Close-Range Human Detection for Head-Mounted Cameras

Dennis Mitzel
mitzel@vision.rwth-aachen.de
Bastian Leibe
leibe@vision.rwth-aachen.de

Computer Vision Group
RWTH Aachen University
Aachen, Germany

Abstract

In this paper we consider the problem of multi-person detection from the perspective of a head mounted stereo camera. As pedestrians close to the camera cannot be detected by classical full-body detectors due to strong occlusion, we propose a stereo depth-template based detection approach for close-range pedestrians. We perform a sliding window procedure, where we measure the similarity between a learned depth template and the depth image. To reduce the search space of the detector we slide the detector only over few selected regions of interest that are generated based on depth information. The region-of-interest selection allows us to further constrain the number of scales to be evaluated, significantly reducing the computational cost. We present experiments on stereo sequences recorded from a head-mounted camera setup in crowded shopping street scenarios and show that our proposed approach achieves superior performance on this very challenging data.

1 Introduction

Robust multi-person detection and tracking is an important prerequisite for many applications. Examples include the use of mobile service robots in busy urban settings or mobile AR applications such as Google's project Glass. In this paper, we address the problem of stereo based person detection from the perspective of a moving human observer wearing a head-mounted stereo camera system. From this viewpoint many pedestrians in a crowded scenario are only partially visible due to occlusions at the image boundaries. In such situations, standard full-body object detectors such as [4, 7] are not well-suited, since they cannot deal with the large degree of occlusion. On the other hand, we can take advantage of the elevated viewpoint of a head-mounted camera, which typically leaves the head-shoulder region of close-by pedestrians well visible.

Taking inspiration from a recently proposed human upper body detector for Kinect RGB-Depth data by [3], this paper proposes an improved stereo depth-template based approach which can quickly and reliably detect close-by pedestrians. Similar to [2] we generate regions of interest (ROIs) based on the stereo data in order to reduce the search space of the detector. Our approach learns a continuous normalized depth template from annotations of human upper bodies and slides this template over the extracted depth ROIs at several scales

in order to compute a normalized similarity distance score. The output of this process are distance matrices whose entries represent the distance between the template and the overlayed segment of the ROI for each scale. After non-minimum suppression (NMS) in the distance matrices we obtain several detections (from different scales) for a person that are pruned to a set of final detections by a second, template-based NMS stage. We systematically evaluate this approach and characterize the effects of its parameters. In addition, we show how it can be integrated into a mobile multi-person tracking framework.

Besides this technical design and evaluation of our proposed detector, a second main contribution of this paper is its empirical demonstration of the somewhat surprising fact that such a relatively simple and fast approach can reach superior detection performance on very challenging outdoor data. This is even more surprising since our approach is based entirely on stereo range data, which is considerably more noisy than the Kinect RGB-D data often used for indoor scenarios.

The paper is structured as follows. The next section discusses related work. After that, Sec. 3 presents an overview of our detection framework. Sec. 4 introduces the integration of the detector with a tracking system. Finally, Sec. 5 presents a detailed experimental evaluation of the full design space of the detector.

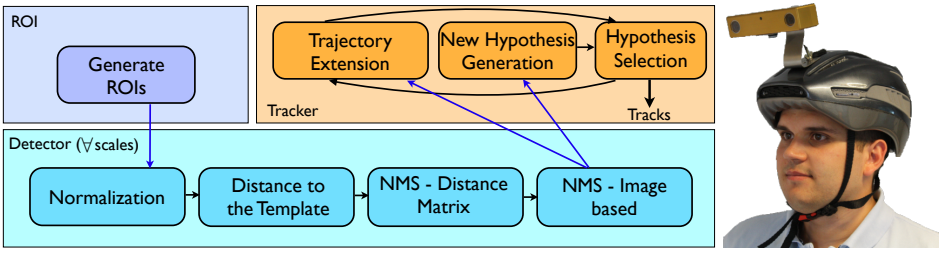
2 Related Work

The ability to reliably detect pedestrians in real-world images is required for a variety of automotive and robotics applications. Nowadays, state-of-the-art object detectors such as [4, 7] yield highly accurate human detection results for fully observed pedestrians. However, highly occluded scenarios still present major problems. Furthermore these approaches are quite expensive to evaluate which limits their deployment for use on autonomous platforms.

For detecting pedestrian that are undergoing partially occlusions, Wojek *et al*. [18] propose a framework where a full-object detector and several object-part detectors are combined in a mixture-of-experts based on their expected visibility. Enzweiler *et al*. [14] learn local body part detectors combined in a mixture-of-experts framework supported by stereo and flow cues. Wang *et al*. [17] perform occlusion handling in a modified SVM framework by combining HOG and LBP features. However, all of these methods suffer from high computational cost of either features or classifiers.

To deal with this problem, several approaches have been proposed to restrict the evaluation of the detector to only few ROIs that are extracted based on, *e.g*., stereo range data [1, 2, 8], motion [5], or scene geometry [11]. In this paper we follow a similar strategy of ROI extraction based on stereo information in order to reduce the search space for the detector. As a result, we reduce the computational effort, but also the number of possible false positives since only image regions are evaluated that are likely to contain target objects.

Several existing approaches incorporate stereo information in order to improve detection accuracy. Gavrila *et al*. [8] employ depth cues for first extracting the ROIs. After generating detection hypotheses by measuring the Chamfer distance between a learned shape contour model and the image input, the hypotheses are verified by cross-correlation between the two stereo images. Walk *et al*. [16] propose a disparity statistics feature which is combined with motion features and HOG to give a large gain in performance. Spinello *et al*. [15] propose a full-body pedestrian detector using dense depth data from an RGB-D Kinect camera. Based on the idea of HOG, the approach introduces Histograms of Oriented Depths as a new feature that follows the same computational procedure as HOG. The approach by Ikemura *et al*. [12]



(a) (b)

Figure 1: (a) Overview of the different modules of the proposed approach. Blue arrows indicate the interaction between the individual modules. Black arrows represent the interaction within the modules. (b) Camera setup used for recording the video sequences.

proposes a histogram based procedure for detection of humans in depth data, by learning a histogram based template that is employed in the detection process. The approach by Choi *et al*. [3] proposes an ensemble of detectors on multiple perceptual cues, where the output of all detectors is fused using an RJ-MCMC tracker. One of their detectors is a manually created binary template that is compared to the depth image of a Kinect RGB-D sensor to measure the distance to observed shape of a human. We take inspiration from this approach, but refine and improve it in several respects. In particular, we propose a continuous normalized-depth template, which is learned from annotated upper bodies of pedestrians in noisy stereo data. As we show experimentally, our detector performs significantly better than the binary one, as it represents the distribution of the depth values of the objects more precisely.

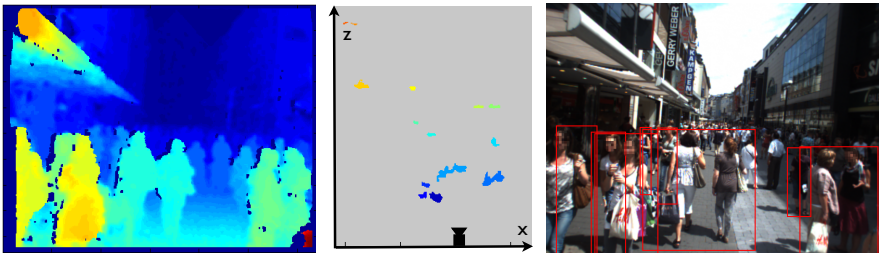
3 Approach

The key objective of our approach is to be able to detect pedestrians in a close range to the camera, where a standard full-body detector will usually fail. In addition, by focusing only on promising ROIs which are likely to contain a target object, we want to significantly reduce the computational cost and the number of false positives of our detector.

In Fig. 1 we illustrate a compact overview of our proposed detection and tracking framework. For each new frame, given the stereo pair and the corresponding depth map, we project the 3D points onto a (automatically estimated) ground plane and extract the ROIs using connected components on the ground projection image. For each extracted 3D ROI, we generate the corresponding ROI in the image plane by backprojecting the ROIs from the ground plane to the image. The 2D ROIs are passed to the detector, which slides the learned upper body template over the ROIs and computes the distance matrix by taking the Euclidean distance between the template and the overlayed, normalized depth image segment. Using a minimum filter on the distance matrix, we obtain possible bounding box hypotheses for the upper bodies. These hypotheses are pruned to a final detection set by using a template based intersection-over-union (IoU) NMS stage, where the detection with the lowest distance is chosen first and all other detections within a certain overlap area are removed iteratively.

3.1 Learning

We first learn a depth template from 600 annotations of upper bodies. To this end, the annotated depth regions are first normalized by the median depth and scaled to a fixed size



(a) (b) (c)

Figure 2: Extraction of the ROIs. (a) Input stereo depth. (b) ROIs - the connected components on the ground plane projection map. (c) The ROIs backprojected to the image plane.

of 150×150 pixels. Before summing up the annotated bounding box contents to a final template, a binary mask of the same size is generated for each annotation, where the invalid pixels (pixels that failed the left-right check of the stereo algorithm) are set to zero and valid pixels to one. Finally, all the annotations and masks are summed up to obtain the template illustrated in Fig. 4 (left) by dividing the summed annotations by the summed masks in a point-wise manner.

3.2 ROI Extraction

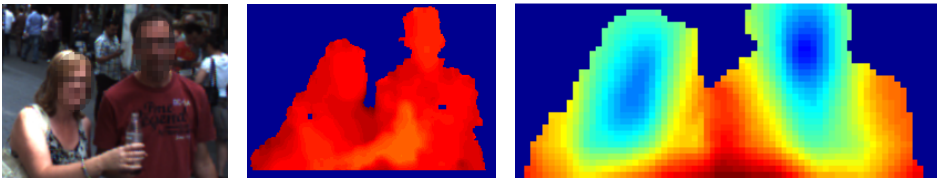
In contrast to classical sliding window object detectors, such as HOG [4], we reduce the search space for the detector to few small ROIs and few scales in the image. In order to achieve this, we exploit information about the location of the objects on the ground plane based on the stereo data, similar to [2].

The ROI extraction process is visualized in Fig. 2. We first project the 3D depth points onto a ground plane divided into bins of $5 \times 5cm$. The number of points that fall into each bin is weighted by the square distance to the camera in order to compensate for the fact that far-away objects have smaller support in the image than closer ones. The final set of ROIs is obtained by connected components, as shown in Fig. 2(b). To generate the corresponding ROIs, which will be scanned by the detector in image space, we backproject the ROIs from 3D to the image plane, as shown in Fig. 2(c). The width of the corresponding bounding box is derived from the width of the ROI, and the height is obtained from the highest point that is enclosed by the ROI.

3.3 Depth Template Detector

The pipeline for the detector consists of the following steps. First, for each ROI in the image plane, we discard the pixels which are not in the depth range of the ROIs in 3D by setting them to zero, as illustrated in Fig. 3(b). Then, starting from an initial template size that is one third of the ROI height, we slide the template over the ROI. At each position, the segment of the ROI that is overlayed with the template is normalized with its median depth and then the distance between the template and the segment is computed. As a final result we obtain a distance matrix that contains for each position of the template the corresponding distance to the segment in the depth image, see Fig. 3(c).

Distance Measure. For the distance measure between the template and the ROI we explore



(a) (b) (c)

Figure 3: (a) Original ROI from the left image. (b) Input depth from the ROI. (c) The resulting distance matrix for the initial scale.

two options: *Hamming Distance* on binarized template and depth as proposed by [8],

$$\sum_{(i,j)\in W} |T_{bin}(i,j) - I_{bin}(x+i,y+j)|_H \tag{1}$$

where T denotes the template and I the ROI, and the *Euclidean Distance* on continuous depth-normalized templates.

$$\sum_{(i,j)\in W} min(d_{max}, (T(i,j) - I(x+i,y+j))^2). \tag{2}$$

In the case that two pedestrians are walking closely behind each other, the resulting ROI will cover both of them. The person in the background will, however, be occluded by the person in front. By overlaying the template over the person behind, some of its pixels will fall in the occluded region. The resulting distance between the template and the overlayed region will increase, leading to a rejection of the detection. To avoid such false negatives, we use a truncated distance that clips the contribution of each pixel to a fixed maximum value d_{max}.

Multi-Scale Handling. Most of the pedestrians in our scenarios are walking close to each other in groups of two or more people. This often leads to ROIs that represent the persons not individually but that cover entire groups of pedestrians. As the pedestrian heights may vary, we need to rescale the template, because the initial scale estimation based on the height of the 2D ROI might not be representative for all pedestrians in the group. Note that here we only need to down-sample the template starting from the initial scale, since the height of the ROI in the image plane is based on the highest 3D point (from the tallest person) that falls inside the 3D ROI, as already mentioned before. Each downscaled version of the template is slid over the ROI again, generating an additional distance matrix.

Non-Minimum Suppression. For each scale of the template, we obtain a distance matrix on which we perform NMS by applying a minimum filter of size 3×3. As a result we obtain few positions in the distance matrix which are the minima in their local neighborhood. The multi-scale approach introduces several additional detections on a person for a number of neighboring scales, as the scale stride is usually small. To reduce this set to only one representative detection for each pedestrian, we additionally perform an image based NMS, where we iteratively select the detected bounding box with the lowest distance and remove all bounding boxes that have a certain overlap with this box. A standard approach for this is to compare the bounding box IoU. However, such a strategy will cause some correct detections to be rejected, as shown in Fig. 4 (b), (*e.g.*, for persons walking behind each other, where the overlap between the boxes is high). To cope with this problem, we propose a more precise intersection criterion which is based on the intersection between the templates (intersected pixels that are non-zero) of the individual detections. In Fig. 4(b,c) we show the comparison between the bounding box based IoU and the template based one. In our

(a) (b) (c)

Figure 4: (a) Depth template learned based on 600 upper body annotations. (b) Illustration of the bounding box based intersection over union. (c) Template based intersection over union.

experiments we used a threshold of 0.2 in IoU for pruning the detections. In the classical bounding box based approach, this leads to rejection of one of the detections, whereas in the template based approach both detections will be kept. Further quantitative experiments (see Sec. 5) underline that this criterion is better suited for our scenario.

4 Integration with Tracking System

For the tracking module we employed a simplified version of the robust multi-hypothesis tracking proposed by [13]. Briefly stated, our tracking framework applies the following procedure. In each frame the detections are transformed to global 3D world coordinates by projecting the middle point of the bounding box onto an automatically estimated ground plane using visual odometry estimate of the camera egomotion. These 3D positions are connected to a set of trajectory hypotheses over a temporal window. The final best representative hypotheses in terms of scene representation are obtained using a model selection framework.

Generation and Pruning of Track Hypotheses. The projected 3D points on the ground plane are associated to long trajectories by employing an Extended Kalman Filter with a constant-velocity motion model. For all detections in the current frame we apply the Kalman filter backwards in time and use the same detections in order to extend existing trajectories. With this strategy, we obtain trajectories that might share an observation, which needs to be resolved, as two persons cannot occupy the same space in 3D. To this end, we prune the over-complete set of trajectories which compete for observations using the model selection framework proposed by [13]. Due to the employed trajectory generation process we usually obtain two hypotheses for one and the same person: the new one that was generated by applying the Kalman Filter backwards in time and the extended one. Since the model selection framework may choose either the extended or the newly created one, we need a process that assigns always a consistent ID to the same person. For this, the ID to the final trajectory in the current frame is assigned based on the overlap to the trajectories from the previous frame. If the overlap of the new trajectory with one of the previous trajectories is above 95% then we propagate the ID from the previous trajectory to the new one. As demonstrated in our experiments, this tracking framework succeeds in grouping the detections into stable tracks.

5 Experimental Evaluation

To assess the performance of our upper body detector we report a detailed experimental evaluation of the full design space of the detector that gives clear guidelines how to apply it for optimal performance and fast evaluation.

Data Set. The evaluation was performed on several very challenging sequences captured from a head mounted camera setup (Bumblebee2), as shown in Fig. 1(b). The images were captured at 15 fps in crowded shopping streets. The evaluation set consist of 2,543 frames, in which we have annotated 19,461 pedestrians. For all frames we have computed visual odometry using the approach proposed by [11]. For stereo depth estimation we used the the fast and robust algorithm presented in [9]. The detection code and all sequences (annotations, stereo depth, visual odometry, and estimated ground planes) will be publicly available at www.vision.rwth-aachen.de/projects/upperbody

For the evaluation in all following experiments we applied the evaluation criteria from [5]. In each frame the detections are compared to manually annotated ground-truth boxes. In case that the IoU overlap of a detection with the ground truth annotation is above 0.5 the detection is assumed to be correct. To investigate the performance of our detector for different distance ranges, we adapted the annotation files by setting the annotations that are beyond the range to *don't cares* (which result in neither false positives nor missing detections).

Overall Performance. In Fig. 5 we present a detailed evaluation in terms of recall *vs*. false positives per image (fppi) over several parameters of our detector pipeline. First, we compare our continuous template to a binary version similar to the one used in [3] for three different distance ranges of 5, 7, and 10 meters. To generate a binary depth-shape template (which was generated manually in [3]), we used our continuous template by simply converting all pixels above zero to one. As shown in Fig. 5(a), our detector performs significantly better (3-5% higher recall) for the close range of up to 7 meter, which shows that our continuous template represents the depth distribution on the object better. Note that the binary template approach profits from the innovations of our ROI extraction, template learning, and template based NMS and would probably perform worse in the original form as presented in [3].

Comparison with DPM. We also run the DPM detector from [7] on our sequences. The results plotted in Fig.5(b) underline the complexity of the dataset, as many persons are occluded by the image boundaries and cannot be detected with a full-body detector. As expected, the DPM detector performs better for the range up to 7 meters than for the range up to 5 meters, since starting from 5 meters the pedestrians become fully visible. Still, the results clearly show our approach's superior performance.

Number of Scales. In the next experiment, we explore how many scale evaluations are necessary in order to reach best possible performance. The results of this test parameter are plotted in Fig. 5(c), showing that we reach the best performance with 1-3 downscalings of the template (the scale stride in this experiment was 1.03). Essentially, the result indicates that the initial scale estimation based on the height of the 2D ROI box is accurate enough and further downscaled versions of the template introduce more false positives.

Scale Stride. The above observation is corroborated by our next experiment, in which we vary the scale stride using 5 scales. As shown in Fig.5(d), increasing the scale stride from a base setting of 1.01 introduces more false positives without significantly increasing recall. In all further experiments, we use a scale stride of 1.03, which is a good compromise between detection accuracy and processing speed.

NMS. Next, we explore how the different NMS approaches described in Sec. 3.3 affect the overall performance of the detector. As expected, the template based NMS performs better (Fig.5(e)), since it takes into account the expected pedestrian shape. We obtain a 3% higher recall at 0.4 fppi. Note for all the experiments we used an overlap threshold of 0.2.

Fixed Size Template. As can be seen in the example results shown in Fig. 6, many pedestri-

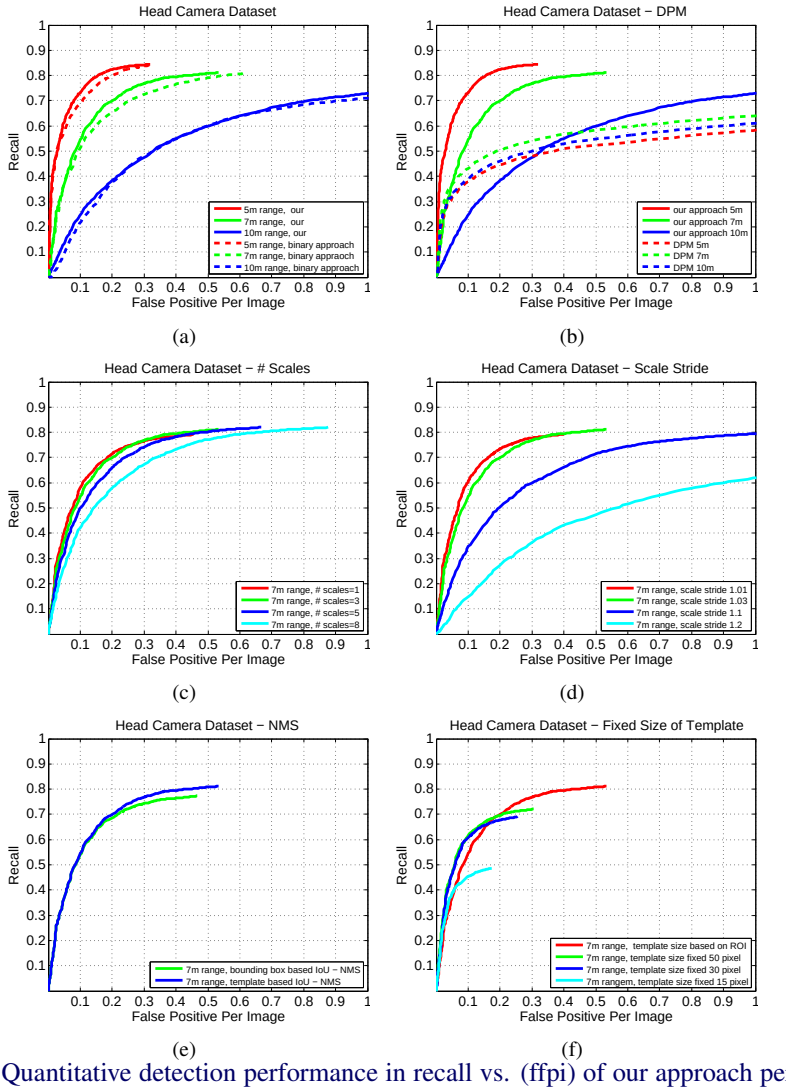


Figure 5: Quantitative detection performance in recall vs. (ffpi) of our approach performed on full design space of the detector. If not explicitly mentioned otherwise, the number of evaluated scales was set to three and the scale stride was set to 1.03. (a) Performance of our detector for different distance ranges in comparison to a baseline approach [3]. (b) Comparison of our detector to the DPM detector [7] for different distance ranges. (c) Analysis of the number of scales. (d) Analysis of employing different values for the scale stride. In this experiment we used 5 scales. (e) Evaluation of different IoU criteria, namely template based and bounding box based, as discussed in Sec. 3.3. (f) Effect of fixing the size of the template to a fixed value. Number of evaluated scales was set to 1.

ans appear close to the camera, which results in large 2D ROI boxes. In a further experiment, we therefore investigate how the resolution of the template affects the performance of the detector and whether we can reduce the initial resolution of the template in order to decrease computational cost. The results in Fig. 5(f) show that by fixing the size of the template to a fixed small value of 50×50 pixels and downsampling the ROI to the corresponding size, we

Figure 6: Experimental detection and tracking results.

still obtain a reasonable performance of about 70% recall at 0.3 fppi.

Computational Performance. Our current system, including ROI candidate generation, object detection, and tracking runs, without any optimization work, with more than 4 fps on a single CPU of an Intel Core2 Quad Q9550 @ 2.83GHz, 8GB RAM. (For this experiment we used 3 scales, scale stride 1.03 and distance range 7 meters). There is still considerable optimization potential, since the ROI comparisons can easily be parallelized.

Qualitative Evaluation. Finally, Fig. 6 shows some qualitative results achieved on our sequences. It can be seen that our system is able to detect nearly all pedestrians correctly with only few false positives. In addition, our approach correctly tracks detected persons over time keeping correct person identities.

6 Conclusion

We have presented a stereo depth-template based detection approach for detecting pedestrians in the close range. The combination of the ROI extraction with the detector allows us to restrict the detector evaluation only to few scales and to few small segments in the image, which reduces the number of false positives significantly. We have shown that even though this approach is simple and fast to apply, we reach superior performance on noisy stereo data

for very challenging scenarios from crowded shopping streets. Additionally, we have shown how this detector can be integrated into an ROI based multi-object tracking framework. For the future, we plan to investigate how we can improve the performance of our upper-body detector in the far-range by combining it with an upper body and a full-body detector.

Acknowledgments. We would like to thank Javier Marín for many fruitful discussions and David Vázquez for providing the evaluation tools. This project has been funded by the cluster of excellence UMIC (DFG EXC 89).

References

[1] M. Bajracharya, B. Moghaddam, A. Howard, S. Brennan, and L.H. Matthies. A Fast Stereo-based System for Detecting and Tracking Pedestrians from a Moving Vehicle. *IJRS*, 28(11), 2009.

[2] M. Bansal, S. H. Jung, B. Matei, J. Eledath, and H. S. Sawhney. A real-time pedestrian detection system based on structure and appearance classification. In *ICRA*, 2010.

[3] W. Choi, C. Pantofaru, and S. Savarese. Detecting and Tracking People using an RGB-D Camera via Multiple Detector Fusion. In *CORP*, 2011.

[4] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.

[5] M. Enzweiler, P. Kanter, and D.M. Gavrila. Monocular Pedestrian Recognition Using Motion Parallax. In *Intel. Vehicles Symp.*, 2008.

[6] M. Everingham, L. Van Gool, C.K.I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes (VOC) challenge. *IJCV*, 88(14), 2010.

[7] P. Felzenszwalb, B. Girshick, D. McAllester, and D. Ramanan. Object Detection with Discriminatively Trained Part-Based Models. *PAMI*, 32(9), 2010.

[8] D. Gavrila and S. Munder. Multi-Cue Pedestrian Detection and Tracking from a Moving Vehicle. *IJCV*, 73(1), 2007.

[9] A. Geiger, M. Roser, and R. Urtasun. Efficient Large-Scale Stereo Matching. In *ACCV*, 2010.

[10] A. Geiger, J. Ziegler, and C. Stiller. StereoScan: Dense 3d Reconstruction in Real-time. In *Intel. Vehicles Symp.*, 2011.

[11] D. Geronimo, A.D. Sappa, D. Ponsa, and A.M. Lopez. 2D-3D-based On-Board Pedestrian Detection System. *CVIU*, 114(5), 2010.

[12] S. Ikemura and H. Fujiyoshi. Real-Time Human Detection using Relational Depth Similarity Features. In *ACCV*, 2010.

[13] B. Leibe, K. Schindler, and L. Van Gool. Coupled Object Detection and Tracking from Static Cameras and Moving Vehicles. *PAMI*, 30(10), 2008.

[14] B. Schiele M. Enzweiler, A. Eigenstetter and D. M. Gavrila. Multi-Cue Pedestrian Classification with Partial Occlusion Handling. In *CVPR*, 2010.

[15] L. Spinello and K. O. Arras. People Detection in RGB-D Data. In *IROS*, 2011.

[16] S. Walk, K. Schindler, and B. Schiele. Disparity Statistics for Pedestrian Detection: Combining Appearance, Motion and Stereo. In *ECCV*, 2010.

[17] X. Wang, T. X. Han, and S. Yan. An HOG-LBP human detector with partial occlusion handling. In *ICCV*, 2009.

[18] C. Wojek, S. Walk, S. Roth, and B. Schiele. Monocular 3D Scene Understanding with Explicit Occlusion Reasoning. In *CVPR*, 2011.