GLOBAL MULTI-VIEW TRACKING UTILIZING COLOR AND TOF CAMERAS BY COMBINING VOLUMETRIC AND PHOTOMETRIC MEASURES

Benjamin Langmann, Klaus Hartmann and Otmar Loffeld
ZESS - Center for Sensor Systems, University of Siegen, Paul-Bonatz-Strasse 9-11, Siegen, Germany
{langmann, hartmann, loffeld}@zess.uni-siegen.de

Keywords: Tracking, Multi-View, PMD, Time-of-Flight, Range data, CONDENSATION

Abstract: In this paper a tracking approach designed to utilize multiple cameras with optional depth information, e.g., ToF cameras, structured light cameras and stereo or multi camera setups, is discussed which combines photometric tracking with volumetric tracking. It is able to work with any number and type of cameras. In order to achieve this objective the tracked object is modeled in 3D with an ellipsoid. To make use of the depth information the density of the observed space is modeled with a set of Gaussian kernels for each line of sight. A proposed target configuration is then evaluated by projecting each observed color image onto the ellipsoid and comparing this projection to the expected appearance. Additionally, the density of the space occupied by the ellipsoid is estimated and compared to the expected density. It is demonstrated that by utilizing the depth information in this way ambiguities due to color similarities can be overcome reliably.

1 INTRODUCTION

Tracking is one of the most important and widely spread video processing steps which has been used since the early days of digital video processing in the fields of motion capture, activity recognition and analysis. Applications range from security and safety systems, assisted living up to gesture recognition and other interaction methods.

The tracking algorithms of choice are the Kalman filter and the later particle filter or CONDENSATION algorithm (Isard and Blake, 1998) which form a widely accepted standard and define mechanisms to locate a given target based on the current observations, i.e., assigning probabilities to proposed target states. Tracking approaches differ in addition to these mechanisms in the way how they describe the target, how they measure the probability of a certain object state, how they adapt the target model over time and so on. These definitions must be adapted to the measurement system, i.e., one or multiple color cameras, Time-Of-Flight cameras, radar and so on.

In this the paper a multi-view and multi-modal tracking approach which is able to utilize any number of color as well as 3D cameras is presented. Particle filters are used as the framework and photometric as well as volumetric measures are utilized to assign a probability to a certain object state. A state is modeled with an ellipsoid in 3D space which can be generated from a set of 3D points with the help of the Maximum Likelihood estimator. The density of the space on the other hand is modeled by a set of Gaussian kernels for each pixel or line of sight more precisely. The difference between the expected intersection density and the actual density is used as the volumetric measure. The photometric measure compares the expected appearance and the observed appearance, which is calculated by projecting the input images onto the ellipsoid.

The photometric measure is evaluated for every color camera whereas the volumetric measure is calculated for each ToF camera in the setup. All models and measures will be described in detail and some experiments will be discussed to confirm the capabilities of the approach presented in this work.

This paper is structured as follows: In section 2 the related work is discussed. Afterwards, the approach of this work is presented in section 3. Some experiments are discussed in section 4 and the paper ends with a conclusion and an outlook in section 5.

2 RELATED WORK

Ghobadi et al. used the CONDENSATION algorithm in (Ghobadi et al., 2008) to track a robot arm and personnel in an industry environment by clustering of depth values gained from a 2D/3D camera and in (Ghobadi, 2010) this approach is compared to other techniques.

In (Kahlmann et al., 2007) depth histograms of persons are tracked also by usage of the CONDENSATION algorithm, whereas in (Bleiweiss and Werman, 2009) objects are tracked based on color and depth histograms using Mean-Shift. A method based on separate color and depth histograms working with the CONDENSATION algorithm was presented in (Sabeti et al., 2008). An different approach that involves tracking was presented in (Grest et al., 2007). Here the pose of an human arm was estimated based on a combination of the silhouette and the 3D position of the arm.

In (Gokturk and Tomasi, 2004) depth measurements from a ToF camera are clustered by k-means and these clusters are compared to a training set of clusters representing head and shoulders. The actual tracking consists of correlation matching which is also used in (Bevilacqua et al., 2006) but here with blobs instead of with clusters. In (Witzner et al., 2008) depth and intensity images from a ToF camera are used to firstly construct a background model. Then the 3D points not belonging to the background are projected on the ground plane and clustered. These clusters are what is actually being tracked.

A multi-view tracking approach was presented by Khan and Shah in (Khan and Shah, 2006). They deal with occlusions and ambiguities by combining the information from different cameras and prevent intersections of objects or people in this case through a global homography constraint. Additionally, in (Bernardin et al., 2006) a multi-view method based on blob-tracking was demonstrated in a smart room environment. The targets were characterized by color histograms and Haar-like features.

The approaches presented in (Kahlmann et al., 2007) and (Bleiweiss and Werman, 2009) are the closest ones to the method introduced in this paper. The main difference is that these approaches treat the distance measurement of a pixel like an additional color channel, whereas in this work the object and the space are explicitly modeled in 3D. Thereby, it is possible to combine multiple cameras of different types globally, i.e., one probability density function describing possible object states is estimated and not several which have to be fused later on.

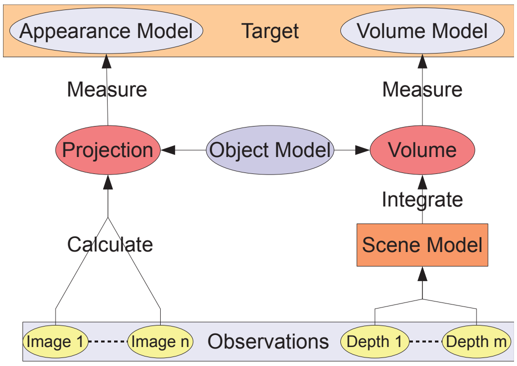


Figure 1: Schematic overview of the approach proposed in this work.

3 Tracking Approach

The tracking method described in this paper is based on the particle filter. It is a technique to estimate a non-Gaussian distribution over a state space. A state X defines a configuration of the tracked object, e.g., the position, size and orientation, and the probability distribution density $p_X(\xi)$ describes the probability of the realization ξ being the true configuration. This distribution is approximated with a set of weighted samples $\{s_i, \omega_i\}$ for $i = 1, \ldots, n$. The weight of a sample is thereby determined by measuring how good a certain configuration $s_.$ explains the current observation. This procedure is called the CONDENSATION algorithm for which an exact description can be found in (Isard and Blake, 1998).

In figure 1 an overview of our approach working with n color and m ToF cameras is illustrated. The likelihood (weight) of a certain configuration $s_.$ is calculated by projecting each color image onto the 3D object model (here an ellipsoid) and by comparing it to the appearance model of the tracking target. The appearance model consists of a photometric model (histogram of the projected image) and a density. In order to determine this density firstly the depth information, gained from the ToF cameras, is used to build up a space model, which consists of a set of Gaussian kernels for each line of sight. The intersection density of an object configuration and the 3D space can then be calculated and compared to the expected density.

In the rest of this section firstly, the object and space models are defined. This section ends with an explanation of the appearance models and the measuring process.

3.1 Object Model

As already mentioned, the tracked object is modeled with an ellipsoid. The reason for this is that a Gaussian distribution $\mathcal{N}(\mu,\Sigma)$ with $\mu \in \mathbb{R}^3$, $\Sigma \in \mathbb{R}^{3\times 3}$ can be created from set of 3D points simply by using the Maximum-Likelihood estimator and when defining a fixed standard deviation or Mahalanobis distance d, this Gaussian describes an ellipsoid.

In the following a distance measure between two samples or states will be explained. The distance between two states described by $\mathcal{N}(\mu_1,\Sigma_1)$ with a volume of V_1 and $\mathcal{N}(\mu_2,\Sigma_2)$ with a volume of V_2 is measured through the Euclidean distance between their centers, their relative difference in volume and with an intersection volume estimate $I(\mathcal{N}(\mu_1,\Sigma_1),\mathcal{N}(\mu_2,\Sigma_2))$ as a fraction of $\max\{V_1,V_2\}$. The intersection volume is estimated by randomly generating N points on the hull of the smaller ellipsoid and checking how many are inside of the larger ellipsoid and, of course, this measure has to be truncated. The similarity measure is then given by

$$\begin{aligned} \varphi((\mu_1,V_1),(\mu_2,V_2)) &= exp\left(-\frac{\|\mu_1-\mu_2\|^2}{2\sigma_{space}}\right) \qquad (1) \\ \cdot\, exp&\left(-\frac{1}{2\sigma_{size}}\log\left(\frac{\min\{V_1,V_2\}}{\max\{V_1,V_2\}}\right)^2\right) \\ \cdot\, exp&\left(-\frac{1}{2\sigma_{is}}\log\left(\frac{I(\mathcal{N}(\mu_1,\Sigma_1),\mathcal{N}(\mu_2,\Sigma_2))}{\max\{V_1,V_2\}}\right)^2\right) \end{aligned}$$

with tuning constants σ_{space}, σ_{size} and σ_{is}.

For simplicity the so-called stochastic diffusion, which is used to generate new samples, is only outlined. The two distributions describing the differences in volume and space are randomly sampled and the ellipsoid is moved and resized accordingly. Afterwards, the ellipsoid is rotated by sampling randomly a Gaussian distribution depending on another constant σ_{angle}.

3.2 Modeling the Density of 3D Space

In this work the density of space is modeled in 3D with the help of depth information. Currently, ToF cameras are used for this purpose which makes it possible to model the observed density for each line of sight. Each of them is assigning a given 3D point a certain probability that it contains mass. This probability is estimated based on the information if a point in space is reflecting light. Then it can be assumed that an object is located there, see figure 2 for plots of the space model for a real video sequence. The space model describes the density of each line of sight with a set of Gaussian kernels. The idea is to represent current and previous observations simultaneously in order to represent the foreground and background accurately.

The variances of the Gaussians provide the means to handle different noise levels, e.g., areas where the depth measurements are extremely noisy the variance will be high and thereby the density of such a line of sight will be assumed to be also high. This will reduce the influence of the volumetric measure since different depth values can not chance the density difference significantly then. For the space behind the deepest Gaussian, i.e., the one the the highest depth value, a minimum density probability of $\gamma_{unknown}$ is assumed.

The probability of mass existing at depth d for a certain line of sight is estimated by

$$p_{xy}(d) = \max_i \{V(d-\mu_i)\} \qquad (2)$$

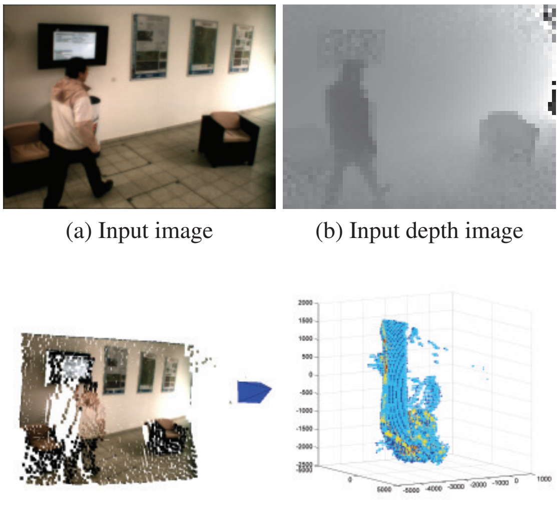
$$V(\tilde{d}) = \begin{cases} \exp\left\{-\frac{\tilde{d}^2}{2\sigma_i^2}\right\} & \tilde{d} < 0 \\ 0 & 0 \le \tilde{d} \le \beta_{width} \\ \exp\left\{-\frac{(\tilde{d}-\beta_{width})^2}{2\sigma_i^2}\right\} & else \end{cases} \qquad (3)$$

with β_{width} being the assumed depth of objects, since only their fronts are observable.

The usual online clustering approach is used to construct the set of kernels $\{\{\mu_1,\sigma_1,\omega_1\},\ldots,\{\mu_n,\sigma_n,\omega_n\}\}$ characterized by their means $\mu_.$, variances $\sigma_.$ and weights $\omega_.$. This procedure is often applied, e.g., for background subtraction, cf. (Stauffer and Grimson, 1999). We consider the background kernels and the kernel belonging to the last observation as being valid. The reason for this is that not only the background or the most often used Gaussian kernels should describe the density of space in the line of sight but also the current observation.

3.3 Measuring and Appearance Models

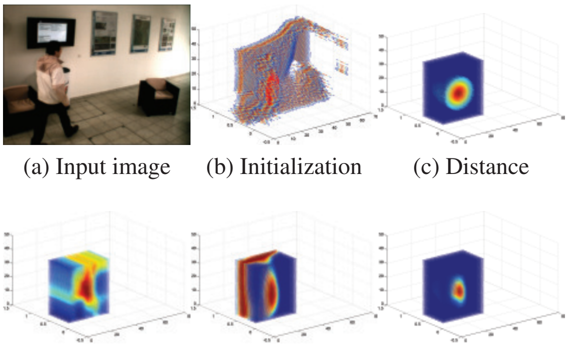
The tracker generates and maintains appearance models which describe the observations which are searched for in consecutive frames. In this work color histograms are used to describe the observations gained from the color cameras (one per camera) and a density model to represent all 3D observations. The weighting measure for a certain configuration (sample) of the object is given by multiplication of the histogram measures and of the volumetric measures. Additionally, a measure based on the distance between the previous or predicted target state and the proposed object configuration can be integrated to smooth the trajectory of the discovered target path and to prevent long jumps or large size changes. This pays respect to the fact that high velocities of the tracked objects are unlikely.



(a) Input image (b) Input depth image

(c) 3D rendering (d) Space model (side view)

Figure 2: 3D space model built from a recorded video using a single MultiCam.



(a) Input image (b) Initialization (c) Distance

(d) Photometric (e) Volumetric (f) Combined

Figure 3: Example scene with per hand initialization of the body of the person and plots of the different measures for different object centers.

3.3.1 Appearance Model

In order to determine how good an actual configuration of the object explains the observations each color image is projected on the object and for every image a histogram is build up from the part of the image which hits the object (to do this efficiently a scanline algorithm or a bounding box of the ellipsoid can be applied). The formulas for this ray casting approach are straightforward and are therefore omitted here.
The comparison of the histogram h_t of the target and the generated one h_g for a proposed target state is done by calculating the sum of squared differences (SSD) of all bins and a similarity measure $\delta(h_t, h_g)$ is given by

$$\delta(h_t, h_g) = \exp\left(-\frac{1}{2\sigma_{hist}} SSD(h_t, h_g)\right) \tag{4}$$

with σ_{hist} being a smoothing constant. Other possibilities to measure the difference between the histograms are the histogram intersection or the Bhattacharyya coefficient. The adaptation over time of the appearance of the target can be achieved with a convex combination of the histograms.
These measure is evaluated separately for every image or camera respectively.

3.3.2 Density Model

The density model simply consists of a value describing the observed density of the object model or ellipsoid. The intersection density of the object model and the space model is estimated for every ToF camera by a discrete integration along each line of sight which intersects the object. For simplicity let $\{\{\mu_1, \sigma_1, \omega_1\}, \dots, \{\mu_m, \sigma_m, \omega_m\}\}$ be the set of all valid Gaussian kernels for the pixel (x, y). Let further the line of sight through pixel (x, y) have the intersection points $X_1 = (x_1, y_1, z_1)$ and $X_2 = (x_2, y_2, z_2)$ with the ellipsoid of the object with a distance $d = \|X_1 - X_2\|$. Then the density Φ_{xy} along the line of sight through pixel (x, y) using an integration step ε is estimated by

$$\Phi_{xy} = \frac{1}{1+\frac{d}{\varepsilon}} \sum_{t=0,\varepsilon,2\varepsilon,\dots,d} p_{xy}\left(\left\|X_1 + t\frac{(X_2 - X_1)}{d} - C_{xy}\right\|\right) \tag{5}$$

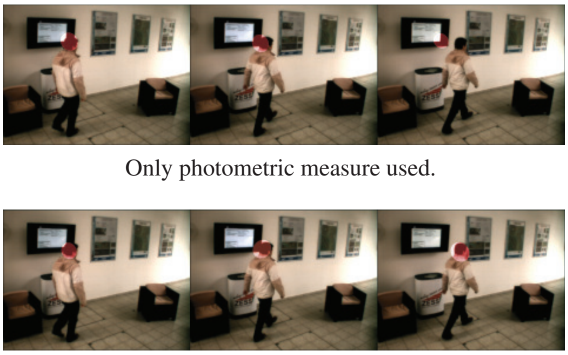
with $p_{xy}(\cdot)$, see eq. 7, being the probability of mass existing at a given point and C_{xy} being the position of the camera. The density of the whole ellipsoid is computed by averaging over all densities Φ_{xy} of all pixels (x, y) which have intersection points.
A similarity measure δ_v between the density Φ_t of the target and the density Φ_c of the current configuration of the object is given by

$$\delta_{vol}(\Phi_t, \Phi_c) = \exp\left(-\frac{1}{2\sigma_{vol}} \log\left(\frac{\min\{\Phi_t, \Phi_c\}}{\max\{\Phi_t, \Phi_c\}}\right)^2\right) \tag{6}$$

with another smoothing constant σ_{vol}. Again the adaptation is performed by a convex combination of the densities.

4 Experiments

The experiments were performed using monocular 2D/3D cameras (Ghobadi et al., 2008) and their focus lies in person tracking especially head tracking.



Only photometric measure used.

Photometric and volumetric measures used.

Figure 4: Experimental results for a head tracking. The head is lost when only the photometric measure is used, but with the additional volumetric measure the target is kept.

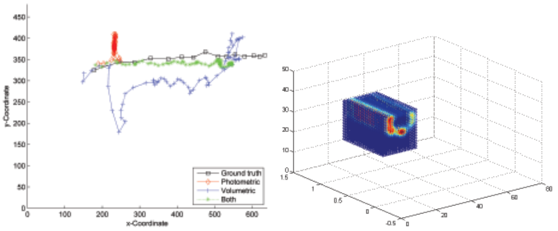


Figure 5: Left: Comparison of the resulting trajectories using only the photometric measure, only the volumetric measure and both measures. Right: Plot of the photometric measure for different object centers.

Figure 3 shows in the first row the color image and an illustration of the 3D space model in which the body of the person was marked per hand. This ellipsoid was moved in the next frame and the photometric, the volumetric, the distance as well as all measures combined were evaluated. The results show only a selected region of the space. For this illustration an orthogonal projection is used and therefore the photometric measure does not change for different depth values. In the plot of the combined measures the center of the body is located with high accuracy.

In figure 4 a few frames of a head tracking experiment using a single 2D/3D camera are shown. Here the initialization was done using a standard face detection method with a subsequent clustering of the 3D points of the head to remove the background. In the first row only the photometric measure was used to weight the particles. Since the head has a color similar to the frame of the LCD TV the head can easily be lost as shown. For a plot of the photometric measure see figure 5, where the measure was evaluated for different centers of the head. Here a position on the TV frame has an equally large likelihood than the true head position. In the second row of figure 4 both measures were used and thereby losing the head can be prevented reliably. A plot showing the resulting trajectories is shown in figure 5.

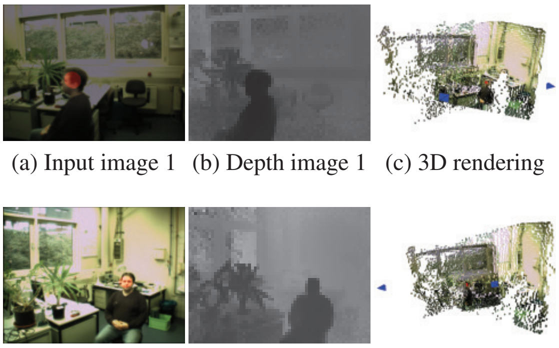
The setup of an experiment utilizing two 2D/3D cameras is illustrated in figure 6. Here the cameras were calibrated and registered using a standard semi-automatic feature based approach. An 3D rendering shows the positions and orientations of the cameras and the scene consists of textured depth measurements. Additionally, the initialization of the tracker is shown which is again based on an standard head detection with subsequent clustering. The video contains fast movements relative to the frame rate and the colors are challenging. On both views the center of the head was marked per hand on all 75 frames of the video. If the (projected) center of the most likely target has a Euclidean distance smaller than 30 pixels to the true center of the head, the target is considered a match. This parameter was variated to ensure the validity of the experiment. In figure 7 the number of matches on any view using different measures and parameters are shown.

Based on this experiment it can be said that volumetric tracking alone does not work reliably. This is quite obvious, since the tracker cannot even distinguish between the body and the head of the person. But on the other hand the volumetric measure is able to enhance the tracking accuracy significantly in conjunction with the photometric measure when compared to a color only based tracking.

The processing time of this tracking approach depends mainly on the number of cameras, the number of particles and the size of the target in pixels. For the head tracking experiment utilizing two cameras with a VGA resolution and 50 particles 5 frames per second can be easily achieved on an standard office computer. Further optimization and possibly the usage of the GPU should allow for real time processing.

5 Conclusion and Future Work

In this paper a rather new tracking approach is presented. It is based on the standard technique of particle filtering and combines photometric with volumetric measures to find the target and is able to utilize multiple color and 3D cameras. The object is represented by an ellipsoid and the appearance of the object is modeled with color histograms. The region of a color image which originated from the object is determined by projecting the image onto the object. Additionally, the density of space is modeled with a set of



(a) Input image 1 (b) Depth image 1 (c) 3D rendering

(d) Input image 2 (e) Depth image 2 (f) 3D rendering

Figure 6: Images made with two registered 2D/3D cameras and 3D renderings with textured depth measurements. The initialization is displayed in 2D and 3D.

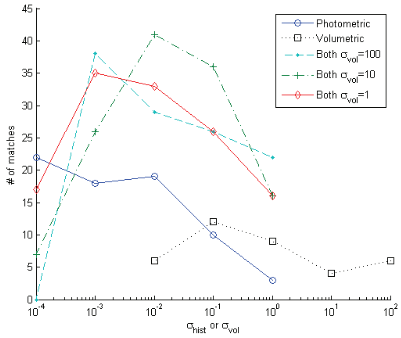


Figure 7: Number of matches for a challenging head tracking experiment using two 2D/3D cameras.

Gaussian kernels for every line of sight which makes it possible to estimate the density of space which is occupied by an object. This density is compared to the expected density of the target in order to benefit from the 3D information.

The general approach was adopted in this work to multiple 2D/3D cameras which combine a color with a PMD chip to gain depth measurements. Additional color or ToF cameras can be incorporated as well without any modifications. So far the control mechanisms which perform the initialization and the termination of the tracker are quite simple, since this work is mainly meant as a proof of concept and not the demonstration of a complete and working system. It is planed in the future to implement more sophisticated management routines and then to use the system for activity recognition. The 3D information may allow for a more precise classification of especially multi agent behavior. Additionally, a global object appearance model is a topic worth researching, since this might improve the tracking in non-overlapping views or the tracking of rotating objects.

REFERENCES

Bernardin, K., Gehrig, T., and Stiefelhagen, R. (2006). Multi-and single view multiperson tracking for smart room environments. *Proceedings of the 1st international evaluation conference on Classification of events, activities and relationships*, pages 81–92.

Bevilacqua, A., Stefano, L. D., and Azzari, P. (2006). People tracking using a time-of-flight depth sensor. *Proceedings of the IEEE International Conference on Video and Signal Based Surveillance*, page 89.

Bleiweiss, A. and Werman, M. (2009). Fusing Time-of-Flight Depth and Color for Real-Time Segmentation and Tracking. *Proceedings of the DAGM 2009 Workshop on Dynamic 3D Imaging*, pages 58–69.

Ghobadi, S. (2010). *Real Time Object Recognition and Tracking*. PhD thesis, Department of Electrical Engineering and Computer Science. to be published.

Ghobadi, S., Loepprich, O., Lottner, O., Weihs, W., Hartmann, K., and Loffeld, O. (2008). Analysis of the Personnel Safety in a Man-Machine Cooperation Using 2D/3D Images. In *IARP/EURON Workshop on Robotics for Risky Interventions and Environmental Surveillance of the Environment*, volume 4, page 59.

Gokturk, S. and Tomasi, C. (2004). 3D head tracking based on recognition and interpolation using a time-of-flight depth sensor. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'04)*, 2:211–217.

Grest, D., Krüger, V., and Koch, R. (2007). Single view motion tracking by depth and silhouette information. *Image Analysis*, pages 719–729.

Isard, M. and Blake, A. (1998). Condensation - conditional density propagation for visual tracking. *International journal of computer vision*, 29(1):5–28.

Kahlmann, T., Remondino, F., and Guillaume, S. (2007). Range imaging technology: new developments and applications for people identification and tracking. *Proc. of Videometrics IX-SPIE-IS&T Electronic Imaging*, 6491(Section 3).

Khan, S. and Shah, M. (2006). A multiview approach to tracking people in crowded scenes using a planar homography constraint. *Computer Vision - ECCV 2006*, 3954/2006:133–146.

Sabeti, L., Parvizi, E., and Wu, Q. M. J. (2008). Visual Tracking Using Color Cameras and Time-of-Flight Range Imaging Sensors. *Journal of Multimedia*, 3(2):28–36.

Stauffer, C. and Grimson, W. (1999). Adaptive background mixture models for real-time tracking. *Computer Vision and Pattern Recognition, 1999. IEEE*, pages 246–252.

Witzner, D., Mads, H., Hansen, S., Kirschmeyer, M., Larsen, R., and Silvestre, D. (2008). Cluster tracking with Time-of-Flight cameras. *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–6.