Multiple People Tracking using Body and Joint Detections

Roberto Henschel[†] Yunzhe Zou[♭] Bodo Rosenhahn[†]
[†]Leibniz Universität Hannover, Germany
{henschel,rosenhahn}@tnt.uni-hannover.de zouyunzhe@qq.com

Abstract

Most multiple people tracking systems compute trajectories based on the tracking-by-detection paradigm. Consequently, the performance depends to a large extent on the quality of the employed input detections. However, despite an enormous progress in recent years, partially occluded people are still often not recognized. Also, many correct detections are mistakenly discarded when the non-maximum suppression is performed. Improving the tracking performance thus requires to augment the coarse input. Well-suited for this task are fine-graded body joint detections, as they allow to locate even strongly occluded persons.
Thus in this work, we analyze the suitability of including joint detections for multiple people tracking. We introduce different affinities between the two detection types and evaluate their performances. Tracking is then performed within a near-online framework based on a min cost graph labeling formulation. As a result, our framework can recover heavily occluded persons and solve the data association efficiently. We evaluate our framework on the MOT16/17 benchmark. Experimental results demonstrate that our framework achieves state-of-the-art results.

Figure 1. Tracking based on people and joint detections. Top row: SDP input detections on MOT17-04. The person with the red bag is not detected anymore after frame 19. Bottom row: The result of our method. Due to the integration of joint detections into the data association, all persons are found and tracked correctly, despite heavy occlusions and missing detections.

1. Introduction

The goal of multiple object tracking (MOT), or specifically multiple people tracking, is to infer the trajectories of all targets that appear in a video sequence. Thus, MOT is an essential component when high-level understanding of a video scene is desired. This may be to study social interactions of humans [38, 2, 1], to understand the environment of a car for autonomous driving or to secure critical areas [11].

The majority of recent approaches focus on the so-called tracking-by-detection strategy. It consists of two parts: first the detection hypotheses, usually bounding boxes, are generated by applying a person detector to each video frame. Then these hypotheses are linked across the video frames according to affinity criteria. This work focuses on the latter task, which is also called the data association problem.

While there has been significant improvements in the performance of tracking-by-detection systems in recent years, the performance of such a tracking system still relies heavily on the employed detector. Further, even with state-of-the-art detectors, people are frequently missed. This may be the case due to heavy occlusions, especially in crowded scenes. Also performing the non-maximum suppression (NMS) results in many missing localizations.

In order to compensate these detector issues, one approach used widely in recent works of the tracking community is that the whole video sequence is taken into account, instead of using just a few frames. In this case, the data association problem is solved with the help of additional information from future frames. Especially sophisticated features based on neural networks [34, 21] allow to re-locate persons even after long occlusions. However, the

integration of long-term considerations is computationally demanding. Hence, they cannot be run online. A complementary approach is to include additional detectors which are trained to locate fine-graded regions of interests of a person. Accordingly these additional detectors are integrated into the data association step, which allows to recover the position of persons that have been missed by the people detector.

Inspired by the latter concept and motivated by the limitation of the first, we build a new framework for MOT which overcomes the detection failure issues and is able to tackle data in a near-online way. As additional information, we employ joint detections. Being very fine-graded, they allow to explain the position of a person in a bottom-up way: Each joint can be related to the full-body detection. This results in very accurate data associations as the additional information are used to ensure spatial and temporal consistencies of body and joint detections.

However, using additional detectors also poses a difficult data association problem: Each person might have multiple detections that need to be associated to each other. This is different to the normal setting where it is assumed that at most one detection at a frame belongs to a person's trajectory. Further, affinities have to be defined not only between body detections but also between joint detections as well as body and joint detections, within the same frame as well as between different frames.

In this work, we design specific affinities between the different inputs and analyze in detail their suitability for MOT. Our experiments show that the fusion of fine-graded detections is very beneficial, leading to state-of-the-art performance.

To summarize, our contribution is three-fold:

- We integrate joint detections into a near-online multiple people tracking system.
- We analyze affinities to fuse people detections and joint detections into a tracking system.
- Our presented tracker is robust against occlusions and sets a new state-of-the-art in tracking.

2. Related Work

Data Association. Most modern multi object tracking systems are based on the so-called tracking-by-detection paradigm [13, 22, 33, 7, 40, 29, 14, 15]. The idea is to split the problem into two subsequent tasks: object detection and data association. There has been significant progress in localizing objects in an image, mainly due to recent advances in deep neural networks [27, 37, 32]. Nonetheless objects of interest are still frequently missed due to occlusions, complex deformations and challenging lightning conditions. The task of the data association is then to compensate the errors of the detector, recover missing positions and to assign correct detections to the corresponding persons. This task has been tackled in a global manner using network flow formulations [40, 22] based on the markov chain assumption. More recently, improvements have been achieved by considering all associations of a trajectory at the same time (instead of only frame-to-frame links), using correlation clustering [33, 34] or graph labeling [13]. While these offline methods perform well, solving the entire data association problem is computationally demanding, especially for long video sequences. With the rise of robust features based on neural networks, many tracking systems employ near-online [7, 20] or online systems [8, 26, 41] as the features are often sufficient to solve the data association correctly without the need of many information from future frames.

Also our work builds upon the near-online framework but generalizes the tracking-by-detection concept by integrating further information, namely joint detections into the decision step, which help especially to obtain robust interpolation as well as extrapolation of body detections.

Integration of additional information. While multiple people tracking based on detection boxes has seen considerable improvements in recent years, many works have recognized that the integration of additional information leads to a performance gain. The work [34] uses joint information to obtain more robust box affinities. However, it still relies on bounding boxes, so that people that have been missed by the people detector are difficult to recover.

Instead of only linking detections by the help of additional information, several works add additional inputs directly into the data association problem, so that these have to be linked too. Common to all these publications is that the additional information are fine-graded, thereby helping the recover false negative detections and to improve the interpolation abilities. Some works add head detections in a network-flow formulation [5] or, more powerfull, in a graph labeling setting [13]. Others successfully add dense point trajectories [18, 15]. Several works consider joints for multi-person pose estimation and tracking. Some methods directly infere pose and temporal consistency on the joint space [16, 17], without considering people detections. Others [36] compute frame-wise poses first and use bounding boxes only to transform a pose to the next frame.

Instead, our work uses all joint information as well as body localizations directly within the data association step, leading to state-of-the-art tracking performance.

3. Method

We introduce our tracking-by-detection system which utilizes joint detections and body detections in a holistic way. We assume that all necessary input detections are already provided by external detectors and focus on the re-

maining data association problem. However in our case, the data association problem differs from the normal setting as there might be several input detections at a frame (people detections and joint detections) that have to be assigned to a person. We cast this problem into a min cost graph labeling problem, where the optimal solution corresponds to an assignment that is consistent in space and time between the different detection types. To make the approach efficient, we embed the graph labeling into a near-online framework. Further, we introduce all necessary affinities.

3.1. Tracking model

A near-online tracker, as introduced in [7], solves the data association problem at frame t using already computed tracking results from the past, within a time window $t - \triangle t_1, \ldots, t - 1$ together with input detections at time $t, \ldots, t + \triangle t_2 - 1$.

Let $\mathcal{D}$ denote the set of detections of the video sequence to be tracked, which decomposes into body detections $\mathcal{D}^{\mathrm{B}}$ and joint detections $\mathcal{D}^{\mathrm{J}}$. Further, $\mathcal{D}_t$ comprises all detections at frame t. A trajectory $T \subset \mathcal{D}$ consists of all detections belonging to a person. Let $\delta(T)$ denote the latest time stamp for which T contains detections and let $\overline{T} := T \cap \mathcal{D}_{\delta(T)}$ contain the corresponding detections at the tail of T. Finally, the set $\mathcal{T}_{t-\triangle t_1,t-1}$ comprises all trajectories T where $\delta(T)$ is within $t - \triangle t_1, \ldots, t - 1$.

Now the tracking task is to find optimal associations between the previously computed trajectories $\mathcal{T}_{t-\triangle t_1,t-1}$ and detections $\mathcal{D}_{t,t+\triangle t_2-1} := \cup_{i=t}^{t+\triangle t_2-1} \mathcal{D}_i$ within the sliding window, which also incorporates to identify newly appeared targets. We note that the sliding window has a size of $\triangle t_1 + \triangle t_2$ and causes a delay of $\triangle t_2 - 1$ frames.

For a sliding window around frame t, we solve the data association problem by finding a min cost graph labeling solution. In particular, we create an undirected weighted graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{C})$, with the vertex set

$$\mathcal{V} := \mathcal{T}_{t-\triangle t_1,t-1} \sqcup \mathcal{D}_{t,t+\triangle t_2-1}. \tag{1}$$

The edge set $\mathcal{E}$ comprises all possible connections between precomputed trajectories $\mathcal{T}_{t-\triangle t_1,t-1}$ and detections $\mathcal{D}_{t,t+\triangle t_2-1}$ as well as between any two detections of $\mathcal{D}_{t,t+\triangle t_2-1}$. Affinity costs c_e for $e \in \mathcal{E}$ reflect how likely an edge connects inputs belonging to the same person. Accordingly, costs c_v for $v \in \mathcal{V}$ reflect how likely an input is a true positive. The association problem can then be formulated as a min cost graph labeling problem [13]:

$$\underset{\mathcal{I} \in \{0,1\}^{P \times |\mathcal{V}|}}{\arg\min} \sum_{l=1}^{P} \sum_{v \in \mathcal{V}} c_v \mathcal{I}_{v,l} + \sum_{e=\{v,v'\} \in \mathcal{E}} c_e \mathcal{I}_{v,l} \mathcal{I}_{v',l} \tag{2}$$

subject to $\sum_{l=1}^{P} \mathcal{I}_{v,l} \leq 1$ for all $v \in \mathcal{V}$. Here, P is an upper bound on the number of persons in the sliding window. If for an indicator variable $\mathcal{I}_{v,l} = 1$ holds, then vertex v is assigned to person l. Accordingly, the aim of (2) is to compute the assignment of the indicator variables $\mathcal{I}_{v,l}$ such that the most plausible data association is selected that is consistent in space and time, with respect to both input detectors.

We solve (2) using the solver introduced in [13], update the trajectories and shift the sliding window one time step forward.

3.2. Features

For the tracking task, we need to define features comparing pairs of detections (d, d'). Thereby, d and d' can be body detections as well as joint detections. Also, d and d' may be from the same frame or between different frames.

We transform features to cost values by training corresponding logistic regression classifiers and use their logit values.

We describe a joint detection j using a vector $\mathbf{d}_j := (x_j, y_j, \mu_j, t_j)$ where $\mathbf{p}_j := (x_j, y_j)$ denotes the detection location; μ_j is the type of the joint and t_j the frame number at which the joint has been detected. Similarly, we define a body detection b: $\mathbf{d}_b := (x_b, y_b, w_b, h_b, t_b)^T$, where $\mathbf{p}_{ul}^b := (x_b, y_b)^T$ is the upper left corner of the box; w_b and h_b are the box width and height, respectively; t_b is the frame number. The middle point of the box is denoted as $\mathbf{p}_m^b := (x_b + w_b/2, y_b + h_b/2)^T$.

In the following we describe various spatial and temporal features that we examined for our tracker. In the experiments section, these features are evaluated and the best performing ones are selected.

Joint-to-Joint spatial. We use the costs defined in ArtTrack [16], which compares for a joint of type μ the expected position with the observed position, in relation to the position of another joint of type $\mu' \neq \mu$.

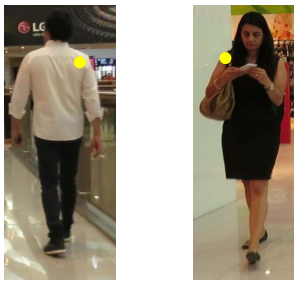


Figure 2. Orientation ambiguity: the right shoulder detection can be located either on the right or left half of the box detection.

Joint-to-Body spatial. When the position of a joint type μ is compared with a bounding box, an orientation ambiguity has to be taken into account, *e.g.* the right shoulder may appear on the left or right half side of a bounding box, depending on the walking direction, see Fig 2. We introduce four different features suitable to decide if a joint and

a body detection may belong to the same person.

(1) Barycentric distance: In order to compare the location of a joint to the position of a detection box, regardless of its size, we employ barycentric coordinates. Consider the positions $\mathbf{p}^b_{\text{ll}}, \mathbf{p}^b_{\text{lr}}$ and $\mathbf{p}^b_{\text{ul}}$ of the lower left, lower right and upper left corner positions of b. Then, the corners $\mathbf{p}^b_{\text{ll}}, \mathbf{p}^b_{\text{lr}}$ and $\mathbf{p}^b_{\text{ul}}$ of b describe a triangle, which we denote by

$$\triangle(b) := \begin{pmatrix} \mathbf{p}^b_{\text{ll}} & \mathbf{p}^b_{\text{lr}} & \mathbf{p}^b_{\text{ul}} \\ 1 & 1 & 1 \end{pmatrix} \in \mathbb{R}^{3\times 3}. \tag{3}$$

With respect to this triangle, we can now consider the barycentric coordinates $\mathcal{B}^b_j$ of a joint j (also outside of b):

$$\mathcal{B}^b_j := \triangle(b)^{-1} \begin{pmatrix} \mathbf{p}_j \\ 1 \end{pmatrix}. \tag{4}$$

The box middle point $\mathbf{p}^b_m$ has barycentric coordinates $\mathcal{B}_m := (0, 0.5, 0.5)^T$. We obtain the euclidean distance between the middle point and the joint position in barycentric coordinates, independent of the box size of b:

$$\text{dist}_{bary} = ||\mathcal{B}^b_j - \mathcal{B}_m||_2. \tag{5}$$

(2) x-y-offset: Also, the offset between the joint position $\mathbf{p}_j$ and the box middle point $\mathbf{p}^b_m$ might be a good affinity. We consider the offset in x-direction and y-direction and normalize by the box width w_b and box height h_b, respectively. We denote the offset from the box middle point $\mathbf{p}^b_m$ to joint location $\mathbf{p}_j$ as $\mathbf{o}_{mj} = \mathbf{p}_j - \mathbf{p}^b_m$.

Due to the orientation ambiguity, for the measure in x direction, we have to compute the absolute value of the offset. Accordingly, these two features are expressed as follows:

$$x_{\text{offset}} = \left| \frac{x_j - x^b_m}{w_b} \right|, \quad y_{\text{offset}} = \frac{y_j - y^b_m}{h_b} \tag{6}$$

(3) Angle in reference box: To compare angles, the orientation ambiguity has to be taken into account. A joint j which is located left to x^b_m is mirrored at the vertical middle line of b, resulting in a position $\mathbf{p}_j$ at the right half side of b. We consider a reference box B_{ref} similar to [13] to maintain the relative position between the joint position and the full-body bounding box with varying sizes (see Fig.3). A joint j is then mapped to position $\mathbf{p}^{\text{ref}}_j$ via barycentric coordinate transformation induced by the triangles of b and B_{ref}:

$$\begin{pmatrix} \mathbf{p}^{\text{ref}}_j \\ 1 \end{pmatrix} = \triangle(B_{\text{ref}})\mathcal{B}^b_j = \triangle(B_{\text{ref}})\triangle(b)^{-1} \begin{pmatrix} \mathbf{p}_j \\ 1 \end{pmatrix}. \tag{7}$$

We learn the mean position of each joint type μ in the reference box from the training data, denoted by $p_{\text{ref}}(\mu)$. With $\mathbf{p}^{\text{ref}}_m$ being the center position of B_{ref}, we compute the offsets $\hat{\mathbf{o}}_{mj} = \mathbf{p}^j_{\text{ref}} - \mathbf{p}^{\text{ref}}_m$ and $\hat{\mathbf{o}}_{m\mu} = \mathbf{p}_{\text{ref}}(\mu) - \mathbf{p}^{\text{ref}}_m$, comparing observed and expected offsets in barycentric coordinates.

Finally, we obtain the angle disagreement via

$$\text{angle}_{\text{ref}} = \arccos\left(\frac{\langle \hat{\mathbf{o}}_{mj}, \hat{\mathbf{o}}_{m\mu} \rangle}{|\hat{\mathbf{o}}_{mj}||\hat{\mathbf{o}}_{m\mu}|} \right). \tag{8}$$

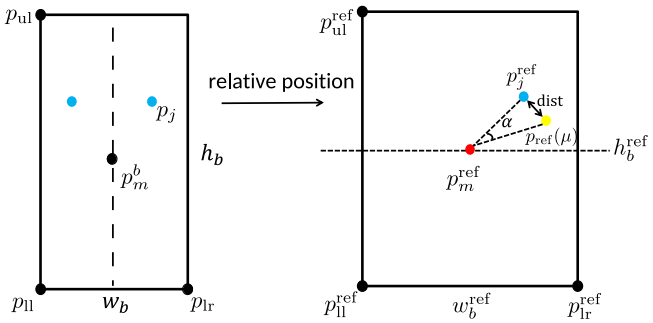


Figure 3. Angle and distance between the expected and observed joint position in the reference box.

(4) Distance in reference box: We also compute the distance between the mean position $\mathbf{p}_{\text{ref}}(\mu)$ and the (possibly) mirrored detection $\mathbf{p}^j_{\text{ref}}$ in barycentric coordinates:

$$\text{dist}^{\text{ref}}_{\text{bary}} = ||\hat{\mathbf{o}}_{mj} - \hat{\mathbf{o}}_{m\mu}||_2. \tag{9}$$

Joint-to-Body temporal. Assuming that there is no large displacements within small temporal distance, we use the same features as those for a joint-to-box spatial pair.

Joint-to-joint temporal As temporal features between two joints j_1 and j_2, we consider euclidean distances:

(5) Euclidean distance: We compute directly the euclidean distance between two joint positions. It is expressed as:

$$\text{dist}^{j_1 j_2}_{eu} = \sqrt{(x_{j_1} - x_{j_2})^2 + (y_{j_1} - y_{j_2})^2} \tag{10}$$

(6) Scaled distance: In case that one joint belongs to an already computed trajectory containing also a detection b, we utilize the scale information provide by b:

$$\text{dist}^{j_1 j_2}_{scaled} = \frac{\text{dist}^{j_1 j_2}_{eu}}{h_b} \tag{11}$$

Body-to-Body spatial. We assume no two bounding boxes should belong to a person (due to the NMS) and set the spatial costs between them to a constant high value.

Body-to-Body Temporal. We employ the Deep Matching (DM) features [35] as temporal measurements.

It provides dense correspondences of pixels between two frames. Given two boxes b_1 and b_2 and the set of DM points M_{b_1} and M_{b_2} inside b_1 and b_2, respectively, we define $MU = |M_{b_1} \cup M_{b_2}|$ and $MI = |M_{b_1} \cap M_{b_2}|$. Inspired by [13, 33], we define the following features: $\frac{MI}{MU}$, $\frac{MI}{|M_{b_1}|}$, $\frac{MI}{|M_{b_2}|}$ and $\frac{MI}{0.5(|M_{b_1}|+|M_{b_2}|)}$. Finally, we stack them in a affinity vector f_{DM}:

$$f_{DM} = \left(\frac{MI}{MU}, \frac{MI}{|M_{b_1}|}, \frac{MI}{|M_{b_2}|}, \frac{MI}{0.5(|M_{b_1}| + |M_{b_2}|)}\right) \tag{12}$$

Affinity between a trajectory and a detection Finally, assume that a trajectory $T \in \mathcal{T}_{t-\triangle t_1,t-1}$ contains body detections $b_1, \ldots, b_m$ and joint detections $j_1, \ldots, j_n$ in its last frame and consider a detection $b \in \mathcal{D}_{t,t+\triangle t_2-1}$. Let

$$\text{Cost}_{\text{body}} = \frac{1}{m}\sum_{k=1}^{m} w_{k,b}, \quad \text{Cost}_{\text{joint}} = \frac{1}{n}\sum_{k=1}^{n} w'_{k,b}, \quad (13)$$

where $w_{k,b}$ and $w'_{k,b}$ denote the costs between b and all body detections (or joint detections) of T, as introduced in Sect. 3.2. Then, we define the cost between trajectory T and detection b as:

$$c_{T,b} = \lambda_{\text{joint}}\text{Cost}_{\text{joint}} + \lambda_{\text{body}}\text{Cost}_{\text{body}}. \quad (14)$$

We set $\lambda_{\text{joint}} = 1$ and λ_{body} to the number of joint types. This compensates unbalances in the number of expected detections that should belong to a person at each frame.

3.3. Post-Processing

Once trajectories fall out of the temporal window, we perform post-processing. Assuming that false alarms usually last not too long, we simply set a threshold for the minimal trajectory length and eliminate too short ones.

The data association may have recovered the position of a person after some frames of missing body detections. It may also contain trajectories which for some frames only have associated joint detections but no full-body detection.

Accordingly, we perform different post-processing methods to localize the position also in those frames where there is no associated body detection.

Consider a trajectory T and let f be a frame that has no associated body detection.

If T contains a body detection in a frame $f' < f$ and $f'' > f$, we recover the position using linear interpolation between the existing body detections. Thereby, we use the smallest frame $f'' > f$ and biggest frame $f' < f$ with body detections.

Assume $f' < f$ is the last frame in which T contains body detections. We extrapolate the position of a bounding bounding box with the help of associated joint detections. Let b' be the body box at frame f'. Further, the sets J^f and $J^{f'}$ comprise all joints of T at frame f and f', respectively. We take the information from any two joints of the same type at the frames f and f'. Thus, let $(j_1, j'_1), \ldots, (j_{N_{\text{com}}}, j'_{N_{\text{com}}}) \in J^f \times J^{f'}$ be all such pairings, so that $\mu_{j_i} = \mu_{j'_i}$ for $i = 1, \ldots, N_{\text{com}}$. Assuming that the offset between box middle point and the mean joint position stays constant withing a short temporal distance, we set up the following equation:

$$\frac{1}{N_{com}}\sum_{r=1}^{N_{com}} \mathbf{p}_{j_r} - \mathbf{p}^b_m = \frac{1}{N_{com}}\sum_{n=1}^{N_{com}} \mathbf{p}_{j'_n} - \mathbf{p}^{b'}_m, \quad (15)$$

where b denotes the to be computed body box at frame f. Then, we obtain the box center $\mathbf{p}^b_m$ of b from Eq. 15:

$$\mathbf{p}^b_m = \frac{1}{N_{com}}\sum_{r=1}^{N_{com}} \mathbf{p}_{j_r} - \frac{1}{N_{com}}\sum_{n=1}^{N_{com}} \mathbf{p}_{j'_n} + \mathbf{p}^{b'}_m \quad (16)$$

We compute the coordinates of the upper left corner via

$$\mathbf{p}^b_{\text{ul}} = \mathbf{p}^b_m - \frac{1}{2}(w_{b'}, h_{b'}). \quad (17)$$

To ensure robustness, we perform extrapolation only when $N_{com} \geq 3$. The case $f' > f$ being the first frame with body detections is done similarly.

4. Experiments

Based on the affinities introduced in Sect. 3.2, we analyze the corresponding classifiers regarding joints in Sect. 4.1. The best performing features are then integrated into our near-online tracker, which deals as the basis for all subsequent experiments.

Fig. 1 shows that joint detections help eliminating false positive body detections while recovering missed detections through inter- and extrapolation. To quantitatively analyse the advantage of fusing joint detections, we compare in Sect.4.3 the results achieved by our tracker using both detection types against body detections only. The influence of the size of the sliding window is discussed in Sect. 4.4. Finally, we compare the performance of our method with other reported results on the MOT benchmark [25] in Sect. 4.5.

(a) DPM Body detections (b) Our tracking result

Figure 4. MOT17-02 sequence, frame 2. (a): Body detections from DPM with a false positive. (b): The result by our tracker. It removes the false positive that does not match the joint detections.

Filtering. After the detections in the sliding window are associated, we filter out false-positive body detections. We revoke a body-detection, if there are at least 3 joints of the person in that frame that are not connected to the box (positive pairwise cost to the box). This helps in particular to remove wrong double detections of a person (see Fig.4).

4.1. Feature Evaluation

We evaluate the performance of all features proposed in Sect. 3.2 which regard joint detections.

Training data. In order to train the joint classifiers, we employed the PoseTrack [3] dataset. This benchmark consists of 500 video sequences, with 20K frames and 150K body pose annotations. The video sequences contain multiple people in different scales engaging in various dynamic activities, where body part occlusions appear frequently. As the dataset contains only labels for the joint detections, we infer corresponding ground-truth bounding boxes by taking the minimal enclosing bounding box around all joints of a person (which we slightly enlarge by 15% in width and height). Further, we filter out poses that are unlikely to be seen on a street sequence.

For the false positive samples, we randomly group joint and body detections of different persons. In total we create a training subset with an equal number of true positive and false positive pairs. For temporal associations between two detections from different video frames, we set the maximum temporal distance to 9 frames. In order to avoid that the samples with smaller temporal distance dominate the training data, we make sure that the number of samples per frame distance remains constant. All classifiers are learnt from the features of Sect. 3.2 using a logistic regresion model [9]. Finally, we split the created dataset into equally sized training and validation sets.

Regression metric. The evaluation of the different classifiers is conducted on the validation set, using the standard accuracy performance indicator:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

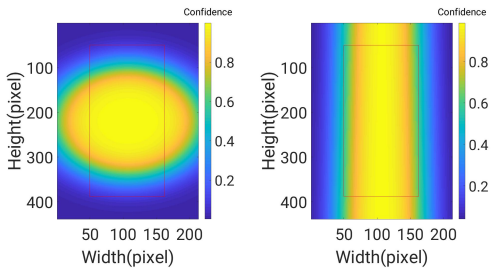
where TP, TN, FP, FN denote the number of true positives, true negatives, false positives and false negatives, respectively.

We analyze the classifiers regarding joints.

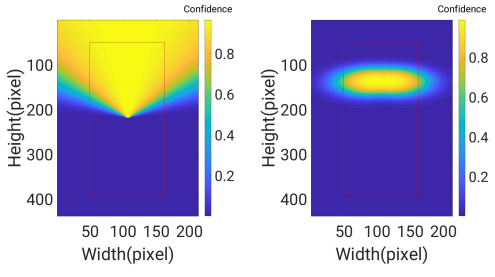
Joint-to-box spatial features We employ the joint detector [16] and use all provided joint types except for neck and head, making 12 joint types in total. 4 features are designed to spatially describe joint-to-box associations: (1) barycentric distance, (2) x-y-offset, (3) angle in reference box, (4) distance in reference box.

First, we train a classifier for each feature individually and validate the performance using a testing box. We compute a confidence map which evaluates for each possible joint position the likelihood that the joint and the test bounding box belong to the same person. Exemplary, we show the results for the right shoulder, in Fig. 5(a)-(d).

Although we want to encode the relative y position in feature (2), Fig. 5(b) reveals it is faulty in this direction. The reason is that the data described by feature (2) is not linearly separable, so that encodings like above or below the middle point cannot be used. The classifier based on



(a) feature (1) (b) feature (2)



(c) feature (3) (d) feature (4)

Figure 5. Confidence map of the right shoulder described by different joint-to-box spatial features.

feature (1) provides a very coarse localization of the shoulders. Only classifiers based on feature (3) and (4) seem to output reasonable predictions. Therefore, we finally build our joint-to-box spatial classifier based on feature (3) and feature (4). The resulting classifier performs well on the validation data (see Table. 1) and outputs meaningful confidence maps similar to Fig. 5(d), but with less variance in implausible directions.

Joint-to-box temporal features Assuming that the joints do not move severely, we use features (3) and (4) also to train a joint-to-box temporal classifier up to 9 frames. Accuracies are shown in Table. 2 and Fig. 6 shows the accuracy plotted over the frame distance for one joint type.

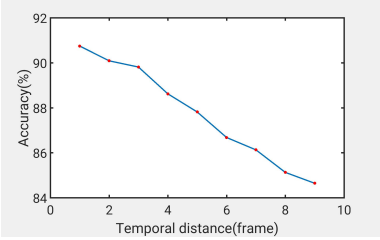


Figure 6. Accuracies of the joint-to-box temporal classifier for the right shoulder.

The classifier achieves over 85% mean accuracy on all types of joints on the validation data. From Fig. 6 we see that the accuracies decrease with frame distance on the validation data. The classifier still gets over 80% accuracy at

	Right ankle	Right knee	Right hip	Right wrist	Right elbow	Right shoulder
Accuracy(%)	93.37	93.50	93.41	88.62	91.74	94.66

Table 1. Training accuracy of the final joint-to-box spatial classifier.

	Right ankle	Right knee	Right hip	Right wrist	Right elbow	Right shoulder
Accuracy(%)	89.05	88.89	88.55	88.81	88.74	90.45

Table 2. Training accuracies of the joint-to-box temporal classifier.

distance of 9 frames, which is acceptable.

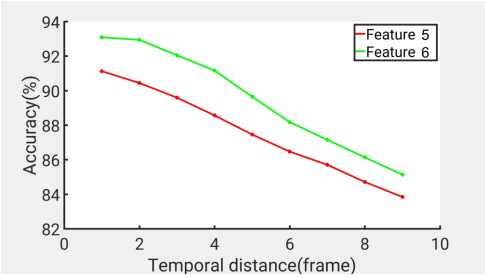


Figure 7. Accuracies of the joint-to-joint temporal classifiers.

Joint-to-joint temporal features Finally, we evaluate the joint-to-joint temporal features (5) and (6). The validation accuracies are shown in Table. 3 and Fig. 7 plots the impact of time. The results show that knowing the scale of the detection box (feature (6)) brings roughly 2% improvement.

4.2. Evaluation details

Finally, we evaluate the tracking performance of the proposed system. We use the classifiers as introduced in Sec. 4.1. To obtain joint detections we utilize [16]. Then, we evaluate the tracking performance using the state-of-the-art MOT17 benchmark [25]. It comprises sequences filmed from various viewpoints, with crowded scenarios and different lighting conditions. In total, there are 7 training and 7 test sequences. Each sequence is provided with 3 sets of detections: DPM [10] with a low recall and precision, Faster-RCNN [27] which performs better and SDP [37], which shows the best detection performance.

Tracking metrics. The MOT benchmark employs the CLEARMOT metrics [4]: MOT accuracy (MOTA), the main metric used to compare MOT approaches, comprises false positives (FP), false negatives (FN) and identity switches (IDS). Another important metric is IDF1 [28] which measure the identity consistency of each trajectory. Further, mostly tracked (MT), mostly lost targets (ML) and track fragmentations (FM) are taken into account [23].

Further, we filter out too short trajectories (length ≤ 3), as they are very likely false positive detections.

4.3. Tracking with additional joint detections

We investigate the impact of adding joint detections to a tracking system. We compare our tracker using body detections only (Body) against body and joint detections (Body+Joints). We compare on all MOT17 training sequences using all 3 full-body detectors. For the temporal sliding window, we set $\triangle t_1 = 4$ and $\triangle t_2 = 5$ in order to emphasize the effect of interpolation and extrapolation. We present the evaluation in Table. 5.

We observe that the MOTA score increases by using both detections. Among the 3 detectors, MOTA increases the most for DPM by more than 10% and the least for SDP by 0.2%. The number of IDs decreases significantly, which is more than halved for Faster-RCNN and SDP. Detectors with a high precision rate (FRCN and SDP) allow to securely revoke remaining wrong detections, either by the solution of the data association problem (2) or by using the aforementioned filter strategy. Accordingly, we observed a decrease of the FP score. Using a detector with a low recall rate (DPM) showed a significant improvement in recovering missing detections, mainly due to the stable interpolation and extrapolation guided by the joints. At the same time, the placement of body detections of DPM show an unstable behavior, making it difficult to revoke wrong detections with the proposed affinities.

4.4. Sliding window size

We investigate the impact of different sliding window sizes on the MOT17 training set. As introduced in Sec. 3.1, the size of the sliding window $\triangle t$ depends on $\triangle t_1$ and $\triangle t_2$ as it is the sum of these two terms: $\triangle t = \triangle t_1 + \triangle t_2$. We keep $\triangle t_1 = 4$ and change $\triangle t_2$ from 1 to 5 frames. Fig. 8 shows the corresponding MOTA values. The MOTA scores increase and reach the maximum at a window size of 3 frames for all three detectors. A major decrease happens at a window size of 5. This means that the added joint feature is robust in small temporal distance and the tracking system benefits from the near-online data association.

4.5. Evaluation on MOT17 test

Finally, the results of our proposed tracker on the MOT17 test dataset are shown in Table. 4 together with the best 10 performing methods.

As we can see, our method outperforms other approaches

Accuracy(%)	Right ankle	Right knee	Right hip	Right wrist	Right elbow	Right shoulder
feature (5)	86.67	87.80	88.57	88.14	87.16	87.92
feature (6)	88.05	89.63	90.49	89.70	88.36	89.86

Table 3. Training accuracies of the joint-to-joint temporal classifiers.

Method	Type	MOTA↑	IDF1↑	MT↑	ML↓	FP↓	FN↓	IDS↓	FM↓
Ours	NO	**52.6**	50.8	19.7	35.8	31,572	**232,659**	3,050	3,792
eHAF17 [31]	B	51.8	54.7	**23.4**	37.9	33,212	236,772	1,834	2,739
FWT [13]	B	51.3	47.6	21.4	35.2	24,101	247,921	2,648	4,279
jCC [18]	B	51.2	54.5	20.9	37.0	25,937	247,822	1,802	2,984
MOTDT17 [24]	O	50.9	52.7	17.5	35.7	24,069	250,768	2,474	5,317
MHT_DAM [19]	B	50.7	47.2	20.8	36.9	22,875	252,889	2,314	2,865
TLMHT [30]	B	50.6	**56.5**	17.6	43.4	22,213	255,030	**1,407**	**2,079**
EDMT17 [6]	B	50.0	51.3	21.6	36.3	32,279	247,297	2,264	3,260
MTDF17 [12]	O	49.6	45.2	18.9	**33.1**	37,124	241,768	5,567	9,260
HAM_SADF17 [39]	O	48.3	51.1	17.1	41.7	**20,967**	269,038	3,410	6,351

Table 4. Results for our method on the MOT17 test set. The ten best performing trackers of MOT17 (accessed on 22/03/2019) are listed. The best values are in bold. Second column: method type with O standing for online, NO for near-online and B for batch.

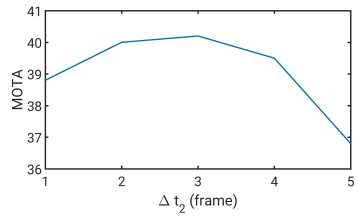
Detectors	Features	MOTA	MT	ML	FP	FN	IDs
DPM	Body	23.9	35	263	**12381**	70780	839
DPM	Body+Joint	**36.8**	**103**	**210**	13346	**55734**	**710**
F-RCNN	Body	49.6	**135**	121	3952	**50652**	1086
F-RCNN	Body+Joint	**51.4**	115	**180**	**1590**	51634	**428**
SDP	Body	62.9	**190**	**128**	3563	**34510**	1883
SDP	Body+Joint	**63.1**	168	134	**1475**	38732	**528**

Table 5. Feature evaluation on the MOT17 training sequences.

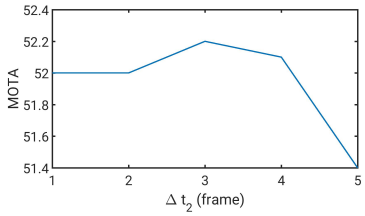
with the best MOTA score and the lowest number of false negatives (FN) on the benchmark. Also the mostly lost (ML) metric shows that our method achieves better or comparable results comparing with most of these approaches. This indicates that many positions have been recovered due to the help of joint detections via interpolation and extrapolation. As our tracker utilizes only a few frames, the IDS score is relative high and IDF1 a bit worse than some other methods. However, comparing with the other online methods, the number of track fragmentations (FM) is considerable lower. This indicates that using a sliding window yields more robust trajectories than only considering the information from the current frame.

5. Conclusions

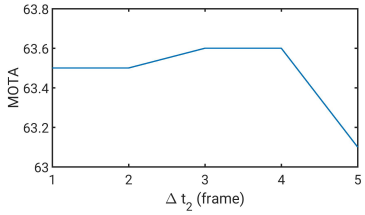
The integration of joint detections to a MOT framework is a difficult task. In this work we proposed suitable affinities that allowed to consider body detections together with joint detections for the purpose of multi object tracking. Further, we embedded these information into an efficient near-online tracking framework. In our experiments, we validated several possible affinities. Using the best performing features, we investigated the impact of adding joint



(a) DPM



(b) FRCNN



(c) SDP

Figure 8. Impact of the sliding window on the MOTA score.

detections, which showed a clear improvement over staying in the traditional tracking-by-detection paradigm, even with hand-crafted features. Finally, the experiments showed that the proposed system outperforms state-of-the-art on the challenging MOT17 benchmark.

References

[1] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese. Social lstm: Human trajectory prediction in crowded spaces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[2] A. Alahi, V. Ramanathan, K. Goel, A. Robicquet, A. A. Sadeghian, L. Fei-Fei, and S. Savarese. Learning to predict human behavior in crowded scenes. In *Group and Crowd Behavior for Computer Vision*, pages 183–207. Elsevier, 2017.

[3] M. Andriluka, U. Iqbal, E. Ensafutdinov, L. Pishchulin, A. Milan, J. Gall, and B. Schiele. PoseTrack: A benchmark for human pose estimation and tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

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

[5] V. Chari, S. Lacoste-Julien, I. Laptev, and J. Sivic. On pairwise costs for network flow multi-object tracking. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[6] J. Chen, H. Sheng, Y. Zhang, and Z. Xiong. Enhancing detection model for multiple hypothesis tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2017.

[7] W. Choi. Near-online multi-target tracking with aggregated local flow descriptor. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015.

[8] Q. Chu, W. Ouyang, H. Li, X. Wang, B. Liu, and N. Yu. Online multi-object tracking using cnn-based single object tracker with spatial-temporal attention mechanism. In *Proceedings of the IEEE Conference on Computer Vision (ICCV)*, 2017.

[9] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874, 2008.

[10] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645, 2010.

[11] M. Fenzi, J. Ostermann, N. Mentzer, G. Payá-Vayá, H. Blume, T. N. Nguyen, and T. Risse. ASEV-automatic situation assessment for event-driven video analysis. In *Proceedings of the IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, 2014.

[12] Z. Fu, F. Angelini, J. Chambers, and S. M. Naqvi. Multi-level cooperative fusion of gm-phd filters for online multiple human tracking. *IEEE Transactions on Multimedia*, 2019.

[13] R. Henschel, L. Leal-Taixé, D. Cremers, and B. Rosenhahn. Fusion of head and full-body detectors for multi-object tracking. In *CVPR Workshop on Joint Detection, Tracking, and Prediction in the Wild (CVPRW)*, 2018.

[14] R. Henschel, L. Leal-Taixé, and B. Rosenhahn. Efficient multiple people tracking using minimum cost arborescences. In *German Conference on Pattern Recognition (GCPR)*, 2014.

[15] R. Henschel, L. Leal-Taixé, B. Rosenhahn, and K. Schindler. Tracking with multi-level features. *arXiv preprint arXiv:1607.07304*, 2016.

[16] E. Insafutdinov, M. Andriluka, L. Pishchulin, S. Tang, E. Levinkov, B. Andres, and B. Schiele. Arttrack: Articulated multi-person tracking in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[17] U. Iqbal, A. Milan, and J. Gall. Posetrack: Joint multi-person pose estimation and tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[18] M. Keuper, S. Tang, B. Andres, T. Brox, and B. Schiele. Motion segmentation & multiple object tracking by correlation co-clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2018.

[19] C. Kim, F. Li, A. Ciptadi, and J. M. Rehg. Multiple hypothesis tracking revisited. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015.

[20] C. Kim, F. Li, and J. M. Rehg. Multi-object tracking with neural gating using bilinear lstm. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[21] L. Leal-Taixé, C. Canton-Ferrer, and K. Schindler. Learning by tracking: Siamese cnn for robust target association. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2016.

[22] L. Leal-Taixé, G. Pons-Moll, and B. Rosenhahn. Everybody needs somebody: Modeling social and grouping behavior on a linear programming multiple people tracker. In *ICCV Workshop on Modeling, Simulation and Visual Analysis of Large Crowds (ICCVW)*, 2011.

[23] Y. Li, C. Huang, and R. Nevatia. Learning to associate: Hybridboosted multi-target tracker for crowded scene. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[24] C. Long, A. Haizhou, Z. Zijie, and S. Chong. Real-time multiple people tracking with deeply learned candidate selection and person re-identification. In *ICME*, 2018.

[25] A. Milan, L. Leal-Taixé, I. Reid, S. Roth, and K. Schindler. MOT16: A benchmark for multi-object tracking. *arXiv:1603.00831 [cs]*, Mar. 2016. arXiv: 1603.00831.

[26] L. Ren, J. Lu, Z. Wang, Q. Tian, and J. Zhou. Collaborative deep reinforcement learning for multi-object tracking. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[27] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems (NIPS)*, 2015.

[28] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *European Conference on Computer Vision Workshop on Benchmarking Multi-Target Tracking (ECCVW)*, 2016.

[29] E. Ristani and C. Tomasi. Features for multi-target multi-camera tracking and re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[30] H. Sheng, J. Chen, Y. Zhang, W. Ke, Z. Xiong, and J. Yu. Iterative multiple hypothesis tracking with tracklet-level association. *IEEE Transactions on Circuits and Systems for Video Technology*, 2018.

[31] H. Sheng, Y. Zhang, J. Chen, Z. Xiong, and J. Zhang. Heterogeneous association graph fusion for target association in multiple object tracking. *IEEE Transactions on Circuits and Systems for Video Technology*, 2018.

[32] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[33] S. Tang, B. Andres, M. Andriluka, and B. Schiele. Multi-person tracking by multicut and deep matching. In *ECCV Workshop on Benchmarking Multi-Target Tracking (ECCVW)*, 2016.

[34] S. Tang, M. Andriluka, B. Andres, and B. Schiele. Multiple people tracking by lifted multicut and person reidentification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[35] P. Weinzaepfel, J. Revaud, Z. Harchaoui, and C. Schmid. DeepFlow: Large displacement optical flow with deep matching. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2013.

[36] B. Xiao, H. Wu, and Y. Wei. Simple baselines for human pose estimation and tracking. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[37] F. Yang, W. Choi, and Y. Lin. Exploit all the layers: Fast and accurate cnn object detector with scale dependent pooling and cascaded rejection classifiers. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2016.

[38] S. Yi, H. Li, and X. Wang. Understanding pedestrian behaviors from stationary crowd groups. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[39] Y. Yoon, A. Boragule, Y. Song, K. Yoon, and M. Jeon. Online multi-object tracking with historical appearance matching and scene adaptive detection filtering. In *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–6, Nov 2018.

[40] L. Zhang, Y. Li, and R. Nevatia. Global data association for multi-object tracking using network flows. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.

[41] J. Zhu, H. Yang, N. Liu, M. Kim, W. Zhang, and M.-H. Yang. Online multi-object tracking with dual matching attention networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.