Downloaded from orbit.dtu.dk on: Nov 07, 2026

DTU Library

TOF Imaging in Smart Room Environments towards Improved People Tracking

Guðmundsson, Sigurjón Árni; Larsen, Rasmus; Aanæs, Henrik; Casas, Josep Ramon; Pardàs, Montse

Published in:
Computer Vision and Pattern Recognition Workshops, 2008

Link to article, DOI:
10.1109/CVPRW.2008.4563154

Publication date:
2008

Document Version
Publisher's PDF, also known as Version of record

Link back to DTU Orbit

Citation (APA):
Guðmundsson, S. Á., Larsen, R., Aanæs, H., Casas, J. R., & Pardàs, M. (2008). TOF Imaging in Smart Room Environments towards Improved People Tracking. In *Computer Vision and Pattern Recognition Workshops, 2008: CVPR Workshop on Time of Flight Camera based Computer Vision* (pp. 1-6). IEEE. https://doi.org/10.1109/CVPRW.2008.4563154

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

TOF Imaging in Smart Room Environments towards Improved People Tracking

Sigurjón Árni Guðmundsson,
Rasmus Larsen and Henrik Aanæs.
Technical University of Denmark
Informatics and Mathematical Modelling
Richard Petersens Plads, Building 321
2800 Kgs. Lyngby, Denmark
{sag,rl,haa}@imm.dtu.dk

Montse Pardàs and Josep Ramon Casas.
Technical University of Catalonia (UPC)
Department of Signal Theory and Communications
Jordi Girona 1-3, Edifici D5
08034 Barcelona, Spain
{montse,josep}@gps-tsc.upc.es

Abstract

In this Paper we present the use of Time-of-Flight (TOF) cameras in Smart-rooms and how this leads to improved results in segmenting the people in the room from the background and consequently better 3D reconstruction of the people. A calibrated rig of one Swissranger SR3100 Time-of-flight range camera and a high resolution standard camera is set in a smart-room consisting of 5 other standard cameras. A probabilistic background model is used to segment each view and a shape from silhouette 3D volume is constructed. It is shown that the presence of the range camera gives ways of eliminating regional artifacts and therefore a more robust input for higher level applications such people tracking or human motion analysis.

1. Introduction

The task we address in this paper is 3D multi-person tracking in smart-rooms. Smart-rooms are multi-modal environments where the audible and visible actions of the people inside the rooms are recorded and analyzed automatically. Typically these systems have to run in real-time which is a large task baring in mind multiple high frame rate cameras, microphone arrays etc. Therefore it is essential to incorporate methods that are both robust and inexpensive computationally
3D People tracking is an important aspect in the over all smart-room concept and is usually performed using many cameras spread around the room.

Multi-camera tracking is dealt with using different approaches, some are based on stereo reconstruction methods [2] that compute correspondences across images and then recover the 3D structure by triangulation and surface fitting. Unfortunately, for operation these techniques demand the camera views to be very close for effective correspondence calculation. This also means that a huge number of points have to be matched and fused into a consistent model, which is a slow and expensive task.

Shape from Silhouettes (SfS) is another approach which has the advantage over most reconstruction methods of being fast and can be performed with relatively few cameras. On the other hand it gives a rather crude 3D result only defining a volume that the object is known to inhabit. If the goal is not photorealistic reconstruction but fast people tracking in 3D, the SfS method can be a good candidate.

SfS is based on binary segmenting several camera views of an object into foreground and background. The foreground silhouettes, the camera positions and projection matrices are then used to define the volume of the foreground. The quality of the resulting volume is thus strongly linked to the quality of the segmentation in each view.

TOF range cameras give depth information per pixel which make them ideal for binary segmentation as in general the depth defines the object from background much better than intensity images; where colors, lighting, reflections and shadows have effect on almost every normal scenario.

In this paper we will focus on the first steps towards people tracking in such multi-camera scenarios i.e. the 3D reconstruction of the persons in the room and show how the use of Time-of-flight (TOF) imaging can help solve some of the more challenging aspects of this problem.
The main contributions can be summed up as: A 2-modal TOF-standard camera rig is installed in a smart-room giving robust segmentation in one view. This error free result has the affect on the reconstruction that regionally the foreground volume is freed from common artifacts that plague methods that rely solely on standard RGB images.

In the next section a description of the setup and the data is given. Section 3 presents the foreground segmentation methods incorporating the TOF data and the SfS 3D reconstruction using this segmentation. In Section 4 we look at some experimental results and finally we draw our conclu-

sions in Section 5.

2. Data and Setup

The presented setup comprises of a 2-modal camera rig that has been installed in a smart-room which has 5 standard RGB cameras. The 2-modal camera rig consists of a Swissranger SR3100 TOF range camera [1] and an additional higher resolution gray scale camera (600×800 pixels) mounted closely beside it (see Figure 1). The smart-room cameras are situated in each upper corner and a zenithal camera catching the birds-eye view.

2.1. 2-modal data fusion

The Swissranger is designed to be a cost-efficient and eye-safe range image solution. Basically it is an amplitude modulated near infrared light source and a specialized 144×176 two dimensional sensor built in a miniaturized package [4]. The sensor simultaneously detects the amplitude and phase of the reflected signal and from this produces respectively an intensity image and a range image.

Because of the low resolution of TOF cameras and very limited intensity images for practical purposes, some have suggested fusing the TOF data with standard high quality images in a calibrated setup. Huhle *et al*. [7] projected the TOF 3D points to the standard cameras image plane and then refined these results using a Markov random field approach. Lindner *et al*. [12] used a computer graphics approach and used biquadratic filtering for refinement. Guðmundsson *et al*. [6] converted the the range-image into a disparity estimate which was then refined in a hierarchical stereo algorithm using a pair of standard cameras.

Here a projection technique is used; the two cameras are calibrated using Bouguet's [3] calibration toolbox and then the TOF measurement points are projected to the high resolution image plane using the projection matrix and information on the distortion, translation and rotation between the cameras. This projection is imprecise due to both the stereo camera calibration, which is subject to error mainly due to the TOF cameras low resolution, and also due to errors in the depth measurements. Kahlmann *et al*. [14] and Lindner [11] *et al*. present methods to calibrate TOF cameras both laterally and also the depth measurements. These methods involve taking many measurements using sophisticated equipment and finding the correct offset for each measurement distance(and also taking into account camera parameters such as integration time). This is typically done only for the center pixels and for one calibration object that has a given reflectance property.

For our purposes a simpler method was developed. As the camera is in a fixed position in a room a 'best scale' is found that makes the 3D measurements fit the high resolution image as well as possible. As the measurements typically get worse when the objects are further away and towards the edges of the field of view, a higher offset is needed in these regions. Using this assumption, a linear scaling triplet of the x,y and z coordinates is estimated and the TOF measurement points are scaled prior to the projection. The points are bilinearly interpolated to 4×4 points and projected to the standard camera image plane. Any missing points are filled with a morphological closing operation using a small 3×3 structure element.

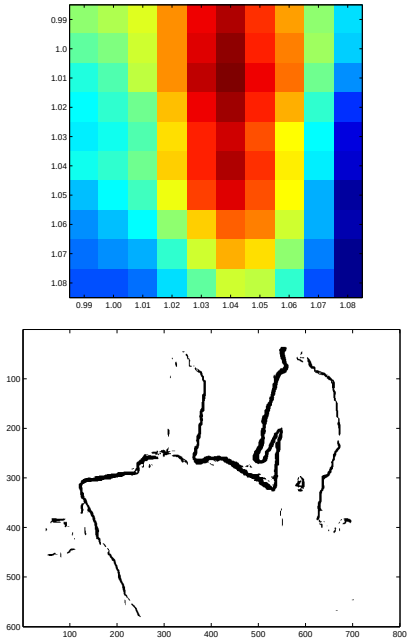

Figure 2. Finding the best alignment. *Above:* Candidates for the best scaling of the x and y coordinates, for a given scale on z, are found where the alignment measure peaks (dark red). *Below:*Co-aligning edges for the the best scales: $(1.04\ 1.01\ 1.04)$

To measure the performance of a scaling triplet, the edges in the resulting range image and standard image are compared (see Figure 2). To find this best scale a fitness measurement is introduced: The two figures strongest edges are found and if the sum of the product of the two edge images is a high number the two images co-align to a high degree. This was tested over a wide range values on a test frame (1000 combinations of scaling from $0.99-1.08$). The best candidates were visually inspected on different scenes for a final estimation. This method produced data that co-aligned well and worked robustly for our purposes.

2.2. The Multi-camera environment

Additionally to the 2-modal TOF-rig we use 5 RGB cameras. All cameras are in fixed locations and run synchronously at 15 fps. The extrinsic parameters of all the cameras is found by using a large calibration object at a known location in the room (see Figure 3). The extrinsic pa-

Figure 1. The 2-modal TOF-rig and it's data. *Left:* The Swissranger SR3100 and standard camera mounted side by side. *Middle:* The 600×800 gray scale image. *Right:* TOF data fused to the standard image.

Figure 3. The smart-room setup. *Above:* The zenithal view of the smart-room. The TOF-rigs position is marked with *red*, camera 1 is in the top left corner . *Below:*View from camera 1

rameters are estimated using Bouguet's [3] calibration toolbox and the results then verified manually.

3. 3D Reconstruction Methods

The 3D reconstruction system we use in the smart-room is a variant of SfS. SfS can be roughly divided into two steps; first a binary foreground segmentation step in each camera and secondly the actual SfS where the segmentation results define the volume or the visual hull [10].

3.1. Foreground Segmentation

Real-time foreground segmentation methods usually perform pixel-based background modelling. The popular Stauffer and Grimson method [13] utilize adaptive Mixture of Gaussian (MoG) background model so that it gains robustness toward background clutter (the MoG model) and slow changes in lighting and displacement of objects (adaptiveness). Typically objects that are moved and then stop get absorbed into the background using such a method.

The degree of adaptiveness must be chosen, regarding the incorporation of foreground pixels into the background for objects that become static for a long period of time. If the application needs to consider these static objects as background, then the background model needs to be updated with the static foreground pixels. For other applications it is better not to include these objects into the background. This is the case of the smart-room scenarios, where people can remain static for long periods, but we wish to continue to track them. In these cases, background models should not be updated with the foreground values' information. This is how our model performs in the experiments in this paper.

Our method is a single level method where we model the background using a probabilistic gaussian model. We only use one gaussian as there is no clutter in our scenario. For the RGB case the model is then:

$$P(X_t) = \frac{1}{(2\pi)^{\frac{3}{2}}\sigma^3} e^{-(\frac{(X_t-\mu_t)^2}{2\sigma^2})}$$

where μ and and σ^2 are the mean and variance of the pixel values and X_t is the input pixel value. Adapting this to the 2-modal rig is straight forward were the gray level and range data can be seen as two channels in the gaussian or only one can be used. Here a background model based on purely the range data from the TOF rig proved very robust. This model is learned in a few frames where no people are present. Foreground is then detected for each frame t on pixel level by a threshold:

$$|X_t - \mu_{i,t}| > k\sigma_{i,t}$$

Where k is a threshold which is scaled by the local variance $\sigma_{i,t}$.

When a pixel value matches this model, the probabilistic

model is updated in order to adapt it to progressive image variations. The update for each pixel is as follows:

$$\mu_t = (1-\rho)\mu_{t-1} + \rho X_t,$$

$$\sigma_t^2 = (1-\rho)\sigma_{t-1}^2 + \rho(X_t - \mu_t)^2$$

where ρ is the update rate.

In intensity image sequences a common false foreground detection is when cast shadows and specular reflections are detected. To reduce this effect a shadow suppression scheme is conducted in a similar manner as suggested by Xu *et al*. [8]. Here these effects are detected by evaluating the variability in both brightness and color distortion between the foreground pixels and the background.This procedure is not as effective in cases where the objects of interest have similar colors as the presumed shadows. To correct this, an assertion process comparing the gradient / textures similarities of the foreground pixels and corresponding background is incorporated. These processing steps, effectively remove cast shadows, but also delete some object pixels and distort object shapes. Therefore, a morphology-based conditional region growing algorithm is employed to minimize this effect.

We will show that this shadow problem does not exist in the foreground detection in the range image sequences and the presence of the camera eliminates some of the shadows in the reconstructed 3D shapes without any shadow removal operations in the foreground segmentation step.

3.2. SfS Reconstruction

Shape from Silhouettes is very intuitive 3D reconstruction method that is computationally effective and widely used in multi camera 3D tracking. Novel approaches of the SfS have emerged in recent years ([5],[9]), however here we use a basic SfS algorithm which illustrates clearly the strengths and weaknesses of our approach.

First the room is divided into voxels: $1.5\ \mathrm{cm}$ each edge and in all $270 \times 370 \times 160$ voxels. Each voxel is then projected to each camera view, if it is projected onto all of the foregrounds, in all the cameras that have this voxel in their field of view, the voxel is marked as foreground. For our method, using a robust depth segmentation in one view, this will have strictly regional impact i.e. outside of the TOF rig's field of view it will not help. Regions that are occluded in this view can also give artifacts. As the current TOF camera technology does not allow multiple cameras in a room, the position of the camera has to be chosen carefully and can be used to solve a specific problem in some application.

4. Results

The SfS reconstruction was performed using three different approaches. First using the 5 smart-room cameras without any shadow reduction scheme, then with shadow reduction on these views and finally using the range image segmentation as a 6th view and no shadow reduction. Evaluation of our approach is done by comparing the resulting SfS volume of the the two latter approaches and see how effectively they have removed the artifacts of the first approach.

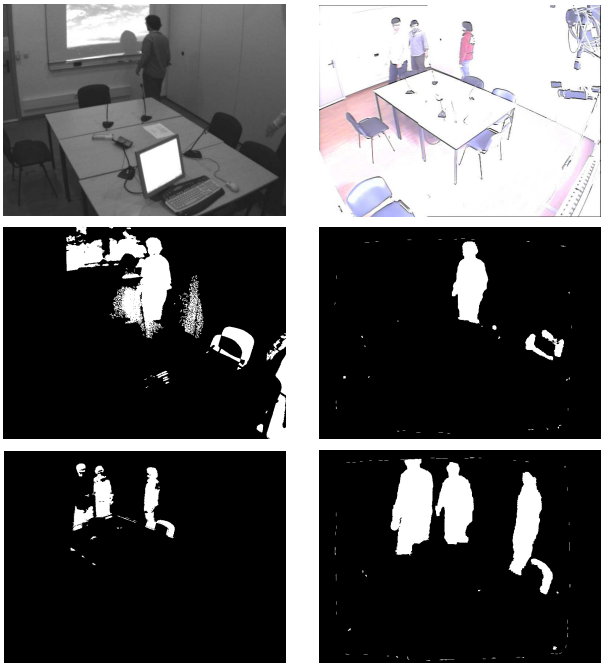

Figure 4. *Top row:* Two test scenarios, with a beamer (Tof -rig view) and sun shining into the room (seen from camera 1). *Middle row*: A beamer causes false foreground detection in the intensity sequence while the TOF sequence is invariant. *Bottom row:* Highlights and shadow suppression cause the bright colored clothes to get fused to the background. The TOF background model proved invariant to lighting, colors and background clutter.

The results of the foreground segmentation on the TOF-rig are shown in Figure 4. The results of the TOF view is totally free from the artifacts the intensity sequences suffer, with and without shadow suppression. Foreground segmentation on the TOF data proved in our experiments totally invariant to light changes, colors of foreground/background and background clutter. We further more provide segmentation videos on the projects homepage [1].

Figure 5 illustrates how the resulting SfS volumes in the first two cases have problems. Without the shadow suppression the volume is far too big and when the suppression is on, the shadows are successfully reduced but some foreground pixels are confused with the background, due to the color similarities. This can be observed in the area of the shirt of one of the persons, thus slicing the person into many

[1] http://www2.imm.dtu.dk/~sag/segmentation.html

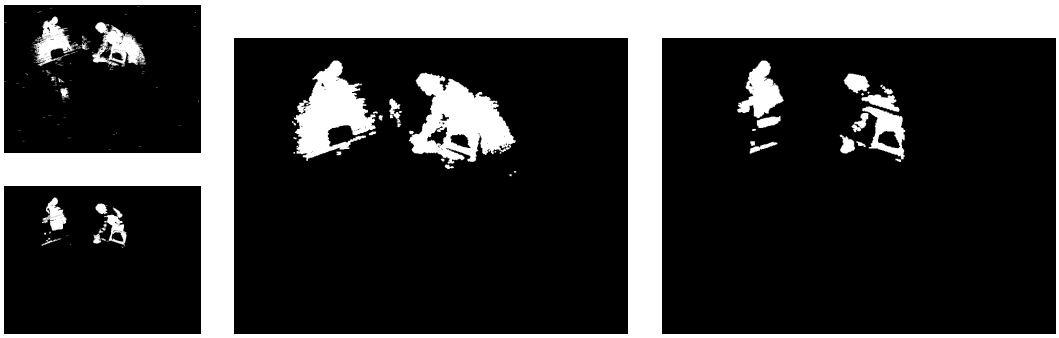

Figure 5. Results of the SfS reconstruction without and with shadow suppression. *Left:* Segmentation results without suppression(above) and with shadow suppression(below). *Middle:* The backprojection of the SfS volume to camera 1 using 5 cameras and no suppression. *Right:* The volume when using the suppression.

parts.

Figure 6 shows again how robust the TOF segmentation is in the presence of shadows. In the far right image the volume is projected onto the same camera view as in Figure 5. This shows how the shadows are removed without distorting the shape of the persons. It is also clear how the camera does not help in regions where it has no coverage as in the gray region.

The SfS Volume seen from the zenithal view is shown in Figure 7. Here it can be seen that the suppression algorithm has reduced the blobs to a smaller size than when using the TOF-view, inspecting Figure 3 however show that the suppression algorithm has deleted object pixels. The occluded regions in the TOF-rigs field of view are defined by the other views without the shadow suppression.

5. Conclusion

We have shown that the incorporation of TOF data in smart-room environments increases the reliability of the 3D reconstruction in smart-rooms. The results show that by using the robust foreground segmentation difficult scenarios can be more effectively handled. The promising results give way for ideas for further development of algorithms. A simple mixed method should perform better than those discussed here above. Where shadow suppression is used to suppress in some views, restricting the result where the TOF camera has no coverage. Many challenging scenarios can be pointed out where our approach would be very supportive. For instance in a meeting scenario where there is a beamer and a standing presenter: the TOF camera can handle the segmentation of the frontal view; efficiently segmenting the speaker without errors caused by the beamer. Other applications could include gesture recognition, motion analysis and gait analysis where e.g. the most interesting view is on the side of the walking person. In such scenarios the perpendicular and top views can then restrict the shadow areas in the TOF view using moderate shadow suppression. A more sophisticated SfS method could also be used to handle the restrictions in an optimal way. E.g. a Bayesian method such as Shape from Inconsistent Silhouettes (SfIS [9]); where the inconsistent foregrounds are managed by choosing the most probable solution where false detections are likely. Here the TOF view could carry a prior so that its' contribution to the resulting volume would have a heavier weight in the regions it has coverage. Our presented framework is a ground on which such more advanced SfS methods can build upon leading to better 3D reconstruction and tracking, this we will investigate in our future work.

Acknowledgements

This work was in part financed by the Graduate School for Biomedical Optics and New Laser Systems(BIOP) under a grant from the Danish Ministry of Science, Technology and Innovation. Also by the ARTTS project which is funded by the European Commission (contract no. IST-34107) within the Information Society Technologies (IST) priority of the 6th Framework Programme. This publication reflects the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

References

[1] http://www.swissranger.ch. *MESA imaging AG*.

[2] T. R. P. Black, J.and Ellis. Multi view image surveillance and tracking. *Proceedings of the Workshop on Motion and Video Computing*, 2002.

[3] J. Y. Bouguet. The matlab calibration toolbox. *http://www.vision.caltech.edu/bouguetj/calib_doc*.

[4] 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. *CSEM, Swiss Center for Electronics and Microtechnology*.

[5] G. K. M. Cheung, T. Kanade, J. Y. Bouguet, and M. Holler. A real time system for robust 3d voxel reconstruction of hu-

Figure 6. Results using the TOF data and no shadow suppression. *Left:* The foreground segmentation of the range data. *Middle:* The projection of the SfS volume to the TOF-rig view. *Right:* The volume projected to camera 1. The gray region is outside of the TOF-rig view.

Figure 7. Results of the three SfS methods projected to the zenithal view. *Left:* No suppression on the shadows. *Middle:* With suppression. *Right:* With the TOF camera.

man motions. *IEEE Conf. on Computer Vision and Pattern Recognition*, 2:714–720, 2000.

[6] S. A. Guðmundsson, H. Aanæs, and R. Larsen. Fusion of stereo vision and time-of-flight imaging for improved 3D estimation. *Int. J. on Intell. Systems and Techn. and App. (IJISTA), Issue on Dynamic 3D Imaging*, 2007.

[7] B. Huhle, P. Jenke, and W. Straßer. On-the-fly scene acquisition with a handy multi-sensor system. *Int. J. on Intell. Systems and Techn. and App. (IJISTA), Issue on Dynamic 3D Imaging*, 2007.

[8] J.-L. Landabaso, M. Pardàs, and J. R. Casas. Shadow removal with blob-based morphological reconstruction for error correction. *Acoustics, Speech, and Signal Processing, ICASSP*, 2005.

[9] J.-L. Landabaso, M. Pardàs, and J. R. Casas. Shape from inconsistent silhouette. *Journal of Computer Vision and Image Understanding*, 2008.

[10] A. Laurentini. The visual hull concept for silhouette-based image understanding. *IEEE Trans. Pattern Anal. Mach. Intell.*, 16(2):150–162, 1994.

[11] M. Lindner and A. Kolb. Lateral and Depth Calibration of PMD-Distance Sensors. 2:524–533, 2006.

[12] M. Lindner, M. Lambers, and A. Kolb. Data Fusion and Edge-Enhanced Distance Refinement for 2D RGB and 3D Range Images. *Int. J. on Intell. Systems and Techn. and App. (IJISTA), Issue on Dynamic 3D Imaging*, 2007.

[13] C. Stauffer and W. E. L. Grimson. Adaptive background mixture models for real-time tracking. *Computer Vision and Pattern Recognition, 1999. IEEE Computer Society Conference on.*, 2, 1999.

[14] F. R. T. Kahlmann and H. Ingensand. Calibration for increased accuracy of the range imaging camera swissranger. *ISPRS Commission V Symposium 'Image Engineering and Vision Metrology*, 2006.