Person Tracking-by-Detection with Efficient Selection of Part-Detectors

Arne Schumann, Martin Bäuml, Rainer Stiefelhagen
Karlsruhe Institute of Technology, Institute for Anthropomatics, 76131 Karlsruhe
{arne.schumann, baeuml, rainer.stiefelhagen}@kit.edu

Abstract

In this paper we introduce a new person tracking-by-detection approach based on a particle filter. We leverage detection and appearance cues and apply explicit occlusion reasoning. The approach samples efficiently from a large set of available person part-detectors in order to increase runtime performance while retaining accuracy. The tracking approach is evaluated and compared to the state of the art on the CAVIAR surveillance dataset as well as on a multimedia dataset consisting of six episodes of the TV series The Big Bang Theory. The results demonstrate the versatility of the approach on very different types of data and its robustness to camera movement and non-pedestrian body poses.

1. Introduction

Person tracking has gained a lot of interest over the last years [1, 6, 4, 10, 11]. The focus usually lies on *pedestrian* tracking, for example in the context of safety & security within camera networks or in cars to increase pedestrian safety. In such scenarios, people are assumed to be in an upright pose, which simplifies both detection and tracking. On the other hand, there are vast amounts of data which do not fulfill these conditions, for example multimedia data such as movies and TV series, as well as personal videos or videos on social media sites. In such data, non-upright poses are much more prevalent and occlusions are very common, for example in close-up shots where only the upper body of a person is visible.

In this paper, we present a part-based tracking approach that works equally well in different scenarios. We use the poselet detector [3] as our underlying detector due to its flexibility and robustness over a wide range of human poses. However, one of the main problems of the poselet detector is its high computational demand. We therefore propose a dynamic part selection method with the goal of significantly speeding up the detection procedure by reducing the number of poselets which have to be evaluated each frame.

We evaluate our approach on two very different data sets: 1. The CAVIAR dataset, which consists of typical surveillance-type data, and 2. the first 6 episodes of the TV series *The Big Bang Theory*. We will show in our experiments that the proposed tracking approach is suited for both of these very different scenarios. Further, we will show significant speed-ups with our proposed dynamic part selection technique, while maintaining a strong level of tracking accuracy. An exemplary tracking result is depicted in Fig. 1.

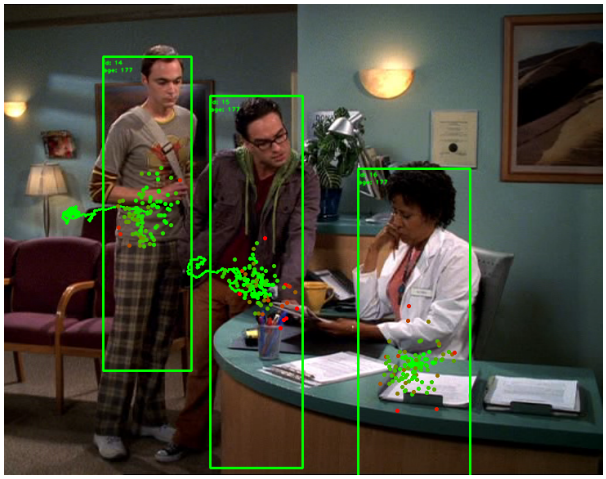

Figure 1. Tracks on *The Big Bang Theory*. The part-based approach can deal with partly occluded persons and persons in non-pedestrian poses (*e.g.*, sitting).

The paper is structured as follows: We first give an overview over related work in Sec. 1.1, and briefly summarize the underlying poselet detector in Sec. 2. In Sec. 3, we describe the proposed tracking approach and extend it by a dynamic way to select detectors in Sec.4. Finally, we present experimental results in Sec. 5.

1.1. Related Work

One key component for person tracking is to reliably locate persons in an image, *i.e. person detection*. The state-of-the art person tracking methods all rely on an underlying person detector not only to initialize tracks, but also to continually evaluate the presence of a person while tracking, which coined the term tracking-by-detection [1]. Different

person detectors have been proposed to be used as underlying detector, for example based on pictorial structures with discriminative part detectors [1], Edgelet-based part detectors [6], the Implicit Shape Model [8, 4] or a Histograms of Oriented Gradients (HOG) detector [4]. We explore the poselet detector [3] as underlying detector, which can by construction deal with a wide range of poses. Unlike previous work, we explicitly take into account the computational demand of the detector and propose a dynamic subsampling technique of the parts depending on the track history (see Sec. 4.1) as well as for track initialization (see Sec. 4.2).

In order to connect detections to tracks, many recent approaches first detect all possible person locations globally in all frames, and then associate them first to short and reliable tracklets, then to longer tracks [6, 10, 11]. Such association-based tracking turns out to be robust against occlusions and mismatches, however, it requires knowledge of all frames in advance and therefore is only suitable in offline settings. This is a strong requirement which cannot always be met, for example for real-time tracking in a camera network. In order to perform online-tracking, a popular choice is to employ a particle filter [4]. Since we do not want to restrict ourselves to offline settings, we employ a particle filter as well in this paper. However, in contrast to [4], we not only use it for the purpose of tracking, but also as a means to restrict the number of required evaluations of the detector in order to improve runtime (see Sec. 3.1).

2. Person Detection

For person detection, we build on the part-based poselet approach by Bourdev *et al*. [3]. A poselet part detector is a HOG feature-based linear support vector machine classifier. Training data for a specific part is selected by the actual underlying 3D-pose of the person in the selected partial view of the training image. As such, parts which might be visually similar, but do not stem from the same pose, will not end up in the same part classifier. Part detectors can actually be for very specific poses, *e.g*. a "right arm crossing a torso". Each poselet detection casts a vote towards the full bounding box of the person. These votes are clustered and confidence-thresholded before being accepted as final person detections. We employ a set of more than 1000 poselet part detectors[1]. The large number of body part detectors results in a good robustness to partial occlusions, pose and orientation variation, but of course results in a high computational cost which is roughly linear to the number of parts.

3. Tracking Approach

In our tracking approach we apply a particle filter in combination with poselet detector as follows. For a new track (see Sec. 3.4), a random set of particles is initialized with uniform weights. In each time step, particles are propagated through a system model into a new state. Each particle's state is evaluated by an observation model which updates the particle weights accordingly. For the next timestep a new set of particles is sampled from the old set according to their weights. For further details on particle filtering please refer to [7].

State We model a track i at time t with a three-dimensional state, containing the location within the image as well as a scale value that represents the size of the track:

$$state_t^i = (x_t^i, y_t^i, s_t^i)^\top. \tag{1}$$

The scale value is related to the size of a person detection through $s_t^i = w_t^i / w_{base}$ with the width of the track's person hypothesis w_t^i and the base width of the person detector w_{base}.

Propagation A simple noise-based motion model is used for particle propagation:

$$(x_t^i, y_t^i) = (x_{t-1}^i + \varepsilon_{loc}, y_{t-1}^i + \varepsilon_{loc}) \tag{2}$$

$$s_t^i = s_{t-1}^i + \varepsilon_{scale} \tag{3}$$

with noise terms ε_{loc} and ε_{scale} which are randomly drawn from zero mean Gaussian distributions $\mathcal{N}_{loc}(0, \sigma_{loc}^2)$, $\mathcal{N}_{scale}(0, \sigma_{scale}^2)$. While σ_{scale} is a fixed parameter over the course of a sequence, the standard deviation for the location within the image is individual to each track and computed as $\sigma_{loc} = \sigma_{loc\_base} * f(s_{t-1}^i)$. This follows the intuition that larger person detections are closer to the camera and their movements have a greater effect in terms of image coordinates. We explicitly do not use a velocity-based motion model results in order to increase robustness in cases where a tracked object abruptly changes direction. Note that these changes in track direction can also be caused by camera movement, when the camera is not fixed as is often the case in multimedia data.

We use two observation models for scoring the particles in order to include both evidence from the detector and appearance cues into each particle's weight.

3.1. Detector Observation Model

The detector observation model updates the weight w_j^i of the jth particle of track i based on poselet detection scores. Given a particle state and a poselet detector one can compute the expected location of a potential positive poselet response. A poselet detection $[x_{pl}, y_{pl}, w_{pl}, h_{pl}]$ with votes $[v_x, v_y, v_w, v_h]$ and scale $s_{pl} = w_{pl}$ votes for a person detection $[x_{ps}, y_{ps}, w_{ps}, h_{ps}] = [x_{pl} + v_x s_{pl}, y_{pl} + v_y s_{pl}, v_w s_{pl}, v_h s_{pl}]$. In reverse – starting from the person

[1] We use the pre-trained models from
`http://www.cs.berkeley.edu/~lbourdev/poselets/`

detection – we can compute the location of an expected corresponding poselet detection as:

$$s_{pl} = w_{ps}/v_w \tag{4}$$

$$[x_{pl}, y_{pl}, w_{pl}, h_{pl}] = [x_{ps} - v_x s_{pl}, y_{ps} - v_y s_{pl}, w_{ps}/v_w, h_{ps}/v_h]\,. \tag{5}$$

In this way, we can determine the position of a potential contributing part-detection, which is faster than scanning a bigger area around the current track position with all part detectors and does not sacrifice much accuracy. The poselet detector is then evaluated at that location and the resulting poselet scores d_k contribute to the particle weight:

$$wd_j^i = \sum_{k=0}^{\#detectors} f(d_k), \quad f(d_k) = \begin{cases} d_k & d_k > 0 \\ 0 & \text{else} \end{cases} \tag{6}$$

While in many cases parts of the features can be shared among particles and need not be recomputed, the large number of detectors makes this particle scoring computationally expensive. On the other hand, leaving out the wrong part detectors may lead to low particle weights and inadvertent termination of tracks. We describe an approach to handle this situation in Sec. 4.

3.2. Appearance Observation Model

In addition to the detector observation model we use an color-based appearance model. Appearance information is complementary to the detections and can keep a track alive if for some reason the detector fails but the person is still visible. It also helps to terminate tracks more quickly if a person disappears but background clutter causes part detectors to still support some particles.

We model the appearance of a track through an RGB color histogram. Histograms are computed for each particle H_j^i and matched against that of the track H^i to compute the particle weight:

$$wa_j^i = 1 - D_B(H^i, H_j^i). \tag{7}$$

D_B is the Bhattacharyya distance. Particle histograms are computed over an area $[x, y, w, h] = [0.33, 0.2, 0.33, 0.3]$ of each particle's bounding box in order to primarily capture upper body clothing colors.

The final weight of a particle j of track i is then determined as a weighted sum:

$$w_j^i = wd_j^i + \alpha * (1 + \beta) * wa_j^i \tag{8}$$

$$\beta = \max_{t_k \neq t_i} \frac{area_{t_i} \cap area_{t_k}}{area_{t_i} \cup area_{t_k}}. \tag{9}$$

with a fixed weight α that determines the influence of appearance on the track. The appearance is also weighted by the maximum degree that track t_i overlaps with other tracks. This ensures that the influence of appearance on particle evaluation increases when tracks start to overlap and their detection confidence becomes a less reliable information.

Taking into account the observation models' scores, a final track hypothesis for the current timestep is computed as the weighted average of all particle states. The track's appearance histogram is then updated using the new track hypothesis.

3.3. Occlusion Handling

Occlusions can be detected by determining when two tracks t_1 and t_2 start to overlap. Once such a situation occurs it must be determined which of the overlapping tracks is the one that gets occluded. We use two cues to make this determination - the difference in scale and changes in appearance. We compute an occlusion term

$$occ(t_1, t_2) = \frac{d_{app}(t_1)}{d_{app}(t_2)} \cdot \gamma \frac{s_{t_2}}{s_{t_1}}. \tag{10}$$

with the difference $d_{app}(t_1) = |app(t_1) - app_{avg}(t_1)|$ between the current appearance of a track $app(t_1)$ and its average appearance over the recent past $app_{avg}(t_1)$ and the scale s_{t_1} of a track. A value $occ(t_1, t_2) \geq 1$ hints towards either the scale of track t_1 being smaller than that of t_2, the appearance change in track t_1 being more significant or both. Consequently we assume track t_1 to be occluded for $occ(t_1, t_2) \geq 1$ and t_2 otherwise. A weight γ is used to bias this decision towards the scale or appearance cue. In our implementation we set γ to 1.

Once a track is considered occluded, its motion model is switched from a noise-based propagation to velocity-based propagation. We use velocity in cases of occlusion in order to increase the chance of picking the track up correctly once it reappears. Unless the occluding track has similar size and appearance, the occluded track's particles receive only small weights. A track becomes un-occluded once its bounding box no longer overlaps with the occluding track and its particle weights recover. Tracks that remain occluded for more than three seconds are terminated. While a track is occluded, its appearance does not get updated.

3.4. Track Initialization and Termination

Without assuming any prior knowledge about the scene and especially in the case of non-stationary cameras, new persons can appear anywhere within the image. This is in contrast to other approaches where entry- and exit-zones of the scene are explicitly modelled (*e.g.*, [6]). Accordingly, we scan the entire image in regular time intervals for new tracks. New detections d_n are matched against existing tracks and disregarded if any of the following is true:

$$\max_{track} \frac{area_{t_i} \cap area_{d_n}}{area_{t_i} \cup area_{d_n}} \geq thr_{match} \text{ or} \quad (11)$$

$$\exists t_i : area_{t_i} \cup area_{d_n} == area_{t_i}. \quad (12)$$

The second case is based on the observation that new persons cannot first appear in full occlusion.

If a remaining detection has a high score, a track is created from it. If its score is too low but more detections are found in the same area over multiple consecutive timesteps, their scores are accumulated. When this accumulated score becomes high enough, a new track is created as well. In our implementation we require detections accumulate a score of at least thr_{score} over a maximum of five frames. The score threshold depends on the detection scale s_d and is set to $thr_{score} = 50d_s$. Note that poselet person detection scores are computed as the sum of all contributing poselet scores. Therefore high score values are common.

Tracks are terminated if they either leave the image area or do not get sufficient support from their observation models over a period of one second.

4. Dynamic Part Sub-Selection

We propose a dynamic approach to reduce the number of part-detectors required in the observation model and during track initialization.

4.1. Track-specific Part Selection

In order to speed up the particle evaluation in the observation model while retaining the important detectors and thus the tracking accuracy, we keep an individual set of detectors with each track. The set is divided into two sub-sets, a *core set* and a *dynamic set*.

The core set is meant to contain detectors that are expected to have high relevance for the track. The detectors in the dynamic set are randomly selected at each timestep from those that are not in the current core set. That way we ensure that each available detector is used every once in a while.

For each detector d_k, we keep the number of times c_k^u it was used. A second value c_k^h denotes the "usefulness" of the detector for the current track. This usefulness is determined as the sum of scores of all detections that supported particles of the track. The ratio $r_k^h = c_k^h / c_k^u$ is the relative usefulness of the detector. Each time the observation model evaluates a track's particles, the core set is first filled with those detectors that have the highest r_k^h. The dynamic set is then filled with a random selection of the remaining detectors. r_k^h is computed over a recent history of the track to prevent detectors that were initially supporting the track strongly but do not any more from remaining in the core set for too long. The number of frames in the history is set to correspond to two seconds of video.

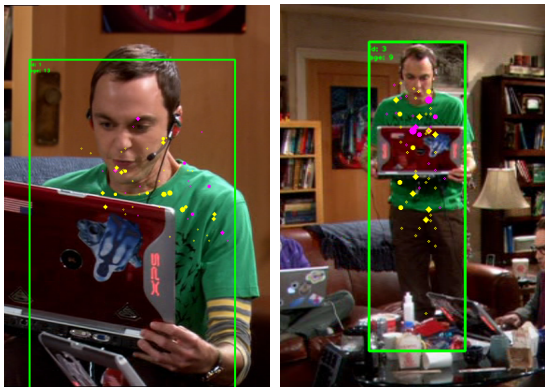

Figure 2. The detectors of the core subset (hits in yellow) provide stronger hits than those of the dynamic subset (hits in magenta). Higher detection scores correspond to a larger radius.

Using this dynamic detector management runtime performance can be increased while the benefits of a large number of part detectors remain intact. This approach works well together with the pose-based nature of the poselet detectors. A standing person may be well detected by a pedestrian poselet and as the person sits down it may go through a range of poselets that correspond to the current body pose. In this way, the detector core set gathers knowledge specific to the track it belongs to represented by the types of poselets it contains.

Fig. 2 depicts detector hits from the core and the dynamic subset. In a sitting and occluded position, the core subset mostly contains detectors that focus on the face but for a standing person a much wider range of detectors is in the core set. Note that this approach to select from a larger detector pool could also be used in combination with holistic person detectors that may be trained for different angles or viewpoints.

4.2. Part Selection for Initialization

The track initialization step can also be sped up significantly by using a subset of detectors. Starting from a random selection of detectors, the set is regularly repopulated by those detectors that performed best in the past. Detectors are considered to perform well, if they contributed to detections that lead to new tracks being created or caused detections that matched existing tracks. The set of part detectors that lead to a new track initialization is used as an initialization of the new track's core subset (see Sec. 4.1). While the track detector subsets contain information specific to each track, the initialization detector subset gathers global information about the scene. In this way we are able to leverage track specific information and scene specific information using the same approach without having to make any prior assumptions.

5. Experimental Results

5.1. Datasets

We evaluate our approach on two datasets. The CAVIAR dataset[2] consists of surveillance videos captured by a static camera in a corridor of a shopping mall. This dataset has a very low resolution (384×288) and contains several scenes with occlusions. We created a modified set of groundtruth annotations where we fixed some annotation oddities such as person bounding boxes that only surround the foot of a person if the rest is occluded. We will make this corrected groundtruth available online[3]. In order to provide comparability with previous work, we report results on both the original and the corrected groundtruth.

We further evaluate on a dataset consisting of the first 6 episodes of the TV show *The Big Bang Theory* (BBT). This multimedia dataset has a higher resolution of 1024×576, moving cameras, many different angles, non-pedestrian-like poses (*e.g.*, sitting), inter-object occlusions and persons are frequently only partially visible because they are cut off by the camera. Each episode is about 20 minutes in length and contains around $30,000$ frames. We labelled every 10th frame as groundtruth.

5.2. Evaluation Methodology

We use Multiple Object Tracking Accuracy (MOTA) [2] as evaluation metric:

$$MOTA = 1 - \frac{\sum_t MISS_t + FP_t + MM_t}{\sum_t GT_T} \qquad (13)$$

where $MISS_t$ is the number track misses, FP_t the number of false positive tracks, MM_t the number of track switches (mismatches) and GT_t the number of groundtruth tracks at time t. For the CAVIAR dataset we report the average MOTA over all sequences and for the BBT dataset the average MOTA over all cuts of an episode.

5.3. Results

For our experiments on both datasets we chose a fixed number of 100 particles. We set the base value for propagation in location σ_{loc_base} to 1. The larger person sizes in the BBT dataset adjust this value automatically. Unless otherwise specified we use all available part detectors in our experiments.

CAVIAR On the CAVIAR dataset our approach achieves a MOTA of 60%. On the modified groundtruth the number of false positive tracks and track misses decrease slightly. In Table 1 we compare our performance to one of the best results on the dataset by Huang *et al*. [6] who achieved 80% accuracy. Note that their approach is not directly comparable to ours because it relies on automatically determined scene knowledge such as a groundplane and information about entry-exit zones. We chose not to use such information because our approach is intended to work even on data with moving cameras where groundplanes and entry-exit zones cannot be automatically determined. Another difference that factors into the comparison of the two results is that while Huang *et al*. optimize their trajectories over the entire length of a scene, our approach is an online approach and does not require global knowledge of all frames.

	Huang *et al*.[6]	Ours	Ours MGT
MOTA	80.00%	60.26%	61.32%
Misses	20.00%	34.22%	32.58%
FP rate	0.025	0.198	0.192

Table 1. Results on the CAVIAR dataset on original and modified ground truth (MGT) compared to a state of the art approach that uses additional scene information.

Lastly, experiments show that we loose most accuracy in those sequences where many very small persons appear and stay in the background. In these cases, the part-based person detector generates few low-score detections that do not allow for reliable track initialization.

The Big Bang Theory dataset On the *Big Bang Theory* dataset we compare our approach to our own implementation of an association-based tracker [9] which is similar in approach to [6]. Our online approach achieves a higher accuracy due to a lower number of mismatches which can be explained by our use of an appearance model. Results are shown in Table 2. Both tracking approaches used the same set of detections.

Dynamic Part Selection In order to evaluate the effect of the detector part selection, we run our approach with different subset sizes on both datasets. The results are depicted in Fig. 3. The subset size in the plots refers to the entire detector subset which we split evenly into core and dynamic set. When using detector subsets only for the track observation model, we observe a moderate overall speedup at a small cost in accuracy. The speedup gained during the particle weighting gets masked by the more cost intensive person detection during the track initialization step which takes about 10 seconds per image. The tracking time is the same on both datasets, because input images are scaled to a uniform size before tracking.

Using all available detectors, the track weighting initially takes approximately 410ms per track. Using a subset of 600 detectors, scoring time is reduced to 230ms and with 200 detectors the required time lies below 100ms. This corresponds to a speedup of 4.2 for the track scoring step. The loss in MOTA is caused by a larger number of misses which

[2] http://homepages.inf.ed.ac.uk/rbf/CAVIARDATA1
[3] http://cvhci.anthropomatik.kit.edu/projects/pri

Episode	Association-based tracker [9]				Ours			
	MOTA	Misses	False Positives	Mismatches	MOTA	Misses	False Positives	Mismatches
S01E01	71.72%	22.40%	2.53%	3.35%	74.34%	20.67%	4.25%	0.75%
S01E02	63.89%	31.58%	1.76%	2.76%	66.75%	28.78%	3.90%	0.57%
S01E03	66.78%	25.23%	4.75%	3.23%	67.57%	24.09%	7.57%	0.77%
S01E04	62.97%	29.49%	3.99%	3.54%	67.05%	24.95%	7.11%	0.90%
S01E05	60.25%	30.10%	6.06%	3.59%	62.79%	26.55%	9.96%	0.70%
S01E06	56.46%	30.95%	7.29%	5.30%	57.43%	28.21%	12.93%	1.42%
Mean	63.68%				65.99%			

Table 2. Results on the *Big Bang Theory* dataset compared to an association-based tracker. Both approaches use the same set of detections.

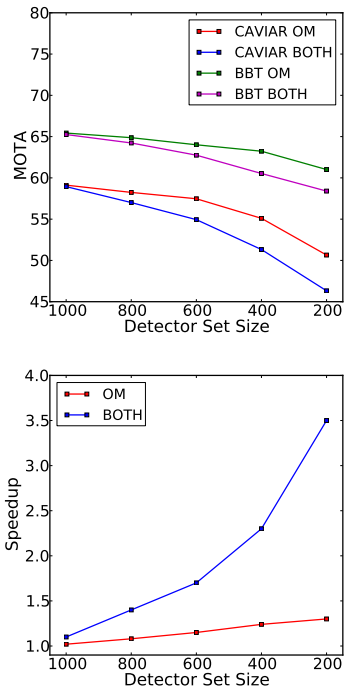

Figure 3. Decrease of accuracy with smaller detector sets (top) and the corresponding increase in speedup (bottom). Detector subsets are either used only for the observation model (OM) or for both observation model and track initialization (BOTH).

happen when tracks are terminated earlier than they should be because the smaller number of detectors does no longer produce sufficient detections. The decrease in MOTA with fewer detectors is more noticeable in the CAVIAR dataset due to the lower level of detail in the images.

When using subsets for both observation model and track initialization, the effect on runtime becomes much more visible. Starting from an average overall time per frame of 11s, the performance increases to 6.3s at subsets of 600 detectors and finally 3.2s at subsets of size 200. The speedup is not quite inversely proportional to the number of detectors, because the costly HOG feature pyramid computation cannot be sped up by lower numbers of detectors.

	OccApp	NoApp	NoOcc	None
CAVIAR	60.26%	59.34%	53.28%	51.12%
BBT	65.99%	64.85%	56.87%	53.47%

Table 3. Average MOTA on both datasets using the full approach (OccApp), leaving out the appearance model (NoApp), leaving out the occlusion handling (NoOcc) or neither occlusion nor appearance information (None).

Again, we observe a comparatively low drop in MOTA that is more noticeable on the lower resolution CAVIAR dataset. The additional accuracy loss can be explained by additional misses resulting from tracks being initialized later when using smaller sets of detectors. However, on the BBT dataset we also observe that the number of false positives is reduced with fewer detectors, because some erroneous tracks do not get initialized.

Occlusion and Appearance To study the influence of the appearance model and occlusion reasoning on the tracking accuracy, we performed experiments with and without each of those components. Results can be seen in Table 3. Leaving out any occlusion reasoning results in a steep drop in MOTA for both datasets. The main cause for this is an increase in track misses and false positive tracks which can both be explained by an increase in cases where two tracks start following the same person after an unhandled occlusion. In some cases the appearance model will be able to detect such cases and terminate the track whose appearance does not match the wrongly tracked person.

Leaving out the appearance model does not reduce tracking accuracy much. This is largely due to the way track switches are counted when computing the MOTA. A track switch only counts as a single error in the frame when it happens. Cases when two tracks attempt to follow the same person due to a lack of appearance information will still be caught by the occlusion reasoning which is why the number of misses and false positives does not change significantly.

Finally, leaving out occlusion reasoning and the appearance model results in the lowest MOTA, because there are no mechanisms in place to prevent track switches or multi-

	Poselet Detector	Felzenszwalb *et al*. Detector
CAVIAR	60.26%	56.86%
BBT	65.99%	42.78%

Table 4. Tracking accuracy (MOTA) of the proposed approach using the poselet person detector compared to using the person detector from [5].

ple tracks following the same person.

Person Detector In order to validate our choice of person detector, we modified the proposed tracking approach to support another state-of-the-art object detector by Felzenszwalb *et al*. [5]. Results are shown in Table 4. Due to the much smaller number of parts in this detector, we used the full set of part detectors in the observation model of the modified approach. Particles are scored based on their distance to the closest precomputed person detection and its detection score. Occlusion reasoning and appearance model as well as all other aspects of the approach remain the same.

Similar to the poselet detector, the detector from [5] has problems detecting the small persons far in the background of the CAVIAR dataset. However, persons which are partially occluded or cut off by the lower border of the image are less frequently detected by [5] than by the poselet detector. This leads to tracks terminating earlier or more frequently, *i.e*. a higher rate of track misses and correspondingly a slightly lower MOTA. On the Big Bang Theory dataset, the difference in tracking accuracy is much more significant. The detector form [5] provides fewer or only weak detections in the many cases where only the upper third of a person is visible or persons are in non-pedestrian poses. These weak detections often do not suffice to initialize or sustain a track. Lowering the corresponding thresholds leads to a large number of false positive tracks on background objects, such as plants or lamps and tracks surviving for a long time without presence of a person.

6. Conclusion

We present a tracking-by-detection approach that combines a particle filter with a part-based person detector. We use both detection and appearance cues for scoring each track and conduct explicit occlusion reasoning. The large number of available part detectors is handled efficiently by a dynamic selection approach which manages a subset of detectors for each track and for track initialization. This approach achieves a significant speed-up while retaining good tracking accuracy. We demonstrate consistently high accuracy on two challenging datasets from different domains. Our approach does not rely on any scene knowledge and deals well with moving cameras and non-pedestrian body poses.

Acknowledgments This work was supported by the German Federal Ministry of Education and Research (BMBF) as part of the MisPel program under grant no. 13N12063; and as part of the Quaero Program, funded by OSEO, French State agency for innovation, The views expressed herein are the authors' responsibility and do not necessarily reflect those of OSEO or BMBF.

References

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

[2] K. Bernardin and R. Stiefelhagen. Evaluating multiple object tracking performance: the clear mot metrics. *EURASIP Journal on Image and Video Processing*, 2008.

[3] L. Bourdev and J. Malik. Poselets: Body part detectors trained using 3d human pose annotations. In *ICCV*, 2009.

[4] M. D. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, and L. Van Gool. Robust tracking-by-detection using a detector confidence particle filter. In *ICCV*, 2009.

[5] P. Felzenszwalb, D. McAllester, and D. Ramanan. A discriminatively trained, multiscale, deformable part model. In *CVPR*, 2008.

[6] C. Huang, B. Wu, and R. Nevatia. Robust object tracking by hierarchical association of detection responses. In *ECCV*, 2008.

[7] M. Isard and A. Blake. Condensation – conditional density propagation for visual tracking. *IJCV*, 1998.

[8] K. Jüngling and M. Arens. Detection and tracking of objects with direct integration of perception and expectation. In *Int. Workshop on Visual Surveillance (VS)*, 2009.

[9] M. Roth, M. Bäuml, R. Nevatia, and R. Stiefelhagen. Robust multi-pose face tracking by multi-stage tracklet association. In *ICPR*, 2012.

[10] B. Yang, C. Huang, and R. Nevatia. Learning affinities and dependencies for multi-target tracking using a CRF model. In *CVPR*, 2011.

[11] B. Yang and R. Nevatia. Online Learned Discriminative Part-Based Appearance Models for Multi-Human Tracking. In *CVPR*, 2012.

Figure 4. Results of our approach on both datasets. Persons in the background of the CAVIAR dataset are too small to be detected by the part-based person detector. The approach deals well with occlusions, camera motion and non-pedestrian poses.