Detection and Tracking of Occluded People

Siyu Tang
tang@mpi-inf.mpg.de
Mykhaylo Andriluka
andriluka@mpi-inf.mpg.de
Bernt Schiele
schiele@mpi-inf.mpg.de

Computer Vision and
Multimodal Computing
Max Planck Institute for Informatics
Saarbrücken, Germany

Abstract

We consider the problem of detection and tracking of multiple people in crowded street scenes. State-of-the-art methods perform well in scenes with relatively few people, but are severely challenged by scenes with many subjects that partially occlude each other. This limitation is due to the fact that current people detectors fail when persons are strongly occluded. We observe that typical occlusions are due to overlaps between people and propose a people detector tailored to various occlusion levels. Instead of treating partial occlusions as distractions, we leverage the fact that person/person occlusions result in very characteristic appearance patterns that can help to improve detection results. We demonstrate the performance of our occlusion-aware person detector on a new dataset of people with controlled but severe levels of occlusion and on two challenging publicly available benchmarks outperforming single person detectors in each case.

1 Introduction

Single people detectors such as the powerful deformable part models (DPM, [10]) have shown promising results on challenging datasets. However, it is well known that current detectors fail to robustly detect people in the presence of significant partial occlusions. In fact, as we analyze in this paper, the DPM detector starts to break already at about 20% of occlusion and beyond 40% of occlusion the detection of occluded people becomes mere chance. Several methods, i.e. tracking and 3D scene reasoning approaches, have been proposed to track people even in the presence of long-term occlusions. While these approaches allow to reason across potentially long-term and full occlusions they still require that each person is sufficiently visible at least for a certain number of frames. In many real scenes however, e.g. when people walk side-by-side across a pedestrian crossing (see Fig. 1), a significant number of people will be occluded by 50% and more for the *entire* sequence.

To address this problem this paper makes three main contributions. First we propose a new double-person detector that allows to predict bounding boxes of two people even when they occlude each other by 50% or more, and propose a new training method for this detector. This approach outperforms single-person detectors by a large margin in the presence of significant partial occlusions (Sec. 3). Second, we propose a joint person detector, that is jointly trained to detect single- as well as two-people in the presence of occlusions. This joint detector achieves state-of-the-art performance on challenging and realistic datasets (Sec. 4).

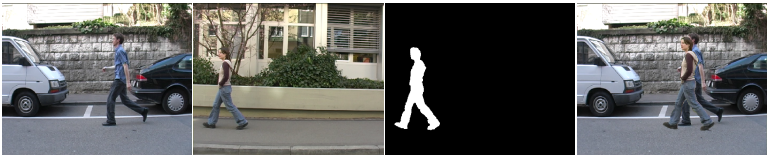
Figure 1: Detection results at equal error rate obtained with the approach of [4] (top) and our joint detector (bottom) on the TUD-Crossing [1] dataset. False-positive detections are shown in red and missing detections in green. One of the two bounding boxes predicted from the two-person detection is shown with the dotted line.

Last, we integrate the above joint model into a tracking approach to show its potential for people detection and tracking (Sec. 5).

2 Related Work

Recent methods to track people [3, 6, 11, 18] employ people detectors to generate initial tracking hypotheses, and often include elaborate strategies to link people tracks across occlusion events. However, they typically fail to track people that remain significantly occluded for the entire sequence. To overcome this limitation we propose a people detection approach that can detect and predict the position of even severely occluded people. State-of-the-art approaches to people detection [7, 10] are able to reliably detect people under a variety of imaging conditions, people poses, and appearance. While being effective when people are fully visible, their performance degrades when people become partially occluded. Various remedies have been proposed, including a combination of multiple detection components [10], large number of part detectors (Poselets) [5], and careful reasoning about association of image evidence to detection hypotheses [9, 12, 16]. [12] proposed an approach that first aggregates evidence from local image features into a probabilistic figure-ground segmentation and then relies on an MDL formulation to assign foreground regions to detection hypotheses. [9] proposed a probabilistic formulation of the generalized Hough transform that prevents association of the same image evidence to multiple person hypotheses. These approaches treat partial occlusion as nuisance and perform decisions based on the image evidence that corresponds to the visible part of the person. This makes them unreliable in cases of severe occlusions (i.e. more than 50% of the person occluded). Several works have aimed at improving such weak detections using information from additional sensing modalities [8] or by joint reasoning about people hypotheses and 3D scene layout [17]. In [17], a bank of partial people detectors is used to generate initial proposals that are refined based on the 3D scene layout and temporal reasoning.

Here, we explore an alternative strategy, observing that in crowded street scenes most occlusions happen due to overlaps between people. Instead of using evidence from individual people that becomes unreliable in cases of severe occlusion, we consider the joint evidence of both people. This is possible since overlapping people result in characteristic appearance patterns that are otherwise uncommon. Our approach is related to the "visual phrases" approach [13] in that we train a joint detector for the combination of two object instances. Our



(a) (b) (c) (d)

Figure 2: Procedure to synthetically generate training images for our double-person detector. (a) background person, (b) foreground person, (c) foreground person map, (d) generated synthetic training image.

approach builds on the state-of-the-art people detector of [10], which we extend in two ways. First, we propose a double-person detector that simultaneously detects two people occluding each other and second, we propose a joint detector that can detect both one as well as two people due to joint training. To capture typical appearance patterns of people occluding each other we automatically generate a dataset of training images with controlled and varying degrees of occlusion. In this respect our work is also related to a recent literature that combines real and artificially generated images to train people detectors [13, 14].

3 Double-Person Detector

Our double-person detector builds on the DPM approach [10] arguably one of the most powerful object detectors today. The key concept of our double-person model is that person/person occlusion patterns are explicitly used and trained to detect the presence of two people rather than to treat these occlusions as distractions or nuisance as it is typically done. Specifically, our double-person detector shares the deformable parts across two people which belong to the same (two-person) root filter. In that way localizing one person facilitates the localization of the counterpart in the presence of severe occlusions and the deformable parts allow to improve the localization accuracy of both people using the above mentioned occlusion patterns whenever appropriate. For this we build on the DPM framework to detect the presence of two people and to predict the bounding boxes of both people, the occluding person as well as the occluded person. The latent SVM algorithm used to train DPMs is susceptible to local minima. Therefore, proper initialization is crucial, as discussed below. For training, we synthetically generate two-people samples based on the TUD training data [1]. The synthetic images are ideal for training as they come with accurate occlusion-level estimates. We demonstrate experimentally that our double-person detector significantly outperforms a single-person detector in the presence of severe occlusions.

Double-person detector model: In full analogy to DPMs, our double-person detector uses a mixture of components. Each component is a star model consisting of a root filter defining the coarse location of two people and n deformable part filters covering representative parts and occlusion patterns of the two people. The vector of latent variables is given by $z = (c, p_0, \ldots, p_n)$ with c denoting the mixture component and p_i specifies the part's image position and feature pyramid level l_i. The score of a double-person hypothesis is obtained by the score of each filter at the latent position p_i (unary potentials) minus the deformation cost between root position and part position (pairwise potentials). As in [10], the un-normalized score of a double-person hypothesis is defined by $\langle \beta, \Psi(x,z) \rangle$, where vector β is a concatenation of the root and all part filters and the deformation parameters, and $\Psi(x,z)$ is the stacked HOG features and part displacement features of sample x. $\Psi(x,z)$ is zero except for a certain

(a) (b) (c) (d) (e) (f)

Figure 3: Examples of synthetically generated training images. From (a) to (f), levels of occlusion are gradually increased.

component c. Therefore, we obtain the construction $\langle \beta, \Psi(x,z) \rangle = \langle \beta_c, \psi_c(x,z) \rangle$. Detection in the test image is done by maximizing over the latent variables z: $\arg\max_{(z)} \langle \beta, \Psi(x,z) \rangle$.

Model training: Given a set of training examples $D = (\langle x_1, y_1 \rangle, \ldots, \langle x_N, y_N \rangle)$, with $y_i \in \{-1, 1\}$, we learn the model parameters β using latent SVM [10]. This involves iteratively solving the quadratic program:

$$\min_{\beta, \xi \geq 0} \frac{1}{2} \|\beta\|^2 + C \Sigma_{i=1}^{N} \xi_i \quad \text{sb.t.} \quad y_i \langle \beta, \Psi(x_i, z) \rangle \geq 1 - \xi_i \quad \xi_i \geq 0, \tag{1}$$

and optimizing for the values of latent parameters z. We solve the quadratic program with the stochastic gradient descent and employ data-mining of hard-negative examples after each optimization round as proposed in [10].

Initialization: The objective function of the latent SVM is non-convex, which makes the training algorithm susceptible to local minima. Therefore, a good initialization of the model components is crucial for good performance. Instead of relying on the bounding box aspect ratio as in [10], we initialize our model using different occlusion levels. This follows the intuition that degree of occlusion is one of the major sources of the appearance variability, and we capture it by different components. Other sources of appearance variability such as poses of people and varying clothing are then captured by displacement and appearance parameters of each component. In the experiments reported below we rely on the three component double-person model. The components are initialized with the occlusion levels 0%–25%, 25%–55%, and 55%–85%.

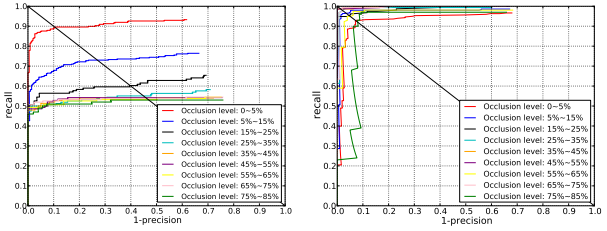
Bounding box predictions: Given a double-person detection we predict the bounding boxes of individual people using a linear regression. The location of each bounding box is modelled as

$$B_i = g_i(z)^{\mathrm{T}} \alpha_c + \varepsilon_i, \tag{2}$$

where B_i is the predicted bounding box for a detection i, c is the index of the DPM component that generated the detection, and $g_i(z)$ is a $2 * n + 3$ dimensional vector that is constructed by the upper left corners of the root filter and the n part filters as well as the width of the root filter. ε_i is a Gaussian noise that models deviations between the predicted and observed location of the bounding box.

The regression coefficients α_c are estimated from all positive examples of the component c. For each of the model components we estimate two separate regression models that correspond to each of the persons in the double-person detection. This procedure allows to accurately localize both people despite severe occlusions as can be seen e.g. in Fig. 5.

Training data generation: As it is difficult to obtain sufficient training data for the different occlusion levels of our double-person detector we synthetically generate it. Fig. 2 illustrates this process. For each person we first extract the silhouette based on the annotated foreground person map. Next, another single-person image is selected arbitrarily and combined



(a) Single-person detector (b) Double-person detector

Figure 4: Detection performance of single- and double-person detectors for different occlusion levels.

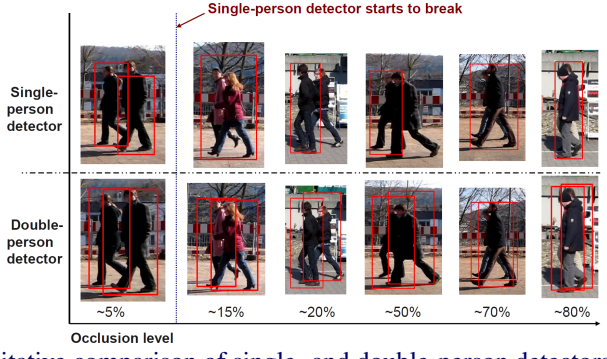


Figure 5: Qualitative comparison of single- and double-person detectors with occlusion.

with the extracted silhouettes. In order to generate a double-person training dataset we randomly select background images, 2D positions and scale parameters. Each synthetic image provides an accurate occlusion ratio estimated from the two persons' silhouettes. For the experiments reported below we generate 1,300 double-person training images from the 400 TUD training images [1]. For the synthetic dataset we uniformly sample occlusion levels between 0% and 85%, and scale factors between 0.9 and 1.1.

Experimental study: In order to explicitly compare single-person and double-person detector performance for person/person occlusion scenarios, we captured several video sequences and constructed a new double-person dataset (850 images) where the images are categorized by different occlusion levels. The occlusion level is estimated from 2D truncated quadrics which are constructed from stick-man annotation[1] .

Single-person detector: Fig. 4(a) shows the performance of the standard DPM single-person detector on our double-person dataset. In case of little partial occlusion (red curve, below 5%), the single-person detector obtains good performance both in terms of recall (up to 90% recall) as well as a high precision. However, the single person detector already misses many people when the occlusion level is increased up to 15% (blue curve, maximal recall below 80%), and further decreases in the presence of more occlusion. When the occlusion level is 35% or more, the achieved recall is only slightly above 50% clearly indicating that in most cases only one of the two people is correctly detected.

Double-person detector: Fig. 4(b) shows the performance of our proposed double-person detector. For almost all occlusion levels the detector allows to reach 100% recall which is clearly a significant improvement over the single person detector. Interestingly, for the lowest occlusion level (red curve, up to 5%) we loose some recall which can be explained by

[1]The training and test datasets are available at www.d2.mpi-inf.mpg.de/datasets

the difficulty to differentiate a single person that does not occlude a second person from the case that a person occludes a second person significantly (e.g. 80%) (see for an example of 80% occlusion Fig. 5). At the same time the precision is very high for all but the highest occlusion level (green line, up to 85%). From this experiment we conclude that our double-person detector is much more robust than the single-person detector and obtains excellent performance both in terms of recall and precision even for the heavy occlusion cases. Single person localization (bounding boxes prediction) is not a trivial task especially for intermediate occlusion level cases ($30\% \sim 60\%$), since we observe fair evidence from both persons, which can be distracting for single bounding box localization. However, the results shows that our double-person detector accurately and robustly predicts the single bounding box for the above mentioned case as well. Fig. 5 shows comparative qualitative results. For the same test examples, our double-person detector correctly detects the position of two persons and predicts their respective bounding box with high accuracy.

4 Multi-Person Detection

The previous section has shown that our double-person detector can indeed outperform a single person detector when people occlude each other by 25% or more. The employed dataset however was somewhat idealistic as it contained exactly two people that occluded each other at various degrees. In realistic datasets we will have both single people that are fully visible and two and more people that occlude each other. This section therefore proposes a detector that combines both single and two-person detectors into a single model that is jointly trained. The model is again built upon the DPM-approach where the role of the different components is now to differentiate both between single and two people as well as between different occlusion levels among two people.

4.1 Joint Person Detector

We jointly train single- and double-person detectors by representing them as different components of the DPM. We allocate 3 components for the double detector and 3 components for the single-person detector, which after mirroring results in a 12 component DPM model. Similarly to Sec. 3 we initialize the double-person components with training examples corresponding to gradually increasing levels of occlusion. For the single-detector components we rely on the standard initialization based on the bounding box aspect ratio. During learning we allow training examples to be reassigned to other components of the DPM model, but prevent assignments of 2-person examples to 1-person components and vise versa. We found this to be important to improve detection of two people in cases of particularly strong occlusion, that are otherwise often incorrectly handled by the single-person components.

Training data: We train our joint detector on the combination of 1-person and 2-person training sets described in Sec. 3, but slightly modify the initial assignment of images to the DPM components. We assign training images with less than 5% occlusion to the single-person training dataset, since in that case the single-person detector already obtains high performance for both people. We initialize the 3 double-person DPM components with images corresponding to occlusion levels: 5%–25%, $25\% - 55\%$, and $55\% - 75\%$.

Non-maximum suppression (NMS): The non-maximum suppression in the joint detector is more complicated than in the standard DPM since we have bounding box predictions from two different types of detections (single and two-person detections) as well as strongly

overlapping bounding box predictions from our two-person components. We thus implement NMS in two steps. The first step is performed prior to bounding box prediction and already removes a large portion of multiple detections on the same person. In this first step two-people detections and single-person detections compete and suppress each other depending on the respective score. The remaining multiple detections are either due to multiple two-person detections in cases when more than two people appear close to each other (e.g. rightmost tree people in the fourth image in Fig. 1) or detections with significantly different bounding box aspect ratios. Given the reduced set of hypotheses after the first round of NMS, we perform bounding box prediction followed by the second round of NMS. This second step corresponds to the NMS typically performed for DPM [10]. The second round is done independently for single-person and two-person components of DPM, as we found that one-person detections may incorrectly suppress two-person detections otherwise. During NMS of detections from the two-person components we additionally prevent two bounding boxes predicted from the same double-person detection from suppressing each other. As an illustrative example, we could correctly detect all three people in the fourth image on Fig. 1 despite strong occlusion of the middle person. In that case the single-person detections where predicted from two double-person detections and multiple detections on the middle person were correctly removed by the second stage of the non-maximum suppression.

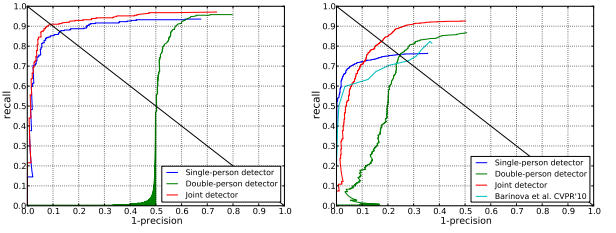
4.2 Results

We evaluate the performance of our joint detector on two publicly available datasets, "TUD-Pedestrians" and "TUD-Crossing", originally introduced in [1]. "TUD-Pedestrians" contains 250 images of typical street scenes with 311 people all of which are fully visible. "TUD-Crossing" contains a sequence of 201 images with 1008 annotated people that frequently occlude each other partially or even fully. To capture the full range of occlusions we extended the annotations of the "TUD Crossing" dataset to include also strongly occluded people, which resulted in 1186 annotated people.

We begin our analysis with the "TUD-Pedestrians" dataset. Detection results are shown in Fig. 6(a) as recall-precision curves. Since this dataset does not contain any occluded people our double-person detector (Sec. 3) generates numerous false positives interpreting each person as a pair of people one of which is severely occluded. As expected the single-person detector performs well on this dataset, achieving an equal error rate (EER) of 87% . The joint detector slightly improves over the single person detector achieving 90.5% EER. This result is a bit surprising because the joint detector is trained to solve a more difficult problem of detecting both fully visible and partially occluded people. We attribute the improvement of the joint detector to the training set that in addition to real images has been augmented with artificial training examples (c.f. Sec. 3). This parallels the recent results on using artificially genertated data for training of people detection and pose estimation models [14, 15].

The evaluation on "TUD Pedestrian" demonstrates that integrating single- and double-person detectors within the same model does not result in a performance penalty in the case when people are fully visible.

In order to assess the joint detector in realistic scenes that contain both occluded and fully visible people we evaluate its performance on the TUD-Crossing dataset. Quantitative results are shown on Fig. 6(b) and a few example images on Fig. 1 (bottom row). First we compare the performance of single and double-person detectors, which achieve approximately the same EER of 76%. The double-person detector achieves higher recall compared to the single-person detector, being able to detect even strongly occluded people. However



(a) TUD-Pedestrians (b) TUD-Crossing

Figure 6: Detection performance on TUD-Pedestrians (a) and TUD-Crossing (b).

the precision of the double-person detector suffers from multiple detections on fully visible people. The single-person detector produces fewer false positive detections, but also fails to detect occluded people, saturating at a recall of 76%. Finally, the joint detector significantly improves over both single and double person detectors achieving an EER of 83%. Note, that while demonstrating overall improvement, the joint detector has a somewhat lower performance in the high precision area compared to the single person detector. Inspecting the false positives of the joint detector with highest scores reveals that most of them correspond to cases when one-person and two-person components of the detector fired simultaneously on the same pair of people, but these detections where sufficiently far from each other to persist through the non-maximum suppression step (e.g. false positive detection in the first image on Fig. 1).

Finally, we compare the performance of our approach with the Hough transform based detector of [4], which is specifically designed to be robust to partial occlusions. The authors of [4] kindly provided us their detector output (in terms of bounding boxes) which allows to compare their result on our full ground-truth annotations making these results directly comparable to the rest of our experiments (Fig. 6(b)). The approach of [4] improves over the single-person detector in terms of final recall, but loses some precision, likely because their local features are rather weak compared to the discriminatively trained DPM model. Our joint model outperforms the approach of [4] by a large margin. Fig. 1 shows a few example frames from the "TUD-Crossing" sequence, comparing our joint detector with the results of [4]. Note that our approach is able to correctly detect occluded people in the presence of very little image evidence (e.g. three pairs of people in the second image), whereas the approach of [4] fails in such cases. At the same time our approach also correctly handles detection of single people (e.g. second and third images).

5 Multi-Person Tracking

This section compares the performances of a single-person and the joint detector (Sec. 4) in the context of multiple people tracking. To that end we employ the people tracking-by-detection formulation of [2]. Given the set of detections in frame j as $\mathbf{h}_j = [h_1, \ldots, h_{N_j}]$, we find people tracks by finding sequences H of hypotheses that maximize the objective:

$$p(H) = p_{det}(h^1_{j_1}) \prod_{k=2}^{T} p_{trans}(h^{k-1}_{j_{k-1}}, h^k_{j_k}) p_{det}(h^k_{j_k}), \tag{3}$$

where $p_{det}(h)$ is the probability of correct detection and $p_{trans}(h_1, h_2)$ is the probability that hypotheses h_1 and h_2 correspond to the same person in subsequent frames. Tracking proceeds by maximizing Eq. 3 subject to the constraint that none of the transition probabilities falls below a predefined threshold τ_{trans}. At each iteration the longest track that does not

Figure 7: Tracking-by-detection results on the TUD-Crossing dataset with single-person detector (top row) and our joint detector (bottom row). Colors and numbers indicate tracks corresponding to different people.

violate this constraint is returned and all detection hypotheses overlapping with the found track are removed from further consideration. The maximization of Eq. 3 is repeated until all hypotheses are removed. Similarly to [2] we model $p_{trans}(h_1, h_2)$ as Gaussian distribution with respect to differences in position and scale of the detections and set $p_{det}(h)$ to the exponent of the score of h returned by the person detector. We keep the tracking parameters as in [2] and set $\tau_{trans} = exp(-5)$, which achieves a reasonable trade-off between obtaining longer tracks while preventing tracks to drift from one person to another. The resulting set of tracks will typically contain a few long tracks corresponding to correct detections of people, but will also include a large number of short tracks which result from spurious detections in background and occasional detections at significantly wrong scale. In order filter of such spurious tracks we remove all tracks of length smaller than 10 from further consideration.

We apply the above tracking-by-detection approach without modifications both to the output of the single-person and the joint detectors. The set of hypotheses in each frame is given by the output of the detectors prior to non-maximum suppression. In the case of the joint detector any hypothesis h_i corresponds either to one person or two people, depending on the detector component. Since the single-person and two-person detection components are trained jointly we expect their detection scores to be comparable, and let the temporal inference decide which component to choose. Given the final set of tracks we then predict people bounding boxes for all two-person hypotheses using the procedure described in Sec. 4.

Fig 7 shows sample frames visualizing the tracking results. Note, that tracker based on the single-person detector is able to recover tracks of people even under significant partial occlusions (e.g. track 5 in the first image and track 6 in the second image). However, it fails when people become strongly occluded as for example the person behind track 28 in the first image or tracks 16 and 39 in the fourth image. The tracker based on the joint detector is able to correctly track people even in such difficult cases clearly showing the potential of using our joint detector as the basis for multi-people tracking in scenes with many people and in the presence of severe occlusions.

6 Conclusion

Occlusion handling is a notorious problem in computer vision that typically requires careful reasoning about relationships between objects in the scene. Building on the state-of-the-art DPM detector [10], we developed a joint model that is trained to detect single people as well as pairs of people under varying degrees of occlusion. In contrast to standard people

detectors that treat occlusions as nuisance and degrade quickly in the presence of strong occlusions, our detector is specifically trained to capture various occlusion patterns. Our joint detector significantly improves over a single-person detector when detecting people in crowded street scenes, without loosing performance on images with one person only. On the challenging TUD-Crossing benchmark our joint detector improves the previously best result of [4] from 73% to 83% EER. Finally, we have demonstrated the effectiveness of our joint detector as a building block for tracking-by-detection. We envision that our approach can be applicable to detection of multiple people in various domains (e.g. surveillance videos or sports scenes) and can be used as a building block for tracking-by-detection, pose estimation, and activity recognition in multi-people scenes.

Acknowledgement: We would like to thank Bojan Pepik for the code and helpful discussions on DPM.

References

[1] M. Andriluka, S. Roth, and B. Schiele. People-tracking-by-detection and people-detection-by-tracking. In *CVPR'08*, .

[2] M. Andriluka, S. Roth, and B. Schiele. Monocular 3d pose estimation and tracking by detection. In *CVPR'10*, .

[3] A. Andriyenko, K. Schindler, and S. Roth. Discrete-continuous optimization for multi-target tracking. In *CVPR'12*.

[4] O. Barinova, V. Lempitsky, and P. Kohli. On detection of multiple object instances using hough transform. In *CVPR'10*.

[5] L. Bourdev and J. Malik. Poselets: Body part detectors trained using 3D human pose annotations. In *ICCV'09*.

[6] M.D. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, and L. Van Gool. Robust tracking-by-detection using a detector confidence particle filter. In *ICCV'09*.

[7] P. Dollár, C. Wojek, B. Schiele, and P. Perona. Pedestrian Detection: A Benchmark. In *CVPR'09*.

[8] M. Enzweiler, A. Eigenstetter, B. Schiele, and D. M. Gavrila. Multi-cue pedestrian classification with partial occlusion handling. In *CVPR'10*.

[9] A. Farhadi and M.A. Sadeghi. Recognition using visual phrases. In *CVPR'11*.

[10] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *PAMI'10*.

[11] C. Huang, B. Wu, and R. Nevatia. Robust object tracking by hierarchical association of detection responses. In *ECCV'08*.

[12] B. Leibe, E. Seemann, and B. Schiele. Pedestrian detection in crowded scenes. In *CVPR'05*.

[13] J. Marin, D. Vazquez, D. Geronimo, and A.M. Lopez. Learning appearance in virtual scenarios for pedestrian detection. In *CVPR'10*.

[14] L. Pishchulin, A. Jain, C. Wojek, M. Andriluka, T. Thormählen, and B. Schiele. Learning people detection models from few training samples. In *CVPR'11*.

[15] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake. Real-time human pose recognition in parts from a single depth image. In *CVPR'11*.

[16] X. Wang, T.X. Han, and S. Yan. An hog-lbp human detector with partial occlusion handling. In *ICCV'09*.

[17] C. Wojek, S. Walk, S. Roth, and B. Schiele. Monocular 3d scene understanding with explicit occlusion reasoning. In *CVPR'11*.

[18] B. Wu and R. Nevatia. Detection and tracking of multiple, partially occluded humans by Bayesian combination of edgelet based part detectors. *IJCV*, 75:247–266, November 2007.