Learning People Detectors for Tracking in Crowded Scenes

Siyu Tang[1] Mykhaylo Andriluka[1] Anton Milan[2]
Konrad Schindler[3] Stefan Roth[2] Bernt Schiele[1]

[1]Max Planck Institute for Informatics, Saarbrücken, Germany
[2]Department of Computer Science, TU Darmstadt
[3]Photogrammetry and Remote Sensing Group, ETH Zürich

Abstract

People tracking in crowded real-world scenes is challenging due to frequent and long-term occlusions. Recent tracking methods obtain the image evidence from object (people) detectors, but typically use off-the-shelf detectors and treat them as black box components. In this paper we argue that for best performance one should explicitly train people detectors on failure cases of the overall tracker instead. To that end, we first propose a novel joint people detector that combines a state-of-the-art single person detector with a detector for pairs of people, which explicitly exploits common patterns of person-person occlusions across multiple viewpoints that are a frequent failure case for tracking in crowded scenes. To explicitly address remaining failure modes of the tracker we explore two methods. First, we analyze typical failures of trackers and train a detector explicitly on these cases. And second, we train the detector with the people tracker in the loop, focusing on the most common tracker failures. We show that our joint multi-person detector significantly improves both detection accuracy as well as tracker performance, improving the state-of-the-art on standard benchmarks.

1. Introduction

People detection is a key building block of most state-of-the-art people tracking methods [3, 22, 23]. Although the performance of people detectors has improved tremendously in recent years, detecting partially occluded people remains a weakness of current approaches [8]. This is also a key limiting factor when tracking people in crowded environments, such as typical street scenes, where many people remain occluded for long periods of time, or may not even become fully visible for the entire duration of the sequence.

The starting point of this paper is the observation that people detectors used for tracking are typically trained independently from the tracker, and are thus not specifically tailored for best tracking performance. In contrast, the present work aims to train people detectors explicitly to address failure modes of tracking in order to improve overall tracking performance. However, this is not straightforward, since many tracking failures are related to frequent and long-term occlusions – a typical failure case also for people detectors.

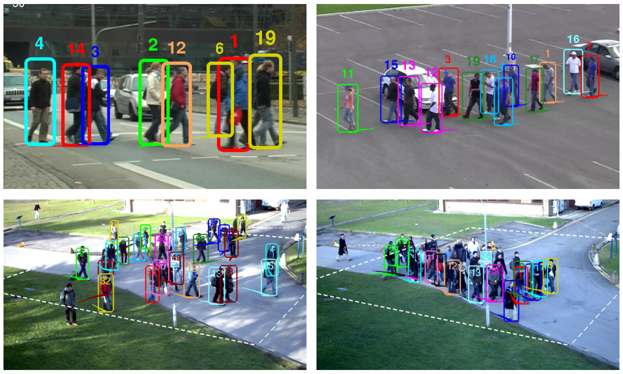


Figure 1. Tracking results using the proposed joint detector on four public datasets: (clockwise) TUD-Crossing, ParkingLot, PETS S2.L2 and PETS S1.L2.

We address this problem in two steps: First, we target the limitations of people detection in crowded street scenes with many occlusions. Occlusion handling is a notoriously difficult problem in computer vision and generic solutions are far from being available. Yet for certain cases, successful approaches have been developed that train effective detectors for object compositions [10, 17], which can then be decoded into individual object detections. Their key rationale is that objects in such compositions exhibit regularities that can be exploited. We build on these ideas, focusing on person-person occlusions, which are the dominant occlusion type in crowded street scenes. Our first contribution is a novel structural loss-based training approach for a joint person detector, based on structured SVMs.

In the second step of our approach, we specifically focus on patterns that are relevant to improving tracking performance. In general, person-person occlusions may result in a large variety of appearance patterns, yet not all of these patterns are necessarily frequent in typical street scenes. Furthermore, not every pattern will possess a discriminative appearance that can be detected reliably in cluttered images. Finally, some of the person-person occlusion cases are already handled well by existing tracking approaches (*e.g.*, short term occlusions resulting from people passing each other). We argue that the decision about incorporating certain types of occlusion patterns into the detector should be done in a tracking-aware fashion, either by manually observing typical tracking failures or by directly integrating the tracker into the detector training.

Our second contribution is to propose and evaluate two alternative strategies for the discovery of useful multi-view occlusion patterns. First, we manually define relevant occlusion patterns using a discretization of the mutual arrangement of people. In addition to that, we train the detector with the tracker in the loop, by automatically identifying occlusion patterns based on regularities in the failure modes of the tracker. We demonstrate that this tighter integration of tracker and detector improves tracking results on three challenging benchmark sequences.

Related work. Many recent methods for multi-person tracking [1, 2, 3, 22] follow the tracking-by-detection paradigm and use the output of people detectors as initial state space for tracking. Although these methods are often robust to false positive detections and are able to fill in some missing detections due to short term occlusions, they typically require successful detection before and after the occlusion events, thus limiting their applicability in crowded scenes. Various solutions to the detection of partially occluded people have been proposed in the literature [4, 9, 14, 18, 19]. Such methods often rely on additional information, such as stereo disparity [9], or 3D scene context [19]. Approaches that operate on monocular images typically handle occlusions by carefully separating the evidence coming from the occluder and the occluded objects, either by reasoning on an image segmentation [14], or by iteratively discarding image evidence corresponding to foreground objects [4, 18]. Recently, [17] proposed a people detector for crowded street environments that exploits characteristic appearance patterns from person-person occlusions. This is motivated by the observation that most of the occlusions in street scenes happen due to an overlap of multiple people, which can beleveraged. In a similar spirit, approaches to object detection using visual phrases [10] and detection of person-object interactions [7] have demonstrated that detecting a constellation of objects can be easier than detecting each object alone.

Most closely related to our work is the approach of [17], which demonstrated the advantages of joint multi-person detection for the simplified case of people seen from the side. We generalize this approach in several ways: First, we reformulate the approach as a structured prediction problem, which allows us to explicitly penalize activations of single-person detector components on examples with two people and vice versa. A novel formulation, in which constraints on the detection type are encoded into the structured loss function, significantly improves detection performance. Moreover, we generalize the joint detection approach of [17] to cope with a variety of viewpoints, not just side views, which is important when using the detector for tracking in more general scenes. Note that varying viewpoints are significantly more complex to handle than side views, because the number of ways people can potentially occlude each other increases considerably. To address this we propose an approach tailored to the requirements of people tracking, and in particular propose to train a people detector based on feedback from the tracker.

Addressing both detection and tracking as a joint problem has been considered in the literature. In [13], the task is formulated as a quadratic Boolean program to combine trajectory estimation and detection. The objective is optimized locally, by alternating between the two components. In contrast, [21] formulate a joint integer linear program and allow data association to influence the detector. However, their approach is based on background subtraction on a discretized grid. Unlike previous work, we here not only consider detection and tracking jointly, but also explicitly adapt the detector to typical tracking failures.

2. Joint People Detection

Before describing our multi-view joint people detector, let us briefly review the deformable parts model (DPM, [11]), which forms the basis of our approach. The DPM detector is based on a set of M detection components. Each component is represented by a combination of a rigid root filter F_0, and several part filters $F_1, \ldots, F_n$, which can adjust their positions w.r.t. the root filter in order to capture possible object deformations $p_1, \ldots, p_n$. The detection score of the DPM model is given by the sum of the responses of the root and part filters, a bias b, and the deformation costs between the ideal and the inferred locations of each part (with parameters $d_1, \ldots, d_n$). The positions of the part filters and the component assignment m are assumed to be latent variables $h = (p_1, \ldots, p_n, m)$, which need to be inferred during training and testing. Given training images with ground truth labels, the parameters $\beta = (F_0, F_1, \ldots, F_n, d_1, \ldots, d_n, b)$ are trained by iterating between finding the optimal position of the latent parts in each training example and optimizing the model parameters given the inferred part locations. At test time the model is evaluated densely in the image and each local maximum is used to generate a detection bounding box hypothesis, aided by the model parts. The initial set of detections is then refined by non-maximum suppression.

(a) Double person outscores single person with Δ_{VOC}

(b) Double person outscores single person with $\Delta_{\text{VOC+DT}}$

Figure 2. Structured training of joint people detectors: Green – correct double-person bounding box. Red – single-person detection whose score should be lower by a margin.

Overview. We now use the DPM model to build a joint people detector, which overcomes the limitations imposed by frequent occlusions in real-world street scenes. In doing so, we go beyond previous work on joint people detection [17] in several significant ways: *(1)* The approach of [17] focused on side-view occlusion patterns, but crowded street scenes exhibit a large variation of possible person-person occlusions caused by people's body articulation or their position and orientation relative to the camera. To address this we explicitly integrate *multi-view person/person occlusion patterns* into a joint DPM detector. *(2)* We propose a *structured SVM formulation* for joint person detection, enabling us to incorporate an appropriate structured loss function. Aside from allowing to employ common loss functions for detection (Jaccard index, a.k.a. VOC loss), this allows us to leverage more advanced loss functions as well. *(3)* We model our joint detector as a mixture of components that capture appearance patterns of either a single person, or a person/person occlusion pair. We then introduce an explicit variable modeling the *detection type*, with the goal of enabling the joint detector to distinguish between a single person and a highly occluded person pair. Incorporating the detection type into the structural loss then allows us to force the joint detector to learn the fundamental appearance difference between a single person and a person/person pair.

Before going into detail on learning occlusion patterns in Sec. 4, let us first turn to our basic structured SVM formulation for joint person detection.

Structural learning for joint detection. We adapt the structured SVM formulation for DPMs proposed in [15] for our joint person detection model. Given a set of training images $\{I_i | i = 1, \dots, N\}$ with structured output labels $y_i = (y_i^l, y_i^b)$, which include the class label $y_i^l \in \{1, -1\}$ and the 2D bounding box position y_i^b, we formulate learning the parameters of the DPM, β, as the optimization problem

$$\begin{aligned} \min_{\beta, \xi \geq 0} \quad & \frac{1}{2}\|\beta\|^2 + \frac{C}{N}\sum_{i=1}^{N} \xi_i \qquad (1) \\ \text{sb.t.} \quad & \max_{h} \langle \beta, \phi(I_i, y_i, h) \rangle - \max_{\hat{h}} \langle \beta, \phi(I_i, \hat{y}, \hat{h}) \rangle \\ & \qquad \geq \Delta(y_i, \hat{y}) - \xi_i, \quad \forall i \in \{1, \dots, N\}, \hat{y} \in \mathcal{Y}, \end{aligned}$$

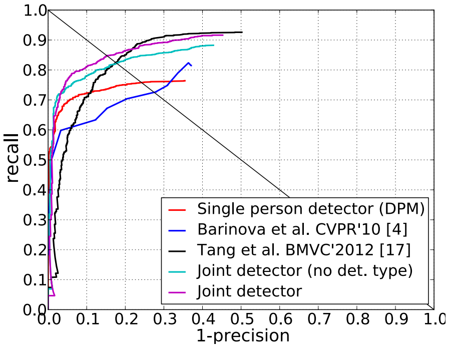


Figure 3. Detection performance on TUD-Crossing.

where ξ_i are slack variables modeling the margin violations. For the loss function Δ, we employ the area of the bounding box intersection $A(y_i^b \cap \hat{y}^b)$ over their union $A(y_i^b \cup \hat{y}^b)$

$$\Delta_{\text{VOC}}(y, \hat{y}) = \begin{cases} 0, & \text{if } y^l = \hat{y}^l = -1 \\ 1 - [y^l = \hat{y}^l] \frac{A(y^b \cap \hat{y}^b)}{A(y^b \cup \hat{y}^b)}, & \text{otherwise,} \end{cases} \tag{2}$$

as it enables precise 2D bounding box localization. The advantage of the proposed structured learning of a joint people detector is that it learns that a detection with larger overlap with the ground truth bounding box has higher score than a detection with lower overlap. Hence, the single person component should also have a lower score than the double person component on double person examples (see Fig. 2(a)).

Introducing detection type. One limitation of the loss Δ_{VOC} for joint person detection is that it does not encourage the model enough to distinguish between a single person and a highly occluded double person pair. This is due to the large overlap of the ground truth bounding boxes, as illustrated in Fig. 2(b). In order to teach the model to distinguish a single person and a highly occluded person pair, we extend the structured output label with a detection type variable $y^{dt} \in \{1, 2\}$, which denotes single person or double person detection. The overall structured output is thus given as $y = (y^l, y^b, y^{dt})$. We can then additionally penalize the wrong detection type using the loss

$$\Delta_{\text{VOC+DT}}(y, \hat{y}) = (1 - \alpha)\Delta_{\text{VOC}}(y, \hat{y}) + \alpha \left[y^{dt} \neq \hat{y}^{dt} \right]. \tag{3}$$

Experimental results. In order to fairly compare our joint detector with the joint detector proposed in [17], we explicitly train a side-view joint person detector using the same synthetic training images[1] and initialize the single and double person detector components in the same way. Fig. 3 shows the benefit of the proposed structured training (*Joint detector, no det. type*). By introducing the detection type loss (*Joint detector*, $\alpha = 0.5$), the joint detector further improves precision and achieves similar recall as [17]. At 95% precision it outperforms [17] by 20.5% recall.

[1] The data is available at `www.d2.mpi-inf.mpg.de/datasets`.

Method	Rcll	Prcsn	MOTA	MOTP	MT	ML
single (DPM)	78.0	94.1	72.1 %	78.5 %	4	0
Tang et al. [17]	79.9	96.5	75.6 %	79.1 %	6	0
Joint det. (no det. type)	81.9	93.2	75.1 %	79.1 %	8	0
Joint detector	82.7	93.9	76.0 %	78.6 %	7	1

Table 1. Tracking performance on TUD-Crossing evaluated by recall *(Rcll)*, precision *(Prcsn)* and standard CLEAR MOT metrics [5], including Multi-Object Tracking Accuracy *(MOTA)* and Tracking Precision *(MOTP)*. MT and ML show the number of mostly tracked and mostly lost trajectories, respectively [20].

3. Multi-Target Tracking

Our proposed detector learning algorithm (Sec. 4) is generic and can, in principle, be employed in combination with any tracking-by-detection method. Here, we use a recent multi-target tracker based on continuous energy minimization [2]. The tracker requires as input a set of person detections in a video sequence, and infers all trajectories simultaneously by minimizing a high-dimensional, continuous energy function over all trajectories. The energy consists of a data term, measuring the distance between the trajectories and the detections, and several priors that assess the (physical) plausibility of the trajectories. We use a fixed parameter setting throughout all experiments. Note that the employed tracking approach does not include any explicit occlusion handling. It is thus important to consider occlusions directly at the detector level, so as to provide more reliable information to the tracker.

Baseline results. Table 1 shows tracking results on the TUD-Crossing sequence [1], using various detector variants as described above. As expected, tracking based on the output of the joint detector shows improved performance compared to the single-person DPM detector. Note that the side-view joint detector of Tang et al. [17] was specifically designed to handle the occlusion pattern prevalent in sequences of this type. Even so, structured learning with a detection type variable slightly increases the multi-object tracking accuracy (MOTA, [5]). This experiment is meant to serve as a proof of concept and demonstrate the validity of the joint people detector. Please refer to Sec. 5 for an extensive experimental study on more challenging datasets.

4. Learning People Detectors for Tracking

So far we have shown that the proposed structured learning approach for training joint people detectors shows significant improvements for detection of occluded people in side-view street scenes. This suggests the potential of leveraging characteristic appearance patterns of person/person pairs also for detecting occluded people in more general settings. However, the generalization of this idea to crowded scenes with people walking in arbitrary directions is rather challenging due to the vast amount of possible person-person occlusion situations. This variation may arise from several factors, such as people's body articulation, or their position and orientation relative to the camera. The number of putative occlusion patterns is exponential in the number of factors. The crucial point here is, however, that not all of them are equally relevant for successful tracking. For example, short term occlusions resulting from people crossing each other's way are frequent, but can be often easily resolved by modern tracking algorithms. Therefore, finding occlusion patterns that are relevant in practice in order to reduce the modeling space is essential for applying joint person detectors for tracking in general crowded scenes.

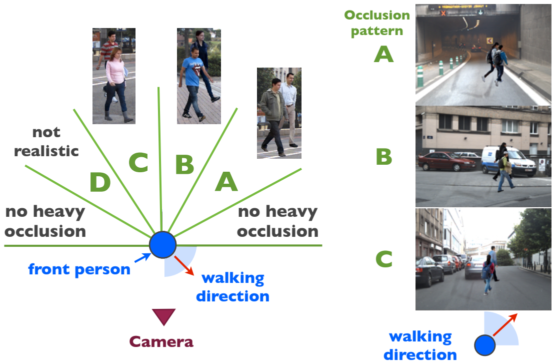


Figure 4. Bird's eye view of occluded person's state space (*left*). Synthetically generated training images for different occlusion patterns and walking directions (*right*).

We now propose two methods for discovering occlusion patterns for people walking in arbitrary directions by *(a)* manually designing regular occlusion combinations that appear frequently due to long-term occlusions and are, therefore, most relevant for tracking (Sec. 4.1); and *(b)* automatically learning a joint detector that exploits the tracking performance on occluded people and is explicitly optimized for the tracking task (Sec. 4.2).

4.1. Designing occlusion patterns

For many state-of-the-art trackers, the most important cases for improving tracking performance in crowded scenes correspond to long-term partial occlusions.

Occlusion pattern quantization. We begin by quantizing the space of possible occlusion patterns as shown in Fig. 4 (left). Given the position of the front person, we divide the relative position of the occluded person with respect to the occluder into 6 equal angular sectors. We consider the full half circle of the sectors behind the occluder, and do not explicitly quantize the space of possible relative distances between subjects; instead we only consider a fixed threshold, below which the second subject is significantly occluded.

In addition to quantizing the relative position, we also quantize the orientation of the front person with respect to the camera. To keep the number of constellations manageable, we use four discrete directions corresponding to four diagonal views. Independent of the orientation of the front person, the first and last sectors shown in Fig. 4 (left, no heavy occlusion) correspond to people walking side-by-side, slightly in front or behind each other. We found that these cases are already handled well by current person detectors. We denote the remaining four sectors as "A", "B", "C" and "D", according to the relative position of the occluded and occluding person. The sector "D" corresponds to a constellation of people walking directly behind each other at close proximity. Although physically possible, this configuration is extremely unlikely in real-world scenes, because people usually tend to leave some space to the person in front when walking. We restrict ourselves to cases in which people walk in the same direction, as they cause long-term occlusions and moreover appear to have sufficient regularity in appearance, which is essential for detection performance in crowded scenes. The occlusion patterns that we consider in the rest of this analysis correspond to a combination of the four walking directions of the subjects and one of the three remaining sectors ("A", "B" or "C").

Joint detector with designed occlusion patterns. Our joint detector uses a mixture of components that capture appearance patterns of either a single person or of a person/person occlusion pair. In case of double person components, we generate two bounding boxes of people instead of one for each of the components' detections. The training procedure in Sec. 2 is based on the optimization of a semi-convex objective, thus susceptible to local minima. Therefore, a meaningful initialization of the detector components is important for good performance. One option is to initialize the double-person components with different degrees of occlusion [17]. However, in the multi-view setting, the same degree of occlusion can result in very different occlusion patterns. Here, we instead initialize the components from the quantized occlusion patterns from above (Fig. 4, left), combining different walking directions with relative positions of the person/person pair; we construct 6 double-person components. The single-person components are initialized with different orientations, clustering appearance into 10 components, and mirroring.

Generating synthetic training examples. Training of our model requires a sufficient amount of training images. As it is very difficult and expensive to collect a representative training dataset with accurate occlusion level annotation for each image, we choose to synthetically generate training data. Most importantly, this allows us to control the data's variation with respect to viewpoint, degree of occlusion, and variability of backgrounds, as opposed to uncontrolled clutter often present in manually collected datasets.

We collect 2400 images of people walking in 8 different walking directions to construct a synthetic training image pool. We mirror the training images to double the training set. For each captured image, we segment the person and use the segmentation to generate a number of training examples by combining the segmented person with novel backgrounds. In a similar fashion, we are able to generate training examples for different occlusion patterns and walking directions by overlaying people on top of each other in a novel image. In our experiments, we use 4000 synthetic images for training the single-person components, and up to 1200 synthetic images for the double-person components. Fig. 4 (right) shows several examples of our synthetically generated training images for different constellations illustrated in Fig. 4 (left).

Occlusion-aware NMS. We perform non-maximum suppression in two rounds: First, we consider single-person detections and the predicted occluder bounding box of double-person detections. If the occluder is suppressed by a single-person detection, then the occludee is also removed. For the second round, we allow the predicted individual bounding boxes to suppress each other, except when two bounding boxes are generated by the same double-person component.

4.2. Mining occlusion patterns from tracking

As we will see in Sec. 5 in detail, carefully analyzing and designing occlusion patterns by hand already allows to train a joint detector that generalizes to more realistic and challenging crowded street scenes. Nonetheless, the question remains which manually designed occlusion patterns are most relevant for successful tracking. Furthermore, it is still unclear whether it is reasonable to harvest difficult cases from tracking failures and explicitly guide the joint detector to concentrate on those. In the following, we describe a method to learn a joint detector specifically for tracking. We employ tracking performance evaluation, occlusion pattern mining, synthetic image generation, and detector training jointly to optimize the detector for tracking multiple targets. The approach is summarized in Alg. 1.

Input: For our study, we use the first half (frames 1–218) of the challenging PETS S2.L2 dataset [12] as our mining sequence. We use the same synthetic training images to train a single-person baseline detector, as we used for training the single-component of our joint detector with manually designed occlusion patterns (see Sec. 4.1). Moreover, we employ a recent multi-target tracker [2], *c.f.* Sec. 3.

Output: A joint detector that is tailored to detect occlusion patterns that are most relevant for multi-target tracking.

Tracking evaluation (step 4): We concentrate on missed targets, which are the main source of failure in crowded scenarios. To that end, we extract all missed targets, evaluated by the standard CLEAR MOT metrics [5] for the next step.

Occlusion pattern mining (step 5): The majority of missed targets are occlusion related. For our mining sequence, the total number of missed targets is 1905, only 141 of them are not caused by occlusions (Fig. 5(a)). Missed tar-

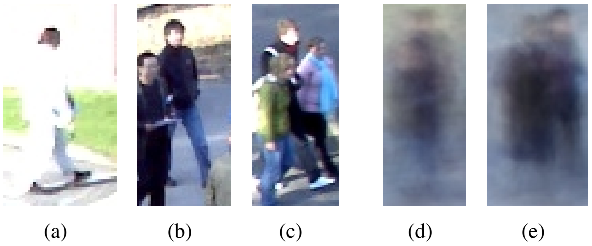


Figure 5. Missed targets from PETS S2.L2 mining sequence and mined occlusion patterns: *(a)* No person nearby; *(b)* interfered by one person; *(c)* interfered by more persons; *(d)* mined occlusion pattern – 1$^{\text{st}}$ iteration; *(e)* mined occlusion pattern – 2$^{\text{nd}}$ iteration.

gets can be occluders and/or occludees for a pair of persons (Fig. 5(b)), or within a group of multiple people (Fig. 5(c)). Here, we concentrate on mining occlusion patterns for pairs of persons and consider the multiple people situation as a special case of a person pair, augmented by distractions from surroundings. Note that our algorithm can be easily generalized to multiple people occlusion patterns given sufficient amount of mining sequences that contain certain distributions of multi-people occlusion patterns. From the missed targets (step 4), we determine the problematic occlusion patterns and cluster them in terms of the relative position of the occluder/occludee pair. We only consider the most dominant cluster. Fig. 5(d) and 5(e) show the dominant occlusion pattern of the first and second mining iteration. Note that we only mine occlusion patterns and no additional image information (see next step).

Synthetic training example generation (step 6): We generate synthetic training images for the mined occlusion pattern using the same synthetic image pool as in Sec. 4.1, which requires the relative position of a person pair, as well as the orientation of each person. To that end, we sample the relative position of a person pair from a Gaussian distribution centered on the dominant relative position cluster from step 5. We further extract a dominant orientation of the mined examples for occluders and occludees. Training image generation, in principle, thus enables us to model arbitrary occlusion patterns in each iteration. We generate 200 images for every new occlusion pattern, which amounts to the same number of training images as we used in the context of manually designed occlusion patterns. The major benefit of learning these patterns is that more training images can be easily generated for the next iteration, specifically for those relevant cases that still remain unsolved.

Algorithm 1 Joint detector learning for tracking

```
Input:
     Baseline detector
     Multi-target tracker
     Synthetic training image pool
     Mining sequence
Output:
     Joint detector optimized for multi-target tracking
 1:  run baseline detector on mining sequence
 2:  run target tracker on mining sequence, based on the detection
     result from baseline detector
 3:  repeat
 4:      collect missing recall from the tracking result
 5:      cluster occlusion patterns
 6:      generate training images for mined patterns
 7:      train a joint detector with new training images
 8:      run the joint detector on mining sequence
 9:      run the target tracker on mining sequence
10:  until tracking results converge
```

Joint detector training with mined occlusion patterns (step 7): The single-person component of the joint detector is initialized with the same training images as the baseline detector. For each iteration, we introduce a new double-person component that models the mined occlusion pattern. Joint training is based on the structured SVM formulation from Sec. 2. Learning stops when the tracking performance does not improve further on the mining sequence.

5. Experiments

We evaluate the performance of the proposed joint person detector with learned occlusion patterns and its application to tracking on three publicly available and particularly challenging sequences: PETS S2.L2 and S1.L2 [12], as well as the recent ParkingLot dataset [16]. All of them are captured in a typical surveillance setting. S2.L2 and S1.L2 show a substantial amount of person-person occlusions, in particular. We employ the first half of S2.L2 (frames 1–218) as our only *mining sequence* and use the remaining data for testing. Note that our mining algorithm only extracts occlusion patterns and no additional image information. Also note that we do not mine on any of the other sequences, and that the results on the second PETS sequence (S1.L2) and ParkingLot allow to analyze the generalization performance of our approach to independent sequences.

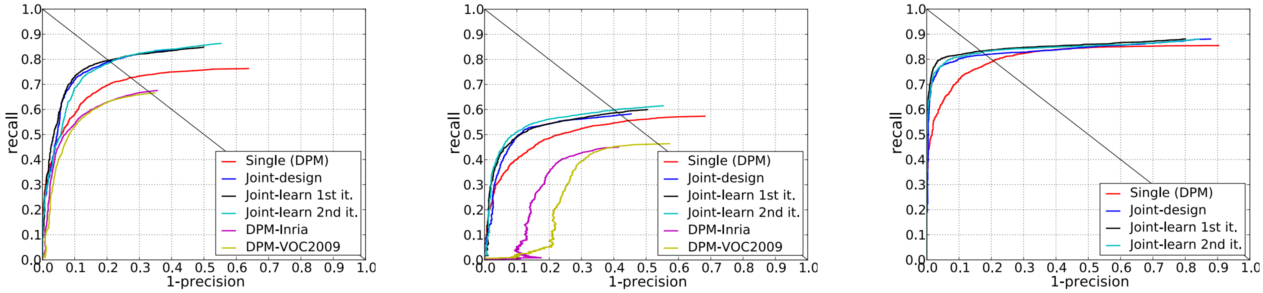
To quantify the tracking performance on the test sequences, we compute recall and precision, as well as the standard CLEAR MOT metrics [5]: Multi-Object Tracking Accuracy *(MOTA)*, which combines false alarms, missed targets and identity switches; and Multi-Object Tracking Precision *(MOTP)*, which measures the misalignment of the predicted track with respect to the ground truth trajectory.

Single-person detector. We begin our analysis with the baseline detector, which is a standard DPM single-person detector [11]. For a fair comparison, we use the same synthetic training images and component initialization as for the joint detector. Note that this already yields a rather strong baseline, with far better performance than DPM-INRIA and DPM-VOC2009 (see Fig. 6). Tracking results

Method	Rcll	Prcsn	MOTA	MOTP
Single (DPM)	60.8	83.8	47.5 %	73.5 %
Joint-Design	65.0	91.7	57.6 %	75.2 %
Joint-Learn 1st	60.6	95.0	56.5 %	75.7 %
Joint-Learn 2nd	64.0	91.7	56.9 %	74.4 %
HOG [2]	51.0	95.5	47.8 %	73.2 %
Particle filter [6]	-	-	50.0 %	51.3 %

Method	Rcll	Prcsn	MOTA	MOTP
Single (DPM)	24.8	90.1	21.8 %	70.6 %
Joint-Design	28.5	86.3	23.0 %	70.8 %
Joint-Learn 1st	28.9	86.2	23.4 %	69.8 %
Joint-Learn 2nd	32.7	86.7	26.8 %	69.3 %
HOG [2]	24.2	83.8	19.1 %	69.6 %

Method	Rcll	Prcsn	MOTA	MOTP
Single (DPM)	90.5	97.7	87.9 %	77.2 %
Joint-Design	91.3	97.5	88.6 %	77.6 %
Joint-Learn 1st	91.0	98.5	89.3 %	77.7 %
Joint-Learn 2nd	91.0	98.0	88.7 %	76.9 %
Part-based [16]	81.7	91.3	79.3 %	74.1 %
GMCP [23]	95.0	94.2	89.1 %	77.5 %



(a) PETS S2.L2 (frames 219–436).

(b) PETS S1.L2.

(c) ParkingLot.

Figure 6. Tracking (*top*) and detection (*bottom*) performance on PETS S2.L2, S1.L2, and ParkingLot: *Single (DPM)*: single-person detector; *Joint-Design*: joint detector with designed occlusion patterns; *Joint-Learn 1st*: joint detector with learned occlusion pattern after the first mining iteration; *Joint-Learn 2nd*: joint detector with learned occlusion pattern after the second mining iteration.

using this baseline detector are also quite competitive and already outperform a state-of-the-art method [2] on S1.L2.

Joint detector with designed occlusion patterns (4.1). Next, we evaluate the performance of our joint detector with manually designed occlusion patterns (see Fig. 6). The joint detector (blue) shows its advantage by outperforming the single-person detector on all sequences. It achieves 10% more recall at high precision for S1.L2 and ParkingLot. For the S2.L2 test sequence, the joint detector outperforms the baseline detector by a large margin from 0.9 precision level. These detection results suggest that the joint detection is much more powerful than the single detector; the designed occlusion patterns correspond to compact appearance and can be detected well.

The performance boost is also reflected in the tracking evaluation. Using the joint detector (Joint-Design) yields a remarkable performance boost on the S2.L2 test sequence (reaching 57.6% MOTA), improving *MOTA* by 10.1% points and *MOTP* by 1.7% points at the same time. It also improves *Recall* by 4.2 and *Precision* by 7.9 compared to the single-person detector (Single DPM). On the S1.L2 and the ParkingLot sequences, the joint detector also outperforms the single-person detector with a significantly higher recall achieved by detecting more occluded targets.

By carefully analyzing and designing the occlusion patterns, we obtain very competitive results on publicly available sequences, both in terms of detection and tracking, which shows the advantage of the proposed joint detector for tracking people in crowded scenes.

Joint detector with learned occlusion patterns (4.2). We report the joint detector performance for one and two mining iterations. As mentioned above, we employ the first half of S2.L2 (frames 1–218) as mining sequence, extracting occlusion patterns, but no further image information.

On the S2.L2 test sequence (frames 219–436), which is more similar to the mining sequence than the other two sequences, our joint detector (black, Joint-Learn 1st, 56,5% MOTA) is nearly on par with the hand-designed patterns after the first iteration, as shown in Fig. 6(a). This is because the most dominant occlusion pattern is captured and learned by the joint detector already. For the second iteration (cyan, Joint-Learn 2nd), we also achieve higher recall on the S2.L2 test sequence, but the precision slightly decreases because the dominant occlusion pattern of the second iteration only contains about 48 missed targets, compared to 5861 ground truth annotations, thus limiting potential performance improvement and introducing potential false positives.

Additionally, we compare our tracking results with [2] and [6] on the S2.L2 sequence, as shown in Tab. 6(a). They report tracking performance for the whole sequence, ours is for the second half of the sequence. After the second iteration of mining, we obtain a tracking performance of 56.9% MOTA, significantly outperforming the other methods[2].

Next, we verify the generalization ability of our algorithm on two more sequences: PETS S1.L2, which is extremely crowded, and the ParkingLot sequence, which contains relatively few occlusions. On PETS S1.L2, the learned joint detector (black) is already slightly better than the Joint-Design detector after the first iteration, as shown in Fig. 6(b). The second iteration (cyan) once again improves the performance, both in terms of recall and precision. The tracking result is also very promising. Directly mining

[2] Note that, for the first half of the S2.L2 sequence where we mine the occlusion patterns, we even achieve 63.8% MOTA.

occlusion patterns from the tracker improves the accuracy (MOTA) with each iteration (from 21.8% over 23.4% to 26.8% MOTA). Note that, similar to the findings above, the tracking performance reaches competitive levels after only one iteration, when compared to manually designed occlusion patterns. This is remarkable, since for the S1.L2 sequence many targets are occluded for long time periods. Our mining algorithm is able to fully recover twice as many trajectories and increase the recall by over 8%.

The ParkingLot sequence contains relatively few occlusions, such that our mining algorithm cannot fully unfold its benefits, and does not improve further after the first iteration. As shown in Fig. 6(c), the joint detector from the first iteration outperforms all other detectors, and reaches similar performance for tracking (Tab. 6(c)). We also compare our method to two other state-of-the-art multi-person trackers [16, 23]. To enable a fair comparison, we compute the performance of [23] using the authors' original results and ground truth. Our joint detector yields state-of-the-art results, both w.r.t. MOTA and MOTP.

Discussion. We observed that the proposed approach converges already after two iterations; further iterations do not lead to an additional performance boost for detection or tracking. We attribute this mainly to the limited size of the mining sequence and its limited diversity. Still, the experimental results on the S1.L2 and ParkingLot sequences suggest that our detector learning algorithm is not limited to particular occlusion patterns or crowd densities. For more complex scenes such as PETS S1.L2, the performance could be further improved by utilizing a more crowded mining sequence. To that end, we plan to build a large dataset of crowded street scenes to mine a more diverse set of occlusion patterns. Another promising future extension would be to learn a joint upper-body detector on extremely dense scenes, yielding specialized upper-body occlusion patterns.

6. Conclusion

We presented a novel joint person detector specifically designed to address common failure cases during tracking in crowded street scenes due to long-term inter-object occlusions. First, we showed that the most common occlusion patterns can be designed manually, and second, we proposed to learn reoccurring constellations with the tracker in the loop. The presented method achieves competitive performance, surpassing state-of-the-art results on several particularly challenging datasets. We make the code of our approach and pre-trained models publicly available[3].

References

[1] M. Andriluka, S. Roth, and B. Schiele. People-tracking-by-detection and people-detection-by-tracking. *CVPR 2008*.

[2] A. Andriyenko and K. Schindler. Multi-target tracking by continuous energy minimization. *CVPR 2011*.

[3] A. Andriyenko, K. Schindler, and S. Roth. Discrete-continuous optimization for multi-target tracking. *CVPR 2012*.

[4] O. Barinova, V. Lempitsky, and P. Kohli. On detection of multiple object instances using hough transform. *CVPR 2010*.

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

[6] M. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, and L. Van Gool. Online multiperson tracking-by-detection from a single uncalibrated camera. *PAMI*, 33(9):1820–1833, 2011.

[7] C. Desai and D. Ramanan. Detecting actions, poses, and objects with relational phraselets. *ECCV 2012*.

[8] P. Dollár, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: An evaluation of the state of the art. *PAMI*, 34(4):743–761, 2012.

[9] M. Enzweiler, A. Eigenstetter, B. Schiele, and D. M. Gavrila. Multi-cue pedestrian classification with partial occlusion handling. *CVPR 2010*.

[10] A. Farhadi and M. Sadeghi. Recognition using visual phrases. *CVPR 2011*.

[11] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *PAMI*, 32(9):1627–1645, 2010.

[12] J. M. Ferryman and A. Shahrokni. PETS2009: Dataset and challenge. *Winter-PETS*, 2009.

[13] B. Leibe, K. Schindler, and L. Van Gool. Coupled detection and trajectory estimation for multi-object tracking. *ICCV 2007*.

[14] B. Leibe, E. Seemann, and B. Schiele. Pedestrian detection in crowded scenes. *CVPR 2005*.

[15] B. Pepik, M. Stark, P. Gehler, and B. Schiele. Teaching 3d geometry to deformable part models. *CVPR 2012*.

[16] G. Shu, A. Dehghan, O. Oreifej, E. Hand, and M. Shah. Part-based multiple-person tracking with partial occlusion handling. *CVPR 2012*.

[17] S. Tang, M. Andriluka, and B. Schiele. Detection and tracking of occluded people. *BMVC 2012*.

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

[19] C. Wojek, S. Walk, S. Roth, and B. Schiele. Monocular 3D scene understanding with explicit occlusion reasoning. *CVPR 2011*.

[20] B. Wu and R. Nevatia. Tracking of multiple, partially occluded humans based on static body part detection. *CVPR 2006*.

[21] Z. Wu, A. Thangali, S. Sclaroff, and M. Betke. Coupling detection and data association for multiple object tracking. *CVPR 2012*.

[22] B. Yang and R. Nevatia. Online learned discriminative part-based appearance models for multi-human tracking. *ECCV 2012*.

[23] A. R. Zamir, A. Dehghan, and M. Shah. GMCP-Tracker: Global multi-object tracking using generalized minimum clique graphs. *ECCV 2012*.

[3] www.d2.mpi-inf.mpg.de/tang_iccv13