$(MP)^2T$: Multiple People Multiple Parts Tracker

Hamid Izadinia, Imran Saleemi, Wenhui Li, and Mubarak Shah

Computer Vision Lab, University of Central Florida, Orlando, USA
{izadinia,imran,wli,shah}@eecs.ucf.edu

Abstract. We present a method for multi-target tracking that exploits the persistence in detection of object parts. While the implicit representation and detection of body parts have recently been leveraged for improved human detection, ours is the first method that attempts to temporally constrain the location of human body parts with the express purpose of improving pedestrian tracking. We pose the problem of simultaneous tracking of multiple targets and their parts in a network flow optimization framework and show that parts of this network need to be optimized separately and iteratively, due to inter-dependencies of node and edge costs. Given potential detections of humans and their parts separately, an initial set of pedestrian tracklets is first obtained, followed by explicit tracking of human parts as constrained by initial human tracking. A merging step is then performed whereby we attempt to include part-only detections for which the entire human is not observable. This step employs a selective appearance model, which allows us to skip occluded parts in description of positive training samples. The result is high confidence, robust trajectories of pedestrians as well as their parts, which essentially constrain each other's locations and associations, thus improving human tracking and parts detection. We test our algorithm on multiple real datasets and show that the proposed algorithm is an improvement over the state-of-the-art.

Keywords: multi-target, tracking, pedestrians, humans, body part tracking, network flow optimization, k-shortest paths.

1 Introduction

Recent years have seen consistent improvements in the task of automated detection and tracking of pedestrians in visual data. The problem of tracking of multiple targets can be viewed as a combination of two intertwined tasks: inference of presence and locations of targets; and data association to infer the most likely tracks. Research in the analysis of objects in general, and humans in particular, has often attempted to leverage the parts that the objects are composed of. Indeed, the state-of-the-art in human detection has greatly benefited from explicit and implicit detection of body parts [1–4]. We observe however, that while human detection has been employed as a first step in many tracking algorithms with very reasonable outcomes [5–7], the presence and positions of human body parts have not been explicitly leveraged to constrain the data association between successive human observations across frames. This is expected because evaluation of temporal persistence of body parts is a severely under-constrained problem, not in the least due to inherent difficulties in detecting them in the first place.

A. Fitzgibbon et al. (Eds.): ECCV 2012, Part VI, LNCS 7577, pp. 100–114, 2012.

In this paper, we propose a framework which attempts to solve the pedestrian tracking problem by simultaneously constraining the detection, temporal persistence, and appearance similarity of the detected humans, as well as the observable or inferable parts they are composed of. In addition to obtaining high quality, high confidence pedestrian trajectories, a byproduct of our framework is explicit part trajectories which are useful for other commonly encountered surveillance tasks, such as action [8] and event or activity recognition [9].

1.1 Overview of Our Approach

A brief outline of our proposed approach is now described. We begin by applying a state-of-the-art human detector [2] to all frames in the input video sequence, to obtain preliminary detections of pedestrians. We then extract some key features from the observed pedestrians. Using the appearance and motion features computed for detected pedestrians, we obtain high confidence, but potentially short trajectories which link these detections. We use part detectors to obtain human body parts detections, while using the location and scale priors obtained from previously run part-based human detector. The dense output of this sliding window based detector is saved, even for very low confidence part detections, and is used to stitch tracklets where a gap exists due to human mis-detections. Normalized cross-correlation in the RGB space, along with motion features are used to connect body parts detections over frames. A model of spatial relationships between detected parts is learned in an online fashion so as to split pedestrian tracklets at points of low confidence. The pedestrian and parts tracklets are merged into final trajectories based on multiple costs that constrain the trajectories according to appearance features of the pedestrian as well as the parts it is comprised of, while maintaining the spatiotemporal relationships between them.

1.2 Related Work and Our Contributions

We formulate the pedestrian tracking problem as that of simultaneous association between pedestrian as well as parts detections, which are obtained from distinct processes. Although quite a few methods have employed the so called 'tracking-by-detection' frameworks [5–7], we propose to expand the notion of 'detection' to include individual body parts as well.

Explicit tracking of detected body parts in order to aid human tracking has been proposed in [10] and [11], but such methods deviate from ours in many respects. First, in these methods, the part detections are treated as binary outputs in terms of presence or absence. Second, the human tracking is posed as a sequential process, whereby data association with holistic human detection is preferred, and part detections are used only in cases of failure to observe the entire object. Finally, the task of part detection is treated as a general, stand-alone object detection problem, where distinct classifiers are learned for each part, e.g., head or torso.

The proposed method on the other hand, not only implicitly leverages the notion of whole versus part in the process of part detection, but also avoids a sequential processing whereby problems in human detection can negatively affect part detection and vice versa. Specifically, we iteratively commit to pedestrian and parts associations, and

finally perform a joint association step for both. Instead of using body parts as detected during human detection by the deformable parts model (DPM) [2], we only employ the confidence of detection and use position and scale priors for part tracking. The parts as well as pedestrian detections are 'committed' to, only when all the evidence from all the frames has been observed, and the final trajectory is the result of an optimal data association that links these observations.

To summarize, in addition to novel proposals related to representation, selective appearance modeling, and data association costs and affinities, our contribution includes explicit, simultaneous tracking of body parts along with the pedestrians, and demonstration of an iterative optimization algorithm to do so, which is theoretically justified.

2 Tracking Model

The problem of data association across frames is central to the tracking task, and our proposed framework employs the general min-cut/max-flow network paradigm inspired by recent successes of [12] and [13]. We pose the problem of simultaneous tracking of pedestrians and their parts as optimization over a flow network. We explicitly show the inter-dependence of node and edge costs for our problem and propose an iterative network construction and optimization framework. These dependencies arise from the fact that a meaningful tracking of detected parts requires at least temporary data association (tracking) of pedestrian detections.

Formally, let us denote the set of all vertices as $\mathcal{V} = \mathcal{H} \bigcup \mathcal{B}$, where $\mathcal{H} = \{h\}$ is the set of all pedestrian or human observations, and $\mathcal{B} = \{b\}$ is the set of body parts detections. In the most general case, the goal is to perform data association over $\mathcal{V}$, such that in the final solution a part is associated with another, only if the pedestrians they belong to, are also associated, and vice versa. In absence of this constraint, each member in the set of all pedestrians $\{h_t\}$ in a frame t, has a link or edge to each member in the set of pedestrians, $\{h_{t+1}\}$, and similarly for the parts. Such a flow network however, poses two significant problems: (1) a part can be associated to another without association of corresponding pedestrians, and (2) there is a combinatorial increase in the number of possible track hypotheses. Moreover, computation of a measure of deformity in part locations (which is an important cue in part detection and tracking), requires at least a temporary committal to pedestrian association.

We begin the description of our proposed framework by explaining the optimization problem from a generative point of view. We want to maximize the joint posterior probability of pedestrian and part tracks $\mathcal{X}$ given dense image observations $\mathcal{Y}$ generated from multiple processes:

$$\mathcal{X}^* = \underset{\mathcal{X}}{\operatorname{argmax}} P(\mathcal{X}) P(\mathcal{Y}|\mathcal{X}), \tag{1}$$

where, the set of k pedestrian tracks is $\mathcal{X} = \{\theta_i\}_1^k$, a track is a sequence of T data points $\theta_i = \{\mathrm{x}_t^i\}_1^T$, and the vector random variable $\mathrm{x}_t^i = \left(\mathrm{p}_{i,t}^h, \mathrm{v}_{i,t}^h, \mathrm{s}_{i,t}^h, \mathbf{G}_{i,t}\right)$, defines a point on the track, with position, velocity, bounding box size, and a set of N body parts $\mathbf{G}_{i,t} = \{g_{i,t}^m\}_1^N$, such that, $g_{i,t}^m = \left(\mathrm{p}_{i,m,t}^g, \mathrm{s}_{i,m,t}^g\right)$, i.e., $\mathrm{p}_{i,m,t}^g$ is the position of the m^{th} part of the i^{th} person in frame t. The number of body parts N is 8 in all our experiments.

The pedestrian and part properties are distinguished by 'h' and 'g' respectively. The set of observations $\mathcal{Y}$ is formally defined later.

Prior: We can expand the first term in Eq. 1 using the first order Markov as,

$$P(\mathcal{X}) = \prod_{i=1}^{k} P(\theta_i) = \prod_{i=1}^{k} P_{init}\left(\mathrm{x}_1^i\right) \left(\prod_{t=2}^{T-1} P\left(\mathrm{x}_t^i | \mathrm{x}_{t-1}^i\right) \right) P_{term}\left(\mathrm{x}_T^i\right), \tag{2}$$

where P_{init} and P_{term} are probabilities of track initialization and termination respectively. Assuming constant velocity motion for the pedestrian, and constant location for its parts relative to the pedestrian, we can write the following simple dynamic model:

$$P(\mathrm{x}_t | \mathrm{x}_{t-1}) = \underbrace{\mathcal{N}\left(\mathrm{p}_{i,t}^h; \mathrm{p}_{i,t-1}^h + \mathrm{v}_{i,t-1}^h, \Sigma_{i,t-1}^h\right)}_{\substack{\text{constant} \\ \text{pedestrian velocity}}} \cdot \prod_{m=1}^{N} \underbrace{\mathcal{N}\left(\mathrm{p}_{i,m,t}^g; \mathrm{p}_{i,t-1}^h + \mu_{i,m,t-1}^g, \Sigma_{i,m,t-1}^g\right)}_{\substack{\text{constant relative} \\ \text{part location}}}, \tag{3}$$

where Σ^h is the pedestrian state covariance, and (μ^g, Σ^g) are the parameters of a 2D Gaussian distribution of 'relative' locations of parts within a pedestrian bounding box. In order to find the 'relative' location of a part however, we obviously require a reference point, which can be chosen as a specific part (e.g., head: $\mathrm{p}_{i,1,t}^g$), or the person location (i.e., p^h). Either way, in order to maintain a formation or spatial relationship over body parts, the dependency on the entire person's state cannot be removed. In other words, for the learning of 2D distributions of relative part locations, it is necessary to commit to a state x first, i.e., assume an object track. It is therefore not possible to construct the general flow network described earlier since the cost of some edges depends on the solution of part of the network.

Likelihood: We assume that $\mathcal{Y} = \{\mathrm{y}_t^x\}$ is a set of observations or measurements from the image such that $\mathrm{y}_t^x = \left(h_t^x, \{b_t^{x,m}\}_1^N\right)$. The goal of the optimization process is to estimate the set of tracks among a number of hypotheses, that best explains the set of observations. The likelihood term for generating observations given tracks of a pedestrian as well as its parts is now written as:

$$P(\mathcal{Y}|\mathcal{X}) = \prod_{i=1}^{k} P(\mathcal{Y}|\theta_i) = \prod_{i=1}^{k} \prod_{t=1}^{T} P\left(\mathrm{y}_t^x | \mathrm{x}_t^i\right). \tag{4}$$

Since observation y_t^x is generated from multiple processes (pedestrian and part detectors), and because this vector is selectively dependent on the state vector x_t, we employ unweighted opinion pooling of the following conditional probabilities, where subscripts 'δ' and 'κ' imply SVM detector and appearance similarity respectively:

Pedestrian Detector Confidence: This is the output classifier score of an SVM based sliding window detector, and indicates how likely the presence of a human at a certain location and scale is. For detection by testing at every spatial and scale window, this can be a dense output and is conditionally independent of the state, i.e., $P_\delta^h\left(h_t^x | \mathrm{x}_t^i\right) =$

$P^h_\delta\left(h^x_t\right)$. The classifier score is converted to the probabilistic confidence by fitting a sigmoid function.

Pedestrian Similarity: The similarity likelihood $P^h_\kappa\left(h^x_t|\mathrm{x}^i_t\right) = P^h_\kappa\left(h^x_t|\mathrm{p}^h_{i,t}, \mathrm{s}^h_{i,t}\right)$, estimates the level of affinity between appearance features of pedestrian detection h^x_t in frame t, and the pedestrian detection associated with the i^{th} track in frame $t-1$ (leading to current state $\mathrm{p}^h_{i,t}$, $\mathrm{s}^h_{i,t}$, etc.). The similarity is computed as intersection between histograms of optical flow, HOG and color intensities, by exponential scaling.

Part Detector Confidence: The body part detection confidence $P^g_\delta\left(b^{x,m}_t|\mathrm{x}^i_t\right)$ is also independent of the state and is equal to $P^g_\delta\left(b^{x,m}_t\right)$. This is similar to pedestrian detector output, but the output of such a detector can vary depending on the method used. On one hand, a purely appearance based classifier will in general have poor performance due to the large degree of variability in parts appearance, but will output a truly stand-alone detection. On the other hand, methods such as the very successful deformable parts model [2], implicitly prunes the output of stand-alone detectors by imposing a deformity minimization constraint. Such a detector confidence will be much better, but implicitly associates multiple parts into a single person detection. We used dense part detections without regards to deformity.

Part Similarity: This is computed by simple normalized cross correlation, since histograms of features are not meaningful due to the small size of parts. The probability $P^g_\kappa\left(b^{x,m}_t|\mathrm{p}^g_{i,m,t}, \mathrm{s}^g_{i,m,t}\right)$ essentially compares the appearance of part $b^{x,m}_t$ to the i^{th} pedestrian's part number m in frame $t-1$.

We can then write the opinion pooled likelihood as:

$$P\left(\mathrm{y}^x_t|\mathrm{x}^i_t\right) = \underbrace{P^h_\delta\left(h^x_t\right)}_{\substack{\text{detector}\\ \text{confidence}}} \underbrace{P^h_\kappa\left(h^x_t|\mathrm{p}^h_{i,t}, \mathrm{s}^h_{i,t}\right)}_{\substack{\text{pedestrian}\\ \text{similarity}}} \prod_{m=1}^{N} \underbrace{P^g_\delta\left(b^{x,m}_t\right)}_{\substack{\text{detector}\\ \text{confidence}}} \underbrace{P^g_\kappa\left(b^{x,m}_t|\mathrm{p}^g_{i,m,t}, \mathrm{s}^g_{i,m,t}\right)}_{\substack{\text{part}\\ \text{similarity}}}. \tag{5}$$

Iterative Optimization: Due to the strict dependence between the state of parts and the pedestrian they belong to (Eq. 3), a truly simultaneous optimization over both is not straightforward. In other words, a detected part needs to 'commit' to a certain pedestrian in order to evaluate its deformation based likelihood. This however, requires the tracker's committal to a pedestrian detection (in the previous frame).

From a different point of view, it can be seen that if we attempt to write the above described general *maximum a posteriori* (MAP) estimate $\mathcal{X}^*$, as an Integer Linear Program, the motion based cost of an incoming part edge, -log $P\left(\mathrm{p}^g_m|.\right)$, requires a state estimate p^h for a pedestrian. The more general and unconstrained way would be to consider all possible hypotheses for p^h, resulting in a combinatorial explosion of hypotheses. The alternative is to limit the number of edges between parts across frames, by reducing the number of pedestrian track hypotheses, i.e., a temporary committal to specific data association.

We therefore propose an approximate optimization process whereby we iteratively perform data association between pedestrians, followed by association between parts. This process follows three key steps:

Step I: Associate pedestrian detections to obtain several short tracklets.
Step II: Associate part detections by computing the likelihood leveraging pedestrian tracklets. Revert pedestrian associations that do not conform to part tracklets.
Step III: Perform simultaneous association between all tracklets jointly.

We employ a successive shortest paths algorithm for each step performing data association [12]. This approach has several important characteristics. First, it includes a non-maximum suppression after finding each shortest path. Second, it takes into account the birth and death for each path by determining the edge weight for source and target nodes, which is especially important in the case of parts tracking due to severe and frequent occlusions and mis-detections. Finally, the use of dynamic programming makes the algorithm much faster allowing us to perform the multiple levels of tracking. The details and our contributions in each of the three steps are outlined below.

2.1 Pedestrian Tracklets

Given a video, we begin by applying a state-of-the-art human detector [2] to detect pedestrians in all frames. For each detection in a frame, we also obtain the detector confidence that is used to compute $P_{\delta}^{h}\left(h^{x}\right)$. We then create a flow network [12] for these detections. The set of pedestrians detections, $\mathcal{H}$, becomes the nodes and potential associations between them are the edges. The cost of each node is -log $P_{\delta}^{h}\left(h^{x}\right)$. We then connect each detection to the detection in previous and next frames. We use the overlap between detections as a gating function, so as to not connect detections that are very far apart across frames. The cost of each edge between observation h_{t}^{x} in frame t, and potential point x_{t}^{i}, on the i^{th} track, is the affinity between detections, i.e., -log $P_{\kappa}^{h}\left(h_{t}^{x}|\mathrm{p}_{i,t}^{h},\mathrm{s}_{i,t}^{h}\right)$. The affinity is based on appearance and uses a $9\times9\times9$ bins color histogram, a HOG feature descriptor for the entire bounding box enclosing the pedestrian, and a motion feature descriptor with optical flow binned into 12 intervals. The presence and properties of any pedestrian parts are ignored in this network.

The successive shortest paths algorithm is then used to obtain tracklets for each set of pedestrian detections. Due to problems inherent in the surveillance task, including mis-detections, merged detections, occlusions, clutter, and false positives, these tracklets are less than ideal. They break at points of occlusion and mis-detections, while merged detections and false positives allow connections between unrelated shorter tracklets. In order to break the wrong associations, we attempt to leverage the temporal persistence of pedestrian parts. We therefore first perform explicit tracking of these parts.

2.2 Part Tracking

For part tracking, we take the detections $\mathcal{B}$, for all parts within the spatiotemporal tubes representing a pedestrian tracklet. Again, we employ the k-shortest path algorithm [12], for which we need the node and edge weights in the flow network of body parts.

To this end, we compute the node cost as the negative log likelihood of part detector confidence (-log $P_{\delta}^{g}\left(b^{x,m}\right)$). This is similar to the part HOG-filter convolution over the HOG features of a frame. This confidence also includes a deformation cost, computed by evaluating all part detections at x, as per a pre-trained score map, centered

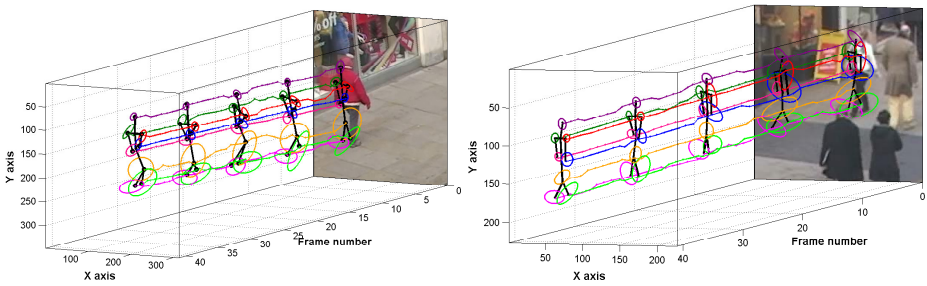

Fig. 1. Two examples showing how spatial models of parts evolve over time. At each frame, the mean relative location of a part is shown as a black dot, and the colored ellipse depicts the covariance of the 2D Gaussian distribution. These models not only change over time, but they are dependent on the camera view point and the pedestrian's specific motion. Such a specific and discriminative dynamic model of parts relationships and positions is difficult to fix apriori. Notice the relatively large variance in relative positions of extremities.

on the potential pedestrian location x. The cost of the edge between parts (i.e., -log $P^g_\kappa\left(b^x|\mathrm{p}^g,\mathrm{s}^g\right)$) is obtained by performing a pixel level correlation in the RGB intensity space, which is analogous to the patch based optical flow computation.

The formulation to obtain part tracklets is essentially an optimization of three costs: the part detection cost $P^g_\delta\left(b^x\right)$, the star-model deformation cost $P\left(\mathrm{p}^g|\mathrm{p}^h+\mu^g,\Sigma^g\right)$, and the temporal persistence cost $P^g_\kappa\left(b^x|\mathrm{p}^g,\mathrm{s}^g\right)$. Once the flow network is established, we apply the successive shortest path algorithm for each part separately. In other words, the pedestrian tracklets allow us to skip exhaustive search for part tracking in the scale and spatial spaces, without which not only will the computation cost be much higher, but the quality of part detection and tracking would be worse (due to lack of deformity cost).

Modeling Relative Part Locations: We also exploit the relationships between detected pedestrian body parts as a measure for evaluating the quality of pedestrian tracklets. For this purpose, in addition to the pre-learned, fixed relationships between the body parts, which Felzenszwalb et al [2] used as a prior for human detection in static imagery, we propose to learn an online model of the relative positions of body parts as they pertain to a specific pedestrian while moving. Instead of computing and modeling an exhaustive set of pairwise relationships, we begin by setting the mean position of head as reference point. This is a meaningful simplification since this part is most likely to be visible across the frames, especially when the pedestrian of interest is under partial occlusion. Moreover, in our experiments, this scheme performed better than the modeling of relative locations between all pairs of parts.

For each part on the pedestrian body, we model the relative location with respect to the reference point as a 2D Gaussian distribution, (μ^g,Σ^g), which becomes part of the dynamic model (Eq. 3). This distribution takes the vector values, $[\mathrm{p}^g_{i,m,t}-\mathrm{p}^g_{i,1,t}]$, as samples, where $m=1$ corresponds to the first part, i.e., the head. At any given time instant during the evaluation of a pedestrian tracklet, we can find the probability of the track of same person in the current frame while allowing a degree of deformability.

(a) (b)

Fig. 2. An illustration of the effect of merged detections on tracking. The figure shows parts of two frames from the Town Center sequence, centered around a target. The large green bounding boxes show the location and scale of pedestrian detection, and the smaller colored bounding boxes show location of parts detections. The similarly colored ellipses represent the online learned model of part locations with respect to the human. In (a), the locations of detected parts are likely as per the model because all parts are within or close to the error ellipses. Figure (b) shows a subsequent frame where two humans are merged into the same detection using the head and shoulders of the incorrect target, thereby producing part detections that are far away from the predicted positions. The initial pedestrian tracklet will therefore be disconnected at this frame.

This probability is computed as an average of the probabilities of observation of each of the parts, as per the Gaussian distribution of their relative locations, i.e., $\frac{1}{N}\sum_{m=1}^{N} \mathcal{N}\left(\mathrm{p}^{g}_{i,m,t}; \mathrm{p}^{h}_{i,t-1} + \mu^{g}_{i,m,t-1}, \Sigma^{g}_{i,m,t-1}\right)$. If this average for a pedestrian i, at frame t, is less than a threshold, we split the pedestrian tracklet at t. Fig. 1 graphically illustrates this model for two pedestrians. It can be noticed that the limbs and extremities obviously have a larger degree of variance in their positions relative to the pedestrian's head. Using the scheme, we are able to split the tracklets at the points in time where the original pedestrian detector made a mistake of putting the wrong parts together to obtain an otherwise high confidence detection (see Fig. 2 for example). Similarly, by learning motion models of individual body parts, we are able to split the wrong tracklet of a pedestrian in cases of merged detection, label switch, or inaccurate detections.

2.3 Merging of Pedestrian and Part Tracklets

Finally, we employ pedestrian parts to merge the tracklets into correct, high confidence trajectories, again by leveraging the flow network optimization algorithm. The entire pedestrian tracklets will now act as nodes, whose costs will be the output of the previous network flow (chosen nodes and edges). We observe that two tracklets to be associated may have a temporal gap due to problems with pedestrian detections within the gap (see Fig. 4). These problems often arise due to mis-detections or merged detections of pedestrians. The parts of such pedestrians however, are often partially visible, with a high confidence (detection scores). For instance, in a part based pedestrian detection method such as [2], it is possible that parts belonging to upper body are fully visible, and detected with high confidence, but do not become part of a high confidence pedestrian detection due to partial occlusion (low overall confidence due to deformity cost). In our proposed framework, we attempt to fully utilize such part detections by trying to associate them over time (rather than to other parts or pedestrians within the frame).

Fig. 3. Three different methods for pedestrian appearance learning. Each figure shows average color of training samples over a short pedestrian tracklet. A simple average over the entire person bounding box is shown in the top row, while the middle row uses detected DPM parts, and our tracked parts are used for examples in the bottom row. Using the entire person bounding box the model is very vague. Since a certain minimum number of parts are always detected in DPM, the model contains background when the person is partially occluded. The handling of occlusion and clutter for parts while temporally aligning them, makes our model more accurate.

Moreover, we notice that due to the previous part detection step (or pedestrian detection using part based models), we have already obtained dense part detections in each frame (even if with low confidence or below detection threshold), and therefore do not require any additional computation for detection. These dense part detections become additional nodes in our final flow network. Consequently, the cost of edges between two pedestrian tracklets is assumed to be a product of three likelihoods: the global motion model of a pedestrian which essentially is a gating function employing constant velocity; the consistency in appearance of pedestrians using a linear SVM model of part based appearances; and part−track confidence, which is averaged over the parts, and is the same as in the previous network except that there is no cost of deformation.

Pedestrian-Specific Appearance Modeling: We use the SVM model of the actual appearance for each part to model each pedestrian. This SVM uses a linear kernel and employs both the edge information and color appearance together. The pedestrian and part detections in each tracklet become the positive training samples for the classifier, where pixel level RGB descriptors for each part are concatenated together. We use an indicator function to zero out the response of parts that are under occlusion, or otherwise mis-detected. Other pedestrian detections within the same temporal window as the tracklet of interest, become the negative training examples. The cost of similarity between two tracklets is the classifier confidence in detecting the pedestrians in the second tracklet. The idea of ignoring certain parts while learning the appearance of a pedestrian is inspired by [14]. Fig. 3 visually illustrates the output of three different schemes for learning such a model, including ours.

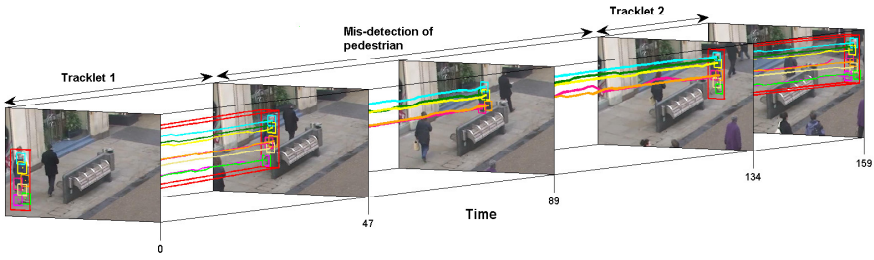

Fig. 4. An example showing the final step of merging pedestrian tracklets. The initial tracklets were obtained due to reasonable performance of the human detector (frames 0-47, and 134-159), but due to occlusion (bench), the pedestrian is mis-detected in a significantly sized temporal window (frames 47-134). The visible parts however, have still been tracked well during the partial occlusion, leading to successful merging of the two tracklets.

Finally, we perform the part tracking by considering the high score part detections within split tracklets which had not previously been part of the pedestrian detections, due to problems like clutter and occlusion. The stand-alone part detections in the occlusion areas are also linked to parts within tracklets, and contribute to overall cost between pedestrian tracklets. These links (edges) are weighted using correlation between part patches, while stand-alone part detections (nodes) are weighted by their corresponding detection scores. The latter obviously do not have any deformation cost.

In the case where the split tracklets are not fully merged despite new detections of stand-alone parts (probably due to complete occlusion), we fill the gap by interpolation. The final result is a pedestrian trajectory, along with the estimated trajectories of their parts. It should be noticed that we never used the parts which are associated with each detection in DPM [2]. We always assume that there are densely detectable, stand-alone parts observable with different, possibly low confidence scores.

3 Experiments and Results

We evaluated our method using four challenging datasets: Town Center [15] and PETS 2009 datasets which are publicly available, and two new datasets: the Parking Lot and Airport sequences. These datasets have a variety of challenges including inter-person and person-scene occlusion, cluttered background, linear and non-linear motion, and crowd scenarios. Our results are evaluated using the standard CLEAR MOT metrics [16]. The Multiple Object Tracking Accuracy (MOTA) is a combined measure which takes into account false positives, false negatives and identity switches. The Multiple Object Tracking Precision (MOTP) measures the precision with which objects are located using the intersection of the estimated region with the ground truth region. The detection precision and recall are also included for comparison with state-of-the-art methods. A brief description of the datasets follows:

• The Town center dataset has a resolution of 1920×1080 at 25 fps. This sequence is semi-crowded with occlusion and the motion of pedestrians are almost linear.

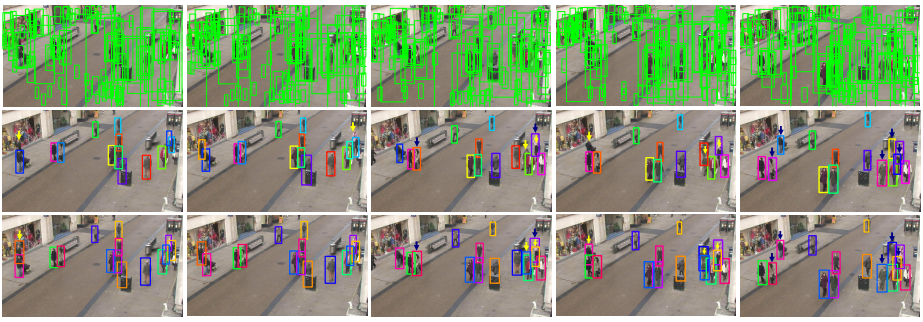

Fig. 5. Handling of common problems shown in the Town Center sequence. Each column depicts a single frame. Top row shows likely pedestrian detections as green bounding boxes, which include several false positives, merged detections, and mis-detections. Middle row shows the results of tracking using k-shortest paths algorithm [12]. We point to mis-detections using yellow arrows ↓, while blue ones ↓ represent id switches. The third row shows corresponding results using our method, with which we are not only able to overcome false negative due to explicit part detection and tracking, but also improve on label switching.

• We used the PETS 2009, monocular tracking sequence L1 (view1), which has a resolution of 768×576, at 7 fps. This dataset is very challenging, because the pedestrians often change direction and groups form and split frequently.
• The Parking Lot video has a resolution of 1920×1080 at 29 fps, and although it has relatively low number of pedestrians, they move in crowded groups causing severe person-person occlusions.
• The Airport dataset has a resolution of 4000×2672 at 5 fps and views the scene from an oblique angle. This is an extremely challenging dataset with severe occlusions by people and static objects. It contains at least a few tens of objects in most frames.

To quantify our method's performance and a fair comparison the following standard annotations were used: for Town Center, annotation provided by [15], who proposed this dataset; for PETS2009, annotation by the TUD GRIS group; and for Airport and Parking Lot datasets, our own annotation (bounding box around each person). For evaluation of part tracking, we manually annotated pedestrian parts in the Town center dataset (Table 2). The detection criteria for pedestrian tracking was the conventional 50% bounding box overlap, and we used 25% overlap for true/false detection of parts following [15].

The use of a linear SVM in appearance modeling of tracklets, and network flow optimization via dynamic programming [12], allows each run of k-shortest paths to take 0.5 seconds for the entire sequence and each run of the linear SVM which is only trained once per tracklet, takes 0.4 seconds. Depending on the number of humans in a video sequence, the run time of our tracking method on a 2.4 GHz desktop running un-optimized code is between 1 to 4 fps. The same parameter settings were used for all datasets, except the overlap condition in network flow construction, where we use a smaller overlap in case of lower frame rate.

Fig. 6. Some qualitative results of tracking. Each row shows three frames each for the Town Centre, PETS 2009, Parking Lot, and Airport video sequences. Notice that the labels of most targets are successfully maintained amid clutter, and severe person-person occlusions.

Fig. 7. Left: a qualitative example of tracking of pedestrian parts for 4 targets. The track for each part is shown in a distinct color as per the legend on the right. The yellow arrow above each track depicts the position of the pedestrian in the image. It can be observed that the proposed method is able to track body parts despite severe person-person, and self occlusions. Right: The location of each of the parts relative to the pedestrian.

Table 1. Quantitative comparison of the proposed pedestrian tracking framework with some of the state-of-the-art techniques, for various datasets

Method	MOTP	MOTA	Prec	Rec	Method	MOTP	MOTA	Prec	Rec
Town Center					**PETS 2009**				
Benfold [15]	**80.3**	61.3	82	79	Breitenstein [17]	59	74	89	60
Yamaguchi [18]	70.9	63.3	71.1	64	Berclaz [13]	62	78	78	62
Pellegrini [19]	70.7	63.4	70.8	64.1	Conte [20]	57	81	85	58
Zhang [21]	71.5	65.7	71.5	66.1	Berclaz [22]	52	83	82	53
Leal-Taixe [23]	71.5	67.3	71.6	67.6	Alahi [24]	52	83	69	53
Our baseline [12]	68.8	63.5	84.9	78.9	Our baseline [12]	73.7	84.6	96.8	93.2
Proposed	71.6	**75.7**	**93.6**	**81.8**	**Proposed**	**76**	**90.7**	**96.8**	**95.2**
Parking Lot					**Airport**				
Our baseline [12]	72.5	83.5	92.6	95.1	Our baseline [12]	67.7	32.7	76.5	54.9
Proposed	**77.5**	**88.9**	**93.6**	**96.5**	**Proposed**	**67.9**	**46.6**	**89.9**	**55.4**

Table 2. Annotation for quantitative evaluation of part tracking (Town Center)

Part ID	1	2	3	4	5	6	7	8	Total
Annotations/frame	14	14	13	13	12	12	11	11	100
Annotations/sequence	2713	2617	2586	2401	2376	2243	2086	2020	19042

Table 3. Quantitative evaluation of the proposed part tracking for Town Center dataset

Method	MOTP	MOTA	Prec	Rec	Method	MOTP	MOTA	Prec	Rec
Part1 (Head)					**Part5**				
HOG detection	45.8	-	35	52.7	HOG detection	26.2	-	6.4	9.4
Benfold et al. [15]	50.8	45.4	73.8	71	Our tracking	52.7	56.2	84.3	69.7
Our tracking	55.4	62.1	87.5	73.1					
Part2					**Part6**				
HOG detection	42.5	-	29.8	43.6	HOG detection	34.9	-	13.7	20
Our tracking	56.8	44.2	76.9	64.2	Our tracking	49.1	50.8	82.6	65.2
Part3					**Part7**				
HOG detection	42.1	-	37.3	54.6	HOG detection	31.7	-	31.9	46.7
Our tracking	61.2	59.9	86.2	72.2	Our tracking	59.4	47.6	80.3	63.9
Part4					**Part8**				
HOG detection	35.5	-	8	11.7	HOG detection	26.5	-	28.4	41.6
Our tracking	45.2	56.6	84.6	70	Our tracking	56	37.5	74.8	57.5

Fig. 5 shows a qualitative overview of a human detector [2] (top row), and a state-of-the-art tracking algorithm [12] (middle row), which uses the same flow network optimization as our proposed framework (bottom row). It can be concluded that the improvement we obtained over the method of [12] is a direct consequence of explicit part tracking, which not only constrains pedestrian tracking hypothesis, but allows data association over and across frames where pedestrian detections are not available due to severe occlusions. Instead of only using pedestrians obtained by a part-based detector, we exploit the notion of high confidence stand-alone parts while the human they belong to, is not fully observable. We therefore are able to better handle the problems of occlusions, mis-detections, merged detections, and therefore, label switching.

Some of the qualitative results for the four video sequences are reported in Fig. 6. It can be observed that most targets are tracked successfully over a large number of frames despite extremely crowded scenarios. The proposed framework performs very well in scenarios of extreme person-person dynamic occlusions, where objects move together in close proximity to each other, resulting in missed and merged human detections.

A comparison of quantitative results of pedestrian tracking for all datasets is shown in Table 1, where the proposed method performs comparably or better than state-of-the-art in human tracking. Since tracking of human parts is a major contribution of our work, unlike other methods which only quantify the effect of part tracking on human tracking, we explicitly ground truthed (see Table 2) and quantified our results. The results of part tracking using the proposed system for the Town Center dataset are shown in Table 3.

4 Conclusion

To summarize our contributions, we propose a method for simultaneous tracking of pedestrians as well as their parts in difficult video sequences involving severe occlusions. We show that by constraining pedestrian tracking by part tracking, we can explicitly improve tracking accuracy over state-of-the-art. Our method implicitly improves detection of humans as well as their parts. We also quantify the performance of our parts tracking algorithm. In conclusion, by building upon recent works in human detection and tracking, we have not only improved results of human tracking, but also proposed a formal method for constrained tracking of human body parts.

Acknowledgement: The research presented in this article is supported by the Pacific Northwest National Laboratory (PNNL) in Richland, Washington.

References

1. Mikolajczyk, K., Schmid, C., Zisserman, A.: Human Detection Based on a Probabilistic Assembly of Robust Part Detectors. In: Pajdla, T., Matas, J. (eds.) ECCV 2004, Part I. LNCS, vol. 3021, pp. 69–82. Springer, Heidelberg (2004)
2. Felzenszwalb, P., Girshick, R., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. PAMI 32, 1627–1645 (2010)
3. Bourdev, L., Malik, J.: Poselets: Body part detectors trained using 3d human pose annotations. In: ICCV (2009)
4. Tian, T.P., Sclaroff, S.: Fast globally optimal 2d human detection with loopy graph models. In: CVPR (2010)
5. Andriluka, M., Roth, S., Schiele, B.: People-tracking-by-detection and people-detection-by-tracking. In: CVPR (2008)
6. Breitenstein, M., Reichlin, F., Leibe, B., Koller-Meier, E., Van Gool, L.: Robust tracking-by-detection using a detector confidence particle filter. In: ICCV (2009)
7. Andriluka, M., Roth, S., Schiele, B.: Monocular 3d pose estimation and tracking by detection. In: CVPR (2010)
8. Lu, W.L., Little, J.: Simultaneous tracking and action recognition using the pca-hog descriptor. In: The 3rd Canadian Conference on Computer and Robot Vision (2006)
9. Li, R., Chellappa, R., Zhou, S.: Learning multi-modal densities on discriminative temporal interaction manifold for group activity recognition. In: CVPR (2009)
10. Wu, B., Nevatia, R.: Tracking of multiple, partially occluded humans based on static body part detection. In: CVPR (2006)
11. Zhao, Q., Kang, J., Tao, H., Hua, W.: Part based human tracking in a multiple cues fusion framework. In: ICPR (2006)
12. Pirsiavash, H., Ramanan, D., Fowlkes, C.: Globally-optimal greedy algorithms for tracking a variable number of objects. In: CVPR (2011)

13. Berclaz, J., Fleuret, F., Turetken, E., Fua, P.: Multiple object tracking using k-shortest paths optimization. PAMI 33, 1806–1819 (2011)
14. Yang, W., Wang, Y., Mori, G.: Recognizing human actions from still images with latent poses. In: CVPR (2010)
15. Benfold, B., Reid, I.: Stable multi-target tracking in real-time surveillance video. In: CVPR (2011)
16. Kasturi, R., Goldgof, D., Soundararajan, P., Manohar, V., Garofolo, J., Bowers, R., Boonstra, M., Korzhova, V., Zhang, J.: Framework for performance evaluation of face, text, and vehicle detection and tracking in video: Data, metrics, and protocol. PAMI 31, 319–336 (2009)
17. Breitenstein, M., Reichlin, F., Leibe, B., Koller-Meier, E., Van Gool, L.: Online multiperson tracking-by-detection from a single, uncalibrated camera. PAMI 33, 1820–1833 (2011)
18. Yamaguchi, K., Berg, A., Ortiz, L., Berg, T.: Who are you with and where are you going? In: CVPR (2011)
19. Pellegrini, S., Ess, A., Schindler, K., Van Gool, L.: You'll never walk alone: Modeling social behavior for multi-target tracking. In: ICCV (2009)
20. Conte, D., Foggia, P., Percannella, G., Vento, M.: Performance evaluation of a people tracking system on pets2009 database. In: AVSS (2010)
21. Zhang, L., Li, Y., Nevatia, R.: Global data association for multi-object tracking using network flows. In: CVPR (2008)
22. Berclaz, J., Fleuret, F., Fua, P.: Multiple object tracking using flow linear programming. In: PETS-Winter (2009)
23. Leal-Taixe, L., Pons-Moll, G., Rosenhahn, B.: Everybody needs somebody: Modeling social and grouping behavior on a linear programming multiple people tracker. In: ICCV Workshops (2011)
24. Alahi, A., Jacques, L., Boursier, Y., Vandergheynst, P.: Sparsity-driven people localization algorithm: Evaluation in crowded scenes environments. In: PETS-Winter (2009)