\mathbf{p}_k^{\mathcal{C}}, \mathbf{C}_k^{\mathcal{C}})$, the radar domain $\mathcal{R} = (\mathbf{p}_k^{\mathcal{R}}, \mathbf{C}_k^{\mathcal{R}})$, and the fusion domain $\mathcal{F} = (\mathbf{p}_k^{\mathcal{C}}, \mathbf{C}_k^{\mathcal{C}}) \wedge (\mathbf{p}_k^{\mathcal{R}}, \mathbf{C}_k^{\mathcal{R}})$. Here, the fusion domain $\mathcal{F}$ uses only associated camera and radar measurements in each time slot of the duration T_F.

A. Data Alignment

So far, each sensor (camera and radar) provides for each detected target, the mean position and covariance $\left(\mathbf{p}_{kX}^{X}, \mathbf{C}_{kX}^{X}\right)$, $1 \le k \le K^X$, $X \in \{\mathcal{C}, \mathcal{R}\}$. But both sensors run at different frame rates. This makes it necessary to define a common time domain to combine the camera and the radar targets. We introduce a fusion time slot of duration T_F. Based on the timestamps of the targets, all targets of the camera and radar appearing in the same time slot are validated in a measurement-to-measurement data association before fusion and tracking.

B. Measurement-to-Measurement Data Association

For the measurement-to-measurement data association, we make use of ellipsoidal gating by computing the Mahalanobis distance

$$d_{ij}^2 = \mathbf{z}_{ij}^{\mathrm{T}}\,\mathbf{C}_{ij}^{-1}\,\mathbf{z}_{ij} \tag{10}$$

with $\mathbf{z}_{ij} = \mathbf{p}_i - \mathbf{p}_j = [x_i, z_i]^{\mathrm{T}} - [x_j, z_j]^{\mathrm{T}}$ and $\mathbf{C}_{ij} = (\mathbf{C}_i + \mathbf{C}_j)/2$ for all pairs of targets $\left(\mathbf{p}_{kX}^{X}, \mathbf{C}_{kX}^{X}\right)$ in each fusion time slot. We obtain the measurement-to-measurement association set

$$\boldsymbol{\Upsilon} = \left\{d_{ij}^2 \le \gamma(\alpha)\right\} \tag{11}$$

leading to $|\boldsymbol{\Upsilon}| \le (K^{\mathcal{C}} K^{\mathcal{R}})$ associated targets. The threshold $\gamma(\alpha)$ is obtained from the inverse χ^2 cumulative distribution function with 2 degrees of freedom at a significance level α.

C. Measurement-to-Track Data Association

For the measurement-to-track data association, we check how close a target is to an existing track. For this purpose, we again make use of the ellipsoidal gating

$$d^2 = (\mathbf{B}\,\mathbf{x}_{n+1} - \mathbf{p})^{\mathrm{T}} \left(\mathbf{P}^{*}_{n+1}\right)^{-1} (\mathbf{B}\,\mathbf{x}_{n+1} - \mathbf{p}) \tag{12}$$

with the Kalman filter state vector prediction $\mathbf{x}_{n+1}$, the system noise input matrix $\mathbf{B}$, the residual error covariance matrix $\mathbf{P}^{*}_{n+1} = \mathbf{M}\,\mathbf{P}_{n+1}\,\mathbf{M}^{\mathrm{T}} + \mathbf{R}_{n+1}$, the measurement matrix $\mathbf{M}$, the state covariance prediction $\mathbf{P}_{n+1}$, and the latest measurement noise covariance matrix $\mathbf{R}_{n+1}$. We use the James Munkres' variant of the Hungarian assignment algorithm to find the lowest cost regarding association of predicted track state $\mathbf{x}_{n+1}$ and given measurement $\mathbf{p} \in \boldsymbol{\Upsilon}$ in order to find the measurement-to-track association set $\boldsymbol{\Omega}$, the unassigned track set $\boldsymbol{\Psi}$ and the unassigned measurement set $\boldsymbol{\Theta}$ between all tracks and associated targets in $\boldsymbol{\Upsilon}$ [18].

D. Track Management and Track Initialization

All association sets, the measurement-to-track association set $\boldsymbol{\Omega}$, the unassigned track set $\boldsymbol{\Psi}$ and the unassigned measurement set $\boldsymbol{\Theta}$ need to be managed in terms of a track update, creation of a new track or a track deletion. Fur this purpose, we apply a finite state machine (FSM) track management which consists of five different states for each track: *candidate*, *tentative*, *confirmed*, *expiring*, and *delete*. The most reliable tracks are found in *confirmed* state and denote the final object representing a detected pedestrian each as shown in Fig. 1.

The initialization of a new track

$$\mathbf{x}_n = \mathbf{M}^{\mathrm{T}}\,\hat{\mathbf{p}} \in \mathbb{R}^4 \tag{13}$$

requires the estimated position $\hat{\mathbf{p}}$ at the latest time instance [19] and further, to fully initialize the Kalman filter, the initial measurement noise covariance matrix

$$\mathbf{R}_n = \frac{1}{|\boldsymbol{\Theta}|} \sum_{\kappa \in \boldsymbol{\Theta}} \mathbf{C}_{\kappa} \tag{14}$$

and the initial estimation error covariance matrix

$$\mathbf{P}_n = \mathbf{M}^{\mathrm{T}}\,\mathbf{R}_n\,\mathbf{M} + \mathbf{B}\,\mathbf{Q}\,\mathbf{B}^{\mathrm{T}}. \tag{15}$$

E. Kalman Filtering and Prediction

As a human motion model we consider a simple constant velocity linear movement model with a random walking velocity. In this case, the state vector at fusion time slot n is $\mathbf{x}_n = [x(n), v_x(n), z(n), v_z(n)]^{\mathrm{T}} \in \mathbb{R}^4$. The motion model is $\mathbf{x}_{n+1} = \mathbf{A}\,\mathbf{x}_n + \mathbf{B}\,\mathbf{q}$ with the state transition matrix $\mathbf{A}$, system noise input matrix $\mathbf{B}$ and the system noise vector $\mathbf{q} \sim N(\mathbf{0}, \mathbf{Q})$. Note that the incoming measurement in our system only contains the (x, z) position of a target (no usage of velocity estimates). We use a standard linear Kalman filter for tracking.

VI. Tracking Evaluation

A. Experiments

The stationary but portable tracking system is placed in height $y = 1.5$ m and horizontally aligned to the floor. The image and radar signal processing is done offline. Each sensor provides a certain amount of frames per second (FPS), the camera with $\mathrm{FPS}_{\mathcal{C}} \approx 48$ and the radar with $\mathrm{FPS}_{\mathcal{R}} \approx 94$ where each radar measurement cycle provides one radar frame. The time slot duration for fusion is $T_F = 0.05$ s yielding 20 FPS.

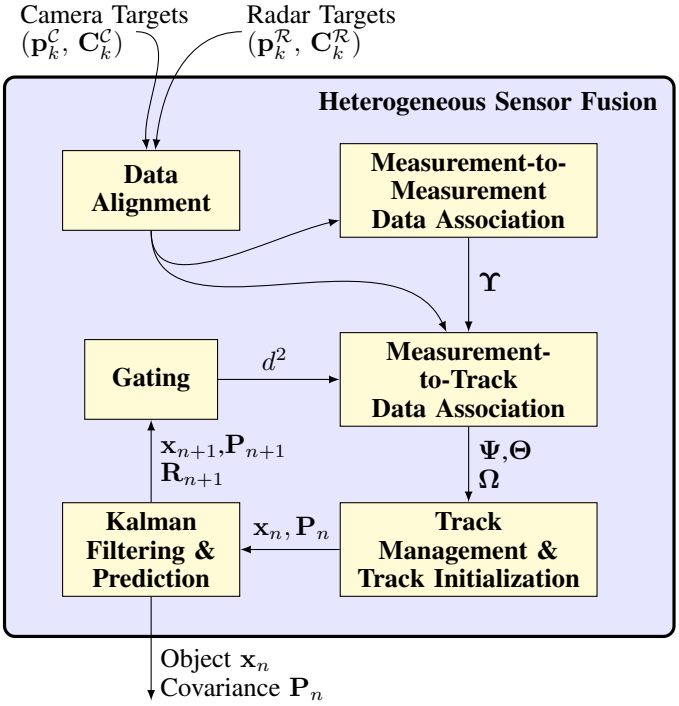

Fig. 7. Overview of the heterogeneous sensor fusion chain.

We collected image and radar signals for four different walking pedestrian scenarios: (a) single pedestrian walking towards and away from the sensors, (b) single pedestrian walking in a circle, (c) single pedestrian walking in a circle around an occluding dummy, (d) single pedestrian walking in a circle and pushing a metal shopping cart, see Fig. 8(a-d). Each scenario has been captured for a duration of 20 seconds.

We compare the results of camera only tracking, radar only tracking with fusion based tracking. In the first two cases, only the targets detected from image or radar signal are used. Each tracking runs independently by using the same Kalman filter and track management. For performance evaluation, only *confirmed* tracks are considered.

For a quantitative performance evaluation, we make use of the 3D-MOTChallenge development kit from [20]. We use the metrics recall, precision, false alarm rate (FAR), the number of mostly tracked (MT)[1], partly tracked (PT)[2] and mostly lost (ML)[3] objects. Furthermore, we calculate the multiple object tracking accuracy

$$\mathrm{MOTA} = 1 - \frac{\sum_{t_F} (\mathrm{FP}_{t_F} + \mathrm{FN}_{t_F} + \mathrm{IDS}_{t_F})}{\sum_{t_F} \mathrm{GT}_{t_F}} \in (-\infty, 1] \tag{16}$$

from the number of false positive (FP) and false negative (FN) detections, the number of identity switches (IDS) and the number of ground truth (GT) objects of all frames with the frame index t_F. Since no real 3D GT data is available, we took the measurement data from both sensors, removed outliers and created manually a trajectory per pedestrian to have quasi GT

[1]A pedestrian is MT if tracked at least 80% of the time being present in consecutive frames.

[2]If tracked between 20% and 80% of the time.

[3]If tracked at most 20% of the time.

for performance evaluation. MOTA = 1 indicates a perfect tracking, while multiple object tracking accuracy (MOTA) can become negative if we have too many FP detections and/or FN detections and IDS.

Additionally, [20] provides the number of interruptions during the tracking of a pedestrian called fragments (FM) and, as a measure of localization precision in the physical 3D space (not in the image), the multiple object tracking precision

$$\text{MOTP}_{\text{3D}} = 1 - \frac{\sum_{i,t_F} d_{i,t_F}}{t_d \sum_{t_F} c_{t_F}} \in [0,1]. \tag{17}$$

The parameter c_{t_F} denotes the number of object matches in frame t_F, d_{i,t_F} is the Euclidean distance between object i to its assigned ground truth object, and $t_d = 1$ m is the distance threshold for pedestrian tracking.

B. Results

Fig. 8 shows the tracking results of our four scenarios and three tracking cases. The tracks provided with track identification (ID), position covariance (ellipse), velocity estimate including direction as an arrow are shown.

In Fig. 8(a) the pedestrian is walking towards and away from the sensors. This is quite difficult to detect in an image since the pedestrian is growing and shrinking in size only. In this case, the track with ID = 2 in Fig. 8(e) does not respond as quickly, as the radar tracking in Fig. 8(i), indicated by the magnitude of the velocity arrow. The radar tracking snapshot shows a good track regarding position and velocity, but it has a larger track ID = 4 indicating several FP detections in the past. The best tradeoff shows the fusion tracking in Fig. 8(m).

Fig. 8(b) shows a scene of a pedestrian walking in a circle, passing the wall closely and walking through blind disparity areas (see Fig. 3(b)). The camera tracking in Fig. 8(f) suffers from wall detections caused by overfitted human upper body detections while the radar tracking in Fig. 8(j) loses the track when the pedestrian moved laterally and in addition directly in front of the wall. The fusion tracking in Fig. 8(n) shows a good tracking result with the lowest track ID and the longest trajectory.

The third scenario in Fig. 8(c) is an extension of the second scenario by placing an occluding dummy in the middle of the scene. The camera tracking in Fig. 8(g) shows two false tracks due to wall detections and since the dummy is detected as well by the human upper body detection. The radar tracking in Fig. 8(k) and the fusion tracking in Fig. 8(o) seem to be robust enough since the dummy is a static object. But the dummy causes fragments in the trajectories due to occlusion of the pedestrian.

Fig. 8(d) plots a scene where a pedestrian is pushing a shopping cart. Such a shopping cart, with its reflecting and semi-transparent wire frame made out of metal, is a challenge in both image and radar signal processing. The camera tracking in Fig. 8(h) tracks both the pedestrian and the shopping cart including 2 false tracks placed at the wall. Considering the radar tracking in Fig. 8(l), it is not clear if the pedestrian is still in focus or the shopping cart. The fusion tracking in 8(p) still provides a single track.

Table I provides the quantitative performance evaluation result for all four walking pedestrian scenarios together. The camera tracking approaches a MOTA of -16.7%, while the radar tracking achieves a MOTA of 38.0%. The best result is obtained via fusion tracking with a MOTA equal to 90.9%. Comparing the MOTP_{3D}, the fusion tracking provides the best result with 90.3%.

So far the quantitative performance evaluation has been done with one pedestrian only since the creation of manually labeled GT data is highly time demanding and difficult, especially in multiple object scenarios where objects occlude each other. In a future study, multiple pedestrian scenarios have to be considered.

VII. Conclusion

We have presented a pedestrian detection and tracking system for indoor environments consisting of a stereo camera, a radar and a fusion unit. Due to specific indoor challenges such as occlusion and rich reflections, either a camera or radar only system may not satisfy detection and tracking performance. Some preliminary experiments have shown that fusing the two heterogeneous sensors can significantly improve the tracking performance. The camera only tracking achieves a MOTA of -16.7%, the radar only tracking reaches an accuracy of 38.0%, while the fusion tracking provides the best MOTA result of 90.9%. Furthermore, the proposed fusion system achieves a MOTP_{3D} of 90.3% compared to 83.4% and 81.8% obtained via camera and radar only tracking, respectively.

The radial velocity from the radar has not been considered in this paper. The next step would be to integrate both scene flow information from the stereo camera and velocity information from the radar into the heterogeneous sensor fusion.

References

[1] A. Simon and J. C. Becker, "Vehicle guidance for an autonomous vehicle," in *Intelligent Transportation Systems, 1999. Proceedings. 1999 IEEE/IEEJ/JSAI International Conference on*, 1999, pp. 429–434.

[2] D. M. Gavrila, M. Kunert, and U. Lages, "A multi-sensor approach for the protection of vulnerable traffic participants the PROTECTOR project," in *Instrumentation and Measurement Technology Conference, 2001. IMTC 2001. Proceedings of the 18th IEEE*, vol. 3, 2001, pp. 2044–2048 vol.3.

[3] Y. Qi, X. Yang, Y. Wang, W. Tan, and W. Hong, "Calibration and fusion of SAR and DEM in deformation monitoring applications," in *EUSAR 2014; 10th European Conference on Synthetic Aperture Radar; Proceedings of*, June 2014, pp. 1–4.

[4] U. Franke, D. Pfeiffer, C. Rabe, C. Knoeppel, M. Enzweiler, F. Stein, and R. G. Herrtwich, "Making Bertha see," in *Computer Vision Workshops (ICCVW), 2013 IEEE International Conference on*, Dec 2013, pp. 214–221.

[5] M. Andres, K. Ishak1, W. Menzel, and H.-L. Bloecher, "Extraction of micro-Doppler signatures using automotive radar sensors," in *Frequenz - Journal of RF-Engineering and Telecommunications*, December 2012, pp. 371–377.

[6] A. Bartsch, F. Fitzek, and R. H. Rasshofer, "Pedestrian recognition using automotive radar sensors," *Advances in Radio Science*, vol. 10, pp. 45–55, 2012. [Online]. Available: http://www.adv-radio-sci.net/10/45/2012/

TABLE I
QUANTITATIVE PERFORMANCE EVALUATION FOR WALKING PEDESTRIAN SCENARIOS

Tracking	Recall (%)	Precision (%)	FAR	GT	MT	PT	ML	FP	FN	IDS	FM	MOTA (%)	MOTP_{3D} (%)
Camera	**94.6**	46.0	1.11	4	**4**	0	0	1617	**79**	6	10	-16.7	83.4
Radar	93.7	63.4	0.54	4	**4**	0	0	789	92	24	10	38.0	81.8
Fusion	91.1	**100.0**	**0.00**	4	3	1	0	**0**	130	**3**	**8**	**90.9**	**90.3**

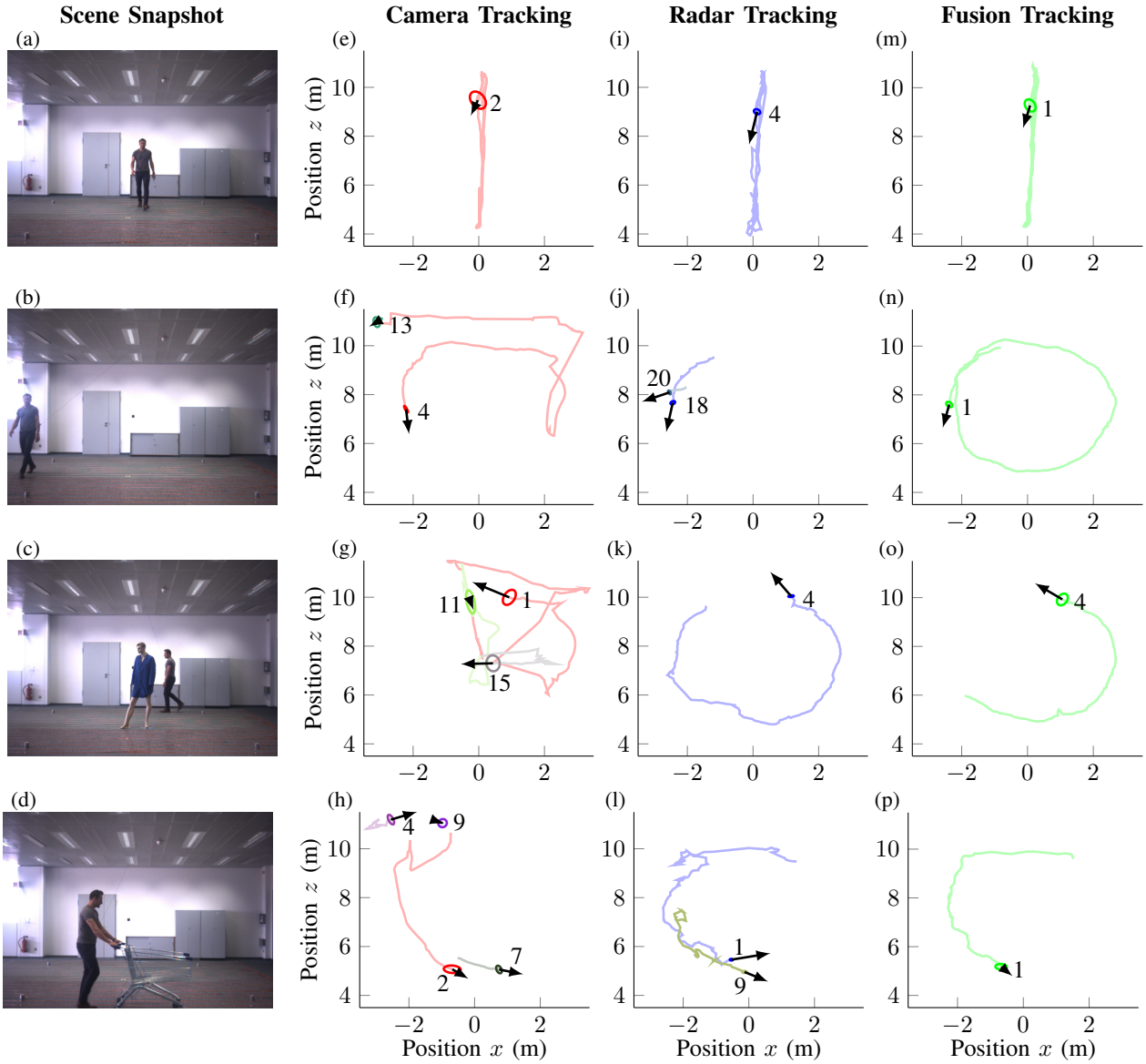

Fig. 8. Tracking snapshots of our four walking pedestrian scenarios

[7] E. Schubert, M. Kunert, A. Frischen, and W. Menzel, "A multi-reflection-point target model for classification of pedestrians by automotive radar," in *European Radar Conference (EuRAD), 2014 11th*, Oct 2014, pp. 181–184.

[8] T. Wagner, R. Feger, and A. Stelzer, "Modification of DBSCAN and application to range/Doppler/DoA measurements for pedestrian recognition with an automotive radar system," in *Radar Conference (EuRAD), 2015 European*, Sept 2015, pp. 269–272.

[9] E. Schubert, F. Meinl, M. Kunert, and W. Menzel, "High resolution automotive radar measurements of vulnerable road users - pedestrians and cyclists," in *Microwaves for Intelligent Mobility (ICMIM), 2015 IEEE MTT-S International Conference on*, April 2015, pp. 1–4.

[10] P. Dollar, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: An evaluation of the state of the art," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 4, pp. 743–761, April 2012.

[11] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, vol. 1, 2001, pp. I–511–I–518 vol.1.

[12] S. Gruenwedel, N. Petrovic, L. Jovanov, J. Nino-Casta-neda, A. Pizurica, and W. Philips, "Efficient foreground detection for real-time surveillance applications," *Electronics Letters*, vol. 49, no. 18, pp. 1143–1145, August 2013.

[13] K. Konolige, *Robotics Research: The Eighth International Symposium*, Y. Shirai and S. Hirose, Eds. London: Springer London, 1998. [Online]. Available: http://dx.doi.org/10.1007/978-1-4471-1580-9_19

[14] M. Ester, H. peter Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise." AAAI Press, 1996, pp. 226–231.

[15] P. Molchanov, "Radar target classification by micro-Doppler contributions." Tampere University of Technology, 2014, PhD-Thesis.

[16] M. Ispir, "Design of moving target indication filters with non-uniform pulse repetition intervals." Department of Electrical and Electronics Engineering, Middle East Technical University Ankara, January 2013, Master-Thesis. [Online]. Available: http://etd.lib.metu.edu.tr/upload/12615361/index.pdf

[17] H. Rohling, "Radar CFAR thresholding in clutter and multiple target situations," *Aerospace and Electronic Systems, IEEE Transactions on*, vol. AES-19, no. 4, pp. 608–621, July 1983.

[18] S. S. Blackman, *Multiple-target tracking with radar applications*, ser. Artech House radar library. Norwood, Mass. Artech House, 1986.

[19] M. Bühren and B. Yang, "Initialization procedure for radar target tracking without object movement constraints," in *Telecommunications, 2007. ITST '07. 7th International Conference on ITS*, June 2007, pp. 1–6.

[20] L. Leal-Taixé, A. Milan, I. Reid, S. Roth, and K. Schindler, "MOTChallenge 2015: Towards a benchmark for multi-target tracking," *ArXiv e-prints*, apr 2015. [Online]. Available: http://adsabs.harvard.edu/abs/2015arXiv150401942L