

People Detection and Tracking with TOF Sensor

Rudolf Tanner, Martin Studer, Adriano Zanolli
University of Applied Sciences, HTW Chur
Pulvermühlestrasse 57
7004 Chur, Switzerland

Rudolf.Tanner@ieee.org, Martin.Studer@fh-htwchur.ch

Andreas Hartmann
GST GmbH
Schlyffistrasse 17f
8860 Bäch, Switzerland

andy.asim@gmail.com

Paper ID 60

Abstract

This paper presents a method to detect, track and count people which pass a confined space using the latest 3D Time-of-Flight (ToF) sensor technology. A slicing algorithm is employed and a novel method is proposed which derives the slicing level based on the probability density function and histogram respectively.

It was found that the algorithm, exploiting the data of one sensor only, works very well for normal situations. However, additional information needs to be made available to the system to solve special cases, such as wheel chairs, push chairs or very closely spaced people.

1. Introduction

Automatic detection and tracking of people from image sequences generated by cameras has been a research topic for some years. Using conventional cameras, either monocular or binocular systems, reliable detection has been found difficult. Recently however, companies like MESA Imaging AG [9], Canesta Inc. [4] or PMD Technologies GmbH [10], have introduced *Time of Flight* (TOF) IR cameras delivering the range information also.

This technology promises to advance the state-of-the-art in detection and tracking of people [2].

An analysis of published algorithms using 3D cameras [1, 11] show that there are a few common steps. Firstly, the background is eliminated. Then the 3D range image is projected onto a 2D plane, and these 2D images are processed using established image processing methods. Finally detected people are tracked through image sequences.

Background elimination in range images can be simple while being a major problem with conventional 2D images. If the camera has a fixed position and the scene is static, a background image can be acquired. Subtracting the static

background image from actual images removes range pixels belonging to the background. Cameras used for surveillance usually fulfill these preconditions. An additional benefit of background elimination is that it reduces the data set for further processing. However, at the cost of, for example, loss of some information which could be otherwise exploited.

The second step involves detecting people in a single range image. Projecting a 3D image onto a 2D plane traces the problem of detection back to the detection of shapes in 2D images. Thus allowing the use of proven algorithms in the field of image analysis and computer vision, like edge detection and image segmentation [5, 3]. 2D images are either generated by slicing the range image at different heights [11] or by converting the height to intensity values in the 2D image [1]. The result of that step is an image with one center point for each detected person.

The input of the third step is a temporal sequence of images each with center points. Tracking of people can be done by connecting center points of consecutive images closest to each other. Accuracy usually is improved by using Kalman [6] or Particle filter [11] to predict the possible position of a person in the next image frame.

Our algorithm basically follows the steps described above also, but using a novel slicing algorithm in step two. Reference [11] slices at different heights above the ground to cater for both, children (cut at $\approx 1\text{m}$ above ground) and adults ($\geq 1.5\text{m}$ above ground). Our approach differs by first working out an optimized slicing level and size using histogram analysis of the original picture.

This paper is organized as follows. Next we describe our methodology in five steps. Then we introduce the test environment and the configuration followed by recent results. Finally, we conclude this paper with a summary and the observations made to date, draw some conclusions about our work, and outline ongoing and future work.

2. Method

The aim of the system is to detect persons and count the number of people crossing the observation area, i.e. camera beam. Such an observation area can look like the one shown in Figure 1. The people flow is from the bottom to a door at the top in the illustration. The walking floor may be further obstructed by tables, chairs, flower pots or other furniture.

The approach taken here has five stages

- picture conditioning,
- threshold estimation,
- 2D picture generation,
- person detection,
- direction identification and counting.

The first phase of this research focuses on processing 2D pictures in order to keep the requirements for computation power low, and hence the demands on the embedded system and cost. Thus the generated 3D picture is converted into a 2D picture with a simple projection method.

2.1. Conditioning

The signal pre-processing stage is concerned with improving the *Signal to Noise Ratio* (SNR). We can also remove in each picture, for example, known static objects such as desks. Figure 1 shows a possible deployment. The wire-grid reflects the observation area of a TOF sensor, mounted under the ceiling. Of course the beam of a single camera is subject to its lens design, and thus has not the shape of a square but more of a cone or tetrahedron. The walls and the table are static and can be used to provide corner points for a plausibility assessment later, when the location and movement is computed. Picture 2 shows the picture resulting from the data of the TOF sensor. To ease further processing, we simply remove the corner (table) of any dynamic picture, by

$$\mathbf{Z}(k) = \mathbf{Z}_{dynamic}(k) - \mathbf{Z}_{static}, \quad (1)$$

where matrix $\mathbf{Z}$ contains the depth data and k denotes the k th image.

2.2. Threshold Setting

Sensible threshold derivation and setting is essential to define the level where the 3D picture is cut into slices to arrive at 2D shapes. This can be carried out "automatically" (like a contour function). Alternatively, one could decide to slice the 3D picture in equally distant instances [11], say every 20 cm, to ensure that one detects children as much as

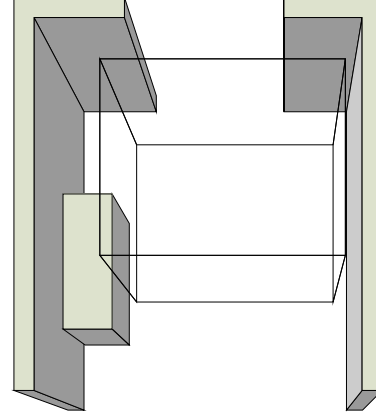


Figure 1. A typical room where the number of people has to be counted through an obstruction free method like TOF.

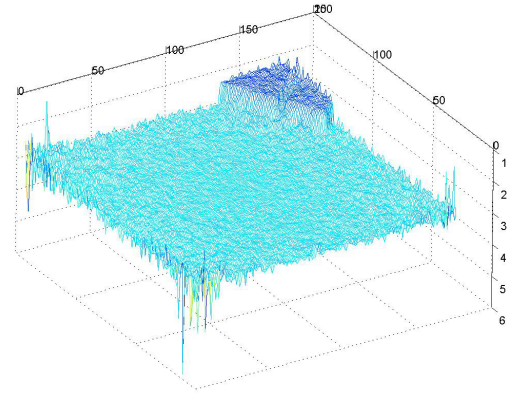


Figure 2. A static picture with the corner of a table visible.

adults. But this causes a new problem as we noted, where we have to discriminate between wheely suitcases and children. Fortunately, suitcases have a square shape.

The presented approach here rests on a statistical analysis of the *probability density function* (pdf) of the range values captured model-free by a histogram. The motivation is to take account as much on the density as on the distance information with the aim to reduce false detections. The static background picture data is analyzed and acts later as a reference. The results of this analysis, the mean $\bar{x}$, local maxima and minima in the histogram, and the variance σ^2 , may be updated regularly to account of changes in the environment. The histogram of a static picture is presented in Figure 3. The original picture has dimensions 176×144 and is re-sized to remove the spikes at the picture edge, the fringes. The slightly smaller picture is then processed, resulting in a mean $\bar{x} = -0.047$ meters. The fact that the mean does not equal zero stems from noise. This error is

within the known range of 5 to 10 cm [7].

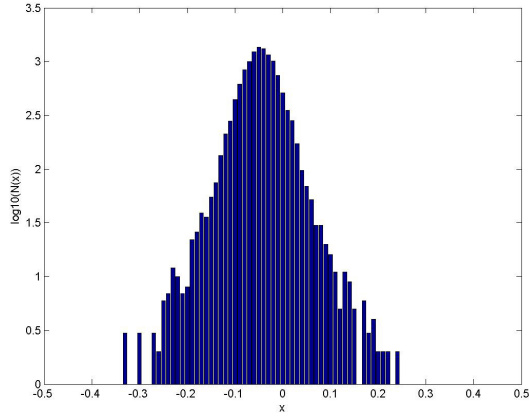


Figure 3. The histogram of a static picture where no person is present.

The situation differs when a person enters the view of the camera. Picture 4 shows the 3D view and one can clearly identify the body and an arm on the left hand side. The re-

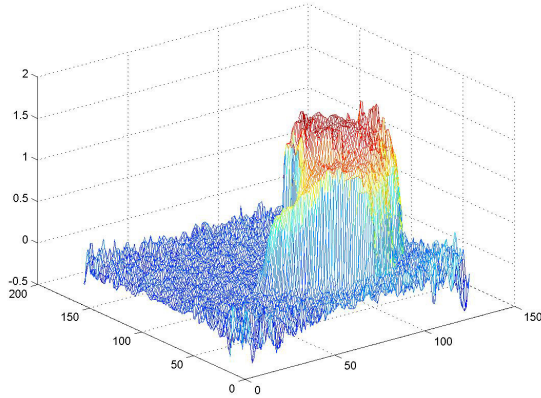


Figure 4. A dynamic picture, the body and an arm (between 50 and 100) can clearly be identified.

sulting histogram of picture 4 is shown in graph 5. Note that we changed the scaling of the y-axis to emphasis the fact that the background noise, i.e. floor, produces a dominant peak. Fortunately, this peak centers around the zero level and hence can be easily removed. What we are left with is the second hump, whose peak is in the range between 1.5 m and 1.8m, which corresponds to the height of the person.

However, graph 5 lets us already note a potential problem. Assume that the picture contains two people, or one adult and a child, how can be safely determine that fact.

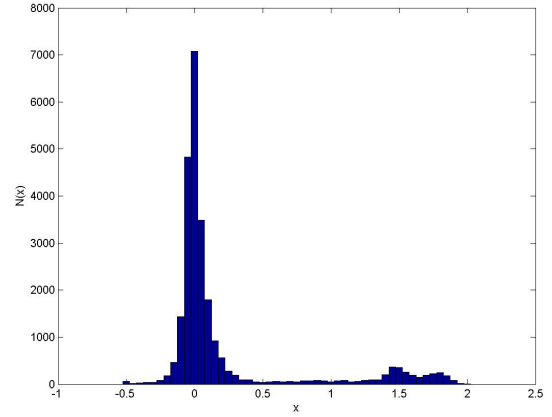


Figure 5. The histogram of picture 4 with the distribution of the noise and a second, smaller hump, from the person in the scene.

We propose that one can identify each hump in the histogram by means of identifying the saddle points. One method is to apply a sliding window, which carries out an interpolation of a few points and determines the local minima in the function, as follows. We start from the basic equation¹ for the parabola:

$$f(x) = w_0 + w_1x + w_2x^2 \quad (2)$$

where w_i denote the coefficients. When $w_2 > 0$ then the parabola is open at the top, otherwise the parabola is open at the bottom. Equation 2 can be applied to three points x_1, x_2, x_3 lying on the x-axis, and their corresponding point y_i on the pdf:

$$\begin{aligned} y_1 &= w_0 + w_1x_1 + w_2x_1^2 \\ y_2 &= w_0 + w_1x_2 + w_2x_2^2 \\ y_3 &= w_0 + w_1x_3 + w_2x_3^2 \end{aligned}$$

which can be written in a convenient matrix form

$$\mathbf{y} = \begin{bmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ 1 & x_3 & x_3^2 \end{bmatrix} \begin{bmatrix} w_0 \\ w_1 \\ w_2 \end{bmatrix} = \mathbf{X}\mathbf{w} \quad (3)$$

from which we can derive the coefficients and conclude where the bottom, or top, of the parabola lies.

$$\mathbf{w} = \mathbf{X}^{-1}\mathbf{y} \quad (4)$$

Equation 4 is applied along the x-axis of the function. Obviously, there is no need to perform the calculation for every 1 cm while a 5 cm to 15 cm interval suffices.

This method can also be expanded and applied to 2D histogram surfaces also to ease the process of identifying

¹Unless stated otherwise, vectors and matrices are denoted in bold.

suitable thresholds for the case when more than one person is present.

We can draw conclusions about the locations of each peak in the pdf, but also, where the distributions of the various humps start to overlap, namely when w_2 changes its sign. Note that the highest peak, which lies around zero and stems from the measurement noise (background, floor), can be ignored when we omit all data below a height of 0.2 m for example.

In this way we can ensure that the slicing threshold is chosen appropriately.

2.3. 3D-to-2D Conversion

The 3D picture is processed using a *slicing* algorithm, similar to the method discussed in [11, 1]. We employed a slicing algorithm to keep the computation simple. First,

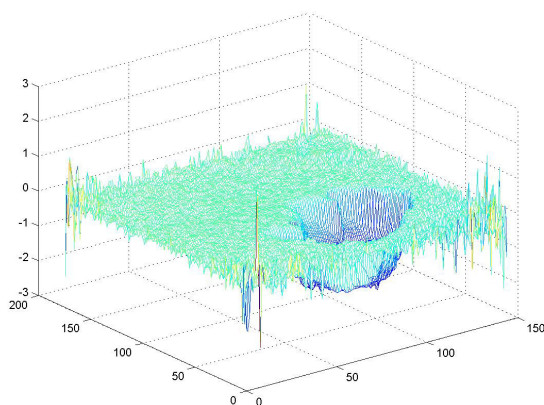


Figure 6. A picture without table but the body of a person.

the data $\mathbf{Z}(k)$ from the previous stage is processed through slicing which leads to a contour. The level at which the cut is performed is provided by the threshold setting algorithm. This is a straightforward process because the picture is re-normalized from the pre-processing, and the base, i.e. the floor, is put at zero meter. Hence, every obstacle detected by the TOF sensor appears as a hump in the negative direction. Thus we only need to identify the pixels and their associated position in 2D at, for example, -1 meter. This yields a single contour plot for the detection stage.

2.4. Person Detection

In the literature, it has been proposed to use, for instance, an ellipse or circle, to identify a person's body or his head respectively. This works well under ideal conditions, however, we do not know if these methods work when people wear hats or carry a backpack. So the lexicon of shapes could become exhaustively large, of which we only show

a few conceivable shapes in Figure 7. Other problems can

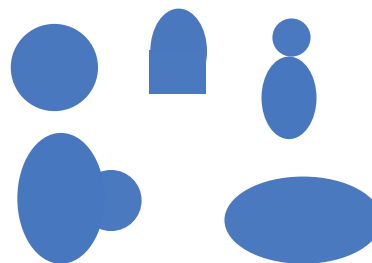


Figure 7. Very idealized shapes to emphasize the shape for detection by means of shape matching: head, person at edge of beam, adult with child, backpacker and body.

be encountered when a person is in a wheelchair or a person pushes a children's push chair. Suddenly, the optimum shape turns into a rather new and different and complex problem, which we cannot address here for brevity. Another problem can arise when the person is not "illuminated" from the very top where the IR TOF sensor is mounted. This is the case when the person walks at the edge of the sensor's beam. Remember, the beam has a certain angle within the infrared light gets emitted. The result is the "body in the pyramid" effect, illustrated in Figure 8. Here the data of the

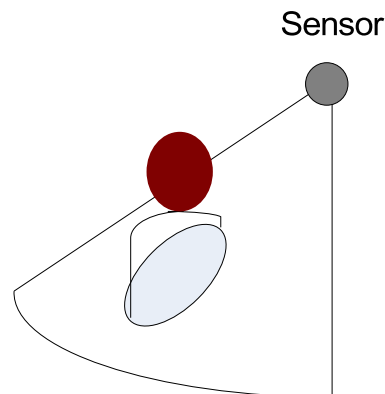


Figure 8. The person is not within the TOF sensor's main beam but at the edge and hence his true height cannot be measured.

TOF sensor shows a body shape which comes closest to the top center shape in Figure 7, because the person is crossing the beam at the edge. One consequence is that every sensor must be placed to yield a view from the top. This can be achieved with several sensors also in large spaces.

Another method is to avoid identifying the body shapes and just concentrate on moving objects in the camera's view. This is easily achieved with a correlation from which we can derive a peak when the picture matches best. This peak is then used for tracking purposes. This has proven to be reliable also and was used in the first phase of this work.

2.5. Direction Identification and Counting

A birth-death vector path tracking mechanism is employed and initially a Kalman filter [8]. However, the Kalman filter expects the error to be Gaussian and this is here not true. The people walk straight to the exit and do not behave like randomly moving particles. Hence we discarded it because there was no benefit but extra computational complexity.

The center of the contour is identified and kept on record. Any available point gets a dynamically allocated path, this is the birth process. As soon as a second point, i.e. co-ordinate, becomes available, then one has a vector. A plausibility logic is then used to check if vectors cross, if vectors move into or come out of a wall, or if the speed of a moving point is unrealistically high. Any unreasonable data points are killed (death process). Equally when an object leaves the observation area at the exit point, i.e. then the track expires. Note that we consider in this specific application one way scenarios.

As soon as a vector has left the observation area, then a person counter gets incremented.

We can also raise an alarm, for example, when an object enters the observation area from the direction defined as exit, i.e. the wrong direction. Then the operator can make use of the cameras normal picture sensor, i.e. not TOF, capability and inspect what is going on. This negates the need for continuous visual observation.

3. System Set-up

A Mesa Imaging TOF sensor with 176×144 pixels was used [7]. The sensor was fixed under the ceiling at a height of 2.8 meter and adjusted to have its sensor parallel to the floor. The floor was a grey-multi-color vinyl floor, no carpet. There was a corner of a table put into the sensors beam, rather to have a reference point as opposed to produce a more realistic scenario. The data got captured and processed with Matlab.

The SNR was estimated and verified using two methods. First, we subtracted one static picture from another static picture, obtained several seconds apart, because it cancels the distance information. Ideally, when no noise is present, then the depth data is zero but this is not the case. Figure 3 shows that there are variations in the distance. This process returns an image SNR of 43.7 dB.

Another method exploits the fact that we know the precise distance from the camera to the floor. This value can be used to offset the distance information in the data set also. This resulted in a calculated SNR of 44.2 dB.

The two methods match and the difference is rather small considering that we operate at an SNR level of around 43 dB.

The second method should of course be less accurate be-

cause the distance is only valid straight under the sensor. We introduce an error which slowly increases as we move away from the center of the picture (floor) to the edge.

4. Results

The first example is presented in Figure 9. A sequence of five pictures are taken and processed. Each (blue) star denotes the center of a picture. We can clearly observe that in this case the system can detect a moving person, work out a reference point, and track the points as the time goes by. In this example, the pedestrian did not walk in the anticipated direction but diagonal.

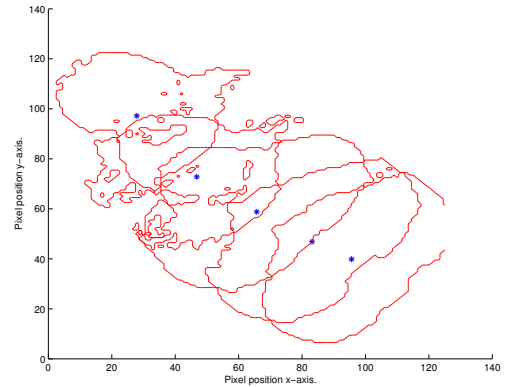


Figure 9. A five picture sequence of a person entering and leaving an observation area.

The second picture, Figure 10, shows two people walking side-by-side. Again, the system could detect and track the path of each person.

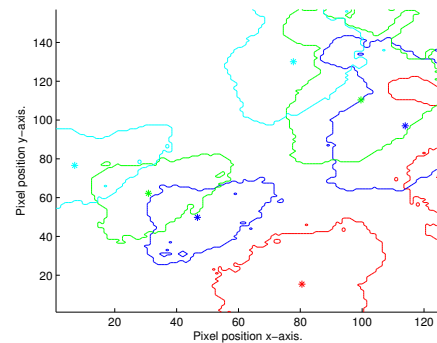


Figure 10. Two four picture sequence of two people walking side-by-side entering and leaving the observation area.

The third picture, Figure 11, depicts the result of two people walking behind each other. Again, the algorithm could track reliably both persons.

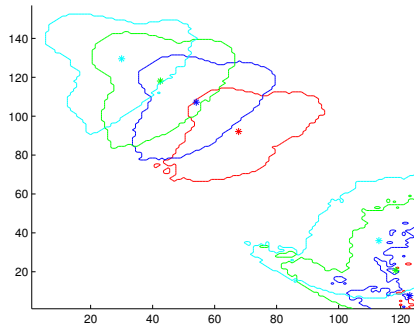


Figure 11. Two four picture sequence of two people in tandem entering and leaving the observation area.

5. Conclusions

This paper presented a TOF based people detection and tracking and counting system which rests on a novel slicing algorithm whose threshold is derived from the histogram of the 3D images.

It was shown that the system could reliably detect people and follow their tracks without the need for computationally complex tracking algorithms. We observed through empirical tests that the detection and discrimination of certain scenarios is a very hard problem. It appears that it is not possible to identify or discriminate people for such scenarios unless we have additional data available, either from a second camera or of a different kind. Such scenarios have not been reported yet to the authors best knowledge, and future results will offer further insight into the capabilities of practical systems.

Current and future work addresses the reliability of the current algorithm for which we need a bigger data set of real-life situations. One objective is, amongst others, to derive the false detection probability.

References

- [1] A. Bevilacqua, L. D. Stefano, and P. Azzari. People tracking using time-of-flight depth sensor. *Proceedings of AVSS06*, 1(1):1, 2006. 1, 4
- [2] B. Büttgen, T. Oggier, M. Lehmann, R. Kaufmann, and F. Lustenberger. Ccd/cmos lock-in pixel for range imaging: Challenges, limitations and state-of-the-art. In *Proceedings of the Range Imaging Research Day*, pages 21–32, 2005. 1
- [3] R. Duda and P. Hart. *Pattern Classification and Scene Analysis*. Wiley, Reading, USA, 2 edition, 2000. 1
- [4] S. B. Gokturk, H. Yalcin, and C. Bamji. A time-of-flight depth sensor - system description, issues and solutions. In *Proceedings of the 2004 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'04)*, page 35, Washington, DC, USA, 2004. IEEE Computer Society.
- [5] R. Gonzales and R. Woods. *Digital Image Processing*. Addison-Wesley, Reading, USA, 1 edition, 1992. 1
- [6] S. Haykin. *Neural Networks*. Prentice Hall, Upper Saddle River, USA, 2 edition, 1999. 1
- [7] T. Kahlmann, F. Remondino, and S. Guillaume. Range imaging technology: new developments and applications for people identification and tracking. *Proceedings of SPIE IS&T*, 6491(1):1, 2007. 3, 5
- [8] T. K. Moon and W. C. Stirling. *Mathematical Methods and Algorithms*. Prentice Hall, Upper Saddle River, USA, 1 edition, 2000. 5
- [9] T. Oggier, M. Lehmann, R. Kaufmann, M. Schweizer, M. Richter, P. Metzler, G. Lang, F. Lustenberger, and F. Blanc. An all-solid-state optical range camera for 3D real-time imaging with sub-centimeter depth resolution (Swiss-Ranger). In L. Mazuray, P. J. Rogers, and R. Wartmann, editors, *Optical Design and Engineering. Edited by Mazuray, Laurent; Rogers, Philip J.; Wartmann, Rolf. Proceedings of the SPIE, Volume 5249, pp. 534-545 (2004).*, volume 5249 of *Presented at the Society of Photo-Optical Instrumentation Engineers (SPIE) Conference*, pages 534–545, Feb. 2004. 1
- [10] T. Ringbeck and B. Hagebecker. A 3d time of flight camera for object detection. In *Proceedings of the 8th Conference on Optical 3-D Measurement Techniques*, page Session 1. ETH Zürich, 2007. 1
- [11] F. Wallhoff, M. Russ, G. Rigoll, J. Göbel, and H. Diehl. Improved image segmentation using photonic mixer devices. *Proceedings of ICIP*, 1(1):53–56, 2007. 1, 2, 4