Person tracking in large public spaces using 3D range sensors

Dražen Brščić, *Member, IEEE,* Takayuki Kanda, *Member, IEEE,* Tetsushi Ikeda and Takahiro Miyashita

***Abstract*—A method for tracking the position, orientation, and height of persons in large public environments is presented. Such information is known to be useful both for understanding their actions, as well as for applications such as human-robot interaction. We use multiple three-dimensional range sensors, mounted above human height to have less occlusion between persons. A computationally simple tracking method is proposed that works on single sensor data and combines multiple sensors so that large areas can be covered with a minimum number of sensors. Moreover it can work with different sensor types and is robust to the imperfect sensor measurements, so it is possible to combine currently available 3D range sensor solutions to achieve tracking in wide public spaces. The method was implemented in a shopping center environment, and it was shown that good tracking performance can be achieved.**

***Index Terms*—person tracking, 3D range sensors.**

I. Introduction

OBSERVING and understanding the actions of persons in public environments is highly valuable. Information on the position and body orientation of persons can be used for the estimation of their attention and intention, and for extracting knowledge on their behavior and social connections. For example, with such knowledge in a museum we can tell which exhibits people consider more attractive, or in a shopping mall we could distinguish whether people are window-shopping or just passing. This information can be very useful for designing and adapting various services for the persons in the space.

Additionally, we believe that in the future robots will be gradually introduced into our daily environments, and a system which can detect and track all the persons in a public space can be very beneficial for the interaction with social robots. The robot could both directly utilize the tracking results or use the analysis of collected long-term data for the design and improvement of services.

In this paper we consider the estimation of the position and body direction of persons in wide public spaces using multiple stationary sensors distributed in the environment. We are interested in a non-wearable solution where persons do not need to attach markers or carry special devices, as we want to observe the unrestricted motion of all the persons in the space. What we aim to achieve is a practical way for tracking of persons in real wide public environments that can reliably and autonomously work for a long period of time. While several solutions for accurate human pose estimation in confined spaces exists, one can only find a limited number of systems capable of doing online long-term large-area tracking of persons. There exist examples of tracking solutions using cameras or laser range finders, but these have also a number of weaknesses for application in large public spaces.

Manuscript received .

The authors are with the Advanced Telecommunication Research Institute International (ATR), Soraku-gun, Kyoto 619-0288, Japan; e-mail: drazen@atr.jp.

This work was supported by JST, CREST.

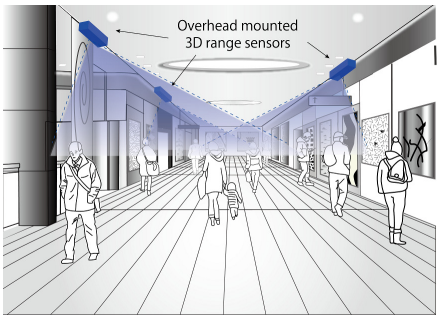


Fig. 1: Illustration of person tracking in public spaces using multiple 3D range sensors

Our proposed setup consists of several overhead mounted three-dimensional range sensors, which provide measurements in the form of distances from the sensor to the objects in the area, Fig. 1. We concentrate on tracking in spaces like museums, shopping malls or galleries, where we expect the persons are mostly walking or standing straight. We also assume that in these spaces the ceiling is high enough to allow a good coverage with overhead mounted sensors.

The requirements for the tracking system in order to be usable is that it has good tracking accuracy, that it can work reliably and robustly on most persons and for a varying number of persons in the covered area, and that it is scalable to wide areas and a large number of sensors. The target test environment in this work is a shopping center, which we believe is representative of the main challenges for people tracking in public spaces.

The paper is organized as follows. In the next section we give a brief overview of the related work. Section III presents the details of the proposed tracking method. After that the evaluation of the tracking is given for two different setups, for a closed laboratory environment in Section IV and for a large public space in a shopping mall in Section V.

II. Related Work

Research on person tracking, in particular using standard RGB cameras, has a long history and the existing literature is very extensive. Detection and tracking of persons using camera images has been surveyed in e.g. [1]–[3]. Recent works present some impressive tracking results, even in very crowded scenes,

such as [4]–[6] to list a few. There also exists a number of computer vision approaches that address the issue of tracking both position and body direction estimation, like [7] or [8].

However, the number of applications of cameras for tracking in a large public space like the environment considered in this work is still small, and they are often also limited to single camera views, relatively simple and static backgrounds, or do not discuss the long-term use of the method with changing illumination. Obtaining a robust and completely autonomous camera based solution for continuous tracking in public spaces appears to still be a hard problem. In addition, although unrelated to the tracking performance, privacy issues can also often pose an obstacle to the introduction of such systems in public areas.

Laser range finders, which use a laser ray to scan the area in a plane, have frequently been successfully applied for human position tracking in wide public spaces [9]–[12]. Although these solutions work very well in practice, generally they are limited only to the estimation of the persons' position. In [11] an extension to body angle estimation was presented but it is sensitive to the occlusion of the view of the person, so it can work reliably only when the number of persons is small. The problem of persons occluding each other in the sensor view is commonly a serious issue for this type of sensors, as it can lead to increased estimation errors due to partial view or complete disappearance of a person from the sensor view.

On the other hand, 3D range sensing has gained more attention recently due to the increased availability of such sensors. Some recent works have applied 3D range sensors for detection and tracking of persons [13]–[17]. In [13] and [14] the problem of detecting pedestrians in complex outdoor scenes was discussed and tracking was applied to the detection results. [15] present methods for explicitly taking into account the possibility that persons are temporarily occluded. Somewhat in the same line of the method proposed in this work, [18] suggest the use of the head and shoulder area to detect humans in 3D range data. An example of the use of 3D range sensors for estimating the person's body orientation was described in [19]. These works all use a horizontal view of the sensors, so similarly to laser range finders they are prone to occlusion between subjects. The use of overhead mounting was proposed in [16] (looking straight down) and [17] (tilted), but they were limited to a single sensor and did not try to detect the body angle from range data.

From the point of view for a long-term implementation in real-world environments range data has an important advantage over RGB data in that the influence of the change of illumination is much smaller and can often be disregarded. Another issue are the changes in the background, which can be easily filtered out in range data by excluding all data outside of the tracking area, but can influence computer vision algorithms. One downside is that range data is generally less informative than RGB when it comes to recognition of persons or distinction between persons and other objects.

Several works also proposed the combination of different sensing modalities, like camera and laser range finders [20], [21] or the use of RGB-D cameras which combine 3D range and camera measurements [22], [23]. However from the practical implementation point of view combining multiple sensor types makes the system more complex, especially for large area tracking. On the other hand, currently available compact RGB-D cameras, like the Microsoft Kinect, still have some weaknesses when applied in public spaces (see discussion in Section III-B).

For these reasons in this work we chose to use 3D range sensors that can be presently found on the market. We propose a simple yet robust estimation technique, which integrates the output from multiple sensors for wide area coverage.

III. Tracking Method

A. Sensor arrangement

As shown in Fig. 1, the considered setup consists of multiple sensors mounted above human height. Based on our personal experience with public space tracking using laser range finders [11], we found that the advantages of using overhead mounting are twofold: (1) it allows a good view of all the persons and minimizes occlusions due to surrounding objects or other persons, which is especially important for situations when the density of persons in the space is high; (2) placing the sensors overhead is essential for their seamless integration in public spaces, as otherwise they can easily become an obstruction and influence the behavior of persons in the space.

The sensors can be mounted so that they either face straight down or at an angle (we refer to this as *observation angle*, where facing straight down corresponds to 0^o). A larger observation angle allows wider coverage with one sensor, but it also increases the possibility of occlusions and incomplete views, so the sensor placement is a matter of compromise between these conflicting factors.

B. Choice of sensors

The sensor choice and the tracking method are both strongly dependent on the characteristics of the 3D range sensors that are used. The three representative types of 3D range sensors currently available on the market are: structured light cameras, time-of-flight cameras and multi-layer laser scanners.

Structured light cameras This type of sensor measures range based on camera views of a projected light pattern. From the point of view of installation in wide public spaces, these sensors have a relatively small usable range: the maximum range at which correct and stable measurements can still be obtained is around 5 meters. The measurements are accurate with low noise, especially for close ranges. But they suffer from a strong influence of external light and from interference between sensors, which can cause missing measurements. The number of missing measurements increases with distance, especially for dark objects like dark hair. The current market price of this type of sensor is very affordable. In this work we use Microsoft Kinect[1] and Asus Xtion PRO[2] models. (The Kinect is actually an RGB-D camera, i.e. it provides both range measurements and RGB camera images, but in this work we use only its range sensing functionality.)

[1] http://www.xbox.com/en-US/kinect

[2] http://www.asus.com/Multimedia/Motion_Sensor/Xtion_PRO/

Time-of-Flight cameras ToF cameras measure the time a projected light needs to travel to an object in order to determine the distance (here we consider ToF cameras that use LEDs as light source). Their usable range is similar to those of structured light cameras, but they in general have much noisier measurements, with the level of noise increasing with the distance. There is also a strong interference between sensors which can result in unstable distance measurements. However, the influence of external light for some of the commercially available ToF cameras is much smaller. The price is in the mid-low range, usually several tens of times more expensive than the available structured light cameras. The sensor used here is Panasonic D-IMager EKL3105[3].

Multi-layer laser scanners These sensors use multiple laser scanning units rotating together in order to obtain a 3D range measurement. The maximum measurement distance for these sensors is much larger than for above types. They are also very accurate with low levels of noise, and virtually no influence from external light or other sensors. Their main weakness is the low resolution of the measurements in the vertical direction as defined by the number of sensing units, which gives a limit on the range up to which they can be used for tracking. The cost of the sensor is considerably higher then for the above 2 types. In the implementation we use the Velodyne HDL-32E[4].

Since all types have both strong and weak points, the decision on which sensor to use will depend on the environment where we wish to do the tracking. It will often be necessary to combine multiple types, as for example in our target environment presented in Section V.

There exist other sensor types that can give 3D range measurements, for example stereo cameras or flash LIDARs (which measure distance based on the ToF principle using laser pulses). However while for the former the range where accurate measurements can be achieved is usually fairly limited, for the latter type the cost of the commercially available solutions at this moment is still prohibitively high for this application.

C. Method overview

The main issues in the considered setup are the large number of sensors that need to be processed and the limitations of the available sensors. We therefore pose the following key requirements on the tracking method: (1) Robust toward measurement imperfections; and (2) Works with single-sensor partial views.

Due to the proposed overhead mounting the persons can be quite far from the sensors, which means they will be used in a range where their weaknesses are amplified. As explained in Section III-B, depending on the sensor type that is being used, the main issues when using 3D range sensors can be: high level of noise, low measurement resolution, influence of external light, and interference with other sensors. The tracking method therefore needs to be robust to noise and missing or sparse measurements. It must also reliably work for a large variety of persons, independent of traits such as hairstyle or clothing.

[3]http://www2.panasonic.biz/es/densetsu/device/3DImageSensor/en/

[4]http://velodynelidar.com/lidar/hdlproducts/hdl32e.aspx

A second requirement is that it should work with single sensor data which only gives a partial view of a person. There are two reasons for this: larger coverage and separate processing. One advantage of using single sensor coverage over multi-sensor full view is that it allows the covering of a much larger space with the same number of sensors, and is hence much more cost effective. But even more importantly, such arrangement is indispensable if we want to avoid mutual interference between sensors, which is characteristic of some of the 3D range sensor types, as explained before. We arrange the sensors in such a way so that they cover separate areas, with partial overlap for a smooth transition of tracking between the areas.

Finally, we choose to separately process the data from each sensor and then combine the processed data. This division makes it possible to have a distribution of processing tasks to several computers, and thus makes the method more scalable for use in large environments. Combining of raw range measurements would be possible in areas where multiple sensor views overlap, but this would also make the processing more interdependent and therefore less scalable.

In light of the above requirements, we propose a method that consists of two independent steps: (1) estimation of position, height and body direction of persons from range data, which is done independently for every sensor; (2) fusion of estimation results from all sensors. In the first step we use a heuristic technique which is designed to be robust to both imperfect measurement data and different person shapes. In the second step the estimates from all sensors are fused using Bayesian filtering.

Figure 2 shows the whole processing pipeline of the proposed method. The details of the single sensor estimation are given in Section III-E whereas tracking using multiple sensors is described in Section III-F.

D. Sensor calibration

In order to use multiple sensors together they need to be calibrated beforehand, i.e. their 3D pose in a common coordinate system needs to be determined. This is crucial for the correct functioning of the tracking. In the experimental room environment in Section IV for each sensor we used a calibration procedure based on markers - flat surfaces on poles - which were placed in the field of view of the sensor. As the position of the surfaces in the coordinate system of the sensor can be easily extracted from range data, by putting the markers on predetermined known locations and using on the known pole heights all 6 sensor pose parameters (3 for position and 3 for angles) are easily calculated.

However in the shopping center setup (Section V) this procedure was not applicable since the previously discussed measurement imperfections did not allow for a precise detection of the markers, and also the measurement of exact marker position in the environment was difficult. We therefore relied on a manual calibration of the sensors, which consisted in visually examining the 3D data. We used a two step procedure: (1) adjusting the height, observation angle and sensor tilt so that floor becomes flat and height of a person is the same

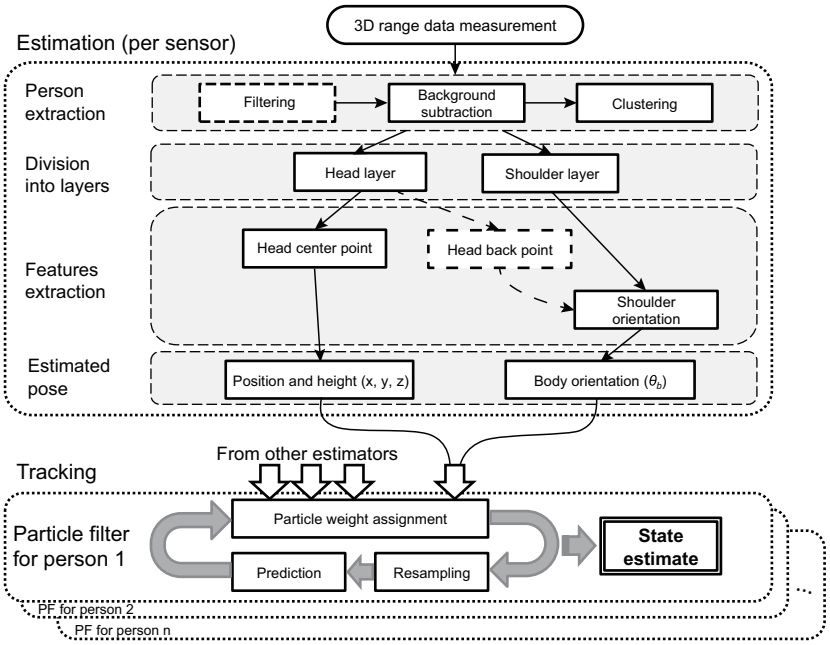


Fig. 2: Pipeline of the tracking method

in the whole tracking area; (2) set the other 3 parameters by comparing and matching the measurements of neighboring sensors as well as to a plan of the environment. This was a quite laborious task, but it only needed to be done once. Obviously this manual procedure introduces an additional calibration error that affects the accuracy of the tracker.

E. Estimation of position and orientation from single sensor data

1) Person extraction: background subtraction and clustering: A 3D range sensor returns measurements in the form of an array, where each value (pixel) corresponds to the distance from the sensor to objects. At first, for noisy range data, it is beneficial to filter out outlier points to prevent their influence on the estimate. This is done by checking the range difference to the neighboring pixels and discarding the ones for which this difference is too large.

Next, we extract the parts of the measurements belonging to each observed person. First, as the sensors are on fixed positions, it is easy to separate the points belonging to moving objects (i.e. persons) from the rest of the scan using background subtraction. The statistic of the"background" measurement – the measurement when there are no moving objects – is learned beforehand and later continuously updated to handle changes in the background. For each pixel the mean range value and the standard deviation are calculated. New measurements are compared with the background and the pixels for which the range difference is larger than 5 standard deviations are extracted.

The extracted points are then converted to 3D points ((x, y, z), with z being the height) and clustered in order to separate points belonging to different persons. The clustering is done point by point, starting from the highest one (i.e. with descending z) to make use of the fact that in general there is better separation between persons in the higher layers where the heads are. A point is assigned to an existing cluster that contains the closest point (based on 3D Euclidean distance). In case the distance to the closest point is larger than a threshold (set to 0.3m in the implementation) a new cluster containing only the examined point is created. This is repeated until all points have been processed. After that we additionally merge clusters that are very close to each other and remove clusters that have a very small number of points. To prevent the merging of clusters of persons standing close or passing nearby, we check the shape of the resulting cluster by calculating the dispersion of cluster points in the (x, y) plane. The clusters are not merged if the obtained standard deviation of the points is larger than a threshold, empiricaly set to 250 mm. We found that this successfully separates persons into different clusters, apart from some difficult situations like children staying very close to parents or being carried in arms.

2) Extraction of head and shoulder regions: From each cluster we extract the points belonging to the top of the head and shoulders. The reason for using these two regions is that due to the overhead positioning of the sensors they are in general always visible, at least partially. Moreover one can expect that the influence of specific hairstyle, facial geometry or clothing on the shape will typically be lower than for other parts of the body.

The extraction procedure we use is based on the division of the cluster points vertically into layers, with the top layer starting at the highest point in the cluster. We used layers of fixed height (10 cm in the experiments). For most occasions when a person is standing straight with hands down, the topmost layer will correspond to the head-top region. However, to make the decision more robust, particularly with respect to noisy points above the head, we also examine the number of points in the layers. Looking at the first two layers we chose

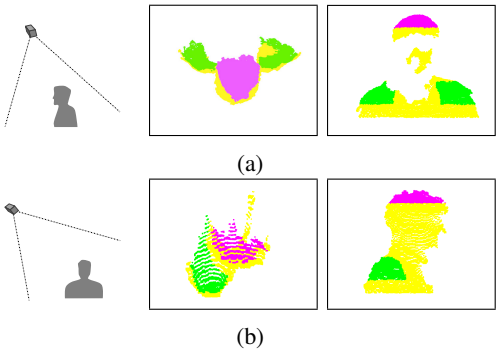


Fig. 3: Example 3D scans of a person with extracted head layer and shoulder layer: (a) small observation angle, human facing towards the sensor; (b) larger observation angle, person facing sideways (top-down and side views are shown).

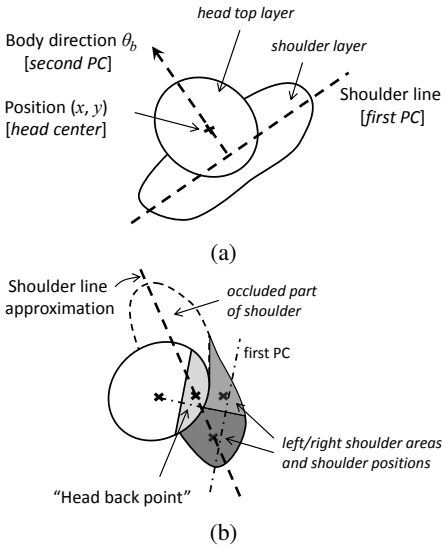


Fig. 4: Estimation of position, body and head angle: (a) Shoulders are not occluded; (b) Correction of shoulder line in case of occlusion. See text for details.

the one with most points as head top layer. The assumption is that the layers with more points contain surfaces which point upwards, in this case the head top.

The shoulder layer is then defined as the layer in the expected range of 2 to 4 layers away from the head top layer, which has the largest number of points, corresponding to the upward facing surface of the shoulders. Finally, the points in the shoulder layer which are close to the head (the (x,y) distance to the head center is small; the head center calculation as explained later) are discarded, to avoid the inclusion of points from the chin or hair, as they could influence the subsequent calculations.

A more direct approach to extract the head-top and shoulder regions would be e.g. to use normals to find flat surfaces. However we found it very hard to obtain a stable calculation of normals from the noisy and incomplete measurements that were obtained from the sensors in our applications. The simplified layer extraction approach presented here generally produced much more consistent results.

In our tests this method gave robust extraction of head and shoulder areas in almost all cases. The most distinctive problem occurs when persons raise their hands high, but this situation is rare.

Fig. 3 shows some typical shapes of extracted clusters containing the upper torso and head of a person, with the extracted head and shoulder areas highlighted. As can be seen from the figure, compared to other body parts like the face, the head and the shoulder areas are more clearly visible for these poses. The pose on the right illustrates common issues when the sensor observation angle is large: (1) self-occlusion: one of the shoulders can be occluded by the head; (2) partial view: often not all points are visible, especially for larger observation angles – note in the figure how only about half of the head area points are seen. The estimation method presented next is designed in such a way so that it is robust to these issues.

3) Position, height and body orientation estimation: We define the human position, height and body direction as follows, see Fig. 4a:

- the *person's position* (x, y) and *height* z are given by the position of the center of the head top area;
- the *body direction* θ_b is the angle of the vector perpendicular to the line connecting the shoulders and directed towards the front of the person;

The head 3D position is calculated as the mean position of the points (x^H, y^H, z^H) belonging to the head top layer and this is used as an estimate of the position and height of the person:

$$(\hat{x}, \hat{y}) = \left(\mu_x^H, \mu_y^H\right) = \left(\frac{1}{N_H}\sum_{i=1}^{N_H} x_i^H, \frac{1}{N_H}\sum_{i=1}^{N_H} y_i^H\right), \quad (1)$$

$$\hat{z} = \mu_z^H = \frac{1}{N_H}\sum_{i=1}^{N_H} z_i^H, \quad (2)$$

where N_H is the number of points in the head layer.

This calculation will obviously slightly underestimate the person's height, but compared to a method such as using the highest point in the layer it proved to be more stable with respect to noise and outliers.

The points (x^S, y^S, z^S) in the shoulder layer are used to calculate the body angle. This is done by applying principal component analysis (PCA) to the ground projection of shoulder layer points (i.e. only the (x, y) values). When the shoulders are clearly visible the first principal component corresponds to the line of the largest extension of points – the shoulder line connecting the left and right shoulder. The line perpendicular to the shoulder line (i.e. the second principal component) defines the line of the body direction. To find out the actual body angle we still need to distinguish the person's back from the front. For that we make use of the fact, which was also confirmed in our tests, that for most persons and typical poses the head center is shifted to the front of the shoulder line (or in other words the projection of the head center estimate $(\hat{x}, \hat{y})$ on the second PC is positive).

$$\hat{\theta}_b = \begin{cases} \angle(\nu_2^S), & \text{if } \left[\hat{x} - \mu_x^S, \hat{y} - \mu_y^S\right] \cdot \nu_2^S \geq 0, \\ \angle(-\nu_2^S), & \text{otherwise,} \end{cases} \quad (3)$$

where ν_2^S stands for the second principal component vector, and (μ_x^S, μ_y^S) is the mean position of the points in the shoulder area.

4) Case of one occluded shoulder: When a person is facing sideways from the sensor often one of the shoulders can be occluded (c.f. Fig. 3), in which case calculating the body angle based on the PCA of the shoulder area points will typically give a biased result. In these cases we use an approximate correction procedure, see Fig. 4b.

First we check if an occlusion happened. After computing the PCA and defining the body front, we divide the shoulder points into the left and right shoulder areas. The body direction line, i.e. the line aligned with the second PC and passing through the head center, is used as the border between the areas. Next, the positions of the left and right shoulders are calculated as mean values of the points in the left and right shoulder areas. We then use two criteria to check if occlusion of one of the shoulders happened, namely: (1) the ratio of number of points on the left and right shoulder areas, which is useful for detecting partial occlusions of one of the shoulders; (2) position of shoulders with respect to the head center, which can indicate a complete occlusion.

If either of these situations is detected we use an approximation for the body angle. It consists of calculating first the "head back point", which is defined as the mean of 1/4 of head top area points which are the most in the back, Fig. 4, where the back is defined with respect to the obtained PCA body angle estimate. Mathematically the calculation of the head back point position (x_{HB} , y_{HB}) can be written as:

$$(x_{HB}, y_{HB}) = \left(\frac{1}{N_H} \sum_j x_j^H, \frac{1}{N_H} \sum_j y_j^H \right), \quad (4)$$
$$\text{s.t. } j \in [1, N_H] : \left[x_j^H - \mu_x^S, y_j^H - \mu_y^S\right] \cdot \nu_2^S < d_{HB}$$

The last expression analyzes the projection of the points in the head top layer on the second principal component of the shoulder area ν_2^S, where d_{HB} is a threshold value defined in such a way so that the number of points in the back area equals 25% of the total number of head area points.

The shoulder line is then set to the line connecting the visible shoulder (the shoulder closer to the sensor) and the head back point, as shown in Fig. 4b, with the body orientation being perpendicular to this line.

This approximation and the chosen number of points to be taken for the head back point calculation are a result of tests with multiple persons and with different sensor poses. In our experiments, for the cases when one shoulder is occluded this corrected calculation for the shoulder direction proved to be more accurate than direct PCA estimation. The value of 25% gave a good estimate while at the same time being robust to noise, partial views of the head, and different head shapes and poses.

F. Multi-sensor Tracking

As explained above, from each sensor s we get an estimate containing the person position, height and body orientation, $\mathbf{z}_s = [\hat{x}, \hat{y}, \hat{z}, \hat{\theta}_b]_s$, for each person detected by that sensor. In order to obtain large area smooth and continuous tracking we combine the estimation results from multiple sensors and fuse them centrally using a set of Sequential importance resampling (SIR) particle filters [24].

One particle filter is used for each tracked person. The state of each particle is given by the position, height and body angle, with the addition of speed v and movement direction angle θ_m of the person.

In the prediction step first for each particle a prediction of the velocity v', motion angle θ'_m and body angle θ'_b states is obtained by adding zero mean Gaussian noise to their previous value. See Table I for the used noise variance parameters: for velocity σ_v^p, motion angle σ_m^p, and body angle σ_b^p (the superscript p stands for "prediction step"). The predicted position is then calculated based on the predicted speed and motion angle: (x', y') = ($x + v'T\cos(\theta'_m)$, $y + v'T\sin(\theta'_m)$), with T being the update time of the filter. The height is estimated separately, as explained later.

In order to assign a weight to a particle m we define the likelihood of the estimate from sensor s, $p\left(\mathbf{z}_s|m\right)$ as a combination of two independent likelihood models:

$$p\left(\mathbf{z}_s|m\right) = p_{xy}\left(\mathbf{z}_s|m\right) p_b\left(\mathbf{z}_s|m\right). \quad (5)$$

The likelihood for the position is defined with a Gaussian function $p_{xy}\left(\mathbf{z}_s|m\right) \sim \mathcal{N}\left(d_{xy}, \sigma_{xy}^l\right)$, where d_{xy} is the Euclidean distance of the estimated and predicted head position $d_{xy} = \sqrt{[(\hat{x} - x')^2 + (\hat{y} - y')^2]}$ and σ_{xy}^l is the variance parameter (the superscript l indicates "likelihood").

As explained in Section III-E3 we decide the forward direction of the person by looking at the position of head center with respect to the shoulder line. For some persons, e.g. due to the influence of hair style, this assumption might be wrong, in which case the estimated body angle will be opposite from the actual one. To take into account this possibility, the body angle likelihood p_b is modeled as being proportional to a weighted sum of two Gaussians, one centered at the estimated body angle value and the other at the opposite direction:

$$p_b(\mathbf{z}_s|m) \sim w_{b1}\mathcal{N}\left(d_b, \sigma_{b1}^l\right) + w_{b2}\mathcal{N}\left(|\pi - d_b|, \sigma_{b2}^l\right), \quad (6)$$

where $d_b = |\hat{\theta}_b - \theta'_b|$ is the absolute difference between the extracted and predicted body angle (normalized to $[-\pi, \pi]$), w_{b1} and w_{b2} are the weights, and σ_{b1}^l, σ_{b1}^l are the corresponding variance parameters.

In cases when the positions of the head or shoulder are close to the edge of the sensor view, we consider the corresponding estimate unreliable and set the associated likelihood to 1.

Likelihoods (5) are calculated for estimates from each sensor and multiplied to obtain the final particle weights:

$$\omega_m = \prod_s p(\mathbf{z}_s|m) \cdot p_{bm}(m). \quad (7)$$

The last term in (7) $p_{bm}(m)$ gives a slight preference to aligned body and motion angles. This is used to reflect the fact that humans are most likely to walk forward. The benefit of adding this term is that even though the motion angle estimate is not reliable when the person's velocity is low, this keeps it aligned to the most probable direction the person might start walking towards. On the other hand, when the person

TABLE I: Used tracking parameters

Prediction parameters

Parameter	Value
σ_v^p	0.2 m/s^2
σ_m^p	0.2 rad
σ_b^p	0.2 rad

Likelihood parameters

Parameter	Value	Parameter	Value
σ_{xy}^l	0.1 m	w_{bm1}	0.2
w_{b1}	1	w_{bm2}	0.8
w_{b2}	0.2	σ_{bm}^l	0.5 rad
σ_{b1}^l	0.5 rad		
σ_{b2}^l	0.35 rad		

is walking this helps filter out the errors in the body angle estimate. This is defined as a weighted sum of a Gaussian and a constant factor:

$$p_{bm} \sim w_{bm1}\mathcal{N}\left(d_{bm}, \sigma_{bm}^l\right) + w_{bm2}, \tag{8}$$

with weights w_{bm1}, w_{bm2} and variance parameter σ_{bm}^l, where $d_{bm} = |\theta_b' - \theta_m'|$ is the absolute difference between the predicted body and motion angle.

One could also include the height in the particle filter likelihood and prediction calculations. However, this would increase the necessary number of particles, which we wish to keep small in order to lower the computation cost. Instead, an estimate of the height value for each particle is obtained as a weighted sum of its current value z_{old} and the estimated value from the sensor, using the likelihood (5) as weight:

$$z = (1 - p(\mathbf{z}_s|m))z_{old} + p(\mathbf{z}_s|m)\hat{z}_s, \tag{9}$$

This is repeated for all sensors s for which the person is visible.

The result of the tracking is calculated at each step as a weighted mean of the particle states. The particles are then resampled using the systematic resampling algorithm [24].

The parameters that were used in the experimental room environment in the following section are listed in Table I. The values were chosen empirically for that specific setup. Small changes in the parameters have only a minor effect on the performance. Nevertheless for different setups or sensor types it might be necessary to re-tune them. For example, for the shopping center implementation (Section V) we changed the following parameters in order to account for the much larger measurement noise: $\sigma_{xy}^l = 0.2$, $\sigma_{b1}^l = 0.7$ and $\sigma_{b2}^l = 0.5$.

During tracking, if a cluster that is not assigned to any particle filter has been observed for several steps in a row it is identified as new person, with an id and a new particle filter assigned to it. The number of steps to wait until starting the tracking is decided based on the situation: e.g. in the setup in Section IV where the measurements were stable the number is set to 5, while in the public space implementation in Section V it was increased to 10, particularly to attenuate the influence of spurious data due to interference between ToF cameras.

For the person deletion we look at the dispersion of the particles in the particle filter – when no cluster is found close to the tracked person it will cause the particles to spread and after the dispersion becomes large (st.dev. of position is larger than 500 mm) the corresponding tracked person will be deleted. In addition, we use a simple reassignment mechanism: in case a new person appears close to a recently deleted one it will be recognized as same person. This enables long time continuous tracking of the persons (i.e. without losing the person's id), as person deletions can often be caused by the person temporarily exiting the tracking area, occlusions in crowds or too unreliable measurements. This was especially true for the shopping center setup (the continuity evaluation results are given in V-B), whereas for the room setup in Section IV we in general did not have problems with tracking continuity.

For each tracked person the number of necessary particles is estimated using KLD sampling [25], with a minimum number of 50 particles. The maximum filter update rate is fixed to 30 Hz. When there are many sensors and tracked persons, like in the setup in Section V, the update rate can drop due to the high computation load, in which case the filter parameters are scaled appropriately.

G. Remarks on the tracking method

The techniques for the detection and tracking presented here might seem overly simplified. However they have been chosen and tested with one basic requirement in mind: they have to work robustly in the targeted real world setup, where the measurements obtained using currently available 3D range sensors are far from perfect, with issues such as noise, missing data and limited resolution. We have tried a number of other more sophisticated methods, such as extraction and tracking of specific face features like nose or chin, matching a simplified body model, or using different 3D feature descriptors. Even though during testing in the laboratory many of these approaches gave satisfactory results, they turned out to be too sensitive to the measurement imperfections and thus not usable in the shopping center setup. In addition, most of them were computationally more expensive than the presented approach.

Obviously, one consequence of using the presented simplified model is that it does not offer an easy way of distinguishing between persons and other objects in the environment. This will be discussed more later in the evaluation results.

Concerning the computational complexity, for the feature extraction the most computationally intensive part is the point clustering algorithm. As it involves checking the distance of each extracted point with all previously assigned points in a straightforward implementation the complexity grows with the square of the number of extracted points. We use a number of heuristics, like stopping the search when a very close point was found, to speed it up. As the complexity of the tracking is clearly proportional to the total number of particles, the complexity of the calculation of both feature extraction and tracking will increase with the number of tracked persons.

IV. Experimental Room Evaluation

We first validate the method in a smaller indoor setup shown in Fig. 5. In this section we wish to: (a) evaluate the tracking performance under different person densities by using an accurate motion capture system; (b) compare the

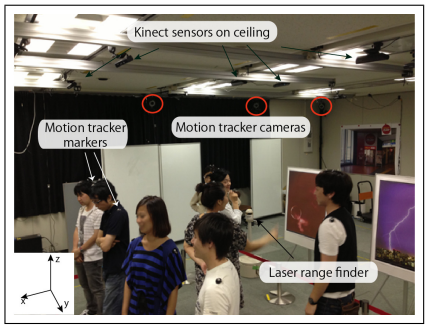


Fig. 5: Experimental room environment.

performance with a laser range finder based system; and (c) show it can be successfully used also in this type of setup as well as work reliably for different persons.

We set up multiple Kinect cameras mounted on the ceiling covering the whole room, Fig. 5. The sensors were at a height of 2.57m with observation angle of approximately 53 degrees. A Vicon motion capture system[5] was installed in the same space and we used markers on the head and shoulders to obtain the ground truth for the position, body direction and height.

Due to the limited coverage of the motion capture system the evaluation was done in the middle of the room, inside a space of approximately 8m^2 (3.5 x 2.3 meters). For the tracking we only used 5 Kinect sensors which completely covered that area. For comparison to a different type of tracking system, we additionally put two laser range finders in opposite corners of the tracking area and independently tracked the persons using the system from [11].

To ensure that the participants behave in a natural way we set up the environment like a photography exhibition (Fig. 5). The participants were asked to move in the space and examine the photos. In total 16 subjects participated in the experiment. We tested the tracking under 3 different conditions: with 2, 4, or 8 participants doing the experiment at the same time (person densities 0.25, 0.5 and 1 person/m^2). For each condition we took multiple trials with different subjects.

The tracking result using the 3D range sensor system for one of the participants is shown in Fig. 6. As seen from the figure, the estimated values follow closely the motion capture system output. The result is smooth with no noticeable problem in the transition between sensors. (For a video of the tracking refer to the supplementary material.)

Occasionally the estimated body direction reversed 180° from the real body orientation, due to the wrong estimation of the front and back. One such example can be seen in Fig. 6d around 5 sec. In all cases the estimate went back to the correct value in a short time.

From the height values in Fig. 6c it is easy to notice the instants when the tracked person leaned forward to take a better look at the picture. This shows an example how useful information can be obtained from the height. The unusual pose affected slightly the height and body angle estimation accuracy.

The result of the performance of both the LRF based tracker

[5]http://www.vicon.com/

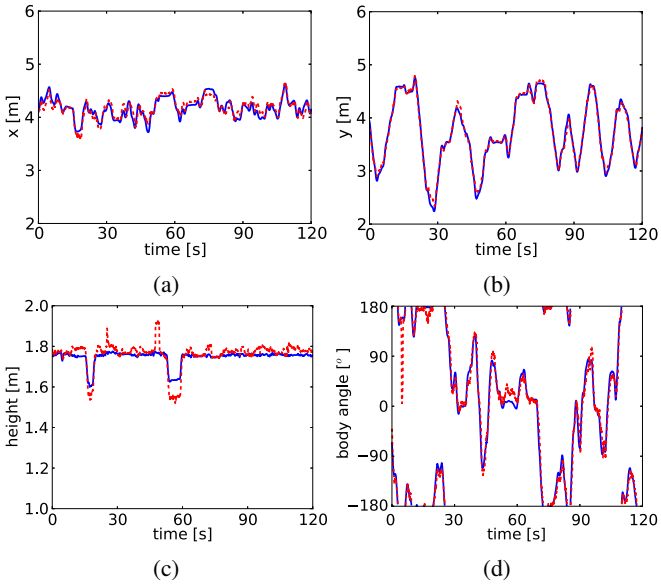


Fig. 6: Sample tracking result – comparison of proposed tracking method (dashed line) with Vicon motion capture system output (solid line): (a) x position, (b) y position; (c) height; (d) body angle.

TABLE II: Room setup evaluation of the tracking with laser range finders and 3D range sensors

Number of persons in area:		2	4	8
LRF	MOTP [mm]	95.17	116.79	124.79
	MOTA [%]	99.83	99.55	97.76
	- miss [%]	0.12	0.26	0.80
	- false pos. [%]	0.05	0.19	1.39
	- ID changes	0	0	29
3D	MOTP [mm]	74.48	82.50	73.60
	height MAE [mm]	26.20	23.95	23.40
	body angle MAE [^{o}]	16.20	21.38	26.57
	MOTA [%]	99.94	99.97	99.88
	- miss [%]	0.04	0.02	0.11
	- false pos. [%]	0.02	0.00	0.00
	- ID changes	1	2	2

and the tracking using multiple Kinects is summarized in Table II. The evaluation was done using the CLEAR MOT metrics [26], which evaluates the multiple object tracking precision (MOTP) as the average error in the estimated position, and the accuracy (MOTA), which takes into account missed persons, false positives and changes in the ID. In addition, for the 3D range sensor system we estimated the mean absolute error (MAE) of the body angle and height estimates.

The results show that with higher density of persons the performance of the laser range finder based tracker gradually decreases. While the precision is around 10cm and the accuracy is close to 100% for 2 persons, the performance degrades with the number of persons, as the number of partial or complete occlusions of persons in the sensor view increases. In the 8 person condition at many instants part of the subjects were completely occluded, which resulted in a drop in the accuracy.

The 3D range sensor system proved to be more robust to the change in person density, with the position and height error as well as MOTA value all showing no significant difference between conditions. However, a gradual increase of the body error angle estimate can be noticed, which can be explained by the increased occlusions in the shoulder area. The tracking precision of the position is also slightly better than for the LRF tracker.

These result show that the proposed system can be successfully used in this setup and that it outperforms a LRF based system. However, note that this environment is quite different from our target shopping center environment presented in the next section. Here there is almost no external light influence and the ceilings are low so it was possible to make use of the Kinect sensors in the measurement range where they have relatively low noise and high accuracy.

V. Shopping Mall Installation

In the following we present the setup and results of the tracking system implementation in a part of the "ATC" business and shopping center in the bay area of Osaka, Japan. We first show the system design for a the real world environment and analyze the achieved performance in terms of tracking accuracy. Second, we provide an example where the obtained large amount of continuous tracking data can provide useful knowledge.

A video of the tracking system in operation can be found in the supplementary materials. Moreover, samples of both the raw sensor data and tracking results are made freely available on our website [27].

A. System setup

A map of the space with the sensor arrangement is shown in Fig. 7. In the west part of the covered area there is a large square, which connects to a long corridor in the east area with several shops on the side. The corridor leads to the train station and to the rest of the shopping center. Going north from the square it is possible to access escalators and elevators which lead to shops, offices and parking lots. Additionally, the corridor on the west leads to a ferry terminal. Once a day a ferry leaves from there.

Over the whole length of the area to the south there is a large window extending from the ceiling to the floor. This lets in much of the external light and gives very uneven lighting conditions during the day.

The tracking system is realized as a combination of different 3D range sensor types. The choice of sensors was based on a compromise between their characteristics (see Section III-B) and the specific requirements for coverage. As the main sensor for close range sensing (e.g. in the corridor) we chose the Panasonic D-IMager, because of its robustness to external light. In total 36 sensors of this type were installed. Due to the operating principle of the sensor, when two sensors using the same operating frequency observe the same space a strong interference occurs. As this sensor model uses 3 operating frequencies it was not possible to obtain a complete coverage of the whole area. Therefore in addition we also

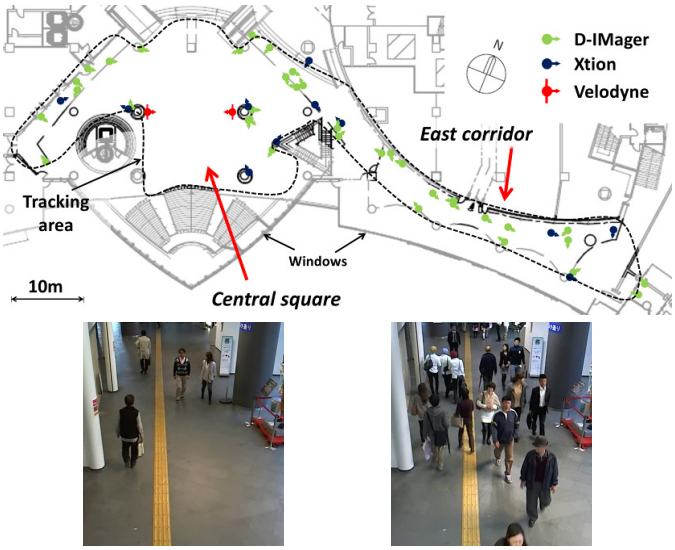


Fig. 7: Tracking area and sensor setup in ATC shopping mall. The dashed line shows the border of the area covered by the sensors. The photos below show the corridor area in the afternoon on a typical weekday (left) and weekend (right).

use 11 ASUS Xtion PRO structured light cameras to cover the remaining parts. Because of their sensitivity to external light we tried to avoid putting them in areas where there was direct sunlight. Finally, the measurement range of both of these sensors was too limited to be used in the square area, so we covered that space with 2 Velodyne HDL-32E rotating laser scanners. The D-IMager and Xtion PRO sensors were mounted on approximately 4 meters height with an observation angle between 40° and 50°, whereas Velodyne sensors were mounted on pillars 8 meters above ground at an angle of 54°. The total area of the space covered by the tracking system is around 900 m^2.

The sensors are connected using USB extensions to a control room. A total of 43 PCs are used to receive the sensor data and forward it to a central processing PC (Intel core i7 CPU, 3.2 GHz, 6 cores, 32 GB RAM), which performs all the processing, including feature estimation and tracking algorithm. We opted to keep all processing centralized as this makes the implementation simpler, and since we plan to use the system for several years it also makes it more reliable and easy to maintain. The implementation is done in Java. The system is able to process the data online with 40 Hz when the number of persons is low, which falls to around 10 Hz when the number of tracked persons approaches 150.

The system started operating in July 2012 and we have been regularly using it to gather pedestrian data and perform experiments with robots.

B. Evaluation of tracking accuracy

Since we did not have a ground truth for the tracking as we did for the experimental room setup, here we focus on the analysis of the tracking accuracy, for which we used manually labeled data. As in the previous section, we use the CLEAR MOT metrics for evaluation. For the labeling we used

TABLE III: Evaluation of shopping center tracking accuracy

	Weekday	Weekend	Combined
Nr. of labeled persons	102	305	407
MOTA [%]	98.63	93.21	94.47
- miss [%]	0.92	4.42	3.6
- false pos. [%]	0.19	1.83	1.46
- ID changes	13	85	98

tracking data from two days in November, one weekday and one weekend day. Four one-minute periods were taken from each day (at 10:00, 13:00, 16:00, and 19:00) and all persons that were inside the tracking area during that time were labeled – a total of 407 persons. In addition to sensors there are also 16 cameras that cover the whole tracking area, so the labeling was done by comparing the view from the cameras and the result of tracking. We asked the labeler to be as precise as possible in evaluating the results, but because of the manual labeling the obtained evaluation result can only be considered approximate. The result of the accuracy evaluation is shown in Table III.

The system performs quite well on weekdays when the density of persons in the space is low. The main source of error are temporarily lost tracks, which happened especially in areas where the coverage was not very good, such as in parts that are far away from all observing sensors. (Even though we planned the setup carefully, due to the interference and noise it was hard to obtain even coverage throughout the tracking area.) In most cases the system recovered the track, but there was also a number of ID changes.

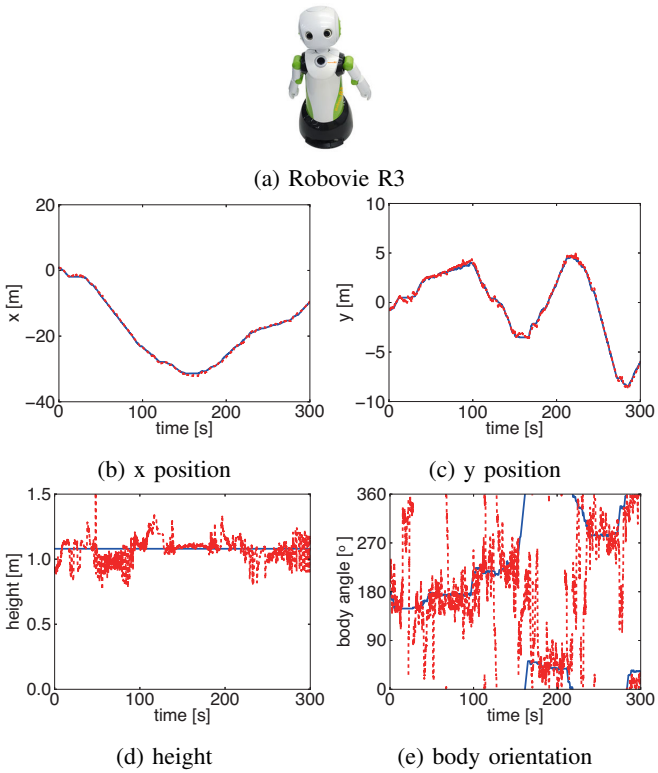
In the evaluation for the weekend most of the analyzed persons were from the afternoon samples (at 13 and 16 hours) when the area is most crowded, so the result gives an insight into the influence of person density on the tracking. There is an increase in all the error measures.

A major cause of the increased number of false positives (about 45% of cases) were objects such as baby carts and suitcases, which are more frequent during the weekend when there are more families and travelers in the center. As we noted before, the tracking method does not explicitly distinguish between human and non-human objects, so these objects are in general being tracked as humans.

In the case of misses and ID changes, more than 60% of them happened on children who are also more frequent during weekends. One reason is that due to their size in particular the small children are more likely to be temporarily lost or occluded by other pedestrians. As children often tend to walk very close to the parents in some cases the system also failed to cluster them correctly.

C. Comparison to robot self-localization

An exact evaluation of the tracking precision is not really possible as no accurate tracking system is available for such a large area, so for an assessment of the tracking precision we present here a comparison of the tracking of a mobile robot with the result of robot self localization. We drove a Robovie R3[6] robot (Fig. 8a) inside the tracking area. It had two laser range finders in front and in the back and it estimated its position and orientation online, using a particle filter based on a grid map of the environment that was built beforehand [25].



(a) Robovie R3

(b) x position

(c) y position

(d) height

(e) body orientation

Fig. 8: Comparison of tracking with robot onboard localization: (a) the robot that was used in the experiment; (b)-(e) result of 3D sensor tracking (dashed line) and robot onboard localization (solid line).

Notice that the result obtained here will necessarily be different from actual person tracking performance, as (1) the tracking was not designed for the robot, whose shape is different from that of a human (Fig. 8a); and (2) the robot's onboard localization is not completely accurate by itself, and the exact error statistics for this environment are not known. Nevertheless, we believe it still provides useful information on the degree of tracking precision that can be obtained with the system.

The result of the comparison during a sample 5 minute run is shown in Fig. 8. The position tracking, Figs. 8b and 8c, was smooth with no large errors. The average position difference between the results was 328 mm. The result in Fig. 8d compares the height estimate from the tracking system with the real robot height, which is 1080mm. The mean absolute difference of the height estimates was 75.3mm. There are also some jumps between sensors, but the accuracy is sufficient e.g. for distinguishing adults and children.

Comparing the body angle estimates is perhaps less useful for the evaluation, since the shape of the robot obviously does not fit with the assumptions we made for people. Nevertheless for completeness we show it in Fig. 8e. In many instants

[6]http://www.vstone.co.jp/english/products.html

(approx. 18% of the time) the angle was inverted due to a wrong differentiation between front and back. If we exclude the data where the angle inversion happened, the obtained statistic of the difference to the onboard localization result gives the estimate mean absolute difference 24.7^o.

D. Example use of the system: extracting statistics

The tracking system allows us to observe the undisturbed movement of persons inside the covered space for extended periods of time. It is therefore possible to gather knowledge on how the space is being used, how this usage varies with time, what are significant areas and points in the environment, what kind of interactions are occurring in the space, etc. Here we illustrate some example statistics that we were able to extract from the tracking results. The used data was taken in the period between mid-August and mid-September 2012, on 10 days in total. The results are shown in Fig. 9.

Fig. 9a highlights the difference between the number of persons inside the tracking area during the day on both working days and weekends. While on weekends there is a large number of visitors, during the week the space is mostly used by the persons working in the building. The variance of the person number is also much larger on weekends, as it is very dependent on parameters such as events and to a certain extent weather conditions. It is also possible to notice a peak in the number of persons before the ferry departure at 19:00, which corresponds to the persons who waited inside the area before boarding.

Other specific information can also be obtained. One example is the staircase south of the square (Fig. 7), which leads outside to a wide space in front of the building. This space is sometimes used for various events, and this can be readily recognized by analyzing the number of persons that enter or exit the area using the stairs, Fig. 9b. This value is considerably larger on event days than on days without any events.

Figs. 9c-e give an insight into the spatial usage of the observed area. These graphs were obtained from all trajectories during 3 hours per day in the afternoon, on 2 weekdays and 2 weekend days. Fig. 9c shows the distribution of the average density of walking persons. The east corridor is the most used part, with its central section having the largest density which is due to the fact that the corridor is narrowest in that part. Compared to the corridors, the density of persons in the square area is much lower. The analysis of the distribution of the walking directions, Fig. 9d, shows very clearly how the persons tend to walk on the left side of the corridor, as is typical in Japan. A similar result was also shown previously for a different environment in [28].

Additionally, 9e shows the positions in the environment where persons often stopped. As could be expected, the result shows that these stopping positions as well as the corresponding body angles are related to specific features in the environment, like information boards or shops.

E. Discussion and observations

The setup presented here shows many problems typical for tracking in public spaces. First, there exist large differences in illumination, ranging from direct sunlight to dim artificial light. Next, the positions where sensors can be mounted are limited due to factors such as the ceiling height or positions of pillars. In addition, since there are both wide and narrow areas, sensing at different ranges is needed. The arrangement of static objects (the background) is also very variable from day to day, and even during one day. For example there is a number of shops in the tracking area which put products in front the shops and regularly change their arrangement, and occasionally there are also events where objects like tables or chairs are set up inside the area.

These are the reasons why public space tracking has been considered difficult, and solutions based on cameras can only be used in limited situations. The described setup uses a combination of different 3D range sensors to overcome these issues, with very good results. The system can operate fully autonomously except for some special cases. For example, for objects that are brought inside the area for an event we rely on an operator to mark them as background, otherwise the system would track them as humans.

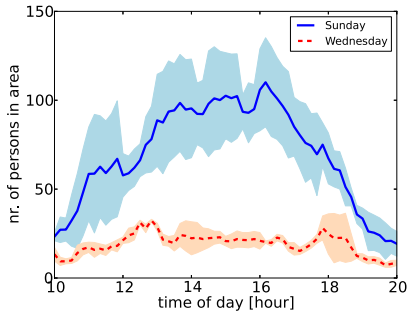
The sensors that were used in the setup could hardly be described as ideal. For example, for the ranges at which we used them, the measurements from a time-of-flight camera are much noisier than in their nominal range, whereas the structured light cameras often have many missing measurements especially when there is much external light. These issues were one of the main limiting factors for achieving good tracking. Different sensor choices could have been possible, for example to use more Velodyne laser scanners, which are very accurate and have long range and low noise. However this would have resulted in a significant and for us prohibitive increase in the cost. With the progress in the development of 3D range sensors we believe that in the near future there will be more devices with good measurement characteristics and affordable prices available on the market, which will make the building of public space tracking systems like this more feasible.

VI. Conclusion

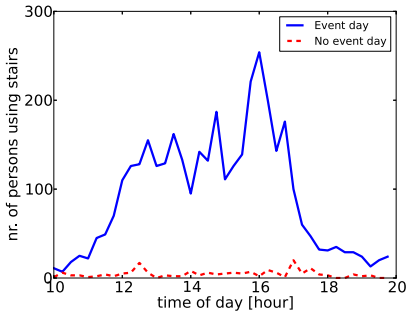
This work considered the use of multiple 3D range sensors for the tracking of persons in a large public space. The sensors are mounted overhead to provide good coverage and minimize the influence of occlusions between persons. We presented a tracking method which allows the real-time estimation of position, height and body angle of all persons inside the space. The proposed method is designed to be simple but robust to imperfect measurements, so that it can be used for wide area coverage of public spaces with multiple 3D range sensors. Evaluation of the method showed that the method performs very well and is more robust to occlusions than a laser range finder based tracker. The shopping center installation demonstrated that reliable and long-term tracking in a wide real environment can be achieved.

Acknowledgments

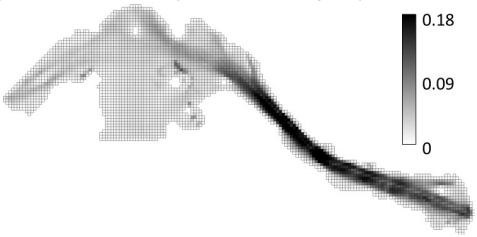
The authors wish to thank Dr. Francesco Zanlungo for his help with the statistical analysis of the collected tracking data.



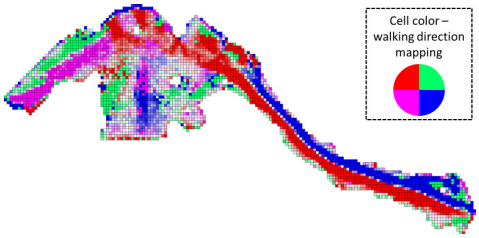
(a) Average value of the total number of persons in the area during weekend (Sunday) and working day (Wednesday).



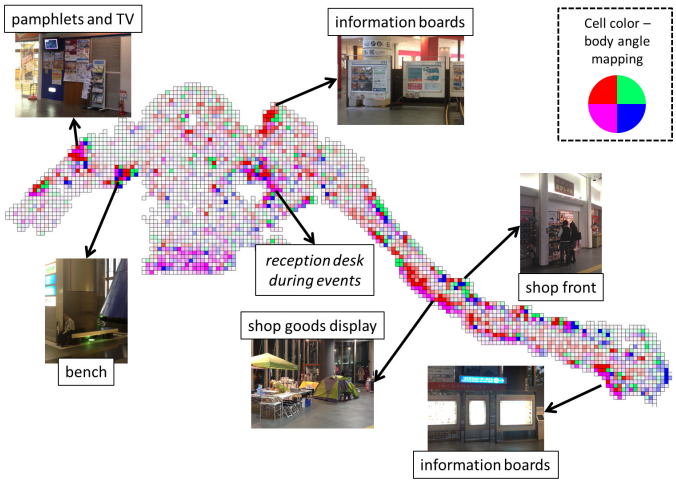
(b) Number of persons that enter/exit from the southern stairs during 15 minutes on a day with and without events.



(c) Grid map of average person density, values are in person/m^2; only moving person data is used (velocity $>$ 0.5 m/s).



(d) Grid map of velocity vector direction (the walking directions are coded by color: moving towards NE - green, SE - blue, SW - magenta; NW - red).



(e) Density map of stopping positions (person velocity $<$ 0.2 m/s) with shown body angles during stopping; the body directions are represented by different colors: looking towards NE - green, SE - blue, SW - magenta; NW - red; the photos show the areas which stand out.

Fig. 9: Example tracked person statistics.

REFERENCES

[1] B. Zhan, D. Monekosso, P. Remagnino, S. Velastin, and L.-Q. Xu, "Crowd analysis: a survey," *Machine Vision and Applications*, vol. 19, no. 5-6, pp. 345–357, Sep. 2008.

[2] T. Moeslund, A. Hilton, and V. Krüger, "A survey of advances in vision-based human motion capture and analysis," *Computer Vision and Image Understanding*, vol. 104, no. 2, pp. 90–126, Nov. 2006.

[3] T. Teixeira, G. Dublon, and A. Savvides, "A survey of human-sensing: Methods for detecting presence, count, location, track, and identity," ENALAB, Yale University, Tech. Rep. 09-2010, Sep. 2010.

[4] G. Shu, A. Dehghan, O. Oreifej, E. Hand, and M. Shah, "Part-based multiple-person tracking with partial occlusion handling," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'12)*, Providence, RI, USA, Jun. 2012, pp. 1815–1821.

[5] I. Ali and M. Dailey, "Multiple human tracking in high-density crowds," in *Advanced Concepts for Intelligent Vision Systems*, ser. Lecture Notes in Computer Science, J. Blanc-Talon, W. Philips, D. Popescu, and P. Scheunders, Eds. Springer Berlin Heidelberg, 2009, vol. 5807, pp. 540–549.

[6] M. Hofmann, D. Wolf, and G. Rigoll, "Hypergraphs for joint multi-view reconstruction and multi-object tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'13)*, Portland, OR, USA, Jun. 2013.

[7] O. Ozturk, T. Yamasaki, and K. Aizawa, "Estimating human body and head orientation change to detect visual attention direction," in *Proceedings of the International Workshop on Gaze Sensing and Interactions*, Queenstown, New Zealand, Nov. 2010.

[8] S. Piérard and M. Van Droogenbroeck, "Estimation of human orientation based on silhouettes and machine learning principles," in *Proceedings of*

the International Conference on Pattern Recognition Applications and Methods (ICPRAM), Vilamoura, Portugal, Feb. 2012, pp. 51–60.

[9] A. Fod, M. Matarić, and G. Sukhatme, "A laser-based people tracker," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA '02)*, Washington DC, USA, May 2002, pp. 3024–3029.

[10] H. Zhao and R. Shibasaki, "A novel system for tracking pedestrians using multiple single-row laser-range scanners," *IEEE Transactions On Systems, Man and Cybernetics – Part A: Systems and Humans*, vol. 35, no. 2, pp. 283–291, 2005.

[11] D. Glas, T. Miyashita, H. Ishiguro, and N. Hagita, "Laser-based tracking of human position and orientation using parametric shape modeling," *Advanced Robotics*, vol. 23, no. 4, pp. 405–428, 2009.

[12] K. Arras, O. Mozos, and W. Burgard, "Using boosted features for the detection of people in 2d range data," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA'07)*, Rome, Italy, May 2007, pp. 3402–3407.

[13] L. Navarro-Serment, C. Mertz, and M. Hebert, "Pedestrian detection and tracking using three-dimensional ladar data," in *Proceedings of the 7th International Conference on Field and Service Robotics*, Cambridge, MA, USA, Jul. 2009.

[14] L. Spinello, M. Luber, and K. Arras, "Tracking people in 3D using a bottom-up top-down detector," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA'11)*, Shanghai, China, May 2011, pp. 1304–1310.

[15] F. Schöler, J. Behley, V. Steinhage, D. Schulz, and A. Cremers, "Person tracking in three-dimensional laser range data with explicit occlusion adaption," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA'11)*, Shanghai, China, May 2011, pp. 1297–1303.

[16] A. Bevilacqua, L. D. Stefano, and P. Azzari, "People tracking using a time-of-flight depth sensor," in *Proceedings of the IEEE International Conference on Video and Signal Based Surveillance*, Sydney, NSW, Australia, Nov. 2006.

[17] D. Hansen, M. Hansen, M. Kirschmeyer, R. Larsen, and D. Silvestre, "Cluster tracking with time-of-flight cameras," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; Workshop on ToF-Camera based Computer Vision*, 2008.

[18] N. Kirchner, A. Alempijevic, and A. Virgona, "Head-to-shoulder signature for person recognition," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA'12)*, Saint Paul, MN, USA, May 2012, pp. 1226–1231.

[19] S. Piérard, D. Leroy, J.-F. Hansen, and M. Van Droogenbroeck, "Estimation of human orientation in images captured with a range camera," in *Advanced Concepts for Intelligent Vision Systems (ACIVS)*, ser. Lecture Notes in Computer Science, vol. 6915. Springer, 2011, pp. 519–530.

[20] J. Cui, H. Zha, H. Zhao, and R. Shibasaki, "Multi-modal tracking of people using laser scanners and video camera," *Image and Vision Computing*, vol. 26, no. 2, pp. 240–252, Feb. 2008.

[21] C. Premebida, O. Ludwig, and U. Nunes, "Lidar and vision-based pedestrian detection system," *Journal of Field Robotics*, vol. 26, no. 9, pp. 696–711, Sep. 2009.

[22] M. Luber, L. Spinello, and K. Arras, "People tracking in rgb-d data with on-line boosted target models," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'11)*, San Francisco, CA, USA, Sep. 2011, pp. 3844–3849.

[23] W. Choi, C. Pantofaru, and S. Savarese, "Detecting and tracking people using an rgb-d camera via multiple detector fusion," in *Workshop on Challenges and Opportunities in Robot Perception (in conjunction with ICCV-11)*, 2011.

[24] A. Doucet, S. Godsill, and C. Andrieu, "On sequential monte carlo sampling methods for bayesian filtering," *Statistics and computing*, vol. 10, no. 3, pp. 197–208, 2000.

[25] S. Thrun, W. Burgard, and D. Fox, *Probabilistic Robotics*. MIT Press, Cambridge, MA, 2005.

[26] B. Keni and S. Rainer, "Evaluating multiple object tracking performance: the CLEAR MOT metrics," *EURASIP Journal on Image and Video Processing*, vol. 2008, 2008.

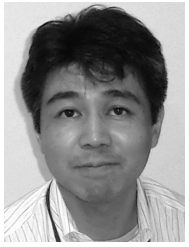
[27] ATC shopping mall dataset. [Online]. Available: http://www.irc.atr.jp/crest2010_HRI/ATC_dataset (accessed Sep 25., 2013)

[28] F. Zanlungo, T. Ikeda, and T. Kanda, "A microscopic "social norm" model to obtain realistic macroscopic velocity and density pedestrian distributions," *PLoS ONE*, vol. 7, no. 12, p. e50720, 2012.

Dražen Brščić (S'06-M'09) received his BSc and MSc degrees in electrical engineering from the University of Zagreb, Croatia, in 2000 and 2004, respectively, and the Dr. Eng. degree in electrical engineering from The University of Tokyo, Japan in 2008. From 2008 to 2010 he worked as a post-doctoral researcher at the Technische Universität München, Germany. In 2011 he joined ATR Intelligent Robotics and Communication Laboratories in Kyoto, Japan, as research scientist. His research interests include person tracking, mobile robotics and human-robot interaction.

Takayuki Kanda (M'04) received the B.Eng, M.Eng, and Ph. D. degrees in computer science from Kyoto University, Kyoto, Japan, in 1998, 2000, and 2003, respectively. From 2000 to 2003, he was an Intern Researcher with the Advanced Telecommunications Research Institute International (ATR) Media Information Science Laboratories. He is currently a Senior Researcher with the ATR Intelligent Robotics and Communication Laboratories, Kyoto, Japan. His research interests include intelligent robotics, humanrobot interaction, and vision-based mobile robots. Dr. Kanda is a member of the Association for Computing Machinery, the Robotics Society of Japan, and the Information Processing Society of Japan.

Tetsushi Ikeda received the M.Eng in information science from Kyoto University, Kyoto, Japan, in 1997. From 1997 to 1999, he was with the Advanced Technology R&D Center at Mitsubishi Electric Corporation. In 2000, he entered the Doctoral Program in the Department of Adaptive Machine Systems, Graduate School of Engineering, Osaka University, which he finished in 2003. From 2003 to 2009, he was with Osaka University, where he worked as a Research Associate and a Research Assistant Professor. He is currently a Researcher at the Intelligent Robotics and Communication Laboratories, Advanced Telecommunications Research Institute International (ATR), Japan.



Takahiro Miyashita received his B.S., M.S., and Ph.D. degree in engineering for computer-controlled machinery from Osaka University, Japan in 1993, 1995, and 2002, respectively. From 1998 to 2000, he was research fellow of the Japan Society for the Promotion of Science. From April to September 2000, he was researcher of Symbiotic Intelligent Group at ERATO Kitano Symbiotic Systems Project of Japan Science and Technology Agency. From October 2000 to July 2002, he was assistant professor at Wakayama University. Now, he is a group leader and a senior research scientist at Intelligent Robotics and Communication Laboratories (IRC), Advanced Telecommunications Research Institute International (ATR). He is also a member of the Robotics Society of Japan, the Japan Society of Mechanical Engineers, the Japanese Society for Artificial Intelligence and the Institute of Electronics, Information and Communication Engineers.