People counting with re-identification using depth cameras

D. Hernandez, M. Castrillon, J. Lorenzo *

SIANI, Universidad de Las Palmas de Gran Canaria, Spain

Keywords: people counting, multisensor detection, depth cameras, re-identification, motion analysis

Abstract

Low cost real-time depth cameras offer new sensors for a wide field of applications apart from the gaming world. Other active research scenarios as for example surveillance, can take advantage of the capabilities offered by this kind of sensors that integrate depth and visual information.

In this paper, we present a system that operates in a novel application context for these devices, in troublesome scenarios where illumination conditions can suffer sudden changes. We focus on the people counting problem with re-identification and trajectory analysis.

Automatic people counting offers different application contexts related to security, safety, energy saving or fraud control. Here we go one step further and give hints to extract useful information using depth cameras. The processing of that information allows us to analyze the individuals behavior, particularly if they go away from the typical trajectory, and the problem of re-identifying people.

1 Introduction

The recent introduction in the mass market of real-time depth cameras [17] has increased the interest about this kind of sensors and their possibilities [16]. Even if they were initially conceived for human-machine interaction, the number of applications and scenarios is not yet closed. Additionally, depth information is accompanied with RGB images, providing roughly aligned data that have attracted researchers attention, as it simplifies some tasks, adding robustness for example against lightning changes.

Automatic surveillance is indeed a current topic of interest. People counting is one of the abilities most demanded, providing information useful in different applications, as for example: 1) the number of passengers getting in/out of a public transport is necessary for control and management, 2) retailers make use of visitors information to analyze their marketing strategy, 3) pubs and discos evacuation protocols are designed according to the building capacity and it must not be exceeded, and 4) the presence control can be used for implementing energy saving politics. In this paper we propose a system to investigate the possibilities that low cost depth sensors can offer to the solution of the people counting problem.

*This work has been partially funded by project TIN2008-06068 from Spanish MICINN.

2 Related work

So far, the different solutions proposed to solve the people counting problem are based on two main technologies: Computer Vision and light beams. On the one hand, Computer Vision techniques has been successfully applied to more and more areas in the recent years (see references in next subsection). This process is favored by the introduction of lower-cost higher-performance hardware and the improvements in the reliability of detection methods. On the other hand, laser sensors have also evolved in the same directions, so that smaller and lighter units are available at a reasonable cost vs. precision ratio.

2.1 Computer Vision Methods

In the literature, we can find many examples of Computer Vision based systems with cameras located both in zenithal and non zenithal position. However for some applications where privacy preserving is a crucial matter, the use of vision-based systems with non zenithal cameras is not permitted.

Chan et al. [2] proposed a method based on analyzing a crowd and making use of mixture of dynamic textures to segment the crowd into different directions; after a perspective correction, some features are computed on each segment and with a Gaussian Process the number of people per segment is obtained. Bozzoli et al. [1] introduced a method for people counting in crowded environments as bus or train gates. The proposal is based on the computation of a running average-like background model applied to edge images in order to avoid the influence of sudden lighting condition changes. Foreground edges are filtered and with the remaining ones the optical flow image is computed. Finally each movement vector is assigned to a segment and all the movement vectors assigned to the same segment can be used to estimate the people passing in each direction.

Vision based techniques are well suited for large, wide and open areas, like train platforms or commercial areas besides gates or corridors, provided that lightning conditions are kept under control.

2.2 Range Laser Methods

Katabira et al. [9] proposed a system based on a sensor mounted on the ceiling of a passage. From the range data acquired by the sensor human shapes can be obtained by transforming the data to $X - Z$ plane. The proposed method detects a passing pedestrian when a prominent object is detected.

Mathews and Poigné [13] introduced a system based on a set of passive infrared beacons. The detection of people is done with an Echo State Network which is trained with a set of motion patterns obtained with a simulator.

Light beams based systems have the advantage of privacy preserving, and are best suited for small areas.

2.3 Hybrid Methods

In order to come together the advantages of light beam and vision based systems, some authors have proposed to fusion laser and camera data [15].

Gwang et al. [11] make use of a laser beam as a structured light source. In this way, 3D estimation can be done in an area by means of the integration of consecutive images. When people cross the area, the obtained pattern allows to count the number of people and also the direction of the movement.

Cui et al. [4] describe a method that fuses data from a laser and a visual tracker. The laser module is based on the integration of several laser readings to detect pair of legs and later tracked using a Kalman filter to estimate the position, velocity and acceleration of both feet. A calibrated camera allows to perform visual tracking with color information which feed a mean-shift tracker. Finally, the results of both tracking process are fused with a Bayesian approach.

The asynchronous combination of range detection and visual detection mechanisms described in [7] compensate some specific problems of each method in scenarios with changing light conditions, exhibiting promising robust results.

2.4 Re-Identification

Once the individuals have been detected and counted, automatic systems can also tackle the task of re-identifying them and observing their space-temporal behavior [10, 14]. The use of depth information in this context is new.

3 System Description

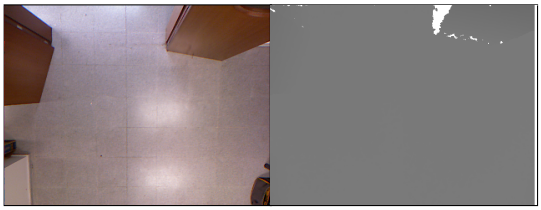
The system is organized in two layers: a main people counting module and a secondary analysis module.

3.1 People counting module

The main purpose of our system is to count the number of persons leaving and entering a space. Other requirements include that the camera position will be fixed, once the initialization is performed. There is no door opening/closing in the camera field of view, and illumination is not controlled, combining areas with natural and artificial light.

The people counting module comprises the following stages: blob extraction, data filtering and crossing event detection.

A) Blob extraction: No motion information in the image space is used to detect the blobs of interest. This approach has the advantage of being less sensitive to visual changes due to illumination. Instead, the system bases its regions of interest on the detection of minima present in the depth image.



(a) (b)

Figure 1. (a) RGB and (b) depth images. Observe in the depth image the artifacts depicted in white.

Once the camera is installed, it is necessary to obtain the background model, that serves additionally to remove shadows. In the depth image those areas where the sensor does not get a correct reading are considered as shadows (white areas in figure 1).

The scene background model (bg) is computed from the initial depth frames. The simple and fast solution adopted here models the background by computing the mean depth value. If the scenario is static, illumination changes will not affect the model as it happens with typical image background models under medium and severe illumination changes.

To detect the regions of interest in the depth images, background subtraction techniques can be applied. Here, an approach similar to the one proposed by Heikkila and Silven [6] where the foreground is computed based on a defined threshold. Those pixels closer to the sensor will be considered foreground, i.e. region of interest. Thus, a pixel (i,j) is marked as foreground if

$$depth(i,j) - bg(i,j) > \tau \tag{1}$$

Unlike in [6], the value of the foreground pixels is kept for further analysis obtaining the thresholded depth image, $s(i,j)$, as

$$s(i,j) = \begin{cases} depth(i,j) & \text{if pixel (i,j) is foreground} \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

The threshold τ in (1) is defined based on the scenario. This value is used to remove from the depth image any area with higher values,i.e. too far values. Some segmentation results are presented in Figure 2 for situations with one or more individuals, and different poses. Depth information provides not just the location but also the height information of each individual.

The background, bg, is computed as the average of the first k depth images as

$$bg(i,j) = \frac{\sum_{l=1}^{k} depth^{(k}(i,j)}{k} \tag{3}$$

where $depth^{(l}(i,j)$ is the pixel (i,j) of the lth depth image from the sequence. Closer pixels are darker, however the depth image may present different artifacts, due to shadows, occlusions and closeness, see Figure 1. When these effects appear, the white color is used to depict the pixel. To avoid the influence of those singularities in the background, a saturation phase is introduced in the background computation as

$$background(i,j) = \begin{cases} bg(i,j) & \text{if } bg(i,j) < \overline{bg} \\ \overline{bg} & \text{otherwise} \end{cases} \quad (4)$$

where $\overline{bg}$ is the average depth value of the background depth image

$$\overline{bg} = \frac{\sum_{i=1}^{width} \sum_{j=1}^{height} bg(i,j)}{width \times height} \quad (5)$$

Assuming that most of the pixels represent the scene floor, this process will force that any pixel with larger values, i.e. clearer, is thresholded to the mean background value.

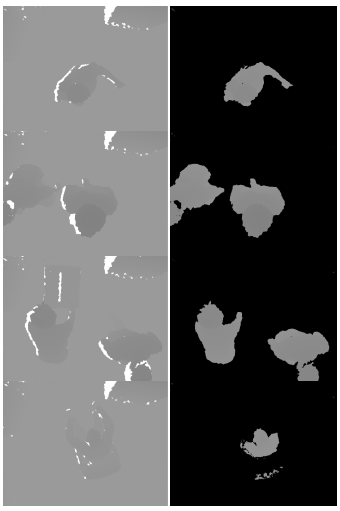


Figure 2. Four samples presenting the depth image and the resulting segmentation, using $th = 0.9$. With one or more individuals, the last row depicts an individual in a wheelchair.

On the segmented area for each big enough component the minimum value (not considering of course the black background) is taken as the person highest point. Figure 3 presents a situation with multiple individuals crossing. The left most column describes some selected RGB frames, the second column represents the corresponding depth images, the segmented images are depicted in the third column (the background are depicted here in white to ease visibility). On each segmented component, the minimum is searched and plotted as seen in the fourth column where previous minima are depicted darker.

B) Data filtering: Once the blobs are detected in the segmented image and the peaks located, they feed a data filtering stage based on an multi-hypothesis Extended Kalman filter (EKF) framework [3]. For each blob b_k, the image location and depth estimation is mapped into 3D world coordinates via a calibration matrix ($b3D_k = MapXYZ(b_k, M_c)$). This calibration matrix includes both a pixel-value to depth conversion and a projective transformation.

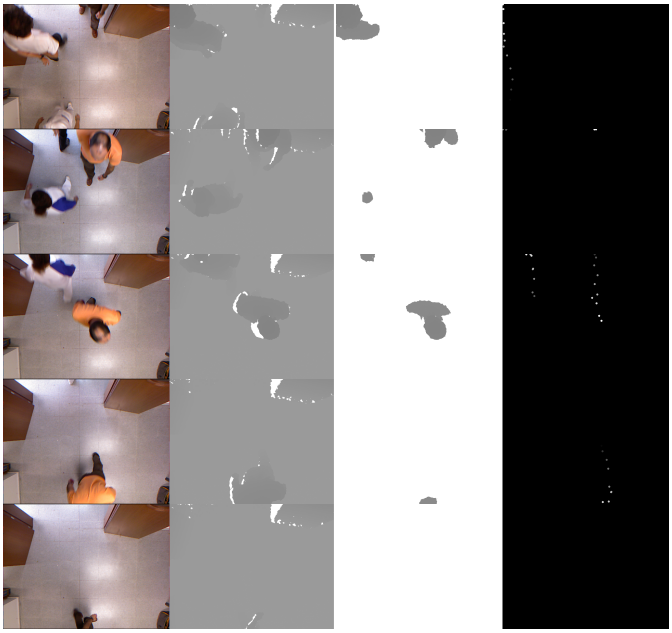


Figure 3. Multiple individual event. Frames 620 to 660 including the RGB, depth, segmented, an trajectory every 10 frames

The set of $b3D_k$ mappings are used to generate object trajectory hypothesis $OTC_c(t)$. These hypothesis are updated on the basis of the $k-th$ detected blob b_k on the current frame, according to the following gate augment/reject rule:

$$\begin{cases} \text{mD} = \min_{i=1\ldots n_c} Dist3D(k,i) & \\ \text{if } (mD < G_R) & \text{EKF update, } OTC_i(t) \\ \text{if } (mD > G_A) & n_c = n_c + 1, \text{ EKF init, } OTC_{n_c}(t) = b3D_k \\ \text{otherwise} & \text{discard} \end{cases} \quad (6)$$

where $Dist3D(k,i) = |b3D_k - OTC_i(t-1)\|$, n_c is the number of current active objects, and G_A and G_R are the gate augment and gate reject thresholds, respectively.

Each EKF filter operates on a four state vector representing the X, Y and Z coordinates of the tracked object and its angle of motion on the XY plane. Initialization considers the vertical position occupied in the image by the seed blob to assign the starting angle for the newly created object hypothesis. If the v blob coordinate is detected on the borders of the image the initial angle points towards the center of the image, otherwise two EKF filters are generated in opposite directions.

C) Crossing event detection: The EKF filters keep integrating data until an object trajectory is not updated with new observations for a given period of time. Then a validation process is applied in order to diagnose if the people in/out counters should be incremented. The factors considered for the decision are the following: trajectory length, number of updates, mean

slope, distribution of XYZ values along the trajectory and uncertainty ellipses.

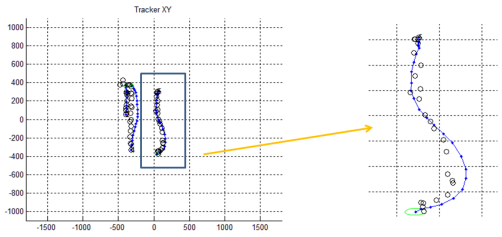


Figure 4. Detail of EKF filtering results from detected image blobs

Figure 4 shows a detail of the EKF blob data processing. It corresponds to the sequence of images shown in Figure 3.

3.2 Analysis module

The analysis module executes two different processes: trajectory analysis and re-identification.

A) Trajectory analysis: As the system evolves in time, additional descriptors can be obtained for the detected in/out events. Our system is able to identify abnormal trajectories on the basis of a density function over metric space. Unusually short or long trajectories are also easily detected using a similar scheme considering now the temporal dimension.

B) People re-identification: People counting is an interesting ability that can be enriched using mechanisms to re-identify the crossing people of application in camera networks where the whole area is not completely covered by the sensors.

Color histograms have been used to model individuals in camera networks in different forms [12, 5], in order to re-identify or track them among the different views. In the approach presented here, we will apply a simple approach as there is no network of sensors, and to evidence that it is enough in some scenarios. Additionally we will observe the use of features of other kind that are provided by the depth camera. Indeed, the camera will give us a top view, with information related to the distance of the individual to the camera.

If the user is in a normal pose standing, the head will be easily extracted, to focus just on the clothes model. Indeed the depth image makes simple to detect the head area (closer to the camera, i.e. darker) and to obtain a mask to establish the area of interest in the RGB image, see Figure 5.

Other elements, such as the user height, can be considered in soft biometrics. The single use of this feature will hardly be able to re-identify people crossing, that are not necessarily previously registered. However, the use of soft biometrics traits can be adopted to model and track a featured individual each time he/she crosses the area [8]. They lack the distinctiveness to identify uniquely an individual, but provide some evidence that can be used to support or discard a hypothesis.

The masked area of the RGB image is used to compute the individual histogram. This histogram is thresholded to avoid noisy pixels. For new images with blobs detected, the back projection is computed for each blob. The ratio of bright pixels for each blob in relation to its size is used as a hint to identify the blob with the histogram previously modeled. The initial model is created in the closest point of the individual trajectory in relation to the image center.

(a) (b)

Figure 5. (a) An individual mask after removing the head, (b) the corresponding RGB image.

4 Experiments

An experimental set have been tested at our laboratory, installing the depth camera on the ceiling next to the entrance. A sequence of 14000 frames has been recorded and manually annotated for validation. A total of 250 crossing events are present in the image set, 120 corresponding to in events and 130 to out events. Among them, 80 events overlap partially or totally in time, including same and opposite trajectory directions. Additionally, 40 events reflect atypical situations such as running, standing, lateral crossing, colliding and occluding.

4.1 People counting and trajectory analysis

Our algorithm was able to detect correctly 241 events with no false positives. Figure 6 shows the detection error and performance evolution along the experiment. The positive axis correspond to entering events and the negative axis for the leaving events. Discarding the first oscillations due to the small number of detections, the system performance is stabilized around 95%.

The nine no detected crossing events correspond to hard to identify situations such as fast and crowded crossings, partially occluded trajectories and erratic movements. In Figure 7, for example, two persons cross the door laterally and the system fails to count them.

Regarding trajectory analysis, the Figure 8 shows the density function obtained during the experiment and used to further characterize events. In the same figure, two examples of standard and atypical trajectories are also depicted.

4.2 Re-identification

As explained above, we have also analyzed the possibilities provided by the sensor to re-identify individuals. For example, we have annotated the crossings of the individual depicted in Figure 5b.

The coincidence of annotation and detections is depicted in Figure 9. In that figure it is suggested the coincidence among

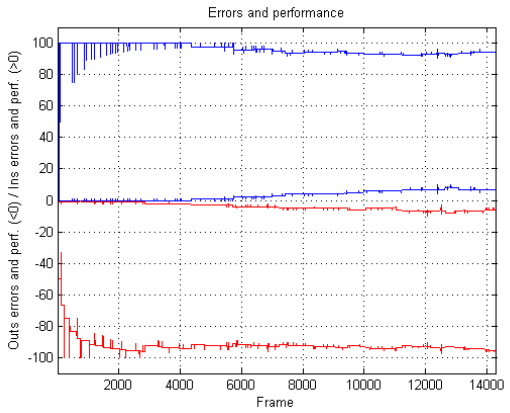


Figure 6. Detection errors (central curves) and performance (top and bottom curves) for in events (positive) and out events (negative)

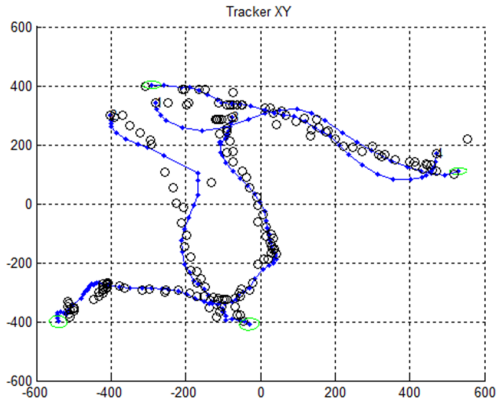


Figure 7. Examples of ambiguous trajectories (horizontal) that the system is unable to mark as crossing events

the ratio assigned to be an individual and the presence of this individual in the scene.

Using only the information combining the resulting back-projection image, the decision threshold to be defined in the period plotted should tackle different difficult to classify situations. On the other side, if that rate is weighted by the height difference of the detected blob and the model height, the threshold is easier to define. If it were fixed to 0.25, re-identifying the individual in 97% of the crossings, and never confusing him with other individuals. The false negative is produced by the entrance of the individual being partially out of the image. Being his head out, the weight introduced by the height difference is incorrectly applied as the individual height could not be correctly measured by the sensor.

5 Conclusions

A simple, robust and low-cost solution to people counting applications based on depth cameras has been described and evaluated. The camera is placed zenithally, configuration that is better suited for different configurations as occlusions among different individuals are particularly unlike, and privacy is preserved as faces are not registered.

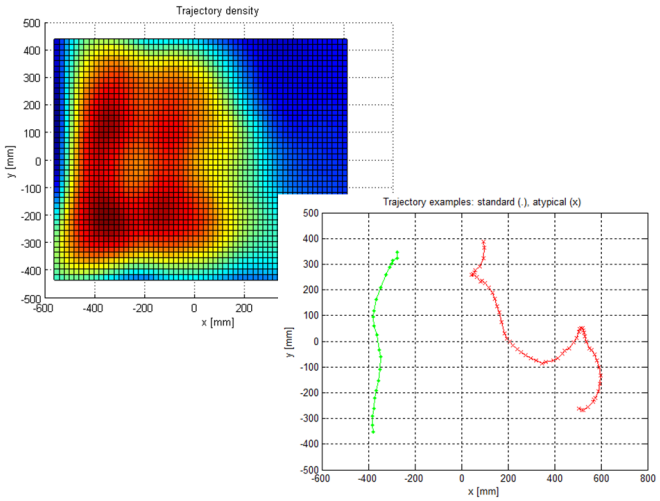


Figure 8. Density function and examples of trajectory classification

A consumer depth camera information is not so precise as a laser sensor, but it gives enough information for the proposed system in the application context considered, with a detection rate greater than 95% and no false detections. The system incorporates an aditional characterization module that enriches the simple crossing event information, providing abnormal trajectory detection and people re-identification.

As future work, this simple approach was robust enough in our scenario, but for more challenging situations such as detection of abandoned objects, an adaptive background model would be needed.

References

[1] Massimiliano Bozzoli, Luigi Cinque, and Enver Sangineto. A statistical method for people counting in crowded environments. In *14th International Conference on Image Analysis and Processing*, 2007.

[2] Antoni B. Chan, Zhang-Sheng J. Liang, and Nuno Vasconcelos. Privacy preserving crowd monitoring: Counting people without people models or tracking. In *Computer Vision and Pattern Recognition*, pages 1 – 7, 2008.

[3] I. J. Cox and Eds G. T. Wilfong, editors. *The Kalman Filter: An Introduction to Concepts Autonomous Robot Vehicles,*. Springer-Verlag, 1990.

[4] Jinshi Cui, Hongbin Zha, Huijing Zhao, and Ryosuke Shibasaki. Multi-modal tracking of people using laser scanners and video camera. *Image and Vision Computing*, 26(2):240 – 252, 2008.

[5] Angela D'angelo and Jean-Luc Dugelay. People re-identification in camera networks based on probabilistic color histograms. In *Visual Information Processing and Communication, SPIE Electronic Imaging*, volume 7882, San Francisco, California, USA, 23-27 January 2011.

[6] J. Heikkila and O. Silven. A real-time system for monitoring of cyclists and pedestrians. In *IEEE Workshop on Visual Surveillance*, pages 82–90, Fort Collins, Colorado, June 1999.

[7] Daniel Hernández-Sosa, Modesto Castrillón Santana, and Javier Lorenzo-Navarro. Multi-sensor people counting. In *IbPRIA*, pages 321–328, 2011.

[8] A. K. Jain, S. C. Dass, and K. Nandakumar. Soft biometric traits for personal recognition systems. In *ICBA2004*, 2004. `http://www.springerlink.com/content/5gg1c23821cevbnk/`.

[9] Kyoichiro Katabira, Katsuyuki Nakamura, Huijing Zhao, and Ryosuke Shibasaki. A method for counting pedestrians using a laser range scanner. In *25th Asian Conference on Remote Sensing (ACRS 2004)*, Thailand, November 22 - 26 2004.

[10] L Kratz and K. Nishino. Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[11] Gwang-Gook Lee, Hyeong ki Kim, Ja-Young Yoon, Jae-Jun Kim, and Whoi-Yul Kim. Pedestrian counting using an IR line laser. In *International Conference on Convergence and Hybrid Information Technology 2008*, 2008.

[12] Christopher Madden, Eric Dahai Cheng, and Massimo Piccardi. Tracking people across disjoint camera views by an illumination-tolerant appearance representation. *Machine Vision and Applications*, 18:233–247, 2007.

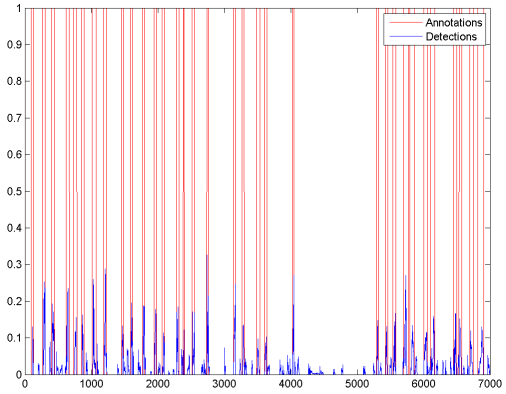
[13] Emi Mathews and Axel Poigné. Evaluation of a "smart" pedestrian counting system based on echo state networks. *EURASIP Journal on Embedded Systems*, 2009:1–9, 2009.

[14] B.T. Morris and M.M. Trivedi. A survey of vision-based trajectory learning and analysis for surveillance. *IEEE Transactions on Circuits and Systems for Video Technology*, 18(8):1114–1127, August 2008.

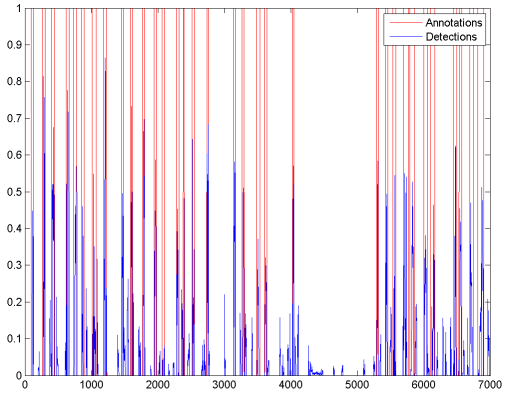
[15] M. Scheutz, J. McRaven, and Gy. Cserey. Fast, reliable, adaptive, bimodal people tracking for indoor environments. In *Proceedings. 2004 IEEE/RSJ International Conference on Intelligent Robots and Systems, 2004. (IROS 2004).*, volume 2, pages 1347– 1352, 2004.

[16] Jamie Shotton, Andrew Fitzgibbo, Mat Cook, and Andrew Blake. Real-time human pose recognition in parts from single depth images. In *CVPR*, 2011.

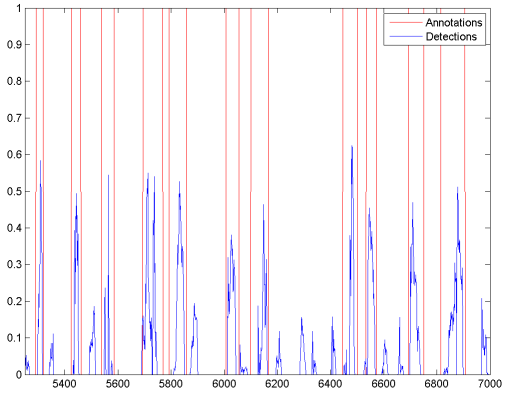
[17] Microsoft Corp. Redmond WA. Kinect for xbox 360.



(a)



(b)



(c)

Figure 9. Annotations vs detections using (a) only the color histogram backprojection, (b) weighted by the individual height difference with the model, and (c) zoomed view of (b).