Int J Comput Vis
DOI 10.1007/s11263-013-0664-6

Detection and Tracking of Occluded People

Siyu Tang · Mykhaylo Andriluka · Bernt Schiele

Received: 14 February 2013 / Accepted: 7 October 2013

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.

Keywords Pedestrian detection · Tracking · Multiple people tracking · Occlusion handling

1 Introduction

Single people detectors such as the powerful deformable part models (DPM, Felzenszwalb et al. 2010) 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 fail 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. Although these approaches allow us 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 us to predict bounding boxes of two people even when they occlude each other by 50 % or more as well as a new training method for this detector. This approach outperforms single-person detectors by a large margin in the presence of significant partial occlusions (Sect. 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 (Sect. 4). Last, we integrate the above joint model into a tracking approach to show its potential for people detection and tracking occluded people (Sect. 5).

2 Related Work

Recent methods to track people (Huang et al. 2008; Wu and Nevatia 2007; Breitenstein et al. 2009; Andriyenko et al. 2012) employ people detectors to generate initial tracking

S. Tang (✉) · M. Andriluka · B. Schiele
Computer Vision and Multimodal Computing,
Max Planck Institute for Informatics, Saarbrücken, Germany
e-mail: tang@mpi-inf.mpg.de

M. Andriluka
e-mail: andriluka@mpi-inf.mpg.de

B. Schiele
e-mail: schiele@mpi-inf.mpg.de

Published online: 08 November 2013

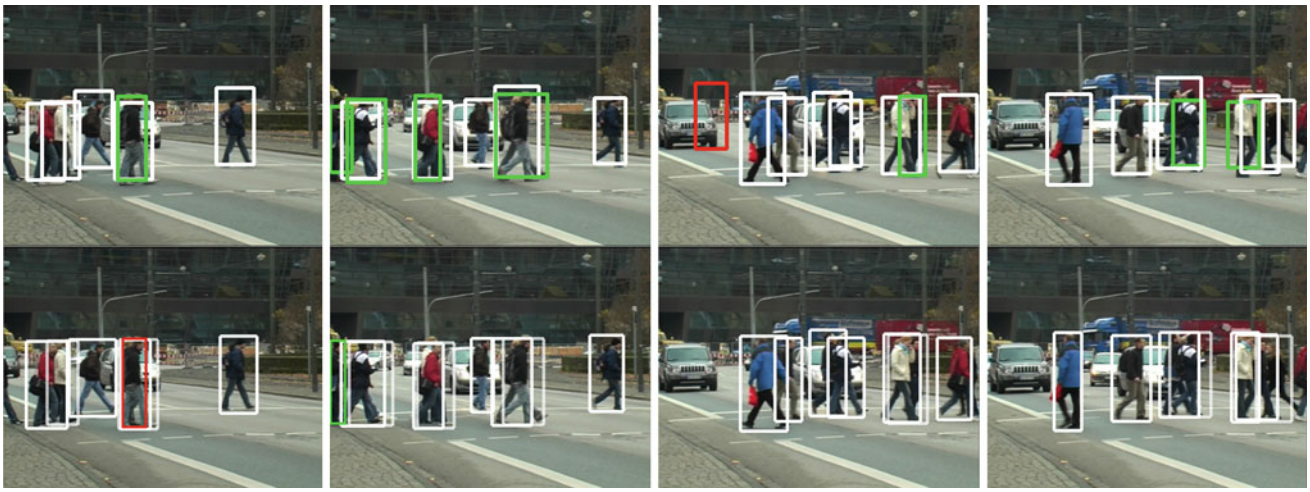


Fig. 1 Detection results at equal error rate obtained with the approach of Barinova et al. (2010) (*top*) and our joint detector (*bottom*) on the TUD-Crossing Andriluka et al. (2008) dataset. False-positive detections are shown in *red* and missing detections in *green*. One of the two bounding boxes predicted from the two-person detection is shown with the *dotted line*

hypotheses, and they 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 (Dollár et al. 2009; Felzenszwalb et al. 2010) are able to reliably detect people under a variety of imaging conditions, people poses, and appearance. Although they are 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 (Felzenszwalb et al. 2010), using a large number of part detectors (Poselets) (Bourdev and Malik 2009), detection of interactions between persons and objects (Desai and Ramanan 2012), and careful reasoning about association of image evidence to detection hypotheses (Leibe et al. 2005; Barinova et al. 2010; Wang et al. 2009). Leibe et al. (2005) 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. Barinova et al. (2010) 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 (Enzweiler et al. 2010) or by joint reasoning about people hypotheses and 3D scene layout (Wojek et al. 2011). In Wojek et al. (2011), 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 because overlapping people result in characteristic appearance patterns that are otherwise uncommon. Our approach is related to the "visual phrases" approach (Farhadi et al. 2011) in that we train a joint detector for the combination of two object instances, and to Desai and Ramanan (2012) that trains mixtures of detectors with some of the mixture components representing appearance of typical occluders. Our approach builds on the state-of-the-art people detector of Felzenszwalb et al. (2010), 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 recent work combining real and artificially generated images to train people detectors (Marin et al. 2010; Pishchulin et al. 2011).

Following the conference version of this paper (Tang et al. 2012), several recent publications proposed approaches for detection and tracking of occluded people (Pepik et al. 2013; Ouyang and Wang 2013; Tang et al. 2013). These approaches differ in the way they identify the occlusion pat-

terns and in the way they represent appearance of occluding and occluded persons. While in Tang et al. (2012) we define occlusion patterns based on the level of occlusion, (Pepik et al. 2013; Ouyang and Wang 2013) and our recent publication Tang et al. (2013) propose to mine occlusion patterns by clustering pairs of people in the training data. In Tang et al. (2012) we combine double and single-person detections by relying on a two-stage non-maximum suppression. As an alternative, Ouyang and Wang (2013) propose a probabilistic approach that allows us to incorporate evidence from any single-person detector, whereas Pepik et al. (2013) directly incorporates the appearance of an occluding person as a part of the single-person detector. In each of these publications the experimental results suggest that incorporating appearance of the occluder into the person detector can significantly improve the detection of the occluded people.

This paper extends our conference paper (Tang et al. 2012) in several ways. As a first extension we integrate our joint detector into two state-of-the-art people tracking approaches (Andriyenko and Schindler 2011; Pirsiavash et al. 2011) and evaluate its performance in the context of people tracking in crowded scenes. In our evaluation we use the standard metrics for multi-target tracking (Bernardin and Stiefelhagen 2008), which permits direct comparison to prior work. Our analysis shows that our joint people detector significantly improves recall of both tracking systems and also results in improved tracking accuracy. As a second extension over (Tang et al. 2012) this paper includes a set of additional experiments that compares the performance of the joint detector to various baseline methods which rely on the detection of single people. Specifically we compare (1) to single-person detectors with different non-maximum suppression parameters, (2) to a detector that predicts positions of occluding and occluded people from the positions of a single-detector bounding box, and (3) to a detector that is composed of the single-person detection components of the joint detector. We show that our joint person detector improves over all these baselines. Finally, we also extend the conference version of the paper with additional illustrations and clarifications that provide more insight into the workings of our approach.

3 Double-Person Detector

Our double-person detector builds on the DPM approach (Felzenszwalb et al. 2010), 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 us to improve the localization accuracy of both people using the above mentioned occlusion patterns whenever appropriate (cf. Fig. 2). 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.

Fig. 2 Visualization of the deformable parts of the double-person detector. (**a**, **c**) are the test images from MPII-2person dataset. (**b**, **d**) are the visualization of the parts locations

3.1 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 that defines the coarse location of two people and n deformable part filters that cover 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 specifying the image position of the part 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 Felzenszwalb et al. (2010), 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 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: $\arg\max_{(z)} \langle \beta, \Psi(x, z) \rangle$ (Fig. 3).

3.2 Model Training

Let $D = (\langle x_1, y_1 \rangle, \ldots, \langle x_N, y_N \rangle)$ denote a set of positive and negative training examples, with x_i corresponding to a bounding box enclosing either a pair of people or a background region and $y_i \in \{-1, 1\}$.

Given this training set we learn the model parameters β using latent SVM (Felzenszwalb et al. 2010). This involves iteratively solving the quadratic program:

Fig. 3 Examples of synthetically generated training images for different levels of occlusion: 5–10 % (**a**), 20–30 % (**b**), 40–50 % (**c**) and 70–80 % (**d**)

$$\min_{\beta,\xi\geq 0} \frac{1}{2}\max_{c}\|\beta_c\|^2 + C\ \Sigma_{i=1}^{N}\xi_i$$

$$\text{sb.t. } y_i\langle\beta, \Psi(x_i, z)\rangle \geq 1-\xi_i \qquad \xi_i \geq 0, \tag{1}$$

and optimizing for the values of latent parameters z. The optimization objective in Eq. 1 includes a regularizer that has been proposed in Girshick et al. (2010) and is slightly different from the one in Felzenszwalb et al. (2010). Instead of penalizing the norm of the whole parameter vector, it only penalizes maximum over the norms of the parameters of each component. The purpose of such regularization is to prevent one single component from dominating the model, and to make the scores of individual components more comparable. We solve the quadratic program with stochastic gradient descent and employ data-mining of hard-negative examples after each optimization round as proposed in Felzenszwalb et al. (2010).

3.3 Initialization

The objective function of the latent SVM is non-convex, which makes the training algorithm susceptible to local minima. Instead of relying on the bounding box aspect ratio as in Felzenszwalb et al. (2010), we initialize our model using different occlusion levels, which we found to produce slightly better results compared to standard initialization. This follows the intuition that the degree of occlusion is one of the major sources of the appearance variability and can be captured 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 use a three component double-person model. The components are initialized with the occlusion levels 5–25, 25–55, and 55–85 %. The percentage of occlusion is defined as a percentage of the occluded pixels in the person segmentation.

3.4 Bounding Box Predictions

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

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

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 component c. For each of the model components we estimate two separate regression models that correspond to the two people in the double-person detection. This procedure allows us to accurately localize both people despite severe occlusions, as can be seen e.g. in Fig. 4.

3.5 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. Figure 5 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 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 (Andriluka et al. 2008). For the synthetic dataset we uniformly sample occlusion levels between 0 and 85 %, and scale factors between 0.9 and 1.1.

3.6 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 (MPII-2Person) where the 850 double-person images are categorized by different

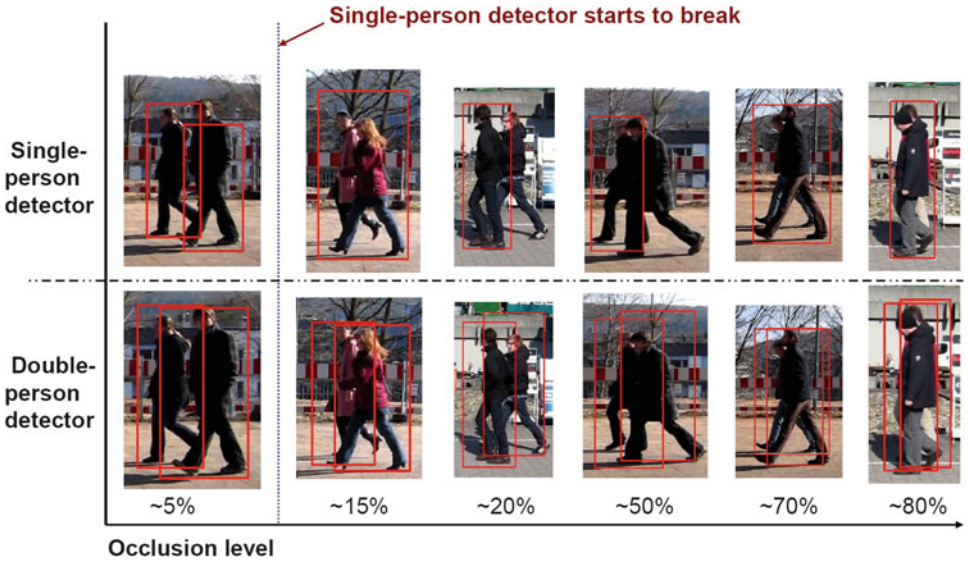


Fig. 4 Qualitative comparison of single- and double-person detectors for different occlusion levels

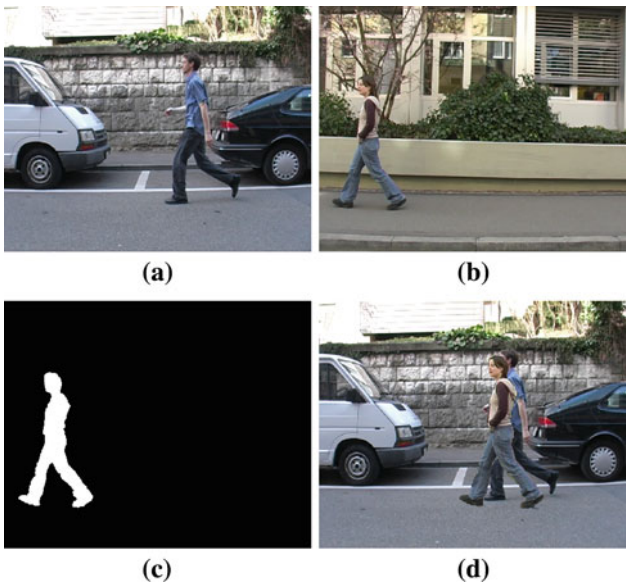


Fig. 5 Procedure to synthetically generate training images for our double-person detector. (**a**) background person, (**b**) foreground person, (**c**) foreground person map, (**d**) generated synthetic training image

occlusion levels[1] (see Fig. 6). The person segmentation and occlusion level are estimated from 2D truncated quadrics which are constructed from stick-man annotation.

3.6.1 Single-Person Detector

Figure 7a 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) and 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 %, indicating that in most cases only one of the two people is correctly detected.

3.6.2 Double-Person Detector

Figure 7b shows the performance of our proposed double-person detector. The detector reaches nearly 100 % recall with very few false positives, which is a significant improvement over the single-person detector. Interestingly, the performance for the lowest occlusion level (red curve, up to 5 %) is lower than for the levels with more occlusion, which can be explained by 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 %) (for an example of 80 % occlusion see Fig. 4). Overall the detection precision is very high for all but the highest occlusion level (black dashed line, up to 85 %).

We now compare the double-person detector with two baselines that rely on the single-person detector. The first baseline is obtained by varying the threshold τ used in the non-maximum suppression (NMS) step. This parameter determines the minimum value of the "intersection over union" ratio required for one detection bounding box to suppress the other. The results of this experiment are shown in Fig. 8. For each detector we plot the area under the recall-precision curve (AUC) for the range of occlusion levels. For low occlusion levels, the detectors with low NMS thresholds perform reasonably well, however, their performance

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

Fig. 6 Example images from the MPII-2Person dataset. The levels of occlusion in (**a**–**d**) are 30, 50, 70 and 80 % respectively

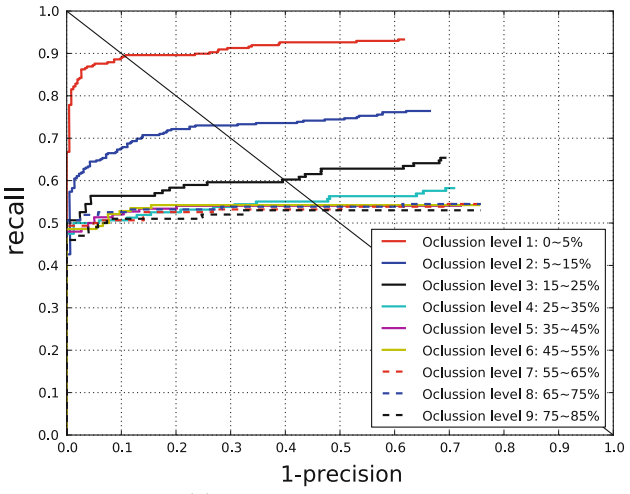
degrades quickly for higher levels of occlusion. Increasing the NMS threshold improves performance for the higher occlusion levels because the larger number of candidate detections survive NMS, but the performance for the low occlusion levels drops due to an increased number of false positives. The first observation from this experiment is that there is no single NMS threshold which works equally well for all levels of occlusion. The second observation is that our two-person detector (blue dashed line) outperforms all single-person detectors above.

Our second baseline is obtained by predicting the detection bounding boxes for two people based on the output of the single-person detector. To that end the bounding box of the second person is randomly generated in the vicinity of the single-person detection. We purposefully choose a small value of non-maximum suppression parameter $\tau = 0.3$ to prune the detections close to each other and to prevent conflicts between generated and detected bounding boxes. The result of this experiment corresponds to the "Predict double from single" curve in Fig. 8. The performance is similar or better than single-person detectors for a full range of NMS thresholds. Recall that the MPII-2Person dataset used in this experiment contains only images of two people walking close to each other, and good performance of the second baseline is not surprising. The performance of the second baseline however drops on images with small amounts of occlusion (less than 15 %). Note that our double-person detector also clearly improves over the second baseline.

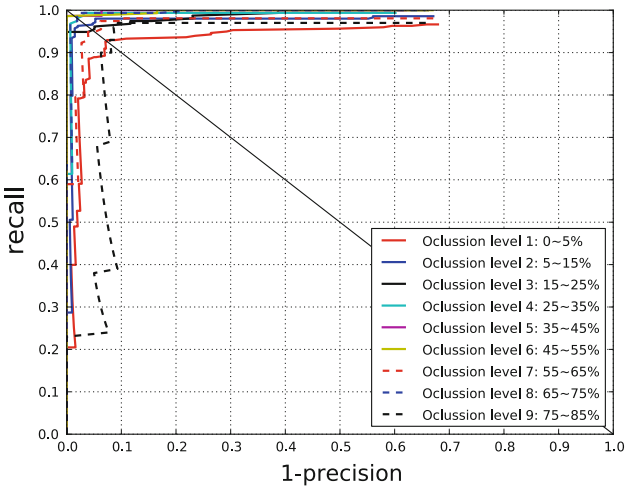
From these experiments we conclude that our double-person detector is much more robust than the single-person detector and obtains excellent performance both in terms of recall and precision, even for the heavy occlusion cases. Single person localization (bounding boxes prediction) is not a trivial task, especially for intermediate occlusion level cases ($30 \sim 60$ %), because we observe fair evidence from both persons, which can be distracting for single bounding box localization. However, the results show that our double-person detector accurately and robustly predicts the single bounding box for the above mentioned case as well. Figure 4 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. However, the employed dataset 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



(a) Single-person detector



(b) Double-person detector

Fig. 7 Detection performance of single- and double-person detectors for different occlusion levels on the MPII-2Person dataset

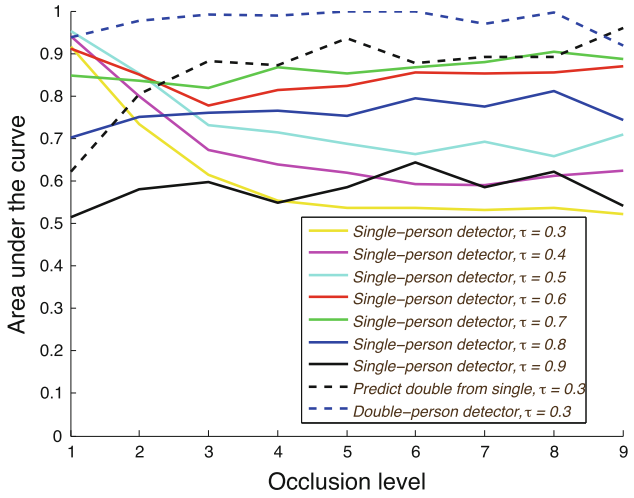


Fig. 8 Comparison of the double-person detector with various baselines based on the single-person detector on the MPII-2Person dataset. See Fig. 7 for the definition of occlusion levels (x axis)

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 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 three components for the double detector and three components for the single-person detector after mirroring results in a 12 component DPM model. Similarly to Sect. 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 two-person examples to one-person components and vice 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.

The performance of the joint detector strongly depends on its ability to distinguish between single and double-person hypotheses, which requires the scores of single and double person components to be comparable to each other. To achieve such comparability we jointly optimize the parameters of all detection components. The optimization procedure used for learning the DPM parameters described in Sect. 3 couples the training of each component in several ways. The components are jointly regularized by penalizing the maximum over the norms of the component parameters (cf. Eq. 1). In addition the training examples can be reassigned between components after each optimization round, and hard negative mining and optimization stopping criterion depends on the full model and not on an individual component. Even though we fix the assignment of training examples to single and double-person components, the other coupling mechanisms remain. The empirical evidence suggests that such joint training makes the output scores of each component comparable (Girshick et al. 2010). In this paper we follow this standard practice, but refer to our recent work (Tang et al. 2013) where we further address this issue by reformulating our joint detector using structural SVM framework and modifying the loss function to penalize detection of single people with double-person components and vice versa. In Fig. 9 we visualize the root and part filters of the joint detector. Note the substantial differences between the filters of the single and double-person components.

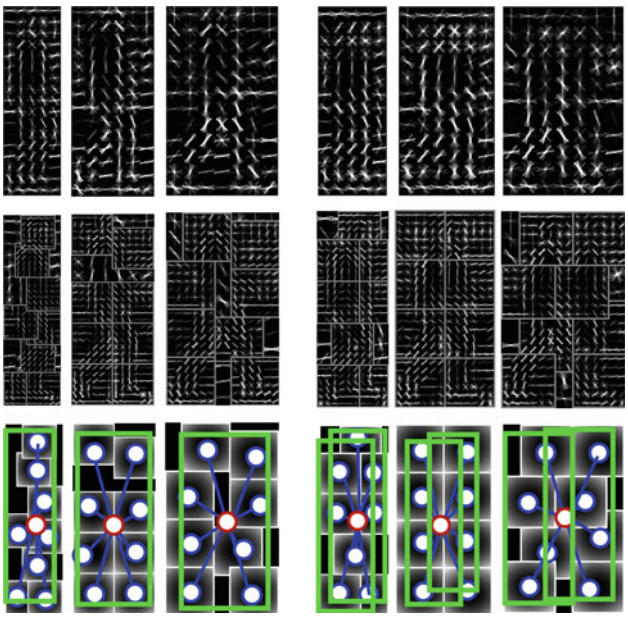


Fig. 9 Visualization of the root filters (*first row*), part filters (*second row*) and mean part locations and detection bounding boxes (*third row*) of the joint person detector. The *first three columns* correspond to the single-person and the *last three columns* to the double-person components

4.1.1 Training Data

We train our joint detector on the combination of 1-person and 2-person training sets described in Sect. 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, because in that case the single-person detector already obtains high performance for both people. We initialize the three double-person DPM components with images corresponding to occlusion levels: 5–25, 25–55, and 55–75 %.

4.1.2 Non-maximum Suppression (NMS)

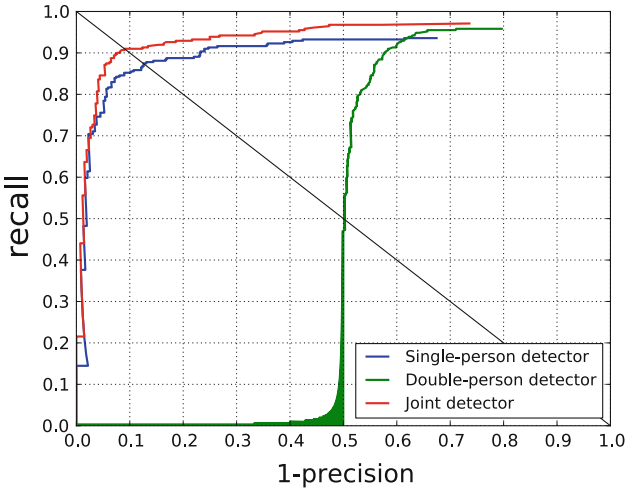
The NMS in the joint detector is more complicated than in the standard DPM since we have bounding box predictions from two different types of detections (single and two-person detections) as well as strongly overlapping bounding box predictions from our two-person components. We thus implement NMS in 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 three 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 (Felzenszwalb et al. 2010). 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 were 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

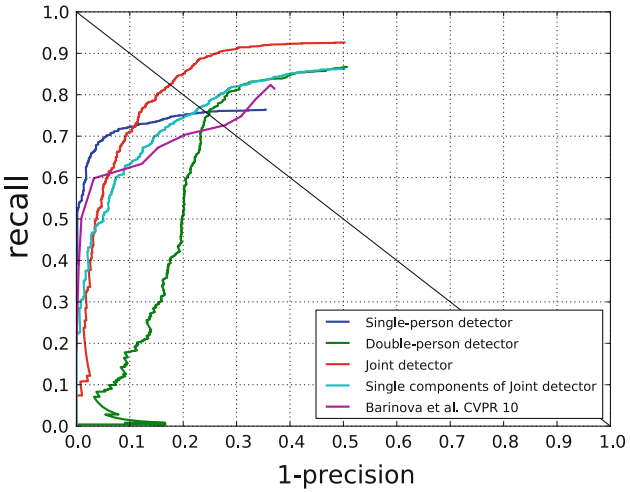
We evaluate the performance of our joint detector on two publicly available datasets, "TUD-Pedestrians" and "TUD-Crossing", originally introduced in Andriluka et al. (2008). "TUD-Pedestrians" contains 250 images of typical street scenes with 311 people all of which are fully visible. "TUD-Crossing" contains a sequence of 201 images with 1,008 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 1,186 annotated people.

We begin our analysis with the "TUD-Pedestrians" dataset. Detection results are shown in Fig. 10a as recall-precision curves. Since this dataset does not contain any occluded people our double-person detector (Sect. 3) generates numerous false positives, interpreting each person as a pair of people in which one of the persons 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. Sect. 3). This parallels the recent results on using artificially generated data for training of people detection and pose estimation models (Shotton et al. 2011; Pishchulin et al. 2011).

The evaluation on "TUD Pedestrian" demonstrates that integrating single- and double-person detectors in the same model does not result in a performance penalty in the case



(a) TUD-Pedestrians



(b) TUD-Crossing

Fig. 10 Detection performance on TUD-Pedestrians (**a**) and TUD-Crossing (**b**)

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. 10b and a few example images in 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, the precision of the double-person detector suffers from multiple detections of 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 %. In order to gain further insight into the workings of our approach, we conduct an additional experiment in which we measure the performance of the detector composed of the single-person components of the joint detector. The results are also shown in Fig. 10b. The single-components detector performs slightly better than the single-person detector (76 vs. 77 % EER), but does not reach the performance of the complete joint detector (77 vs. 83 % EER).

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 Barinova et al. (2010), which is specifically designed to be robust to partial occlusions. The authors of Barinova et al. (2010) 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. 10b). The approach of Barinova et al. (2010) 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 Barinova et al. (2010) by a large margin. Figure 1 shows a few example frames from the "TUD-Crossing" sequence, comparing our joint detector with the results of Barinova et al. (2010). 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 Barinova et al. (2010) 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

In this section we compare the performance of the single-person and the joint detectors (Sect. 4) in the context of multiple people tracking. To that end we rely on two recently proposed tracking approaches (Andriyenko and Schindler 2011; Pirsiavash et al. 2011). Both of them employ the tracking-by-detection strategy and require output of the person detector as a prerequisite for tracking. In the following we first introduce these approaches and then discuss the experimental results. The approach of Andriyenko and Schindler (2011) formulates tracking as a continuous energy minimization problem.

Given a set of person detections in each frame it recovers tracks of people by minimizing an objective function of the form

$$E(\mathbf{X}) = E_{obs} + \alpha E_{dyn} + \beta E_{exc} + \gamma E_{per} + \delta E_{reg}, \qquad (3)$$

where $\mathbf{X}$ is a set of tracks, E_{obs} is a data term that encourages tracks that align well with the person detections, and the terms E_{dyn}, E_{exc}, and E_{per} encode prior assumptions on the tracking trajectories that encourage smooth and persistent trajectories without collisions. The term E_{reg} is a regularizer that penalizes the total number of trajectories. All terms in Eq. 3 depend on $\mathbf{X}$, and we omit explicitly stating this dependency for the brevity of notation. We refer to Andriyenko and Schindler (2011) for the detailed description of the terms in Eq. 3.

The approach of Andriyenko and Schindler (2011) is particularly suited for our task of evaluating different detectors in the context of tracking-by-detection because it relies on a clean formulation that directly accepts object detections as input, and only depends on a handful of free parameters. The only adaptation needed to integrate a particular object detector into the tracking system is to estimate the parameters α, β, γ and δ in Eq. 3. In our evaluation we rely on the publicly available implementation provided by the authors[2], but re-estimate the parameters of the objective function by performing a grid search independently for each of the detectors.

As second tracking approach in our experiments we use the multi-person tracker from Pirsiavash et al. (2011). Similarly to (Andriyenko and Schindler 2011) this approach recovers tracks of multiple people by minimizing the joint objective function that combines the people detection likelihood with the smoothness prior on the track locations. The optimization is performed using an iterative greedy shortest-path algorithm. At each iteration it finds the best track and removes its hypotheses from the search space. The procedure is repeated as long as the newly found tracks have a negative cost and therefore decrease the value of the overall objective function. The objective function optimized in Pirsiavash et al. (2011) is conceptually similar to the one used in Andriyenko and Schindler (2011), but differs in the details of the likelihood and motion smoothness terms. The approach of Pirsiavash et al. (2011) directly links the people detections across frames, whereas the approach of Andriyenko and Schindler (2011) has a soft constraint that pulls the tracks towards detections but permits slight deviations. Moreover, the approach of Andriyenko and Schindler (2011) relies on a constant velocity prior that is more suitable for tracking walking pedestrians compared to constant position prior used in Pirsiavash et al. (2011). Finally, Andriyenko and Schindler (2011) explicitly discourage multiple explanations of the image detections by several tracks via the exclusion term, whereas Pirsiavash et al. (2011) achieves this using non-maximum suppression. Note that the tracker of Pirsiavash et al. (2011) is also conceptually similar to the tracker used in the conference version of this paper (Tang et al. 2012). In both cases the tracking can be interpreted as a MAP estimation in the hidden Markov model, which is performed by iterative greedy procedure that finds one track at a time. The tracker in Pirsiavash et al. (2011) is somewhat more advanced, as it incorporates occlusion handling by allowing tracks that skip several consecutive frames with low detection likelihood. In our experiments we rely on the publicly available implementation of Pirsiavash et al. (2011) and use the default tracking parameters provided by the authors[3].

We quantify the tracking performance using the CLEAR MOT metrics (Bernardin and Stiefelhagen 2008). The tracking results are evaluated with respect to the following characteristics: recall, precision, multi-object tracking accuracy, multi-object tracking precision, and the number of mostly tracked and mostly lost targets. Recall and precision are computed in the same way as in the evaluation of the detection performance, but using the ground truth targets and the tracker outputs. Multi-object tracking accuracy *(MOTA)* is the combined metric that takes missed targets, false alarms and identity switches into account. Multi-object tracking precision *(MOTP)* is computed using the average distance between the predicted track and the ground truth trajectory. *MT* is the absolute number of mostly tracked trajectories, and *ML* is the absolute number of mostly lost trajectories. The hit/miss threshold is 50 % overlap between the ground truth targets and the tracker outputs in 2D.

We evaluate the full system composed of either our single-person or our joint detector and one of the tracking algorithms (Andriyenko and Schindler 2011; Pirsiavash et al. 2011) on the TUD-Crossing dataset. The results are shown in Table 1.

First, we present the results obtained with the tracker of Andriyenko and Schindler (2011). The single-person detector significantly improves over the result of Andriyenko and Schindler (2011) that was obtained using a detector from Walk et al. (2010) based on the HOG and optical flow features. The best result is obtained using our joint detector, that improves over the single-person detector both in terms of recall, and with respect to MOTA/MOTP tracking metrics. Figure 11 shows several example frames visualizing the tracking results. Note that the tracker based on the joint detector is able to track people even under significant partial occlusions (*e.g* track 2 in the first three images), and is able to track subjects for longer periods of time (*e.g* track 10 of the joint detector (third row) corresponds to two tracks of the single-frame detector (second row)). Tracking based on the output of the joint detector also results in fewer identity switches (16

[2] http://www.gris.tu-darmstadt.de/~aandriye

[3] http://people.csail.mit.edu/hpirsiav

Table 1 2D tracking evaluation on the TUD-Crossing dataset

Method	Recall	Precision	MOTA (%)	MOTP (%)	MT	ML
Approach of Andriyenko and Schindler (2011)	69.8	92.4	63.0	75.5	7	1
Our single-person detector, tracking method of Andriyenko and Schindler (2011)	79.9	96.2	75.2	77.7	7	0
Our single-person detector, tracking method of Pirsiavash et al. (2011)	68.3	**98.4**	63.3	76.3	5	0
Our joint detector, tracking method of Pirsiavash et al. (2011)	**77.7**	96.2	**70.7**	**77.1**	6	0

The best results in competition are highlighted in bold

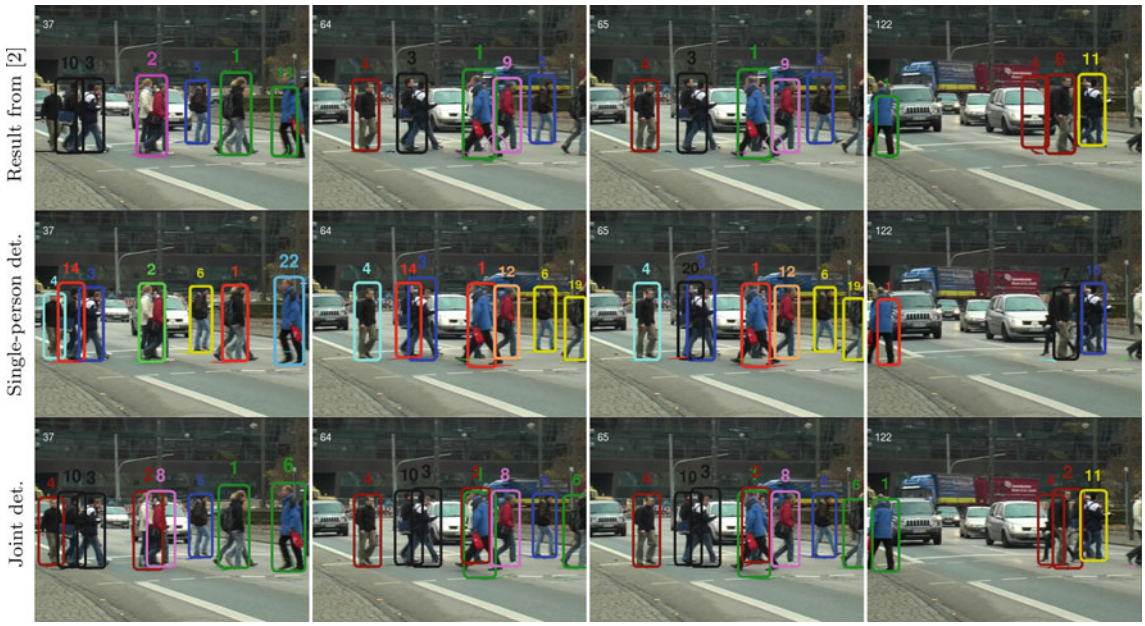


Fig. 11 Tracking results on the TUD-Crossing dataset obtained with the approach of Andriyenko and Schindler (2011) (*top row*), our single-person detector (*middle row*) and our joint detector (*bottom row*). *Colors* and *numbers* indicate tracks corresponding to different people

for the single-person detector vs. 11 for the joint detector). Inspection of the output of the single-person detector reveals that in the case of strong partial occlusions the detection output often jumps between occluder and occluded subjects, which results in frequent identity switches in corresponding track. In contrast the joint detector typically includes detections of both subjects into the hypotheses set, which facilitates more consistent tracking.

Note that although the joint detector achieves the best result, the improvement over the single-person detector is only 3.2 % of MOTA. This is somewhat surprising given the large improvement of the joint detector on the detection task (cf. Fig. 10). This result could be due to the particular choice of the objective function which contains the term E_{exc} which explicitly penalizes tracks which collide with each other in the image space. In the case of strong partial occlusions tracks of both subjects might be rather close to each other, where this exclusion term is likely to be suboptimal. The tracking algorithm does not take advantage of the additional information contained in the output of the joint detector that is able to explicitly label detections as a pair of occluded and occluding people. We envision that a more careful integration of the joint detector into the tracking framework could lead to larger performance gains and leave such integration to the future work.

Next, we evaluate our proposed detectors in combination with the tracking algorithm of Pirsiavash et al. (2011). The results are shown in the last two rows of the Table 1. The tracking results obtained both with single and joint-person detectors are somewhat lower than with the tracker of Andriyenko and Schindler (2011). The large difference in tracking recall is particularly striking. For example, in the case of the single-person detector we obtain 79.9 % for the tracker of Andriyenko and Schindler (2011) and 68.3 % for the tracker of Pirsiavash et al. (2011). The difference could be due to a more sophisticated design of the objective function in Andriyenko and Schindler (2011) that explicitly encourages longer tracks by incorporating the persistence term. Importantly, for both trackers we achieve noticeable improvement from substituting the single-person with the

joint-person detector. The improvement for the tracker of Pirsiavash et al. (2011) is particularly pronounced. For example, the joint detector is able to improve the aggregated tracking accuracy measure MOTA from 63.3 to 70.7. We hypothesize that the improvement for Pirsiavash et al. (2011) is larger because it operates by linking a discrete set of detection hypotheses over time and is therefore more sensitive to missing detections. In contrast the tracker of Andriyenko and Schindler (2011) only uses detections as observations for tracking and explicitly reasons about continuous trajectories, which allows it to better handle gaps in detections.

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 (Felzenszwalb et al. 2010), 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 losing performance on images with one person only. On the challenging TUD-Crossing benchmark our joint detector improves the previously best result of Barinova et al. (2010) from 73 to 83 % EER. Finally, we 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.

Acknowledgments The authors are thankful to Bojan Pepik for the code and suggestion on DPM and to Anton Andriyenko for the help with the multi-people tracking evaluation.

References

Andriluka, M., Roth, S., & Schiele, B. (2008). People-tracking-by-detection and people-detection-by-tracking. In *CVPR'08*.

Andriyenko, A., & Schindler, K. (2011). Multi-target tracking by continuous energy minimization. In *CVPR'11*.

Andriyenko, A., Schindler, K., & Roth, S. (2012). Discrete-continuous optimization for multi-target tracking. In *CVPR'12*.

Barinova, O., Lempitsky, V., & Kohli, P. (2010). On detection of multiple object instances using hough transform. In *CVPR'10*.

Bernardin, K. & Stiefelhagen, R. (2008). Evaluating multiple object tracking performance: The CLEAR MOT metrics. *Image and Video Processing, 1*, 1–10.

Bourdev, L., & Malik J. (2009). Poselets: Body part detectors trained using 3D human pose annotations. In *ICCV'09*.

Breitenstein, M., Reichlin, F., Leibe, B., Koller-Meier, E., & Van Gool, L. (2009). Robust tracking-by-detection using a detector confidence particle filter. In *ICCV'09*.

Desai, C., & Ramanan, D. (2012). Detecting actions, poses, and objects with relational phraselets. In *ECCV'12*.

Dollár, P., Wojek, C., Schiele, B., & Perona, P. (2009). Pedestrian detection: A benchmark. In *CVPR'09*.

Enzweiler, M., Eigenstetter, A., Schiele, B., & Gavrila, D. M. (2010). Multi-cue pedestrian classification with partial occlusion handling. In *CVPR'10*.

Farhadi, A., & Sadeghi, M. A. (2011). Recognition using visual phrases. In *CVPR'11*.

Felzenszwalb, P. F., Girshick, R. B., McAllester, D., & Ramanan, D. (2010). Object detection with discriminatively trained part-based models. In *PAMI'10*.

Girshick, R. B., Felzenszwalb, P. F., & McAllester, D. (2010). *LSVM-MDPM Release 4 Notes*.

Huang, C., Wu, B., & Nevatia, R. (2008). Robust object tracking by hierarchical association of detection responses. In *ECCV'08*.

Leibe, B., Seemann, E., & Schiele, B. (2005). Pedestrian detection in crowded scenes. In *CVPR'05*.

Marin, J., Vazquez, D., Geronimo, D., & Lopez, A. M. (2010). Learning appearance in virtual scenarios for pedestrian detection. In *CVPR'10*.

Ouyang, W., & Wang, X. (2013). Single-pedestrian detection aided by multi-pedestrian detection. In *CVPR'13*.

Pepik, B., Stark, M., Gehler, P., & Schiele, B. (2013). Occlusion patterns for object class detection. In *CVPR'13*.

Pirsiavash, H., Ramanan, D., & Fowlkes, C. C. (2011). Globally-optimal greedy algorithms for tracking a variable number of objects. In *CVPR'11*.

Pishchulin, L., Jain, A., Wojek, C., Andriluka, M., Thormählen, T., & Schiele, B. (2011). Learning people detection models from few training samples. In *CVPR'11*.

Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A., & Blake, A. (2011). Real-time human pose recognition in parts from a single depth image. In *CVPR'11*.

Tang, S., Andriluka, M., Milan, A., Schindler, K., Roth, S., & Schiele, B. (2013). Learning people detectors for tracking in crowded scenes. In *ICCV'13*.

Tang, S., Andriluka, M., & Schiele, B. (2012). Detection and tracking of occluded people. In *BMVC'12*.

Walk, S., Majer, N., Schindler, K., & Schiele, B. (2010). New features and insights for pedestrian detection. In *CVPR'10*.

Wang, X., Han, T. X., & Yan, S. (2009). An hog-lbp human detector with partial occlusion handling. In *ICCV'09*.

Wojek, C., Walk, S., Roth, S., & Schiele, B. (2011). Monocular 3d scene understanding with explicit occlusion reasoning. In *CVPR'11*.

Wu, B., & Nevatia, R. (2007). Detection and tracking of multiple, partially occluded humans by Bayesian combination of edgelet based part detectors. In *IJCV'07*.