Learning People Detectors for Tracking in Crowded Scenes

Siyu Tang[1] Mykhaylo Andriluka[1] Anton Milan[4]
Konrad Schindler[3] Stefan Roth[2] Bernt Schiele[1]

[1]MPI Informatics, Saarbrücken, Germany [2]Department of Computer Science, TU Darmstadt
[3]Photogrammetry and Remote Sensing Group, ETH Zürich [4]The University of Adelaide, Australia

1. Introduction

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. We argue that for best performance one should explicitly train people detectors on failure cases of the tracker output instead. To that end, we first propose a joint people detector that combines a state-of-the-art single person detector with a detector for pairs of people, explicitly exploiting common patterns of person-person occlusions across multiple viewpoints that are frequent failure cases for tracking in crowded scenes (Sec. 2). To explicitly address remaining failure cases of the tracker we explore two methods. First, we analyze typical failures of trackers and train a detector using synthetic images explicitly on these cases. And second, we train the detector with the people tracker in the loop, focusing on the most common tracker failures (Sec. 3). 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.

2. Joint People Detection

The key rationale of our approach is that person-person occlusions, which are the dominant occlusion types in crowded scenes, exhibit regularities that can be exploited.

We propose to train a joint person detector using structural loss-based training approach, based on structured SVMs. We go beyond previous work on joint people detection in several ways: *(1)* Our related approach [5] 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.

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

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 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 to force the joint detector to learn the fundamental appearance difference between a single person and a person/person pair.

Experimental results. We show the benefit of the proposed structured training for joint people detection in Fig. 3(a). At 95% precision it outperforms [5] by 20.5% recall.

3. Learning People Detectors for Tracking

As our second contribution, we propose and evaluate two alternative strategies for the discovery of useful multi-view occlusion patterns.

***(a)* Designing occlusion patterns.** We manually define relevant occlusion patterns using a discretization of the mutual arrangement of people.

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

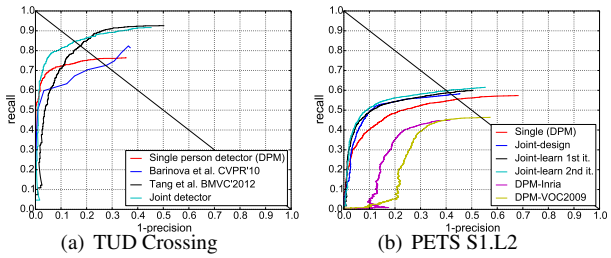
***(b)* Mining occlusion patterns from tracking.** We train the detector with the tracker in the loop, by automatically identifying occlusion patterns based on regularities in the failure modes of the tracker. The approach is summarized in Alg. 1. For our study, we use the first half (frames 1–218) of the challenging PETS S2.L2 dataset [1] as our mining sequence, and we employ a recent multi-target tracker based on energy minimization [2]. The ouput of the method is a joint detector that is tailored to detect occlusion patterns that are most relevant for multi-target tracking.

In a nutshell, we employ tracking performance evaluation (step 4), occlusion pattern mining (step 5), synthetic image generation (step 6), and detector training (step 7) jointly to optimize the detector for tracking multiple targets.

The majority of missed targets extracted from the tracking result (step 4) are occluders and/or occludees for a pair of persons (Fig. 2(b)), or within a group of multiple people (Fig. 2(c)). We determine the problematic occlusion patterns and cluster them in terms of the relative position of the occluder/occludee pair. Fig. 2(d) and 2(e) show the dominant occlusion pattern of the first and second mining iteration.

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

Figure 2. Missed targets from PETS S2.L2 mining sequence and mined occlusion patterns.



(a) TUD Crossing (b) PETS S1.L2

Figure 3. *(a)* Detection performance comparison with [5] on TUD-Crossing sequence; *(b)* Detection performance on PETS S1.L2 sequence.

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 %
Milan et al. [2] (HOG)	24.2	83.8	19.1 %	69.6 %

Table 1. Tracking performance on PETS S1.L2

4. Experiments and Conclusion

We evaluate the proposed joint person detector and its application to tracking on three challenging sequences: PETS S2.L2, S1.L2 [1], and ParkingLot [3]. By analyzing and mining occlusion patterns, we obtain very competitive detection results both in terms of recall and precision, as shown in Fig. 3(b). Furthermore, directly mining occlusion patterns from the tracker improves the tracking accuracy (MOTA) with each iteration (from 21.8% over 23.4% to 26.8% MOTA) (Tab. 1), which shows the advantage of the proposed joint detector for tracking people in crowded scenes. Please refer to our paper for further details [4] [1].

We presented a joint person detector specifically designed to address common failure cases during tracking and a detector learning approach that explicitly optimized for the tracking task. The presented method surpasses state-of-the-art results on several particularly challenging datasets.

References

[1] J. Ferryman and A. Shahrokni. PETS2009: Dataset and challenge. *PETS Workshop*, 2009.

[2] A. Milan, S. Roth, and K. Schindler. Continuous energy minimization for multitarget tracking. *PAMI*, 36(1):58–72, 2014.

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

[4] S. Tang, M. Andriluka, A. Milan, K. Schindler, S. Roth, and B. Schiele. Learning people detectors for tracking in crowded scenes. *ICCV*, 2013.

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

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