RESOLVING PARTIAL OCCLUSIONS IN CROWDED ENVIRONMENTS UTILIZING RANGE DATA AND VIDEO CAMERAS

Dejan Arsić, Björn Schuller, Benedikt Hörnler and Gerhard Rigoll

Institute for Human Machine Communication
Technische Universität München, Germany
arsic - schuller - b - rigoll@tum.de

ABSTRACT

Video surveillance systems are omnipresent in our daily life, but still suffer from some drawbacks, which hardens the integration of fully automated systems. Currently standards CCD sensors are used to monitor public and private spaces. These are not yet able to resolve revere occlusions in narrow environments. Therefore we suggest the integration of 3D sensors, in particular a photonic mixture device, into current frameworks, in order to support the reliable detection and segmentation in dense situations. We propose the use of basic techniques to segment persons in range data, to guarantee real-time processing capabilities. With a reliable foreground segmentation and the computation of depth gradients the segmentation performance will drastically rise.

Index Terms— Object detection, PMD, range data, image segmentation, heterogeneous data fusion

1. INTRODUCTION

Most common object detection methods, either based on foreground segmentation [1] or a trained object detector [2], suffer of discriminative abilities in case two or more objects occlude each other. Besides the insufficient detection abilities, tracking is also hardened in case of merging or separation objects. Most 2D tracking approaches, such as tracking of feature graphs [3] or the KLT tracking [4], require an initialization of the objects present in the scene. Otherwise reliable ID maintenance is almost impossible. To cope with this, Elgammal and Davis [5] presented a general framework which uses maximum likelihood estimation and occlusion reasoning to obtain the best arrangement for people. However, a single view often is not sufficient to detect and track objects due to severe occlusion, which as a fact usually requires the utilization of multiple camera views [6]. This approach is usually only applicable in large spaces as it requires overlapping supplementing views. Furthermore person localization usually fails if the feet are not included within the foreground blob. Consequently the feet localization is mapped incorrectly to the ground plane due to the arising problem of plane parralax [7]. Hence it is necessary to evaluate other methods for the detection and segmentation task in images. With new emerging technologies it is possible to overcome these problems. In order to overcome the partial occlusion problem, this work suggests the use of a photonic mixture device [8], based on time of flight. This camera creates a 3D view of the scene and determines the distance between the camera an the object. Fig. 1 displays an exemplary view of the PMD sensor, showing two persons approaching a sofa located in the center of a smart room. Both range and textural information can be captured by the camera, and are perfectly aligned to each other. Though objects in the range image are still occluding each other, it is possible to segment these, by analysis of the distances between the persons present in the scene. In this work we will show a simple, yet effective, processing methods for a reliable segmentation of partially occluding persons. The aim for real time capability requires this simplicity, as surveillance tasks should be also able to prevent possible threats, than just detect threats after they appeared. Therefore we suggest a segmentation in two stages: First the general foreground region is estimated with a reliable method. The resulting foreground mask is subsequently examined for large gaps in between the foreground pixels, which indicate two occluding objects.
This paper is structured as follows: In sec. 2 we will shortly introduce the function principle of the utilized PMD sensor, followed by a brief description of the application scenario, namely smart room, in sec. 3. A foreground segmentation technique based on Gaussian mixture models is presented in sec. 4. This is the prerequisite to limit the following segmentation task in sec. 5 to the regions of interest. We will present that the use of depth gradients is sufficient for the segmentation task, thus being simple and efficient, in the subsequent evaluation in sec. 6. It will be proved, that this approach is successfully segmenting the persons in the scene and provide aligned boundaries. These can subsequently be used to also segment persons in the standard CCD view, even if it is not aligned, see sec. 7. Finally we will present a short conclusion and show some additional possible improvements for future

This work has been funded by the EU funded FP7 PROMETHEUS project. Thanks to A. Rangelov for part of the implementation.

Fig. 1. 2 Persons approaching a sofa in a smart room. From left to right: NIR image of the scene with texture information, the aligned range image, and the segmented foreground region

research in this area in sec. 8

2. THE PMD SENSOR

The image and depth information acquisition principle of the Photonic Mixer Device (PMD) is basing on the run-time difference of a light impulse directly send to the detector and the reflected light from the surface of objects in the environment. In fig. 2 the simplified so-called time-of-flight measurement principle for smart pixel is shown. With utmost precise counters, emitters and receivers the distance between the camera pixel and the object can be approximated by $d = \frac{d}{2}C$, where t represents the measured turnaround time between the start of a light impulse and its return to the receiver. The variable C represents the speed of light. The measurement of the flight-time is carried out using the phase shift of modulated infrared light pulses [9]. By combining several smart pixels in a two dimensional structure an image sensor with fully parallel operating cells arises, allowing the 3D surface reconstruction of the scene. Since this measurement paradigm is directly implemented in the detectors hardware there is no additional computational effort, such as that arising from stereo cameras.. The refresh rate for one measurement loop allows between 5 and 50 frames/second.

To overcome the problem of background illumination, which superposes the running pulse, various further techniques, such as optical filters and active circuits are implemented on the chips cite. The sensors usage of the suppression of background illumination makes it even possible to suppress the effects of bright ambient light thus this measurement becomes independent from existing lighting conditions. The emitted infrared light has a wavelength of 870nm.

By integrating the received light impulses over a certain interval the PMD camera could further serve as a NIR infrared camera. However, there are still some drawbacks that the employed measurement technique is suffering from. Range measurement problems occur in conjunction with highly reflective surfaces that are too close to the sensor. By the mirroring effect of the infrared diodes on the materials surface pixel distances become too large. On light adsorbent materials the depth values are very noisy.

3. THE SMART HOME SCENARIO

Within the EU funded PROMETHEUS project [10] several scenarios have been recorded aiming at the integration of heterogeneous sensor networks for surveillance and security related application. Additionally a so called "smart home" scenario has been chosen, which demonstrates that these techniques could also be implemented in other fields of daily life. It aims to provide help to elderly and disabled persons in their private home and detect possibly dangerous situations. These might be a person fainting, falling down, dropping objects or simply forgetting to switch off the iron. Therefore a smart room, see fig. 3 has been build and scenarios, including up to three persons, have been recorded. As can be seen the room is quite small and therefore severe occlusions frequently appear. For trials two cameras, a thermal sensor and a PMD sensor have been used, which should be able to provide sufficient visual information for behavior interpretation. Almost three hours of data have been recorded during these sessions. Due to temporal reasons not all the data could yet be annotated, hence only 1000 non consecutive frames have been evaluated, which should be a good indicator for the algorithms performance.

4. FOREGROUND SEGMENTATION IN RANGE DATA

For object detection in range data we apply a common adaptive foreground segmentation method, based on works presented by Stauffer and Grimson[11]. Each pixel of the image is modeled by K Gaussian mixtures. This seems reasonable, as each pixel's variance due to noise can be modeled. With $K = 3$ we compute a model for background, foreground and shadow separately. The probability density function for each

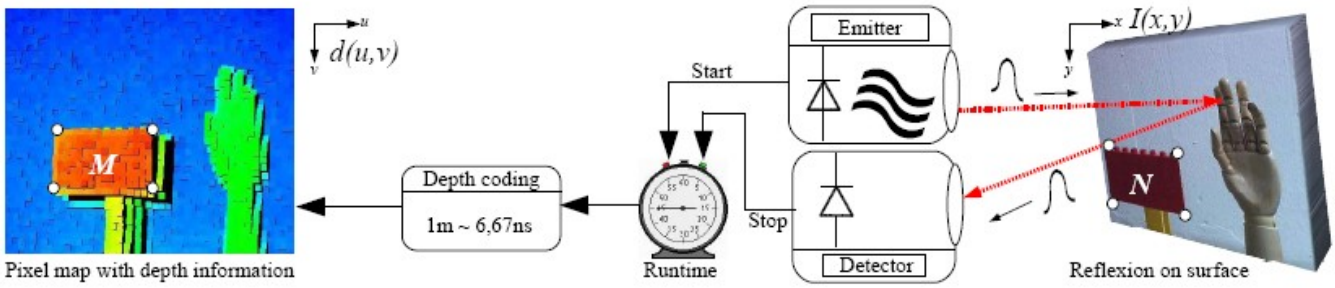

Fig. 2. Time-of-flight measurement principle and calibration body from both sensors perspectives [8].

Fig. 3. Exemplary view of the smart room with the sofa as center of attention and three actors.

pixel is given by:

$$f_{X|K}(X|k,\theta_k) = \frac{1}{(2\pi)^{\frac{\pi}{2}}|\sum_k|^{\frac{1}{2}}} e^{-\frac{1}{2}(X-\mu_k)\sum_k^{-1}(X-\mu_k)}$$

Where X is a vector containing the pixel's range values. Each mixture is defined by $\theta_k = \mu_k, \theta_k$. Taking the probability ω_k into account we get the set of parameters: $\Phi = \omega_1, ...\omega_k, \theta_1, ..., \theta_k$. This way each image pixel is represented by:

$$f_X(X,\Phi) = \sum_{k=1}^{K} P(k) f_{X|k}(X|k,\theta_k)$$

In order to assign the observed pixel to the correct kind of surface the term $P(k,|X,\Phi)$ is maximized by applying the bayes rule and the maximum a posteriori criteria:

$$P(k|X,\Phi) = \frac{P(k) f_{X|k}(X|k,\theta_k)}{f_X(X|\Phi)}$$

$$\hat{k} = \arg\max_k P(k|X,\Phi)$$

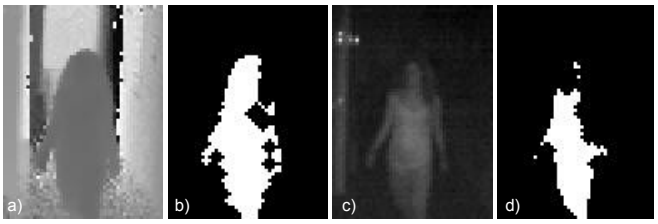

Fig. 4. Comparison of foreground segmentation in range data [a)+b)] and textural information [c)+d)]. Evidently the segmentation is more exact in the range image.

Up to now each pixel has been modeled independently of its neighborhood and some false positives have been produced due to image noise. By applying morphologic operations such as opening and closing [12], noise is eliminated and holes within foreground areas are filled. After this step the original range image is masked so that only the relevant region is visible in the image, to provide more robustness to the subsequent detection step. Connected component analysis [13] is not required, as the blob separation is not yet performed.
The parameter f_α denotes the update time in frames and has to be set carefully. Experiments have shown that a long update time is required, especially if the persons are next to stationary background objects these would be modeled as a background object after a while.
This approach can basically be applied both for the range data and the NIR data, which will not influence the further processing chain. Thus experiments have shown, that the foreground segmentation performs by far better in the range information. Due to the reflective properties of the materials in the scene, foreground and background are frequently confused, as illustrated in fig.4. This of course is not necessary happening in every sensor setup, and therefore should be adopted in each sensor setup.

5. SEPARATION OF HUMANS IN 3D DATA

In the first step, the segmentation of foreground regions in the sensor data, the foreground can be further evaluated. The re-

sulting binary foreground mask FG is now multiplied with the depth image DI, leaving only the region of interest behind. As can be seen in fig. 1, there is still range information left to be utilized, although it is basically only one big blob.
Now that the depth information is limited to the foreground region, it is possible to separate occluding objects. Therefore a so called *depth-gradient*

$$G_z = \frac{\delta f}{\delta z} \tag{1}$$

is computed for the remaining foreground. As the data is basically represented by a two dimensional matrix, the gradient can be computed by computation of the image gradients G_x and G_y

$$\nabla f = \sqrt{G_x^2 + G_y^2} = \sqrt{(\frac{\delta f}{\delta x})^2 + (\frac{\delta f}{\delta y})^2} \tag{2}$$

The resulting gradients G_x, Gy, and ∇f are visualized in fig. 5. Similar to an image gradient, which detects boundaries and high contrasts, the depth gradient detects detects gaps in the rnge data. Obviously the gradients with the highest intensity have been detected mainly at the object boundaries, while some gradients with a smaller absolute value are observed within the objects. As the range image is quite noisy the entire foreground blobs show gradients with values larger than zero. In order to extract the area of interest, here the two separated blobs, it has been decided to threshold the gradient image and remove all gradient values larger than a predefined threshold

$$S(x,y) = \begin{cases} 0 \text{ if } G_z(x,y) > \theta_{dist} \\ G_z(x,y) \text{ else} \end{cases} \tag{3}$$

Experience has proved, that this has to be set individually to each application scenario, as the threshold denotes the minimum distance between two objects.
Now that only the inner surface of the persons remains, these can be used to detect blobs within the gradient image with connected components analysis [13]. The detected blob boundaries hence are the real object boundaries, as illustrated in fig. 6. Subsequently the number of objects in the scene is estimated by applying connected components analysis and counting the number of blobs. This method unfortunately creates some noise in the detection process, as it is highly depending on the quality of the foreground segmentation. Hence often small additional objects can be observed in the segmented data, rising the number of detected persons. As a thresholding based on object size is not applicable, objects almost entirely occluded would also be removed, another method had to be found. Examining the output showed, that these small additional segments are only visible for a short period of time and quite rarely appear. Therefore it is sufficient to window the frame wise output and remove short time peaks in the detected image sequence. Experiments have shown that a simple majority vote within a window of ten frames is ideal for most sequences.

Fig. 6. Separated persons in the NIR image

6. EVALUATION

The presented method has the ability to segment objects partially occluding each other in range data, and create an object shape for each individual in the scene. Due to the lack of annotated data an evaluation of localization accuracy has been discarded. For first trials the number of detected objects in an image frame is compared to hand labeled ground truth, containing the correct number of persons present in the scene. Tab. 1 shows the current results for the smart home scenario recorded for the smart PROMETHEUS projects. 1000 frames, with a constant distance of ten frames have been extracted from the first recorded scenario, for evaluation purposes. This way a sequence with a length of 10.000 length has been processed. As can be seen the detected number [#] of persons in all frames is given and the detected number are indicating a low false positive rate. As more meaningful measure the average number of persons in each frame $avg[\#]$ and the accuracy acc are also given. This results show that in 95% of all frames the number of blobs has been correctly detected. Further a gain of twelve percent in accuracy has been recognized after windowing and averaging the detector output. Both approaches show significantly better results than the simple blob detection in the foreground image.

The presented approach has currently been only implemented in MATLAB for evaluation purposes. Even this implementation has been able to process twelve frames per second, which indicates a guaranteed real-time processing speed in an optimized implementation in C++. This should enable an integration into existing surveillance systems without additional requirements on processing capacities.

Fig. 5. From left to right: The segmented foreground region, the detected depth gradients, two foreground blobs after thresholding.

	[#]	avg[#]	acc.%
ground truth	1400	1.4	–
2D data only	1173	1.2	61,3%
gradient only	1803	1,8	84.3%
with averaging	1498	1,5	95.3%

Table 1. Detected number of persons in the smart home scenario after the detection of depth gradients and averaging.

7. FUSION WITH VIDEO DATA

The separation of the foreground blobs can now be further used for both person identification and behavioral analysis in surveillance applications. Due to the low resolution neither textural nor range information can be considered as reliable feature source. Therefore it is necessary to transfer the extracted information into visual data. Though video cameras can be installed parallel to the 3D sensor, a perfect alignment cannot be guaranteed. A small shift in rotation and translation to the world coordinates' origin will always be experienced. A calibration of the 3D sensor with respect to another coordinate system is rather difficult, as this rather new device is hardly explored and tends to show other error sources than standard CCD cameras. Especially the calibration of the depth information seems rather difficult. One main reason for the distance distortion in the PMD camera is a systematic error due to the demodulation of the correlation function. Beside that, there a several factors like the IR-reflectivity and the orientation of the object that may lead to insufficient incident light to a PMD pixel and thus to an incorrect distance measurement [14]. Therefore the alignment been performed in respect to the PMD sensor's coordinate system rather than world coordinates. Hence Tsai's calibration technique [15] has been utilized to compute the rotation and translation between image and PD coordinates. Due to the low spatial resolution of PMD pixels a large error will be observed in the lateral calibration, whereas depth calibration should be sufficient. Using the homographic transformation

$$p' = Hp, \tag{4}$$

the estimated edge in depth can be transformed from PMD coordinates into image coordinates.

8. CONCLUSION

In the present paper we have presented a novel approach for the segmentation of partially occluding persons by utilizing 3D cameras. The region of interest can be extracted with simple foreground segmentation techniques, here Gaussian mixture models. Depth gradients are computed within the detected area and subsequently be thresholded. This way 95.3% of all images in a short sequence form the Prometheus scenario have been correctly segmented in real-time. The segmented reagions could be overlayed with the aligned NIR image without any additional cost. Further the resulting object boundaries can be transformed into other 2D views of the sequence to aid segmentation by applying homography betwwen the image and the 3D view. This way a combined 2D and 3D tracking approach [16] could be further optimezed, as groups could be easily presegmented and object splits and merges can be easily detected. Additional sensors should also enable the cration of an exact 3D model of the scene and help to resolve total occlusions of objects present in the scene. Advances in range technology will provide additional improvements in segtmentation and tracking abilities, as the limiting factor by now is the low resolution of the sensor. The resolution in depth should be sufficient for most applications besides face recognition, but the lateral resolution is still a major drawback. A higher resolution would enable a more detailed scene representation and more exact trandformation into other image spaces. Furthermore this way a more detailed 3D model of an object could be created to enhance tracking and recognition and even provide information for 3D gesture recognition.

9. REFERENCES

[1] Z. Zivković, "Improved adaptive gaussian mixture model for background subtraction," in *Proceedings 17th IEEE International Conference on Pattern Recognition, ICPR'04*, Washington, DC, USA, 2004, pp. 28–31.

[2] Constantine Papageorgiou and Tomaso Poggio, "A trainable system for object detection," *Int. J. Comput. Vision*, vol. 38, no. 1, pp. 15–33, 2000.

[3] F. Tang and H. Tao, "Object tracking with dynamic feature graph," Oct. 2005, pp. 25–32.

[4] Jianbo Shi and Carlo Tomasi, "Good features to track," Tech. Rep., Ithaca, NY, USA, 1993.

[5] A.E. Elgammal and L.S. Davis, "Probabilistic framework for segmenting people under occlusion," July 2001, vol. 2, pp. 145–152.

[6] D. Arsić, N. Lehment, E. Hristov, B. Hrnler, B. Schuller, and G. Rigoll, "Applying multi layer homography for multi camera tracking," in *Proceeedings Second ACM/IEEE International Conference on Distributed Smart Cameras, ICDSC2008, Stanford, CA, USA*, sep 2008.

[7] S.M. Khan and M. Shah, "A multiview approach to tracking people in crowded scenes using a planar homography constraint," in *Proceedings of the 10th European Conference on Computer Vision, ECCV 2006, Graz, Austria*, 2006, pp. 133–146.

[8] F. Wallhoff, M. Ru, G. Rigoll, J. Gbel, and H. Diehl, "Improved image segmentation using photonic mixer devices," in *Proceedings IEEE Intern. Conference on Image Processing (ICIP) 2007, San Antonio, Texas, USA*. 2007, vol. VI, pp. 53–56, IEEE, 16.-19.09.2007, CD-ROM.

[9] T. Kahlmann, F. Remondino, and H. Ingensand, "Calibration for increased accuracy of the range imaging camera swissrangertm," in *Proceedings of the ISPRS Commission V Symposium Image Engineering and Vision Metrology, D. Schneider Editors: H.-G. Maas, Ed., Dresden, Germany, 25-27 September, 2006IEVM06*.

[10] J. Ahlberg, D. Arsić, T. Ganchev, A. Linderhed, P. Menezes, S. Ntalampiras, T. Olma, I. Potamitis, and J. Ros, "Prometheus: Prediction and interpretation of human behavior based on probabilistic structures and heterogeneous sensors," in *Proceedings 18th ECCAI European Conference on Artificial Intelligence, ECAI 2008, Patras, Greece*, 2008, 21.-25.07.2008.

[11] Chris Stauffer, "Adaptive background mixture models for real-time tracking," in *Proc. of IEEE conference on Computer Vision and Pattern Recoginitions,Fort Collins, USA*, 1999, pp. 246–252.

[12] Bernd Jähne, *Digital image processing (3rd ed.): concepts, algorithms, and scientific applications*, Springer-Verlag, London, UK, 1995.

[13] Rafael C. Gonzalez and Paul Wintz, *Digital Image Processing, Second Edition*, Addison Wesley, 1987.

[14] M. Lindner and A. Kolb, "Lateral and depth calibration of pmd-distance sensors," in *In Proceedings International Symposium on Visual Computing, ISVC06*.

[15] R. Tsai, "A versatile camera calibration technique for high-accuracy 3d machine vision metrology using off-the-shelf tv cameras and lenses," *IEEE Journal of Robotics and Automation*, vol. 3, no. 4, pp. 323–344, Aug 1987.

[16] D. Arsić, B. Schuller, and G. Rigoll, "Multiple camera person tracking in multiple layers combining 2d and 3d information," in *In Proceedings Workshop on Multi-camera and Multi-modal Sensor Fusion Algorithms and Applications (M2SFA2), October 12-18, 2008, Marseille, France*, Oct. 2008.