People Detection in RGB-D Data

Luciano Spinello   Kai O. Arras

***Abstract*— People detection is a key issue for robots and intelligent systems sharing a space with people. Previous works have used cameras and 2D or 3D range finders for this task. In this paper, we present a novel people detection approach for RGB-D data. We take inspiration from the Histogram of Oriented Gradients (HOG) detector to design a robust method to detect people in dense depth data, called Histogram of Oriented Depths (HOD). HOD locally encodes the direction of depth changes and relies on an depth-informed scale-space search that leads to a 3-fold acceleration of the detection process. We then propose Combo-HOD, a RGB-D detector that probabilistically combines HOD and HOG. The experiments include a comprehensive comparison with several alternative detection approaches including visual HOG, several variants of HOD, a geometric person detector for 3D point clouds, and an Haar-based AdaBoost detector. With an equal error rate of 85% in a range up to 8m, the results demonstrate the robustness of HOD and Combo-HOD on a real-world data set collected with a Kinect sensor in a populated indoor environment.**

I. Introduction

People detection is an important and fundamental component for many robots, interactive systems and intelligent vehicles. Popular sensors for this task are cameras and range finders. While both sensing modalities have advantages and drawbacks, their distinction may become obsolete with the availability of affordable and increasingly reliable RGB-D sensors that provide both image and range data.

Many researchers in robotics have addressed the issue of detecting people in range data. Early works used 2D range data for this task [1], [2]. People detection in 3D range data is a rather new problem with little related work. Navarro *et al.* [3] collapse the 3D scan into a virtual 2D slice to find salient vertical objects above ground and classify a person by a set of SVM classified features. Bajracharya *et al.* [4] detect people in point clouds from stereo vision by processing vertical objects and considering a set of geometrical and statistical features of the cloud based on a fixed pedestrian model. Unlike these works that require a ground plane assumption, Spinello *et al.* [5] overcome this limitation via a voting approach of classified parts and a top-down verification procedure that learns an optimal set of features in a boosted volume tessellation.

In computer vision, the problem of detecting humans from single images has been extensively studied. Recent works include [6], [7], [8], [9], [10] that either use part-based voting or window scrolling. In the former approach, body parts independently vote for the presence of a person, in the latter a fixed-size detection window is scrolled over different scale space positions of the image to classify the area under the window. Other works address the problem of multi-modal people detection: [11] proposes a trainable 2D range data and camera system, [12] uses a stereo system to combine image data, disparity maps and optical flow, and [13] uses intensity images and a low-resolution time-of-flight camera.

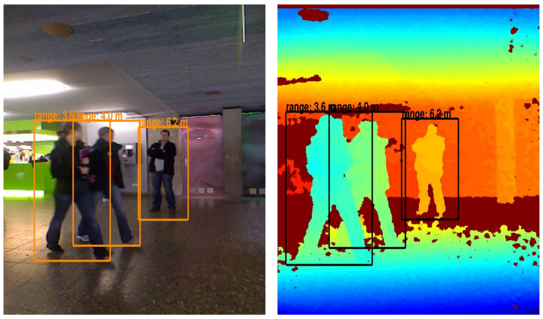

Fig. 1. Detected people in RGB-D data from dense depth data (right) and color image data (left). The method neither relies on background learning nor on a ground plane assumption.

The contributions of this paper to the field of people detection are as follows:

- We develop a robust dense depth person detection called Histogram of Oriented Depths (HOD) that takes inspiration from the method of Histogram of Oriented Gradients (HOG) and from the peculiar depth characteristics of the Kinect RGB-D sensor.
- We perform an informed scale-space search for HOD based on a trained scale-to-depth regression and a novel usage of integral images [14].
- We propose Combo-HOD, a novel principled fusion approach for detecting people in RGB-D data.
- The experiments include a comprehensive comparison with several alternative methods including visual HOG, several variants of HOD, a geometric person detector for 3D point clouds [5], and a Haar-Based AdaBoost detector [15].

Note that the method neither relies on background learning nor on a ground plane assumption.

The paper is structured as follows: the Kinect sensor characteristics are discussed in the next section followed by Section III that presents the HOD detector for dense depth-data and Combo-HOD for detecting people in

All authors are with the Social Robotics Lab, Department of Computer Science, University of Freiburg, Germany. Email: {spinello,arras}@informatik.uni-freiburg.de.

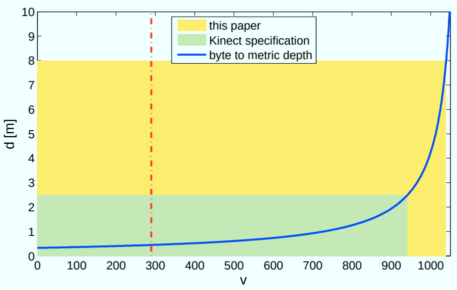

Fig. 2. Kinect depth characteristics. The blue line is the function that relates the byte values of the range image to metric depth, the red line is the sensor's minimal measurable depth. The dark green area indicates the *adequate play space* recommended in the Kinect User Manual, the yellow area is the range considered in this paper for detecting people. Notice that we address the problem of people detection at nearly 4× the suggested working range, where depth resolution is becoming increasingly coarser.

RGB-D data. Section IV describes the data sets, the performance metrics and the comparative experiments. Section V concludes the paper.

II. Kinect Sensor Characteristics

In this section we analyze and discuss the characteristics of the Microsoft Kinect RGB-D sensor used in this paper. The sensor consists in an infrared (IR) camera, an IR projector, and a standard color camera. To measure depth, the sensor follows the principle of structured IR light [16]. The depth image has a 640×480 pixel resolution at 11 bits per pixel. Interestingly, not all bits are used for encoding depth: out-of-range values (e.g. below minimum range) are marked with the value of $V_{max} = 1084$ while the minimum range has been experimentally determined to be $V_{min} = 290$. Thus, only 794 values (10 bits) are used for encoding depth information in each pixel.

The relation between raw depth values v and metric depth in meters d has been experimentally determined to be [17]:

$$d = \frac{8 \cdot B \cdot F_x}{(V_{max} - v)} \tag{1}$$

where $B = 0.075\,m$ corresponds to the distance between the IR projector and the IR camera (the baseline), and F_x is the focal length of the IR camera in the horizontal direction. Negative values of d are discarded. The function 1 is a hyperbolic relationship similar to how depth is determined from point-to-point correspondences in stereo camera systems. Figure 2 shows the relationship and illustrates the *adequate play space* as the space in which the sensor operates reliably specified by the manufacturer [18]. The space is limited at the maximal distance of $2\,m - 2.5\,m$ from the sensor.

In this paper we detect people at $0\,m - 8\,m$ distance, a range that is nearly four times larger than the specification. What makes this task challenging is the loss in depth resolution. 86.9% of the depth values are used to encode the interval between $0\,m$ and $2.5\,m$, leaving just 140 values for describing the remaining $2.5\,m - 8\,m$ range.

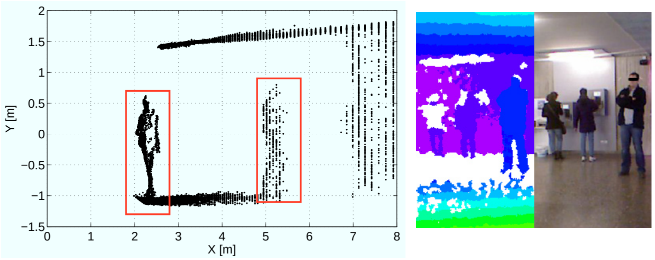

Fig. 3. **Left:** Effects of hyperbolic resolution loss. Side view of two example persons at different ranges from the sensor. Close subjects are accurately described in high detail. Farther away, quantization is becoming increasingly dominant and heavily compromises the shape information on people. Geometric approaches to people detection will perform poorly in such data. **Right:** Example frame to illustrate that IR-absorbing surfaces at large distances lead to blobs of missing depth data (upper body of leftmost subject, white means missing data).

This effect, that follows from the hyperbolic character of Eq. 1, is illustrated in Fig. 3 by the point cloud of two persons at two different distances. While the shape of the person in the foreground at around $2\,m$ is highly detailed, the subject in the background is poorly described with few points at a very coarse range resolution. This makes that the geometrical information content of the 3D data is strongly dependent on range and highly compromised for large distances from the sensor.

Another effect, especially at large distances, is a strong sensitivity on the surface material. Strongly IR-absorbing surfaces cause the projected pattern to be reflected with very low signal strength leading to blobs of missing depth in the image. This effect is shown in Fig. 3, right.

III. Detecting People in RGB-D Data

In this section we introduce the proposed detectors. We first give a summary of the HOG detector for image data. Then we introduce HOD, a novel method for dense depth data that we derive from HOG, and finally present Combo-HOD, that fuses both sensory cues.

A. HOG: Histograms of Oriented Gradients

Histograms of Oriented Gradients (HOG) introduced by Dalal and Triggs [6] is currently one of the most performant and widely used methods for visual people detection [10], [9]. The method considers a fixed-size detection window which is densely subdivided into an uniform grid of cells. For each cell, the gradient orientations over the pixels are computed and collected in a 1D histogram. The intuition is that local appearance and shape can be characterized by a distribution of local gradients without the precise knowledge of their position in the cell. Groups of adjectent cells, called *blocks*, are used to locally normalize the contrast. The descriptor, built by concatenating all block histograms, is then taken for training a linear Support Vector Machine (SVM). For detecting people, the detection window is scrolled over the image at several scales. For each position and scale,

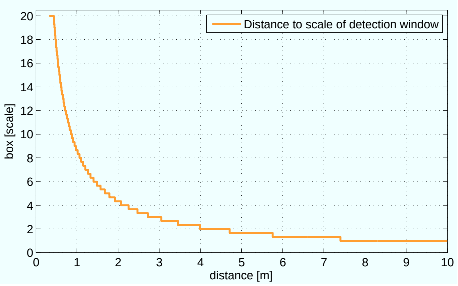

Fig. 4. Quantized regression curve that relates metric depth to the scale of the detection window. The curve is saturated at the maximum scale of 20, to avoid extremely large detection windows.

the HOG descriptors are computed and classified with the learned SVM. For more details, see [6].

B. HOD: Histograms of Oriented Depths

Based on the idea of HOG, we introduce Histograms of Oriented Depths (HOD) as a novel person detector for dense depth data.

1) Operation Principle: HOD follows the same procedure than HOG for the depth image. It considers a subdivision of a fixed window into cells, computes descriptors for each cell, and collects the *oriented depth gradients* into 1D histograms. Four cells also form a block to collect and normalize the histograms to *L2-Hys* [6] unit length and to achieve a high level of robustness with respect to depth noise. The intuition is that an array of local depth changes can robustly characterize local 3D shape and appearance. The resulting HOD features are used for training a soft linear SVM with the same two-stage training method proposed in [6].

2) Depth Image Preprocessing: As discussed in Sec. II, the raw depth image consists of values that very unevenly encode the true metric depth. For far away objects, a difference of one depth value can correspond to a jump in range of 15 cm. This is of particular importance for the HOG/HOD framework since it is known that silhouette blocks at the contour of objects receive very strong weights in this approach. Specifically, these are the *blocks* that correspond to the dimensions of the SVM hyperplane with the highest positive weights. Therefore, we preprocess the raw range image with Eq. 1 to enhance the foreground-background separation. For numerical stability in the gradient computation, we further multiply the resulting metric depth values by M/D_{max} with $M = 100$ being a constant gain and $D_{max} = 20$ the maximum considered range in meters. This preprocessing step resembles in spirit the gamma correction procedure to enhance contrasts in intensity images. Here we can take advantage of the knowledge on the sensor to cancel out the nonlinearity with a physically sound model.

3) Depth-Informed Scale Space Search: Many visual detection methods such as HOG use scale-space search to find objects in an image. In the case of HOD, we can use the depth information to guide this process. The result will not only be a more efficient but also a more accurate search with *informed* scale estimates.

Our idea to improve the search is to create a fast technique to discriminate compatible scales at each position in the depth image. As a first step, the average human height H_m is computed from the training data set, in which ground position and height of each sample is accurately annotated. This information is then used to compute a scale-depth regression as shown in Fig. 4 that follows

$$s = \frac{F_y \cdot H_m}{d} \cdot \frac{1}{H_w} \tag{2}$$

where F_y is the vertical focal length of the IR camera, $H_m = 1.74\,m$ is the measured average height of a person and H_w is the height in meters of the detection window at scale 1. Note that the left term in Eq. 2 represents the image projection of a semi-plane of height H_m, perpendicular to the camera and located at distance d. To limit memory usage, function 2 is quantized each $\frac{1}{3}$ scale. We compute the scale s for each pixel of the depth image to generate a *scale map* from which we derive the list of all used scales $\mathcal{S}$. The list contains only the scales that are compatible with the presence of people in the image. This method avoid the consideration of many scales at a fine resolution which is the case for uninformed search heuristics such as image pyramids.

Given the list of scales $\mathcal{S}$ that is computed once for each image, we can start the informed scale-space search. Only search windows whose depth information corresponds to $\mathcal{S}$ are forwarded to the SVM classifier.

The naive way to address this problem is to select one scale s in the list $\mathcal{S}$ and test if the depth values under the window are compatible with s at each scale-space position. This would involve scanning the entire area under the search window at all positions and test if at least one depth value is compatible with s, a procedure that is computationally expensive especially at large scales.

By using integral images [14], we propose a much faster solution able to test the scale in $O(1)$. Integral images are a technique to efficiently compute the sum of values in a rectangular area of a grid. The value at each image point is the sum of all values above and to the left of the point. The construction of the integral image itself is a $O(N)$ procedure, where N depends on the size of the unscaled original image. The key benefit of integral images is the computation of an area integral with only 4 subtractions. Here we extend the concept to *integral tensors*, multi-layered integral images with as many layers as scales in $\mathcal{S}$ subject to the quantization of Eq. 2. Each layer in the integral tensor is a binary image whose non-white pixels correspond to the layer's scale. This makes it possible to very efficiently test if a given search window contains at least one pixel of a particular scale. The construction of the integral tensor has to be done only once per image.

In the detection phase, a scale s from $\mathcal{S}$ is selected. Then, for each search window position, the test is carried out as an area integral over the search window in the layer of the integral tensor that corresponds to s. If the result is bigger than zero there is at least one compatible depth pixel under the window and HOD is computed. Otherwise the detection window is not considered and the process is continued.

C. Combo-HOD: RGB-D people detector

The two detection approaches described so far consider either image or range data. To take advantage of the richness of RGB-D data, we now propose Combo-HOD, a novel detector that combines the sensory cues. The combination appears promising: depth data are robust with respect to illumination changes but sensitive to low-signal strength returns and suffer from a limited depth resolution. Image data are rich in color and texture, have a high angular resolution but break down quickly under non-ideal illumination.

Combo-HOD is trained separately by training a HOG detector on image data and a HOD detector on depth data. The method fully relies on the informed scale-space search described above: each time a detection window has a compatible scale, HOD descriptors are computed in depth image and HOG descriptors are calculated in the color image using the same window. When no depth data are available, the detector gracefully degrades to the regular HOG detector. A calibration procedure is required to determine the extrinsic parameters that provide the proper correspondence between the two images.

When the HOG and HOD descriptors are classified, the information is ready to be fused. The decision function of a learned SVM is given by the sign of the dot product of the HOD/HOG descriptor with the SVM hyperplane plus the SVM bias. In order to fuse these two pieces of information, we follow the approach by Platt *et al.* [19] and fit a sigmoid function to each SVM output that maps the values onto a probability axis. The probabilities from the HOD detector p_D and the HOG detector p_G are then fused by an information filter

$$p = p_D + k\,(p_G - p_D) \qquad\qquad k = \frac{\sigma_D^2}{\sigma_D^2 + \sigma_G^2}, \qquad (3)$$

where p is the resulting probability of detecting a person, σ_D^2 is set to the ratio of the number of false negatives of the HOD detector divided by the number false negatives of the HOG detector at the equal error rate point of the validation set, and $\sigma_G^2 = 1 - \sigma_D^2$.

IV. Experiments

To evaluate and compare the different detector approaches, we collected a large-scale indoor data set with unscripted behavior of people. The data set has been taken in the lobby of a large university canteen at lunch time. An additional data set has been collected in a visually different university building which is only used for generating background samples. This is to avoid detector bias towards the visual appearance of the canteen lobby, especially since we acquired the data from a stationary sensor. The data set has been manually annotated to include the bounding box in 2D depth image space and the visibility status of subjects (fully visible/partially occluded). A total of 1648 instances of people in 1088 frames have been labeled. The data set are available on the web page of the authors.

As evaluation metrics, we determine precision-recall and the equal error rate (EER). Detections are counted as true positives if the bounding box overlaps with a manually labeled person by more than 40% to account for metric inaccuracies in the annotation and the detection. Adopting the no-reward-no-penalty policy from [9], we do not count true positives or false positives when a detection matches an annotation of a partially occluded person.

The training set for all detectors is composed of 1030 depth data samples of people (that are also mirrored on the horizontal axis) and 5000 negative samples that have been randomly selected from the background data set.

A. Results

We compare the novel HOD detector with other depth-based techniques, visual techniques and the novel multimodal RGB-D detection method Combo-HOD.

Given the importance of depth quantization in Kinect data, we evaluate two HOD variants: HOD11 that considers the full 11 bit range of depths available from the sensor and HOD8 which uses a downscaled 8 bit range. We further compare the HOD detector with other preprocessing techniques than the one described in Sec. III-B. We consider typical techniques from computer vision for contrast enhancement or illumination equalization that include the square root operator, the logarithm operator, and no preprocessing at all.

The experiments in Fig. 5, left, clearly show that HOD11 outperforms HOD8 over the entire precision-recall range: 3 additional bits to encode depth help to disambiguate people from background. This is also true for all preprocessing operations on the depth data (results not shown in Fig. 5). For HOD11, the best preprocessing technique is the one described in Sec. III-B which confirms that a theoretically sound technique outperforms the ad-hoc heuristics. Specifically, the HOD11 has an EER of 83% whereas the best HOD8 variant has an EER of 75%.

A fundamental question in the context of RGB-D data is the contribution of the depth information over purely visual detection techniques. To assess this issue, we consider the performances of the visual HOG detector and a visual Haar-based Adaboost detector (HA) as initially proposed by Viola and Jones [15] that both detect people in RGB images. As can be seen in Fig. 5 left, both methods underperform with respect to HOD11 and Combo-HOD, with EERs of 73% for HOG and only 13% for HA (not shown in Fig. 5). The main reason for these modest results is related to illumination issues. The environment of the data set is not optimally illuminated. Dark areas leads to blurred images of moving people

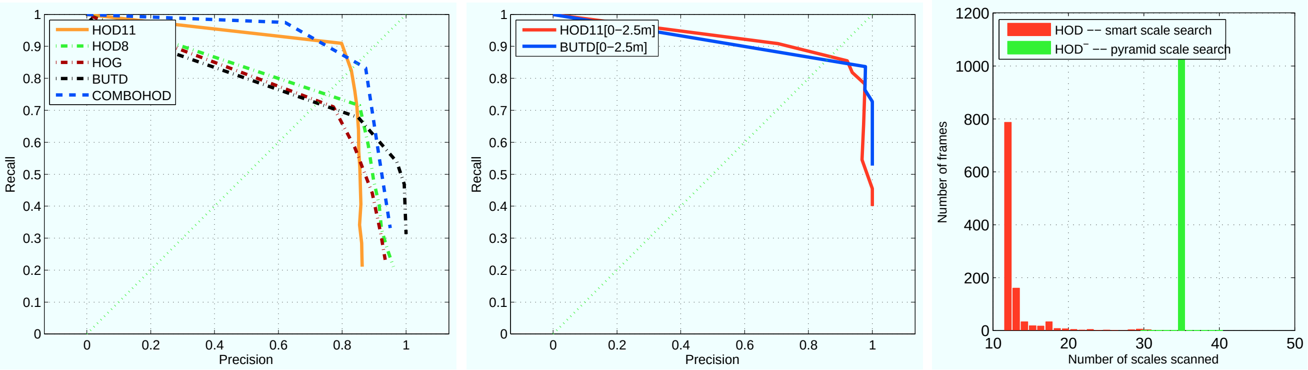

Fig. 5. **Left:** Precision recall curves for depth-based, image-based, and combined RGBD based detection methods. The most performant detector is the combined RGB-D detector, Combo-HOD, that fuses HOD and HOG. HOD is evaluated at different depth discretizations, 8bits and 11bits. HOD11 is the best depth-based detection method. Visual-based HOG detector underperforms due to unfavorable light conditions. BUTD underperforms due to the hyperbolic depth resolution loss of kinect data. **Center:** Evaluation of BUTD and HOD11 in the *Kinect adequate playspace* limited to $2.5\,m$. Both approaches have very similar performance. **Right:** Numbers of scales tested per image with the HOD method and an uninformed HOD method (denoted HOD^{-}). Scale-space search is accelerated by a factor of 3.

as the Kinect RGB camera automatically extends the shutter time to produce brighter images. Background regions with direct sunlight result in saturated image areas and bad contrast. These phenomena also contribute to the failure of the AH method as it uses Haar-wavelets that are not particularly robust to illumination changes. The results demonstrate the need for people detection systems that work in ranges of conditions that are wider than the ones for purely visual detection approaches and motivate the usage of depth information for this task.

Of equal importance is the comparison with geometric approaches in contrast to image-based techniques. We therefore evaluate the HOD11 method with BUTD [5], a 3D person detector for sparse 3D data such as point clouds from a Velodyne sensor. The results are slightly in favor of HOD11 with an EER of 72% (see Fig. 5, center). Note that BUTD still degrades gracefully and can produce a very high precision of 98% at a decent recall of 53%. However, BUTD is a technique that strongly relies on shape information and is therefore compromised by the resolution loss for larger distances from the sensor. Specifically, range image segmentation of BUTD does not work well with coarsely quantized depth data. However, at close range where depth resolution is nearly constant, both detectors perform similarly at an EER of around 86% (see Fig. 5, center). This result demonstrates the appropriateness of shape-based approaches given data of some quality.

The computational performance of the HOD detector is also evaluated and shown in Fig. 5, top. We compare the number of scales that HOD processes per image using the informed scale-space search versus the regular uninformed HOD method (denoted as HOD^{-}). HOD^{-} uses a pyramidal search with a 5% scale increment regardless the image content. This is unlike HOD where scale is a function of depth and changes for each new depth image. We state a nearly three-fold decrease in the number of scales that are searched over all images in the entire data set. This leads to an approximate three-fold acceleration in processing time per image between HOD and HOD^{-}, see Fig. 5, right. The algorithm has been fully implemented on GPUs. The implementation is able to process the RGB-D Kinect data stream ($2 \times 640 \times 480$ pixels at $30\,fps$) in real-time on a Nvida GTX480 card.

Finally, in comparison to all other techniques, the proposed Combo-HOD detector is the winner. Combo-HOD achieves the highest EER of 85% in Fig. 5. This means that the combined use of depth and image information that RGB-D data provide widen the range of conditions under which people detection works reliably. The multi-modality helps to detect people in situations that single-cue detectors cannot deal with.

Qualitative results from the Combo-HOD detector are shown in Figure 6. The figure illustrates several persons detected at different ranges with varying partial occlusions and in different visual clutter conditions.

V. Conclusions

In this paper we introduced Combo-HOD, a novel approach to the problem of detecting people in RGB-D data. We described key insights on the characteristics of Kinect data, the sensor used in the experiments, that guided us in the development of the proposed methods. HOD, that stands for Histogram of Oriented Depths, locally encodes the direction of depth changes and relies on an depth-informed scale-space search that leads to a 3-fold acceleration of the detection process. We then combine the method with visual HOG and propose the Combo-HOD detector that relies on depth and RGB data as sensory cues. The result is a person detector that achieves an Equal Error Rate of 85% in a range of nearly four times larger than the operation space specified by the sensor manufacturer. We have further conducted comparative experiments to analyze the contribution of the depth data over purely visual methods and the performance of shape-based 3D methods. Combo-HOD outperforms all other detection approaches while running at $30\,fps$ on a graphics card implementation.

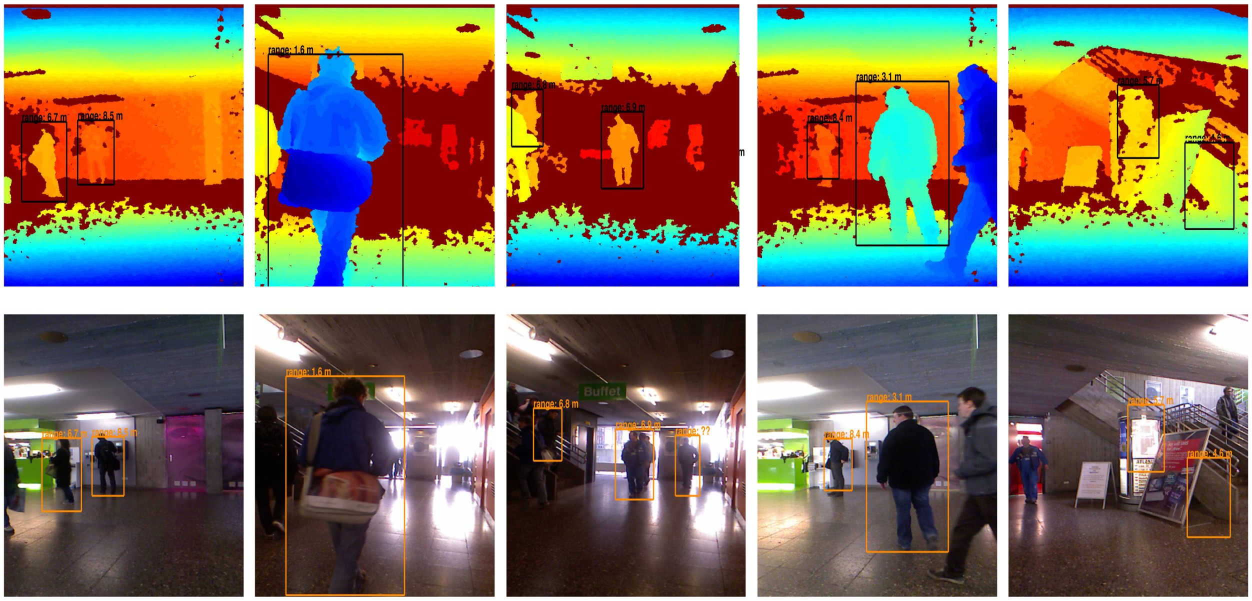

Fig. 6. Qualitative results of people detection in RGB-D data with the Combo-HOD detector. People are detected at several ranges at varying partial occlusions and in different visual and depth clutter. False negatives occur when in both sensor modalities the data are challenging, false positives are found when visual and depth clutter occur simultaneously. In the third column, the detector is able to find a person even though no depth data are available. Note that the method neither relies on background learning nor on a ground plane assumptions.

Acknowledgment

This work has been supported by the German Research Foundation (DFG) under contract number SFB/TR-8.

References

[1] D. Schulz, W. Burgard, D. Fox, and A. Cremers, "People tracking with a mobile robot using sample-based joint probabilistic data association filters," *International Journal of Robotics Research (IJRR)*, vol. 22, no. 2, pp. 99–116, 2003.

[2] K. O. Arras, O. Martínez Mozos, and W. Burgard, "Using boosted features for the detection of people in 2d range data," in *Proc. of the Int. Conf. on Robotics & Automation*, 2007.

[3] L. Navarro-Serment, C. Mertz, and M. Hebert, "Pedestrian detection and tracking using three-dimensional LADAR data," in *Int. Conf on Field and Service Robotics (FSR)*, 2009.

[4] M. Bajracharya, B. Moghaddam, A. Howard, S. Brennan, and L. Matthies, "Results from a real-time stereo-based pedestrian detection system on a moving vehicle," in *Wshop on People Det. and Tracking, IEEE ICRA*, 2009.

[5] L. Spinello, M. Luber, and K. O. Arras, "Tracking people in 3D using a bottom-up top-down people detector," in *Proc. of the Int. Conf. on Robotics & Automation (ICRA)*, 2011.

[6] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. of the IEEE Conf. on Comp. Vis. and Pat. Rec. (CVPR)*, 2005.

[7] B. Leibe, E. Seemann, and B. Schiele, "Pedestrian detection in crowded scenes," in *Proc. of the IEEE Conf. on Comp. Vis. and Pat. Rec. (CVPR)*, 2005.

[8] P. Felzenszwalb, D. McAllester, and D. Ramanan, "A discriminatively trained,multiscale,deformable part model," in *Proc. of the IEEE Conf. on Comp. Vis. and Pat. Rec. (CVPR)*, 2008.

[9] M. Enzweiler and D. Gavrila, "Monocular pedestrian detection: Survey and experiments," *IEEE Trans. on Pat. An. and Mach. Intel. (PAMI)*, vol. 31, no. 12, pp. 2179–2195, 2009.

[10] P. Dollar, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: A benchmark," in *Proc. of the IEEE Conf. on Comp. Vis. and Pat. Rec. (CVPR)*, Miami Beach, USA, 2009.

[11] L. Spinello, R. Triebel, and R. Siegwart, "Multiclass multimodal detection and tracking in urban environments," *Int. Journ. of Rob. Research*, vol. 29, no. 12, pp. 1498–1515.

[12] M. Enzweiler, A. Eigenstetter, B. Schiele, and D. Gavrila, "Multi-cue pedestrian classification with partial occlusion handling," in *Proc. of the IEEE Conf. on Comp. Vis. and Pat. Rec. (CVPR)*, 2010.

[13] S. Ikemura and H. Fujiyoshi, "Real-time human detection using relational depth similarity features."

[14] F. C. Crow, "Summed-area tables for texture mapping," *SIGGRAPH Comput. Graph.*, vol. 18, pp. 207–212, January 1984.

[15] P. Viola and M. Jones, "Robust real-time object detection," in *Int. Journ. of Comp. Vis.*, vol. 57, no. 2, 2004, pp. 137–154.

[16] R. Szeliski, *Computer Vision: Algorithms and Applications*. Springer, 2011.

[17] M. Quigley, K. Conley, B. P. Gerkey, J. Faust, T. Foote, J. Leibs, R. Wheeler, and A. Y. Ng, "Ros: an open-source robot operating system," in *ICRA Wshop Open Source Soft.*, 2009.

[18] *Xbox 360 Kinect Sensor Manual*, Microsoft, Oct 2010.

[19] J. C. Platt, "Probabilities for SV Machines," *Advances in Large-Margin Classifiers*, pp. 61–74, 2000.