This CVPR2013 Workshop paper is the Open Access version, provided by the Computer Vision Foundation.
The authoritative version of this paper is available in IEEE Xplore.

Tracking People across Multiple Non-Overlapping RGB-D Sensors

Emilio J. Almazán and Graeme A. Jones
Digital Imaging Research Centre, School of Computing and Information Systems,
Kingston University, Penrhyn Road, Kingston upon Thames, KT12EE, United Kingdom
{emilio.almazan,g.jones}@kingston.ac.uk

Abstract

This work presents the development of a surveillance system for monitoring wide area indoor spaces using multiple Kinect© devices. The data from these sensors, configured with the widest possible coverage, is integrated into a single coordinate system using a novel calibration technique for non-overlapping range sensors. Moving 3D pixels from each Kinect are transformed into a "plan view" map of activity where the detection and tracking of people is executed. The detection of people is a two step process; data binning and non maxima suppression. The tracking of people is based on the mean-shift algorithm optimized with the prediction step of the Kalman Filter.

1. Introduction

Over the last couple of decades considerable effort has been expended on developing reliable surveillance applications primarily using RGB video cameras. The final goal is the design of systems capable of monitoring public spaces autonomously without the need for human operators. However, this task has turned out being highly complex involving the detection and tracking of objects and the identification of illegal or unusual behaviours. The use of range-based sensors such as the Kinect depth sensor brings two main advantages to these surveillance systems. First, these sensors are highly robust against illumination changes, working even in dark environments. Second, it eases the occlusion problem when tracking. Knowing the depth of the scene, static and dynamic occlusions can be predicted and special measures could be applied.

In the context of video surveillance a tracking algorithm consists of maintaining an awareness of the location of individual people in the scene over time. A typical processing pipeline comprises of an initial detection process followed by a track manager, which keeps a consistent id of the people detected throughout their presence in the sequence. A target or person of interest is normally described by an appearance model of geometric, colour or shape image features which are used later by the track manager to match corresponding objects between frames.

Typically there are two main approaches to tracking. The first approach uses the history of objects to predict their future state, and involves two common processes: segmentation and association of models. Kalman filters and particle filters are two well known algorithms within this category. The second approach searches the neighbourhood of the last target position for a similar model. Such *mode finder* techniques include the mean-shift algorithm.

One of the most important challenges in tracking systems is the change in object appearance over time. Such changes can be due to one or more of the following reasons:

- *Illumination.* Changes of the light in the scene will vary the colour representation of the objects.
- *Position.* Changes in object orientation or distance with respect to the camera.
- *Occlusions* may be *dynamic* where a target is occluded by another moving object, or *static* where a target moves behind scene structure such as desks. Such occlusions may be *partial* or *total*.

In this project, a multi-Kinect system is developed to monitor wide area indoor spaces. The Kinect sensors are placed adjacently but with a minimum overlapping configuration to maximise the area covered and minimise the interference between sensors [2]. A novel calibration technique has been developed for estimating the geometric transformations (rotation and translation) between non-overlapping range sensors. Moreover, two novel formulations of the standard visual surveillance pipeline are presented for the detection and tracking of multiple people in indoor environments using the combined RGB-D information provided by the Kinect sensors. Both algorithms are designed to be robust against appearance changes. The detection of people is defined as a two step process executed on the image plane and on a *plan view* respectively. The first step is a depth-based foreground segmentation methodology where the moving pixels in each Kinect sensor are identified and

projected into the *plan view*. A second step follows to segment blobs (objects of interest) in the scene. The tracking algorithm is executed exclusively on the *plan view* and uses the mean-shift methodology where the position of the search window is determined using a Kalman filter. The main contributions of this paper are:

- A calibration method that recovers the geometric transformation between non-overlapping range cameras.
- The development of a people detection algorithm robust against changes in the pose of the objects and illumination conditions of the scene.
- A tracking methodology that addresses static and dynamic occlusions in crowded scenes.

2. Related work

Whilst very common, RGB camera approaches present clear difficulties when dealing with occlusions, cluttered backgrounds or illumination changes; difficulties that have led researchers to explore alternatives methodologies for the detection and tracking of people based on the use of multiple sensors or using different modalities *e.g*. range sensors.

The detection process has two main classes: feature detection such as HOG and background modelling and pixel differencing. The former offers high performance identifying objects, however it requires the training of a classifier and does not handle specially well the occlusions. The latter, it only detects moving objects in the scene by subtracting a background model with the current frame. Although for object identification the process may require further processing such as the implementation of a classifier, in many applications assuming that the moving object is a person is normally true. In addition, it presents a more robust behaviour when facing occlusions.

Many methodologies have been proposed for tracking objects such as Kalman Filters [17], Particle Filters [14] and Mean-Shift approaches [5]. Mean-shift is a very versatile method and is often found in the literature combined with other techniques; for example using a Kalman filter to predict the next target position[4].

These approaches address the problem of the changing appearance of people over time in different ways. Gradual changes, normally produced by variation in the illumination conditions, are mostly handled by updating the model with new observations. For occlusions in monocular views, it is common to find approaches that continue to predict the position of the occluded object while waiting for it to re-emerge [8]. In addition, if the occlusion is partial, the visible part of the object can still be used to maintain the tracking [6]. Alternatively, as a way to reduce the impact of occlusions, cameras may be placed at higher positions and track primarily heads [12, 7]. Nevertheless, with the continued increase in the speed and power of computers, the tendency among the research community has been the use of multiple overlapping-views to handle occlusions [11]. For instance, Numiaro *et al*. [13] present a collaborative system of multiple cameras where tracking is performed in the 'best' view. Harville [10] uses a stereo system to obtain the 3D information and track on a *plan view* where typically there are less occlusions.

The use of range-based sensors in video surveillance offers an attractive alternative to obtain systems independent to illumination changes and occlusions. As an example, Bevilacqua *et al*. [1] presented a monitoring system based on a time-of-flight sensor, which in fact uses a methodology highly related to the one proposed here, though they use a different tracking method based exclusively in geometric features (position and speed). Until the release of the inexpensive $Kinect^{©}$ depth sensor by Microsoft in 2010, range-based sensors were not widely considered for surveillance purposes. The Kinect sensor is a structured light sensor that provides depth information along with colour video, and has had a significant impact on visual surveillance research as well as more general computer vision research in part because it offers a reasonable accuracy at a very low price. Spinello and Arras [16] proposed a new people detection algorithm called Histogram of Orientated Depths (HOD), inspired by HOG features but using depth gradients instead. Choi *et al*. [3] combine image-based and depth-based detectors in order to obtain a more robust detector. Han *et al*. [9] presented a system for monitoring smart indoor environments that uses a depth-based background subtraction method for people detection, similar to the one proposed in this work, along with a tracker based on depth position of people and their colour appearance.

3. Geometry and Calibration

The proposed wide area monitoring system consists of three Kinect sensors mounted in an adjacent configuration with minimum overlapping as shown in figure 1. Such a configuration allows wide areas to be monitored as the FoV of the device is the aggregation of the FoVs of the three Kinects, while at the same time avoiding the interference that occurs when multiple Kinects project overlapping infra-red dot patterns on the same surfaces within the scene. The device proposed can monitored a room of about 20x25ft (limited in depth by the Kinect range). Optimal

Figure 1: Geometry of the multiple Kinect system

placement for the device are mounted high on the walls.

3.1. Calibration

A novel calibration technique has been developed to enable data generated by each RGB-D sensor to be expressed within a common coordinate system (CS). This calibration process entails the selection of a reference CS (the middle Kinect sensor) and the recovery of the rotation and translation of the other sensors with respect to this reference CS. Typically, the calibration between two CSs is based on the use of corresponding points. Unfortunately this approach is not available for non-overlapping view volumes. Instead the approach presented here exploits the depth capability of the Kinect by using planes as common features. The transformations between CS's can be recovered from the parameters of at least three mutually orthogonal planes extracted in each pair of depth sensors. A calibration tool has been built in order to provide a large over-determined set of such planes. This tool consists of a pole (1.7m length) with two boards (32x18cm) attached at both edges in a way that both boards belong to the same plane. Holding the tool in front of the two cameras, each board can be detected by a different camera creating a pair of corresponding planes.

Rotation

The rotation between a pair of Kinects is estimated by using the normal vectors of a set of corresponding planes[1]. As illustrated in figure 2, the transformation for a pair of corresponding normal vectors $(\widehat{\mathbf{n}}, \widehat{\mathbf{n}}')$ can be modeled with a 3x3 rotation matrix as follows: $\widehat{\mathbf{n}}R = \widehat{\mathbf{n}}'$. However, the following conditions must hold for R: $R^T = R^{-1}$ i.e. R is orthogonal, $det(R) = 1$, and $||R_{i:3}|| = 1$, where the columns R_i of R are unit vectors. This rotation is calculated using the method described by Sorkine [15] which guarantees all these properties.

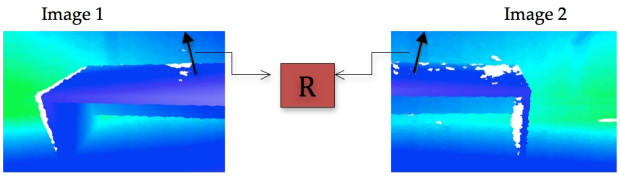


Figure 2: Corresponding Normal Vectors of two depth views from different Kinect sensors

Translation

Based on the rotation obtained, the translation is estimated by error minimization using a set of corresponding points. How these points are obtained is the key innovation of the method. For a plane detected in the non-reference CS (rotated), a unique point can be identified as that point on the plane closest to the origin i.e. $\mathbf{x} = d\widehat{\mathbf{n}}$. This point undergoes an unknown translation to $\mathbf{x}' = d\widehat{\mathbf{n}} + \mathbf{t}$ in the reference CS as shown in figure 3. Since this translated point must lie on the corresponding plane, a constraint on the translation t can be obtained as follows:

$$\widehat{\mathbf{n}}' \cdot (d\widehat{\mathbf{n}} + \mathbf{t}) = d' \quad (1)$$

$$\widehat{\mathbf{n}}' \cdot \mathbf{t} = d' - d\,(\widehat{\mathbf{n}}' \cdot \widehat{\mathbf{n}}) \quad (2)$$

where d and d' are the distances of the plane to both CS origins (local and reference), $\widehat{\mathbf{n}}$ and $\widehat{\mathbf{n}}'$ denote the normal vectors of the plane in both CSs, and $\mathbf{t}$ represents the translation vector between the two CSs. Such a constraint equation is generated for each pair of corresponding planes enabling the generation of the following simple linear estimator of $\mathbf{t}$.

$$N\mathbf{t} = D \quad (3)$$

where

$$N = \left[\widehat{\mathbf{n}}_1'^T, \cdots, \widehat{\mathbf{n}}_M'^T\right]$$

$$D = \left[d_1' - d_1\,(\widehat{\mathbf{n}}_1' \cdot \widehat{\mathbf{n}}_1), \cdots, d_M' - d_M\,(\widehat{\mathbf{n}}_M' \cdot \widehat{\mathbf{n}}_M)\right]^T$$

and M is the number of corresponding planes.

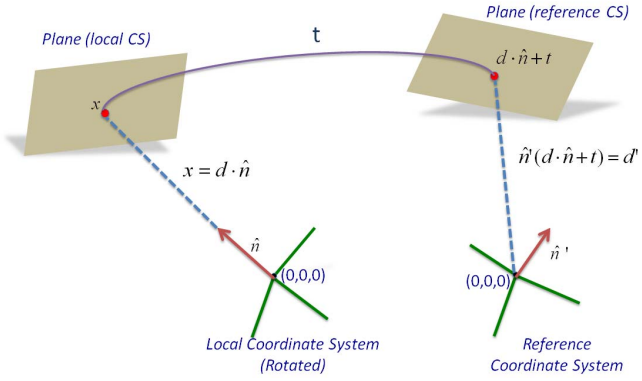


Figure 3: Recovering the translation

4. People Detection

In common with most approaches, the people detection methodology comprises of two sequential steps: *foreground segmentation* to identify moving objects in the scene, and *blob detection* to discriminate people among the innumerable small moving regions detected. The foreground segmentation is applied in the image plane of each camera independently to generate three sets of moving points. Rather than RGB pixels, the technique is applied to the depth information, which makes the algorithm robust against cluttered backgrounds and changing illumination conditions, and works even in dark environments.

[1]The expression "corresponding planes" denotes in this context the same plane represented in different CSs

Foreground pixels are obtained by thresholding the difference between the current depth frame I_t and the depth background model B_t as follow:

$$|I_t - B_t| > \tau$$

where τ is an empirically determined threshold. Pixels with differences larger than τ are classified as foreground and as background otherwise.

The depth-based background model is initialized using the first frames of the sequence on the assumption that there are no foreground objects in the scene. This model is updated selectively over time as follows:

$$B_{t+1} = \alpha I_t + (1 - \alpha) B_t \tag{4}$$

where

$$\alpha = \begin{cases} 0.05 & \text{, if } I_t \notin \text{ background} \\ 0 & \text{, else.} \end{cases}$$

where α is the learning rate. The background regions of the model are not updated as gradual changes are not expected.

The resulting sets of moving pixels, one from each sensor, are then projected into the common CS to create a unified 3D point cloud. This point cloud is projected orthographically onto a *plan view* in a similar manner to Harville [10] where only the highest points in a binned representation are kept. As described below, the final object segmentation and tracking stages are performed on this *plan view*.

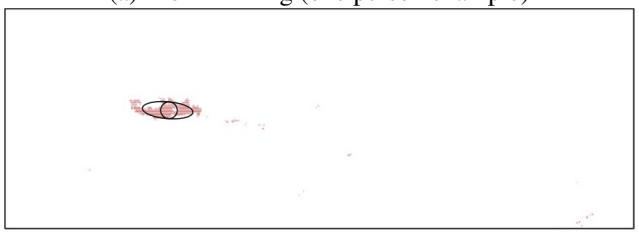
Blob detection

The blob detection process aims to remove the residual noise from the previous step and identify people among the foreground regions detected. The maximum range of the *plan view* is the aggregated field of view of the three Kinect sensors defining the so called *map of activity* (MoA). The MoA is binned to enable modeling of the density of foreground pixels - see figure 4(a). All moving pixels are projected into the MoA and accumulated into their respective bins. A final non-maxima suppression algorithm is applied to identify blobs of people as illustrated in figure 4(b). In more detail, the non-maxima supression finds bins with high values on the MoA, then, centred at each of those bins, a region is defined, whose dimensions have been specified previously according to an average person size on the *plan view*. Finally, a blob is characterized with all the points under the area.

Every detected person is described by a centroid and a set of pixels in the MoA. In addition, each pixel is associated with the colour of the highest of all the points that projected on that particular pixel in the MoA. The colour information will be used in the tracking process.

0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	380	53	9	0	0	0	0	0
0	0	0	0	0	5	0	0	0	0
0	0	0	0	0	0	0	0	0	14

(a) MoA Binning (one person example)



(b) Blob Detection

Figure 4: Processing on the Map of Activity

5. People Tracking

The people tracker presented here combines the mean-shift approach, to search for targets in the current frame, and a Kalman filter to predict the next target location. This combination of techniques was proposed first by Comaniciu and Ramesh [4]. Rather than on the image plane, in this work the tracker is applied on the previously built MoA where occlusions are easier to resolve.

Mean-shift

The mean-shift approach is a probabilistic technique which primarily aims to find modes in density functions but that can also be applied to tracking [5] by searching for a previously built model of a person in the current image. In the proposed implementation, the model, represented by a colour histogram, is built upon the colour distribution of the highest points of a person as captured in the MoA (see equation 5). For each of m bins, the relative probability of a particular bin u is given by

$$p(u) = C \sum_{i=1}^{n} \delta \left[b(x_i) - u \right] \tag{5}$$

where δ is the Kronecker delta function, n is the number of pixels of the person, x_i represents the coordinates of the i^{th} pixel of the person and b is a function that takes the coordinates of an image pixel and returns the corresponding bin in the histogram associated with that colour. The constant C is used to normalize the histogram and is defined as $C = 1/\max(p(u); u = 1 \cdots m)$. This model is searched by mean-shift on the image to determine the new location of the person in the current frame.

Kalman filter

Kalman filtering is a technique that predicts future locations of targets based on their history and it is incorporated into the algorithm as a way to optimize the searching of targets. The predicted position is used as the starting point for mean-shift which normally reduces the number of iterations required to find the target. In addition, Kalman filter predicts a variance in the position, which is used to control the search area for each person in the current frame. Hence, the probabilities are only calculated on the pixels under that area instead of in the entire image.

Other challenges

Independently of the tracking technique used, there are two challenges that any tracking system needs to address: the change in a target's appearance over time (produced by gradual changes in the illumination or object orientation), and the temporal loss of a target due to occlusion or failures in the detection process.

In this work, the first challenge is addressed by updating the model as follows: $p_{t+1} = \alpha d_t + (1-\alpha)p_t$, where p_t and p_{t+1} are the models (histograms) of the person at time t and $t+1$ respectively, d is the colour distribution (histogram) of the target at time t and α is the learning rate at which the model is updated with the new observation. The value of $\alpha = 0,05$ has been determined empirically.

For the second problem, once a target has been labelled as lost[2], the updating of its model is suspended and the algorithm predicts only its location in subsequent frames. The motion model of the target prior to its loss is used in the prediction. Finally, if the target remains lost for more than a certain number of frames, its is automatically discarded.

6. Evaluation

This section describes some qualitative results that illustrate the performance of the detection and tracking in a highly complex environment with a dynamic background and severe occlusions. In addition, it is presented the details of the dataset created for the evaluation of the system.

Dataset

A highly challenging data set has been constructed that represents usual indoor environments. The dataset comprises of 5 sequences of approximately 2 minutes each recorded in a University workshop where students are constantly moving leading to multiple static and dynamic occlusions. The data consists of the RGB-D output of three Kinect sensors set in an adjacent configuration with minimum overlapping at a location of about 2 metres high (see figure 5).

There are two main challenges present in this dataset: a highly dynamic background over a large indoor space. It should be noted that the background is composed of sitting people working at their stations which creates constant unpredictable small movements in the background. The large space is covered by the Kinect sensors with distance up to 10 metres where the depth resolution is very low.

Detection

The detection algorithm shows different performance at different distances. Distant people are more difficult to detect due to the fact that the pixel and depth resolution reduces with distance. Figure 5 illustrates the detection of four people across two of the sensors. In this example, the system failed to detect the person furthest on the right.

In addition, it is important to emphasis that the algorithm continues to successfully detect people when gradual or sudden illumination changes occurred in the scene. It detects people even in dark environments (although this scenario is not considered in this dataset).

Tracking

The tracking of objects in the plan view clearly demonstrates advantages when dealing with all kind of occlusions. The dataset contains many types of static and dynamic occlusions and the tracker successfully resolves most of these even at long distances. Figures 6 and 7 each show a sequence of three frames in time to illustrate the results of the tracking under static and dynamic occlusions respectively. The figures show only the RGB data from the camera where the occlusions occur and the selected region in the MoA.

Empirical Results

Currently an evaluation dataset with accompanying ground truth is in preparation. As a consequence no exhaustive quantitative results are available. However, table 1 presents some empirical values derived from the visual inspection of the algorithm execution on the current dataset videos. The overall performance of the system can be divided according to different ranges of distance regions termed *Near* (0.5-3.5m), *Mid* (3.5-5.5) and *Far* (5.5-10m). Clearly distance has a dramatic impact on performance.

Process	Distance Ranges		
	Near	Mid	Far
Detection	99%	90%	75%
Tracking	95%	80%	60%

Table 1. Empirical results of the detection and tracking processes at different distances.

[2] A target is considered lost in the current frame when the number of points associated with that target, under the search area, is lower than a certain threshold.

Figure 5: Example of Detections in the MoA

Figure 6: Examples of Static Occlusion

7. Conclusions

This paper presents a range-based multi-sensor system for monitoring wide-area indoor spaces. A key step was the development of a novel calibration method based on corresponding planes to allow the data from non-overlapping sensors to be represented in a common CS. Next, a depth-based people detection technique which is robust to varying illumination conditions has been described. Finally, a tracker working on the *plan view* has been described which exploits the range data to resolve static and dynamic occlusions. A very challenging RGB-D dataset is being developed to evaluate the approach which covers a large area indoor space with a extremely complex background. Moreover it includes a significant number of static and dynamic occlusion events at different distances. This dataset is intended to be made public shortly, so researchers will be able to evaluate and compare their results. Due to the increasing popularity of RGB-D sensors in video surveillance applications, it is hoped that the proposed dataset will be an asset and challenge for the wider visual surveillance community.

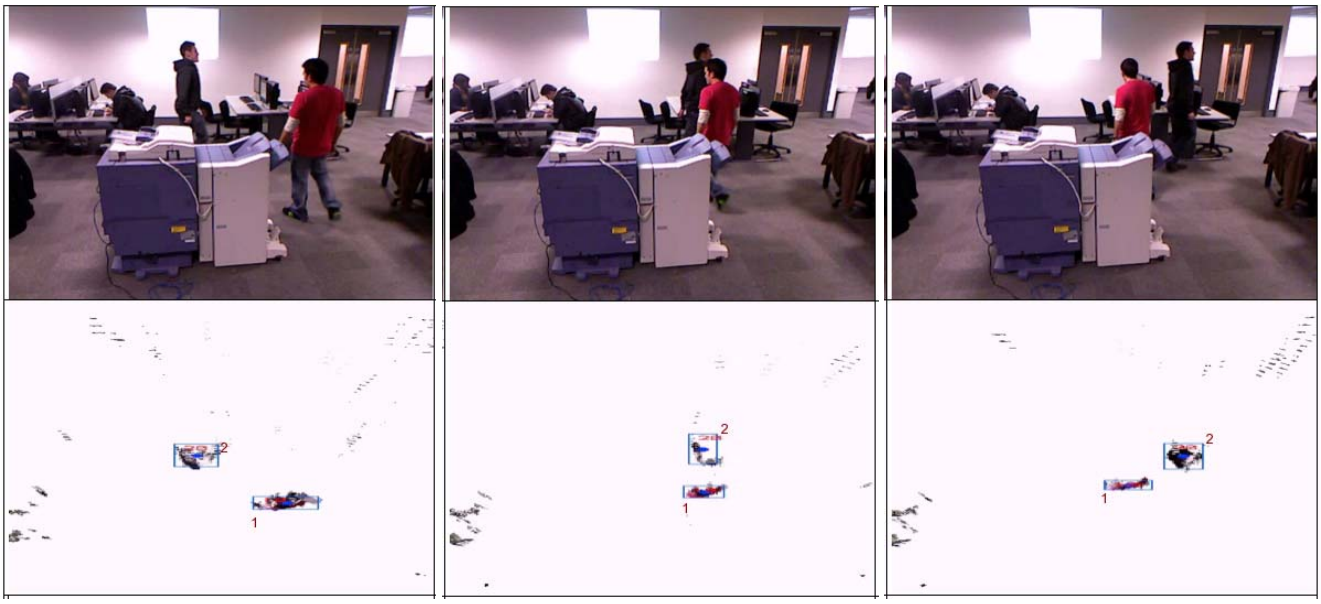


Figure 7: Example of a Dynamic Occlusion

References

[1] A. Bevilacqua, L. Di Stefano, and P. Azzari. People tracking using a time-of-flight depth sensor. In *IEEE International Conference Video and Signal Based Surveillance*, pages 89–89, 2006. 2

[2] D. A. Butler, S. Izadi, O. Hilliges, D. Molyneaux, S. Hodges, and D. Kim. Shake'n'sense: reducing interference for overlapping structured light depth cameras. In *ACM Annual Conference on Human Factors in Computing Systems*, pages 1933–1936, 2012. 1

[3] W. Choi, C. Pantofaru, and S. Savarese. Detecting and tracking people using an rgb-d camera via multiple detector fusion. In *IEEE International Conference on Computer Vision*, pages 1076–1083, 2011. 2

[4] D. Comaniciu and V. Ramesh. Mean shift and optimal prediction for efficient object tracking. In *IEEE International Conference on Image Processing*, volume 3, pages 70–73, 2000. 2, 4

[5] D. Comaniciu, V. Ramesh, and P. Meer. Real-time tracking of non-rigid objects using mean shift. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 142–149. IEEE, 2000. 2, 4

[6] E.Corvee, S.Velastin, and G.A.Jones. Occlusion tolerant tracking using hybrid prediction scheme. *Acta Automatica Sinica*, 29(03):356–369, 2003. 2

[7] R. Eshel and Y. Moses. Tracking in a dense crowd using multiple cameras. *International journal of computer vision*, 88(1):129–143, 2010. 2

[8] L. Fuentes and S. Velastin. People tracking in surveillance applications. In *2nd IEEE International PETS Workshop*, 2001. 2

[9] J. Han, E. J. Pauwels, P. M. de Zeeuw, et al. Employing a rgb-d sensor for real-time tracking of humans across multiple re-entries in a smart environment. *Consumer Electronics, IEEE Transactions on*, 58(2):255–263, 2012. 2

[10] M. Harville. Stereo person tracking with short and long term plan-view appearance models of shape and color. In *IEEE Conference on Advanced Video and Signal Based Surveillance*, pages 522–527, 2005. 2, 4

[11] Y. Kobayashi, D. Sugimura, Y. Sato, K. Hirasawa, N. Suzuki, H. Kage, and A. Sugimoto. 3d head tracking using the particle filter with cascaded classifiers. In *Proceedings of the British machine vision conference (BMVC)*, 2006. 2

[12] R. Mohedano, C. Del-Bianco, F. Jaureguizar, L. Salgado, and N. Garcia. Robust 3d people tracking and positioning system in a semi-overlapped multi-camera environment. In *IEEE International Conference on Image Processing*, pages 2656–2659, 2008. 2

[13] K. Nummiaro, E. Koller-Meier, T. Svoboda, D. Roth, and L. Van Gool. Color-based object tracking in multi-camera environments. *Pattern Recognition*, pages 591–599, 2003. 2

[14] K. Nummiaro, E. Koller-Meier, and L. Van Gool. An adaptive color-based particle filter. *Image and Vision Computing*, 21(1):99–110, 2003. 2

[15] O. Sorkine. Least-squares rigid motion using svd. *Technical notes*, 120, 2009. 3

[16] L. Spinello and K. O. Arras. People detection in rgb-d data. In *International Conference on Intelligent Robots and Systems*, pages 3838–3843, 2011. 2

[17] M. Xu and T. Ellis. Tracking occluded objects using partial observation. *Acta Automatica Sinica*, 29(3):370–380, 2003. 2