

People tracking using a Time-of-Flight depth sensor

Alessandro Bevilacqua

Luigi Di Stefano

Pietro Azzari

ARCES - DEIS (Department of Electronics, Computer Science and Systems)
University of Bologna,
Viale Risorgimento 2 - Bologna, ITALY, 40136

Abstract

Visually track several moving persons engaged in close interactions is known to be a very hard problem, though 3-D approaches based on stereo vision and plan-view maps offer much promise for dealing effectively with major issues such as occlusions and quick changes in body pose and appearance. However, in case of untextured scenes due to homogeneous objects or poor illumination, stereo-based tracking systems rapidly drop their performance. In this work, we present a real time people tracking system able to work even under severe low-lighting conditions. The system relies on a novel active sensor that provides brightness and depth images based on a Time of Flight (TOF) technology. The tracking algorithm is simple yet efficient, being based on geometrical constraints and invariants. Experiments accomplished under changing lighting conditions and involving multiple people closely interacting with each other have proved the reliability of the system.

1. Introduction

Tracking multiple people is a key issue in many advanced vision applications, such as intelligent video surveillance, virtual reality interfaces and human activities understanding. Tracking approaches can be primarily categorized according to the type of device employed, whether they are monocular cameras or “3-D” sensors (usually, stereo cameras). As regards trackers based on 2-D views, they often rely on appearance models, one for each person within the scene [2]. As a matter of fact, the latter suffer from occlusion events and often fail in handling close interactions since evolving the models correctly turns out to be a very hard problem. On the other hand, 3-D approaches based on stereo cameras hold the potential for dealing with some major tracking issues. In fact, yielding the 3-D coordinates associated with image points, stereo allows building an orthographic view of the scene with the projection plane parallel to the ground plane. An orthographic view that has proved to be very effective for reliable multi-person detection and tracking can be obtained by means of plan-view

maps (i.e. occupancy and height maps [5, 6, 1]). Nonetheless, most 3-D stereo-based tracking systems are prone to fail when the scene lacks of texture due to either homogeneous objects or poor illumination.

To the best of our knowledge, the work described in this paper is the first attempt to develop a 3-D multiple people tracking system based on the use of a Time Of Flight (TOF) depth sensor. So far, TOF imaging devices have been used for robot navigation [8], vehicle occupancy monitoring [3] and single head tracking [7]. We believe that, despite the current prototype evolution stage of the technology, TOF depth sensors represent potentially an appealing alternative for the development of 3-D people trackers. In fact, a TOF-based approach requires no computation for 3-D scene reconstruction and turns out to be independent of the degree of texture and the lighting conditions of the scene. In addition, the reduced manufacturing cost (standard CMOS process) makes the approach suitable to the development of compact embedded solutions.

In Section 2 we outline the basic principles of the TOF technology. In Section 3 we describe the tracking algorithm while in Section 4 we discuss some experimental results. Finally, in Section 5 we draw conclusions.

2. TOF Technology Principles

TOF technology represents a well known non-contact measuring technique. It has been employed in a wide spectrum of industrial applications ranging from automatic assembly to quality assurance. Camera-based TOF scanner-less sensors are able to deliver an entire depth image at a time without employing any moving mechanical part. Moreover, depth information is delivered by the solid-state sensor itself, with no need for external circuitry. Typically, these devices consist of a modulated light source such as a laser, a CMOS imager made out of an array of pixels and an optical focusing system.

Two different classes of TOF depth sensors exist, depending on the method they adopt to measure distances and the properties of the transmitted signal. The first class is represented by Pulse Modulation (PM) sensors. Distance is

computed directly from the time of flight using a high resolution timer that measures the delay between signal emission and reception. Depth measures d are simply obtained according to Eq. 1:

$$d = \frac{TOF * c}{2} \quad (1)$$

where TOF is the time of flight and c is the speed of light. The second class is represented by Continuous Wave Modulation (CWM) sensors. Here, the distance is computed from the phase of the modulation envelope of the transmitted infrared light as received at a pixel. Let $s(t) = \sin(2\pi f_m t)$ be the transmitted light, where f_m is the modulation frequency. The amount of light $r(t)$ reflected by the target is given by Eq. 2:

$$r(t) = R \sin(2\pi f_m t - \phi) = R \sin(2\pi f_m (t - \frac{2d}{c})) \quad (2)$$

where ϕ is the phase shift arising when the light returns back to a sensor pixel, R is the amplitude of the reflected light and d is the distance between the sensor and the target. The distance d can be calculated from ϕ as follows:

$$d = \frac{c\phi}{4\pi f_m} \quad (3)$$

A depth image is constructed by measuring the distance d at each pixel. Similarly, a brightness image is built by measuring R at each pixel. At present, the leading manufacturers of CWM sensors are Canesta, CSEM and PMD Tech.

PM devices require a high resolution timer and a large bandwidth signal source to achieve high resolution distance measures. This results in them being more expensive with respect to CWM ones and suitable for long range applications. On the other side, CWM devices are prone to aliasing problems, although this problem can be relieved using multi frequencies scanning. Among the available TOF imaging sensors, Canesta's [4] has been selected for our experiments. This decision has been taken according to scientific and technical reasons. Firstly, this device is a state-of-the-art camera based TOF sensor with quite a wide field of view, high resolution and low latency time. Secondly, the manufacturer provides an evaluation camera packed with software drivers and libraries to develop one's own applications.

3. Algorithm

Several tasks have to be accomplished since acquisition to final tracking. The first task is the segmentation of the scene between background and foreground (moving) regions using depth informations. This is performed by means of a simple background subtraction procedure based on a pixel-wise parametric statistical model. Once constructed, the

background does not need to be maintained over time due to it not being affected by photometric changes. The outgoing foreground pixels with reliable depth are used to build a 3-D point "cloud" in the camera reference frame.

The subsequent task consists in building the plan view maps arising from these points according to the method explained in [5]. A plan view map can be thought of an orthographic projection of the scene as though it were a planar view. Usually, the projected data are collected in bins, like in the occupancy map (the amount of viewed objects area falling into each bin) and the height map (the largest height value in each bin). At the beginning, the camera frame coordinates $(X_{cam}, Y_{cam}, Z_{cam})$ associated with each foreground pixel are computed according to Eq. 4:

$$\begin{aligned} X_{cam} &= (u - u_0) * k_u * (d - f)/d \\ Y_{cam} &= (v - v_0) * k_v * (d - f)/d \\ Z_{cam} &= d \end{aligned} \quad (4)$$

where (u, v) are the image coordinates of the pixel, (k_u, k_v) account for pixel size, d is the measured distance, (u_0, v_0) are the image coordinates of the principal point location and f is the focal length. After that, the camera frame coordinates are transformed into the world space coordinates $[X_W \ Y_W \ Z_W]$ according to Eq. 5, where the Z_W axis is aligned with the perpendicular to the ground plane:

$$[X_W \ Y_W \ Z_W] = -R_{cam} [X_{cam} \ Y_{cam} \ Z_{cam}]^T - \vec{t}_{cam} \quad (5)$$

and the rotation and translation parameters between the two reference frames are determined by a calibration process accomplished off-line. Finally, a proper quantization of the orthographic plan view space yields a good trade-off between computational burden and resolution.

Tracking is performed over the collection of connected components detected within the plan view maps. Basically, tracking works at blob level, that is, the algorithm exploits only a limited set of geometric features associated with each tracked person (i.e. position and speed of the centroid). Besides, for each object only the last centroid is considered in order to infer the current position. The only input to the tracking module is a list of tracked people and the current set of detected blobs within the plan view map. Each object tracked at frame $t - 1$ is represented by a $W_{avg} \times W_{avg}$ bounding-box, given by the averaged head size, whose center is the blob's centroid. In order to trigger an occlusion event, we evaluate the overlap between current and predicted bounding box using a Kalman filter. The existence of an intersection triggers the occlusion handler. In case of occlusion we exploit the following assumptions:

- People can not overlap in the plan-view.
- People should enter in the scene separately.

- Each object keeps the area proportional to the one owned in the last frame before occlusion.

To sort out occlusions we perform an iterative blob segmentation so that each pixel is assigned the temporary ID of the object whose expected centroid has the minimum distance from the pixel itself. The object's area is updated accordingly. In case that the current area keeps lower than the expected one (computed so that respective object's proportions are preserved), this is the ultimate ID. On the contrary, the pixel will be charged with a penalty distance factor in order to increase the probability for the pixel to change its membership in the next iteration step. As soon as the object's expected area is reached, it is segmented and its pixels are excluded from any further computations.

4 Experimental Results

The sensor used for the experiments is the Canesta Electronic Perception EP205 with a field of view of 55° . It is a square array of 64×64 pixels and it is implemented on a single chip using a standard CMOS process. The sensor provides four different modulation frequencies to range up to 12 meters. A proper shutter time can be selected in order to achieve a suitable frame rate. As for the target machine, we use a PC equipped with a Pentium IV 2 GHz processor and 256 MB RAM. Image acquisition is accomplished through a USB connection and the driver primitives provided by the manufacturer.

Our initial experiments have addressed the assessment of the working values of the device's parameters, such as frame rate and modulation frequency. Besides, even environmental parameters, such as the distance of the target, reflectivity of the target and environmental lighting conditions can affect the reliability of the distance to be measured. As a matter of fact, moving the object to be detected away from the sensor yields the integration time to be increased, thus resulting in an unacceptably slow frame rate. A modulation frequency of 52 MHz has been selected to prevent aliasing issues, meanwhile preserving the highest possible distance resolution (it linearly drops as the frequency decreases). As a consequence, we could state that a proper trade-off between real time requirements and precision of the depth measurements could be achieved by placing the objects to be measured at less than 3 meters from the sensor, thus yielding a standard deviation of few centimeters. According to the outcome of the initial experiments, we derived hints in order to determine a suitable system set-up: placing a down-looking sensor at 3 meters from the ground represents a good solution to track people effectively at 12 frame per second (fps).

Then, we have considered two different indoor lighting conditions, i.e. with (Figure 1, left) and without (Figure 2,

right) lighting. The first two rows in both Figures show the brightness images acquired by the 3-D sensor. The second pair of rows shows the depth maps provided by the sensor, which are gray scale encoded - the brighter the pixel the nearer the point. The third pair of rows is the output of the tracking algorithm, with labeled plan view blobs and the associated bounding boxes. The surveyed squared area has a side of 3 meters and consists of the floor of our lab with furniture, that is included into the background during the initial training stage and hence no longer detected as interesting objects. The sequences used in the experiments are some thousands frame in length and up to three persons can move freely within the scene. After having built the 3-D background, the system is ready to perform the background subtraction using the depth maps of Figures 1 2, middle, as current frames. This yields "clouds" of depth points related to moving objects (in camera frame coordinates) which need to be clustered. In order to achieve plan view maps, the outcome is transformed in world frame coordinates and sampled using square bins. Such a sampling procedure acts as an earlier denoising process. As a matter of fact, having a smaller size preserves the details in quite a noisy image, while a larger size improves the SNR. A proper size for our system set-up has proved to be a 10 cm side square. At present, the unreliable depth measures due to a large amount of noise are the main issue to cope with when using this kind of devices. Therefore, the height maps thus obtained are filtered using the occupancy maps as filter masks, as described in [5]. The bins of the plan view maps shown in Figures 1 2, bottom, emphasize the outcome of the sampling process. The quality achieved in these maps has proven to be suitable to track people effectively, as discussed above.

The sequences analysed have been built letting people free to move as they like: they enter and exit the scene, walk aside and run. Also due to the reduced field of view, people engage in very close interactions. The first sequence *S1* has been captured with common indoor lighting conditions. The scene is quite textured, so that a stereo camera could be employed as well. As a matter of fact, the performance of the tracking algorithm as far as the quality is concerned has been compared with that attained using a stereo camera. Basically, they are comparable, since the plan view maps achieved are roughly the same. Almost each tracked person maintains his/her logical identity as long as they remain in the scene. The few errors reported regard splitting or labels jumping between close objects, but they are not due to the use of the 3-D sensor. Rather, the tracking algorithm needs to be refined in order to improve the disocclusion of people showing the same height.

The second sequence *S2* has been taken without any external illumination at all, the TOF camera laser source being the only lighting source. As one can see, the brightness

images (Figure 2, top) is extremely dark and contains very poor information. In such a lighting condition, any passive imaging sensor could not acquire any useful image. On the other hand, the depth maps yielded by the TOF sensor provide nearly the same amount of information as before (Figure 2, middle), thus confirming that different lighting conditions do not affect the output of such an active device. As a consequence, the performance of the tracking system detains the same quality obtained when the scene was illuminated. This achievement allows to have a good people tracker at one's disposal even when a good illumination could not be available, due to environmental conditions or also "voluntary" reasons.

5. Conclusions

We have described a novel proposal pertaining the use of a recently introduced TOF depth sensor in the field of 3-D real time multiple people tracking. The use of depth information and plain view maps allows the system to inherently handle more robustly occlusions with respect to monocular approaches. At the same time the use of a TOF active sensor, instead of a traditional stereo camera, prevents lighting condition to affect system performance and holds the potential to develop more compact embedded solutions for specific applications relying on people tracking. The proposed tracking algorithm permits to handle partial occlusions and the general case of close interactions between several persons. Although the tracking algorithm needs some improvements and the TOF device performance may also be enhanced at the design level, we believe that this work provides clear indications that it is possible to use state-of-the-art TOF depth sensors to track effectively several moving persons over relatively long time spans and in the presence of close interactions.

6. Acknowledgments

We would like to thank DataSensor Spa in the person of CEO Giovanni Fei for having partly granted this research work.

References

- [1] D. J. Beymer. Person counting using stereo. In *Proceedings of the Workshop on Human Motion (HUMO'00)*, 2000.
- [2] R. Cucchiara, C. Grana, G. Tardini, and R. Vezzani. Probabilistic people tracking for occlusion handling. In *Proceedings of International Conference on Pattern Recognition, ICPR'04*, 2004.
- [3] M. Fritzsche, C. Prestele, G. Becker, M. Castillo-Franco, and B. Mirbach. Vehicle occupancy monitoring with optical range-sensors. In *Proceedings of IEEE Intelligent Vehicle Symposium, Parma, Italy*, June 14-17 2004.
- [4] S. B. Gokturk, H. Yalcin, and C. Bamji. A time-of-flight depth sensor - system descriptions, issues and solutions. In *Workshop of Real-Time 3D Sensors and Their Use*, 2004.
- [5] M. Harville. 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 2005, Como, Italy*, volume 1, pages 511-517, September 15-16 2005.
- [6] L. Iocchi and R. C. Bolles. Integrating plan-view tracking and color-based person models for multiple people tracking. In *Proceedings of International Conference on Image Processing, ICIP 2005, Genova, Italy*, September 2005.
- [7] C. Tomasi and S. B. Gokturk. 3d head tracking basaed on recognition and interpolation using a time-of-flight depth sensor. In *Proceedings of International Conference on Computer Vision and Pattern Recognition, CVPR 2004, Washington DC, USA*, 2004.
- [8] J. W. Weingarten, G. Gruener, and R. Siegwart. A state-of-the-art 3d sensor for robot navigation. In *Proceedings of International Conference on Intelligent Robots and Systems, Sendai, Japan*, October 2 2004.
- [9] M. T. Yang, Y. C. Shih, and S. C. Wang. People tracking by integrating multiple features. In *Proceedings of International Conference on Pattern Recognition, ICPR'04*, 2004.

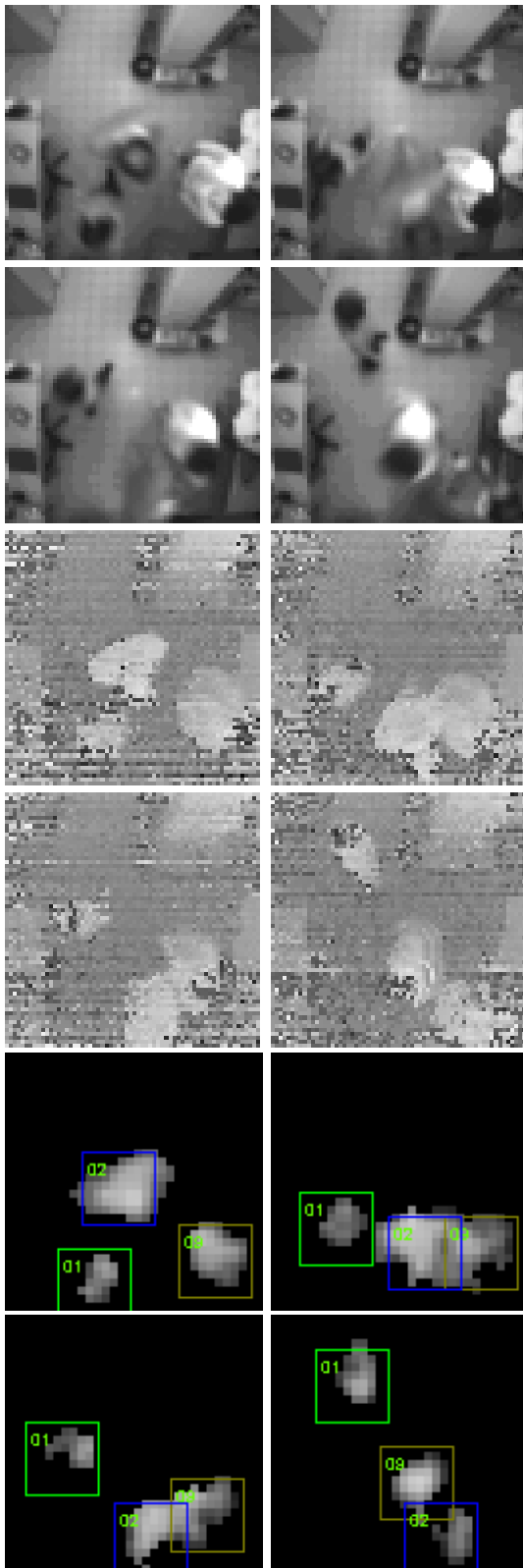


Figure 1: Samples from sequences $S1$: brightness(top), depth images(middle) and plan view maps(bottom)

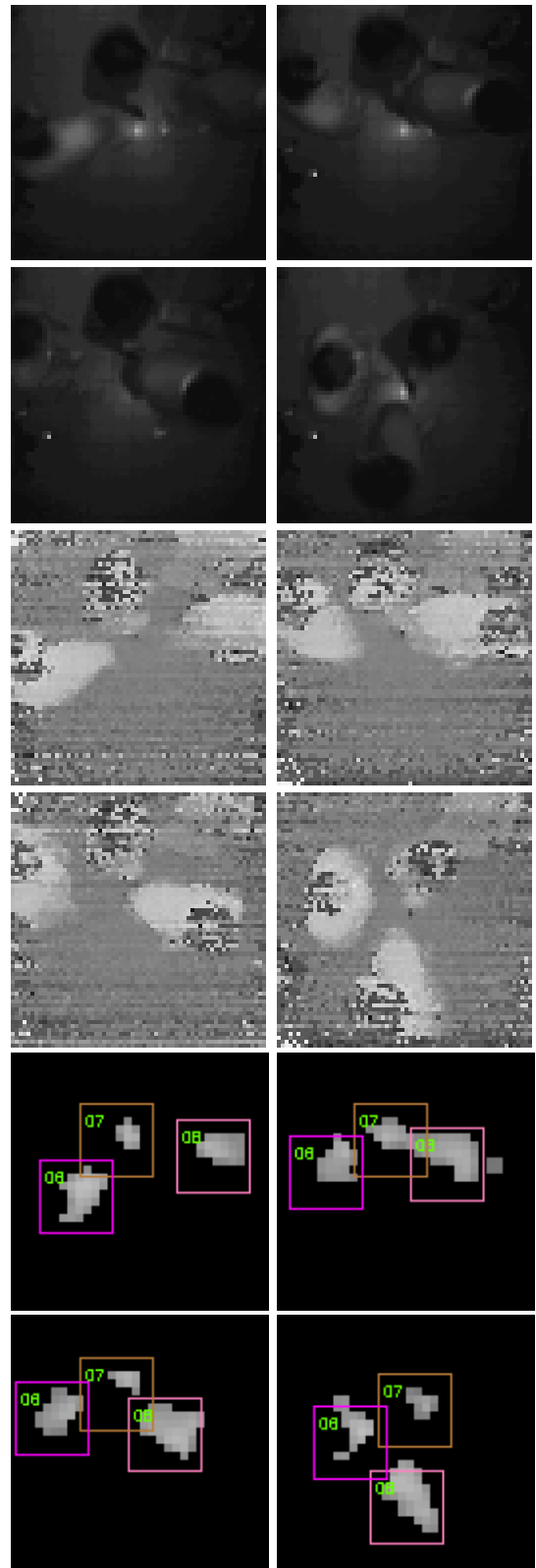


Figure 2: Samples from sequences $S3$: brightness(top), depth images(middle) and plan view maps(bottom)