People Detection and Tracking from a Top-View Position Using a Time-of-Flight Camera

Carsten Stahlschmidt, Alexandros Gavriilidis, Jörg Velten,
and Anton Kummert

Faculty of Electrical Engineering and Media Technologies
University of Wuppertal, D-42119 Wuppertal, Germany
{stahlschmidt,gavriilidis,velten,kummert}@uni-wuppertal.de

Abstract. This paper outlines a method for the detection and tracking of people in depth images, acquired with a low-resolution Time-of-Flight (ToF) camera. This depth sensor is placed perpendicular to the ground in order to provide distance information from a top-view position.

With usage of intrinsic and extrinsic camera parameters a ground plane is estimated and compared to the measured distances of the ToF sensor in every pixel. Differences to the expected ground plane define foreground information, which is used as regions of interest (ROIs). These regions are analyzed to distinguish persons from other objects by using a matched filter on the height-segmented depth measurements of each ROI. The proposed method separates crowds into individuals and facilitates a multi-object tracking system based on a Kalman filter.

Experiments have proven the applicability of the system for different crowding scenarios but also revealed inaccuracies of the detection of people in special cases.

Keywords: people detection, top-view, people tracking, time-of-flight, matched filter.

1 Introduction

Surveillance camera systems in public areas become more and more important in order to increase people safety and security.

Particularly, observation of dense crowds got in the focus of research. Every year, crowd disaster occur many times in different areas of the world [10]. Disasters as the Love Parade catastrophe in Duisburg, Germany, 2010, where 21 people died and more than 500 got injured, show the importance of surveillance systems that help to indicate and avoid potentially dangerous situations. Crowd panics often arise in areas where many people accumulate and form a dense crowd [16]. These areas must be analyzed for risks.

Expert reports regarding the Love Parade disaster outline the importance of crowd monitoring systems [10,16] in order to gain knowledge about the people and analyze the crowding situation. Therefore information about the number of people in an area, flow rates or densities of people within a crowd are useful.

A. Dziech and A. Czyżewski (Eds.): MCSS 2013, CCIS 368, pp. 213–223, 2013.

The proposed system in this paper describes a method for the detection of individuals within a dense crowd, which enables the analysis of every person's movement in contrast to the movement within the entire group. With knowledge about individuals, the crowding situation can be characterized.

A ToF camera, which provides depth and gray-scale images of a scene, is mounted perpendicular to the ground for the detection of people. This enables an easier separation and tracking of people, because people do not overlap much in the dimension normal to the ground [9].

Instead of calculating a plan-view projection of the scene from an eye-level or high-angle positioned camera, the sensor is applied directly to the so called "top-view" position. This saves computational costs and reduces loss of depth information for occluded persons from different camera angles.

The used ToF camera is specified in [12]. In this paper, image processing is restricted to the detection and tracking of human persons merely using depth images. Comparable systems also use Kalman filters for people tracking, but the detection of people in plan-view images varies. Idealized shapes for people detection [17] are as well used as Gaussian blobs [2,3] and adaptive templates with a support vector machine to identify people in depth images [8].

The proposed system in this paper is based on usage of a scaled matched filter in combination with height-segmented foreground information. The advantage of this method is the ability to distinguish individual persons from a dense crowd, in contrast to systems where people need to enter the scene separately [2,17].

Thus method enables a multi-object tracking of persons and their differentiated movement analysis in contrast to the behavior of the crowd.

This paper is organized as follows. Section 2 outlines the basic Time-of-Flight principle. In Section 3 the algorithm is described, partitioned in preprocessing, people detection, and people tracking. Section 4 is used to show and discuss experimental results. Finally, Section 5 draws conclusions.

2 Time-of-Flight Camera Principle

Camera-based Time-of-Flight sensors are relatively new camera systems, which provide depth images of a scene at a high frame-rate. Distance information are captured and provided as optical signals, where each pixel on the CMOS imager describes the distance from the camera to the corresponding point in the real world.

As it is outlined in [14], ToF cameras are based on the principle of pulse modulation. Distances are calculated from a light reflecting object to the sensor by measuring the phase delay between the incoming infrared light and a reference signal directly in each pixel. Therefore, most cameras are equipped with active illumination units. The phase of an incoming signal is calculated by

$$\Phi = \arctan\left(\frac{A_1 - A_3}{A_2 - A_4}\right) . \tag{1}$$

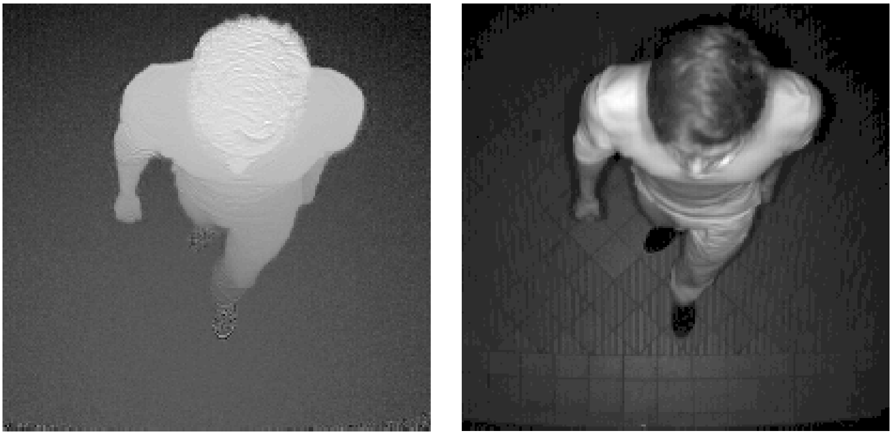
Here, A_1, A_2, A_3, A_4 are four samples of the signal, each shifted by 90°. The distance d is proportional to phase Φ and calculated by using signal frequency f_{mod} and speed of light c by

$$d = \frac{c\Phi}{4\pi f_{mod}}. \tag{2}$$

By measuring d in every pixel of the sensor, a depth signal is generated.

This measurement principle is of importance due to the fact that it defines the maximum distance to be measured without phase ambiguity [7]. This constraint in the maximum distance is depending on modulation frequency f_{mod} and calculated by $d_{max} = \frac{c}{2f_{mod}}$. ToF cameras assume a maximum $\Phi = 2\pi$ that limits every modulation frequency to a d_{max}. For real distances in the scene farther away than d_{max}, this results in a wrong measured distance d. For the proposed system ,the camera was placed below d_{max} in order to avoid phase ambiguity.

Due to the method of illuminating the entire scene with modulated infrared light and measuring the phase delay in every pixel of the PMD sensor individually, rather than using a single scanning laser beam to estimate distances, a much higher frame-rate of gathering depth maps can be accomplished [13]. In contrast to common gray-scale cameras, camera-based ToF sensors measure not only the distances from the camera to an object in the scene, but can also provide gray-scale values of every pixel in an intensity image. In the following sections,



(a) Depth image

(b) Intensity image

Fig. 1. Images provided by ToF camera

$$f(\mathbf{n}) = f(n_t, n_x, n_y) \tag{3}$$

with $\mathbf{n} = (n_t, n_x, n_y)'$ denotes a distance value in pixel (n_x, n_y) at a particular time n_t. $f(\mathbf{n})$ can be interpreted as one depth image from a sequence of depth images, defined by measurements of the ToF camera for $\mathbf{n} \in \mathcal{D}$ with

$$\mathcal{D} = \left\{ \mathbf{n} \in \mathbb{Z}^3 | 0 \leq \mathbf{n} \leq \mathbf{N} \right\}, \tag{4}$$

where $\mathbf{N} = (N_t, N_x, N_y)'$. $\mathbf{N}$ describes the number of points in each dimension. An arbitrarily shaped subset $\mathcal{R} \subseteq I$ of an image, with

$$\mathcal{I} = \left\{ \begin{pmatrix} m_x \\ m_y \end{pmatrix} = \mathbf{m} \in \mathbb{Z}^2 \middle| 0 \leq m_x \leq N_x, 0 \leq m_y \leq N_y \right\} \tag{5}$$

is denoted as region of interest (ROI) within $\mathcal{I}$. In this paper, the signal $f(\mathbf{n})$ describes an $N_x \times N_y$ image containing depth information in N_t frames.

Figure 1 shows the difference between depth and intensity images. Depth images provide range information from camera to real world. Bright pixel indicate reflecting objects located nearer to the camera than dark pixel. Intensity images correspond to $f(\mathbf{n})$ but provide gray-scale information of the scene in each pixel. The corresponding intensity image is of the same size as the depth image.

3 Algorithm

The people detection algorithm is divided into several components regarding preprocessing, people detection and assignment of measurements to tracks as described in the following subsections.

3.1 Preprocessing

Preprocessing of depth data is necessary in order to reduce noise from considerably noisy pixel, estimate regions of interest and segment measured depths into height segments.

Depth measurement errors arise from different sources [4] and affect the posterior height segmentation process. The proposed method uses a straightforward spatial neighborhood filter, sufficient for the system to decrease depth measurement errors. A static ground plane $g(n_x, n_y)$ is defined as described in [5], based on extrinsic camera parameters. Hence, expected distances from camera to ground are specified when no obstacles reflect the camera beam.

The next preprocessing step reduces measured depths $f(\mathbf{n})$ to an image $F(\mathbf{n})$ containing only foreground information, by

$$F(n_t, n_x, n_y) = f(n_t, n_x, n_y) - g(n_x, n_y), \tag{6}$$

where n_t denotes the n_tth image. In other words, measured depths from an empty room lead to an $F(\mathbf{n}) = 0$.

When objects are present, $F(\mathbf{n})$ contains depth values that define the foreground. If one object exists, one associated local region is estimated that defines the ROI of the depth image. In cases of more objects present in $F(\mathbf{n})$, the number of ROIs can alter between one and the number of objects. In cases where several objects are closely spaced and an erosion process fails to differ these, one ROI can contain more than one object.

The depths values of ROIs $\mathcal{R}_i$ of one or more (if overlapping or nearby) objects are sliced into height segments, where $\mathcal{R}_i \subseteq \mathcal{I}$ and $\mathcal{R}_i \cap \mathcal{R}_j = \emptyset$, for $i \neq j$. Individual segmentation of depths in each region $\mathcal{R}_i$ reassigns the measured depth values to segmentation levels $S_{i,l}$, where l denotes the level number and i the index of the belonging region. By using

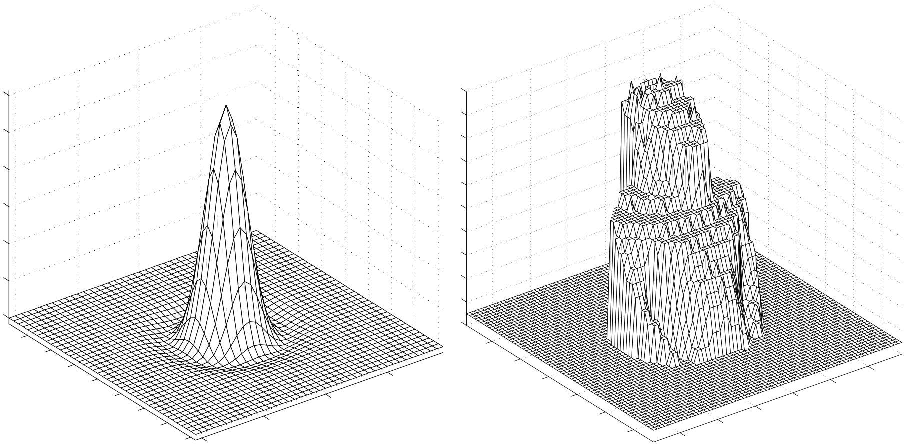
$$S_{i,step} = \frac{F(n_t, \mathbf{m}_{max}) - F(n_t, \mathbf{m}_{min})}{c}, \tag{7}$$

$$S_{i,l} = \left\lfloor \frac{d - F(n_t, \mathbf{m}_{min})}{S_{i,step}} \right\rfloor + 1, \tag{8}$$

$$F(n_t, \mathbf{m}_{max}) = max(F(n_t, \mathbf{m}))\forall \mathbf{m} \in \mathcal{R}_i, \tag{9}$$

$$F(n_t, \mathbf{m}_{min}) = min(F(n_t, \mathbf{m}))\forall \mathbf{m} \in \mathcal{R}_i, \tag{10}$$

increment $S_{i,step}$ and segmentation levels $S_{i,l}$ are determined. d denotes the measured depth of the processed pixel, $F(n_t, \mathbf{m}_{max})$ and $F(n_t, \mathbf{m}_{min})$ denote the extremes. The proposed system clusters depths of every region in a static number c of steps. Depth data is segemented to an image $\widehat{F}(\mathbf{n})$ with



(a) 3D-view of Mexican hat filter

(b) 3D-view of segmented depth image $\widehat{F}(\mathbf{n})$

Fig. 2. Matched filter and segmented depth image

$$\widehat{F}(n_t, n_x, n_y) = \begin{cases} S_{i,l} & \text{, if } \mathcal{R} \in \mathcal{R}_i \\ 0 & \text{, else} \end{cases} \tag{11}$$

for every n_tth image.

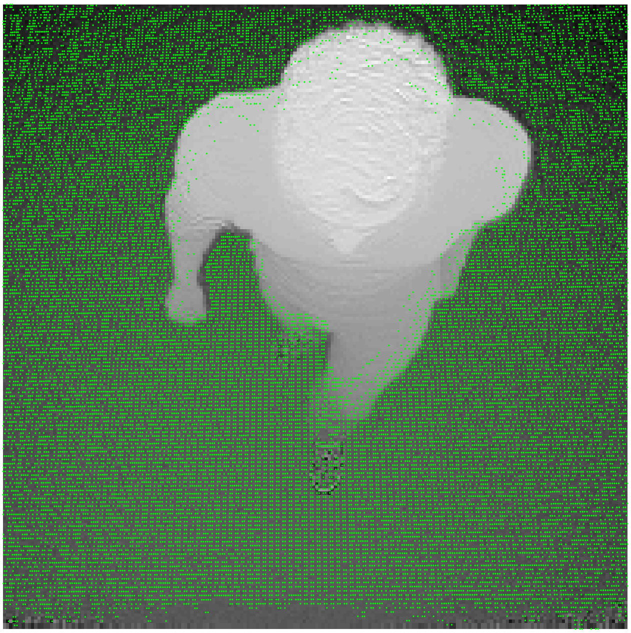


Fig. 3. Object on estimated ground plane

Figure 3 shows the result of the preprocessing steps. Within the depth image, the static ground plane is projected to the image. Green marked pixel indicate areas where the expected distance matches the measured distance from the TOF camera. This concludes to no objects being present at these particular pixel. As the Figure shows, areas where the expected does not match the measured distance, an object occupies the camera beam. Associated regions are here identified as non-ground plane areas and marked as regions of interest $\mathcal{R}_i$ and used for detection of people. Each $\mathcal{R}_i$ is shaped according to the object occupying the camera beam. Image areas, marked as ground plane, therefore cannot contain persons and are disregarded for the matched filter-based people detection.

3.2 Matched Filter-Based People Detection

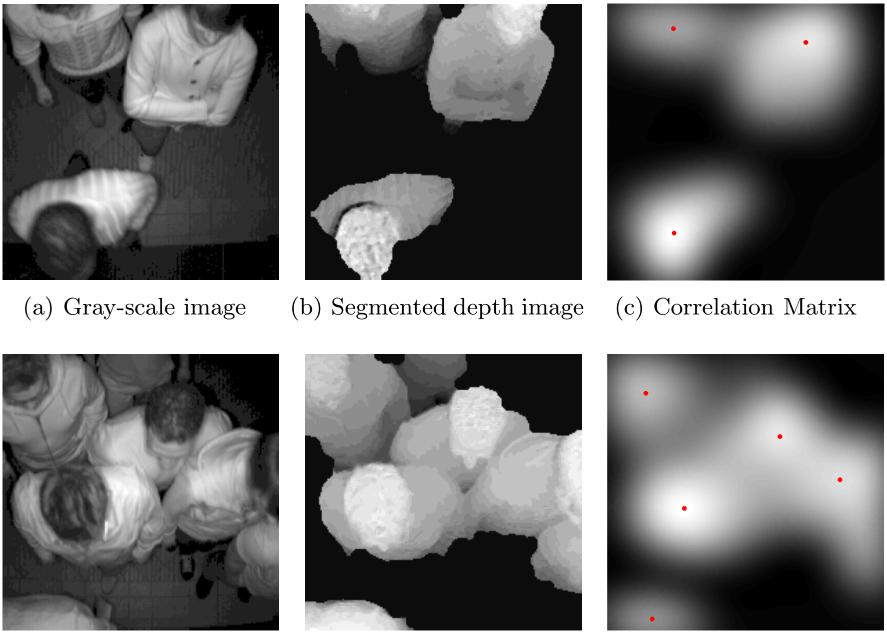
The proposed algorithm uses a scaled Mexican hat wavelet $\Psi(\mathbf{r})$, which is equal to the second derivative of a Gaussian [11]. It is used as a matched filter to distinguish people from different objects in $\widehat{F}(\mathbf{n})$. The normalized Mexican hat wavelet, given by

$$\Psi(\mathbf{r}) = \frac{2}{\sqrt{3\sigma}\pi^{1/4}} \left(1 - \frac{\mathbf{r}^2}{\sigma^2}\right) \exp\left(-\frac{\mathbf{r}^2}{2\sigma^2}\right) \tag{12}$$

with $\mathbf{r} = (x, y)'$. $\Psi(\mathbf{r})$ is scaled according to the expected size of silhouettes of persons, that depends on the positioned height of the camera above ground level.

If the camera is positioned lower above ground level, people appear larger within the image. A high positioned camera results in a larger field of view that results in people appearing smaller.

The Mexican hat wavelet is used as impulse response $\Psi(\mathbf{r})$. Due to its resemblance with upright standing people, it is used here as a generalized depth silhouette for people. This silhouette approaches the depth information of upright standing persons in a way where the orientation of the person in the field of view is of no importance.



(a) Gray-scale image (b) Segmented depth image (c) Correlation Matrix

(d) Gray-scale image (e) Segmented depth image (f) Correlation Matrix

Fig. 4. Correlation procedure visualized by gray-scale image, segmented depth image $\widehat{F}(\mathbf{n})$ and resulting correlation matrix $\mathcal{C}$ with a marker defining the position $\mathbf{z}$ of detected persons for different crowding scenarios

Figure 2(a) shows the 3D-view of a Mexican hat filter used as a generalized depth silhouette for people detection, 2(b) the segemented depth image calculated from Figure 1(a). In a preprocessed image $\widehat{F}(\mathbf{n})$, a detected ROI $\mathcal{R}_i$, containing the foreground object, is segmented in c ascending depth levels. The ground level equals $S_{1,0}$.

The 2D cross correlation, with respect to variables n_x and n_y at time n_t, is calculated by $\mathcal{C}(n_t, n_x, n_y) = \widehat{F}(n_t, n_x, n_y) * \Psi(\mathbf{r})$ of preprocessed depth images $\widehat{F}(\mathbf{n})$ with matched filter $\Psi(\mathbf{r})$. In image processing,

$$\mathcal{C}(n_t, n_x, n_y) = \frac{1}{N_x N_y} \sum_{k_x=0}^{N_x-1} \sum_{k_y=0}^{N_y-1} \widehat{F}(n_t, k_x, k_y) \Psi(n_x + k_x, n_y + k_y) \tag{13}$$

calculates the correlation signal $\mathcal{C}$ [6].

High values in $\mathcal{C}$ denote similarity with matched filter $\Psi(\mathbf{r})$ in a ROI $\mathcal{R}_i$ of image $\widehat{F}(\mathbf{n})$, that may indicate a person. In other words, positions of peaks in $\mathcal{C}$ define the location in $\widehat{F}(\mathbf{n})$ where the matched filter is alike to the silhouette of a human person. Also objects similar $\Psi(\mathbf{r})$ result in high values in $\mathcal{C}$.

The correlation results are independent from the absolute height of a person due to the prior height segmentation.

Taking into consideration that a region $\mathcal{R}_i$ contains more than one person, several peaks in regions are feasible. The combination of a peak finding algorithm using the extended h-maxima transform and a thresholding procedure for a further analyzis of peaks is then used to distinguish people in region $\mathcal{R}_i$ from different objects. Extended h-maxima transform suppresses regional maxima, whose peak is less than a given threshold [15].

Remaining peaks after the thresholding process indicate an upright standing person. Their location is used for the Kalman-filter based tracking procedure. Figure 4 outlines the people detection algorithm for different crowding scenarios. Images 4(a)-4(c) show the detection process for a moderate crowding scene where three persons are present. Images 4(d)-4(f) contain five persons. Depth image 4(b) is partitioned into two regions $\mathcal{R}_i$, one in the upper part containing two persons, the other one in the lower part of the image.

These regions are segmented in height into segmentation levels to normalize the measured depths and achieve independence from the measured height of a person. This is needed for a reliable detection of children or small people.

3.3 Track Assignment

The assignment of measurements $z(n_t)$ to tracks is based on the people detection algorithm that provides $\mathbf{z}_i = [x_i, y_i, d_i]'$. The coordinates (x_i, y_i) of measured depth d_i describe object i. Each image $\widehat{F}(\mathbf{n})$ contains $i = 0, \ldots, O$ detected objects identified as a person, where O defines the number of detections in frame n_t.

A new track is initialized if $\mathbf{z}_i$ cannot be assigned to an existing track. The state of a track is defined as state vector $\mathbf{x}_i(n_t) = [x_i, y_i, \dot{x_i}, \dot{y_i}, d]'$, where $(\dot{x_i}, \dot{y_i})$ is the velocity and used to improve the definition of the state of a person.

Time update and measurement update are defined analogously to [1] as

$$\mathbf{x}(n_t + 1) = \mathbf{A}\mathbf{x}(n_t) + \mathbf{v}(n_t) \tag{14}$$

$$\bar{\mathbf{z}}(n_t + 1) = \mathbf{H}(n_t + 1)\mathbf{x}(n_t + 1) + \mathbf{w}(n_t + 1), \tag{15}$$

where $\bar{\mathbf{z}}$ is the measurement prediction error and $\mathbf{v}(n_t)$ is a sequence of zero-mean white Gaussian process noise with covariance $E[\mathbf{v}(n_t)\mathbf{v}(n_t)'] = \mathbf{Q}(n_t)$. Sequence

$\mathbf{w}(n_t)$ is also of zero-mean white Gaussian measurement noise with covariance $E[\mathbf{w}(n_t)\mathbf{w}(n_t)'] = \mathbf{R}(n_t)$. System matrix $\mathbf{A}$ with assumption of constant velocity is used as

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & \Delta n_t & 0 & 0 \\ 0 & 1 & 0 & \Delta n_t & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0, \end{bmatrix} \qquad (16)$$

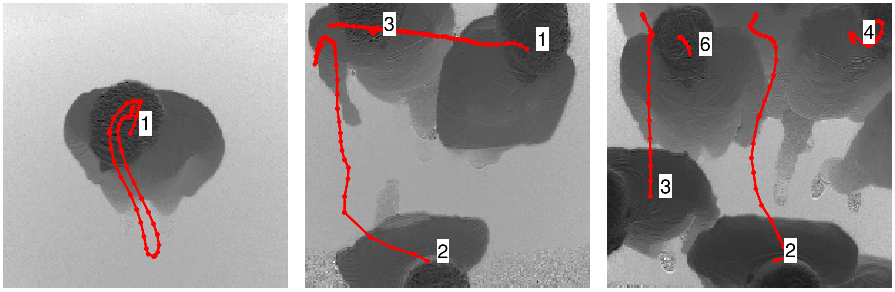
where Δn_t denotes the time difference between two frames.

Assignment of measurements to tracks is performed by usage of a gating process, using the well-known Mahalanobis distance. Tracks taken into account by the previous step selected. This gaiting process is followed by an association process that finally assigns measurements to known or new tracks by applying a nearest-neighbor procedure.

4 System Setup and Experiments

Our test environment is a hallway equipped with a ToF camera, positioned in a hallway normal to entrance of a room at a height of 2.8 meters

The used ToF camera provides a resolution of 200×200 pixel with a frame-rate up to 40 fps, a field of view of 40°×40° and a standard measurement range up to 7 meter. Figure 5 shows images of detected and tracked persons. Red crosses define the positions of measurements from each frame, combined to a tracking path showing the movement of each person. The connectivity of measured positions identifies the successful tracking of a person. Numbers beside the tracking path define the tracking identification number that increases by one for each new track.



(a) Tracking of one person (b) Tracking of multiple persons (c) Tracking of a crowd

Fig. 5. Tracking of people in different crowding scenarios

The proposed algorithm is fully implemented in MATLAB and has been tested for different crowding scenarios. The sequences used for experiments are up to thousand frames in length and consist of up to eight persons simultaneously.

Experiments for different crowding scenarios on real data have demonstrated the applicability of the system as shown in Figure 5.

Problems with people detection occur in cases where a person is not fully present in the camera's field of view, but at the edge. This is called "body in the pyramid effect" [17] that describes the body shape which comes closest to the shape from top-view position. In this case, a person is not fully visible but partially by crossing the camera's beam. A person not fully visible, but detected as a person is shown in Figure 4 in track 3.

If a person walks at the edge of the sensor's beam and alters between providing sufficient and insufficient shape and height information, this person may not be detected certainly while residing in the field of view. As a consequence, the person can be considered a new track multiple times. As long as a person provides sufficient shape and height information it maintains its logical identity in the scene.

As a result of a generalized depth silhouette, experiments have shown that the orientation of a captured person within the image is of no importance. This means, walking direction and angle to the camera do not affect the detection result of a person. Based on the usage of one matched filter, the detection algorithm is adapted to upright standing or walking persons. Currently, the system's reaction for people in wheelchairs, wearing hats or walking ducked has not been tested.

Experiments were conducted offline and show the applicability of the system to detect and track persons from the described top-view position, even in cases with many people forming a dense crowd.

5 Conclusions

This paper has presented a novel approach for the detection and tracking of people from a top-view position using a ToF depth sensor. The presented system contains procedures to determine foreground information based on measured depths, estimated ground-plane and the detection and tracking of multiple-objects recognized as people. A matched filter is applied to the segmented foreground information of measured depth images, scaled as a generalized depth model for the detection of persons. Detected persons are tracked by using a Kalman-filter.

Experiments have shown the applicability of the system for different crowding scenarios of people in real sequences. It has become obvious that the proposed system works reliable in all tested scenarios, which comprise the detection and tracking of one person up to a dense crowd. The system is erroneous when obstacles, e.g. a held-up hand, occupy the camera beam or when people walk at the edge of the sensor's field of view.

We believe that such erroneous people detections can be improved by the fusion with a classifier trained on gray-scale images from this top-view position.

Acknowledgements. The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7 / 2007 - 2013) under grant agreement no. 218086.

References

1. Bar-Shalom, J., Rong Li, X., Kirubarajan, T.: Estimation with Applications to Tracking and Navigation. John Wiley & Sons (2001)
2. Bevilacqua, A., Di Stefano, L., Azzari, P.: People tracking using a Time-of-Flight depth sensor. In: IEEE Fifth International Conference on Advanced Video and Signal Based Surveillance (2006)
3. Beymer, D., Konolige, K.: Tracking People from a Mobile Platform. In: Siciliano, B., Dario, P. (eds.) Experimental Robotics VIII. STAR, vol. 5, pp. 234–244. Springer, Heidelberg (2003)
4. Foix, S., Alenyá, G., Torras, C.: Lock-in Time-of-Flight (ToF) Cameras: A Survey. IEEE Sensors Journal 11(9), 1917–1926 (2011)
5. Gavriilidis, A., Schwerdtfeger, T., Velten, J., Schauland, S., Hohmann, L., Haselhoff, A., Boschen, F., Kummert, A.: Multisensor data fusion for advanced driver assistance systems - the Active Safety Car project. In: International Workshop on Multidimensional Systems (nDs), vol. 7, pp. 1–5 (2011)
6. Gonzalez, R.C., Woods, R.E.: Digital Image Processing, 2nd edn. Prentice Hall International, Boston (2001)
7. Hansard, M., Lee, S., Choi, O., Horaud, R.P.: Time of Flight Cameras: Principles, Methods, and Applications. Springer Briefs in Computer Science. Springer (2012)
8. Harville, M.: Stereo person tracking with short and long term plan-view appearance models of shape and color. In: Proceedings of International Conference on Advanced Video and Signal based Surveillance (AVSS), vol. 1, pp. 511–517. Santa Fe (2005)
9. Harville, M., Li, D.: Fast, Integrated Person Tracking and Activity Recognition with Plan-View Templates from a Single Stereo Camera. In: IEEE Conference on Computer Vision and Pattern Recognition, Washington (2004)
10. Helbing, D., Mukerji, P.: Crowd Disasters as Systemic Failures: Analysis of the Love Parade Disaster. Tech. rep., ETH Risk Center, Zürich (2011)
11. Mallat, S.: A Wavelet Tour of Signal Processing. Academic Press (1998)
12. PMD Technologies, PMD vision CamCube 3.0 Specsheet - High resolution 3D video camera. Tech. rep., PMD Technologies GmbH (2010)
13. Reynolds, M., Dobos, J., Peel, L., Weyrich, T., Brostow, G.: Capturing Time-of-Flight Data with Confidence. In: Proc. of IEEE Conference on Computer Vision and Pattern Recognition, CVPR (2011)
14. Ringbeck, T.: A 3D Time of Flight Camera for Object Detection. In: Optical 3-D Measurement Techniques (2007)
15. Soille, P.: Morphological Image Analysis: Principles and Applications. Springer (1999)
16. Still, G.K.: Duisburg - July 24, 2010, Love Parade Incident, Expert Report. Tech. rep., Bucks New University (2011)
17. Tanner, R., Studer, M., Zanoli, A., Hartmann, A.: People Detection and Tracking with TOF Sensor. In: IEEE Fifth International Conference on Advanced Video and Signal Based Surveillance (2008)