Multiple People Tracking by Lifted Multicut and Person Re-identification

Siyu Tang[1,2*] Mykhaylo Andriluka[1] Bjoern Andres[1] Bernt Schiele[1]
[1]Max Planck Institute for Informatics, Saarbrücken, Germany
[2]Max Planck Institute for Intelligent Systems, Tübingen, Germany

Abstract

Tracking multiple persons in a monocular video of a crowded scene is a challenging task. Humans can master it even if they loose track of a person locally by re-identifying the same person based on their appearance. Care must be taken across long distances, as similar-looking persons need not be identical. In this work, we propose a novel graph-based formulation that links and clusters person hypotheses over time by solving an instance of a minimum cost lifted multicut problem. Our model generalizes previous works by introducing a mechanism for adding long-range attractive connections between nodes in the graph without modifying the original set of feasible solutions. This allows us to reward tracks that assign detections of similar appearance to the same person in a way that does not introduce implausible solutions. To effectively match hypotheses over longer temporal gaps we develop new deep architectures for re-identification of people. They combine holistic representations extracted with deep networks and body pose layout obtained with a state-of-the-art pose estimation model. We demonstrate the effectiveness of our formulation by reporting a new state-of-the-art for the MOT16 benchmark. The code and pre-trained models are publicly available[1].

1. Introduction

Multiple people tracking has improved considerably in the last two years, driven also by the MOT challenges [18, 20]. One trend in this area of research is to develop CNN-based feature representations for people appearance to effectively model relations between detections [14, 17]. This trend has two advantages: Firstly, representations of people appearance can be learned for varying camera positions and motion, a goal less easy to achieve with simple motion models, especially for monocular video due to the complexity of motion under perspective projection. Secondly, appearance facilitates the re-identification of people across long distances, unlike motion models that become asymptotically uncorrelated. Yet, incorporating long-range re-identification into algorithms for tracking remains challenging. One reason is the simple fact that similar looking people are not necessarily identical. To address these challenges, in this paper, we generalize the mathematical model of [28] so as to express the fact that similar looking people are considered as the same person only if they are connected by at least one feasible track (possibly skipping occlusion). More specifically, every detection is represented by a node in a graph; edges connect detections within and across time frames, and costs assigned to the edges can be positive, to encourage the incident nodes to be in the same track, or negative, to encourage the incident nodes to be in distinct tracks. Such mathematical abstraction has several advantages. Firstly, the number of persons is not fixed or biased by the definition of the problem, but is estimated in an unbiased fashion from the video sequence and is determined by the solution of the problem. Secondly, multiple detections of the same person in the same frame are effectively clustered, which eliminates the need for heuristic non-maxima suppression. In order to avoid that distinct but similar looking people are assigned to the same track, a distinction must be made between the edges that define possible connections (i.e., a feasible set) and the edges that define the costs or rewards for assigning

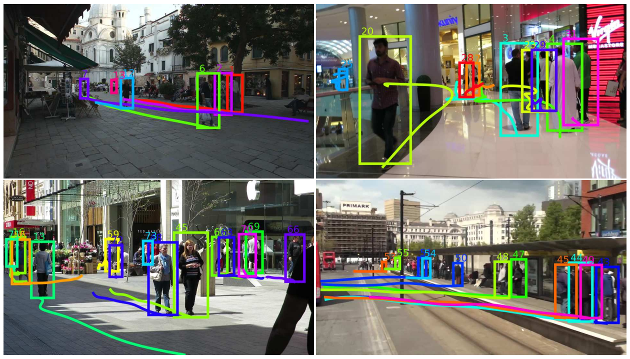


Figure 1. Qualitative results on the MOT16 benchmark. The solid line under each bounding box indicates the life time of the track. The lifted multicut tracking model is able to link people through occlusions and produces persistent long-lived tracks.

[*] This work was done while S. Tang was at Max Planck Institute for Informatics.
[1] http://mpi-inf.mpg.de/multicut_tracking

the incident nodes to distinct tracks (i.e., an objective function). We achieve this, while maintaining the advantages of [28], by casting the multi-person tracking problem as a minimum cost *lifted* multicut problem [1].

Specifically, we make three contributions:

Firstly, we design and train deep networks for re-identifying persons by fusing human pose information. This provides a mechanism for associating person hypotheses that are temporally distant and allows to obtain correspondence before and after occlusion.

Secondly, we propose to cast multi-person tracking as the *minimum cost lifted multicut* problem. We introduce two types of edges (regular and lifted edges) for the tracking graph. The regular edges define the set of feasible solutions in the graph, namely, which pair of nodes can be joint/cut. The lifted edges add additional long range information to the objective on which node should be joint/cut without modifying the set of feasible solutions. Our formulation encodes long-range information, yet penalizes long-term false joints (e.g., similar looking people) by forcing valid paths in the feasible solution in a unified and rigorous manner.

Thirdly, we show that the tracks defined by local optima of this optimization problem define a new state-of-the-art for the MOT16 benchmark [20].

Related Work. Recent works on multi-person tracking focus on the tracking-by-detection approach [21, 36, 35, 29, 30]. Tracking is performed either directly on people detections [21, 23, 34], or on a set of confident tracklets, which are obtained by first grouping detections [6, 27, 33]. Introducing tracklets can reduce the state space; however, such approaches need a separate tracklet generation step, and any mistakes introduced by the tracklet generation are likely to be propagated to the final solution. In this work, our model takes detection as input. As the detections are clustered jointly in space and time, our model is able to handle multiple detection hypotheses of the same target on each frame.

One common formulation for multi-person tracking are network flow-based methods [3, 7, 31]. [3] proposes to model all potential locations over time and find trajectories that produce the minimum cost. [31] extends the work [3] to track interacting objects simultaneously by using intertwined flow and imposing linear flow constraints. [23] shows that their network flow formulation can be solved in polynomial time by a successive shortest path algorithm. A maximum weight independent set formulation followed by hierarchical merging and linking is proposed for the tracking task in [5].

Recently, minimum cost multicut formulation has been proposed to address multi person tracking [13, 27, 28, 25, 15]. [27, 28] propose to jointly cluster detections over space and time. The optimal number of people as well as the cluster of each person are obtained by partitioning the graph with attractive and repulsive terms. [15] proposes to partition the detection graph by considering point tracks, speed, appearance and trajectory straightness. The optimization is performed by a combination of message passing and move-making algorithms. [25] proposes to solve the minimum cost multicut problem by a multi-stage cascade with a temporal sliding window. Our work is different from the previous multicut based works; our lifted multicut formulation introduces additional edges in the graph to incorporate long-range information into the tracking formulation.

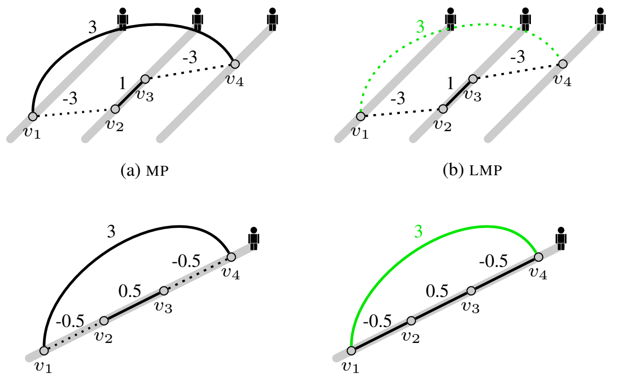
Many works have been proposed to exploit appearance information. [14] proposes a target-specific appearance model which integrates long-term information and utilizes features from a generic deep convolutional neural network. [34] proposes to formulate tracking as a Markov decision process with a policy estimated on the labeled training data and presents novel appearance representations that rely on the temporal evolution in appearance of the tracked target. Recently, [17] proposes to model the similarity between pairs of detections by CNNs. Several architectures have been explored and they present similar findings to our work, that forming a stacked input to CNNs performs the best. Our work additionally incorporates human pose information, which improves the similarity measures by a notable margin.

There are several multi person tracking works that aim to recover people tracks by incorporating longer-range connections between detection hypotheses [35, 21, 7, 33]. [21] employs a simple color appearance model and proposes a continuous formulation, where mutual occlusions, dynamics and long-range trajectory continuity are effectively modeled. [35] proposes a generalized minimum clique formulation which is solved by a greedy iterative optimization scheme that finds one track at a time. In [7], their target appearance model is learned online, and it relies on a heuristic procedure to determine which track segment is valid and the creation/termination of tracks. [33] relies on first grouping detections into tracklets, and then in the subsequent stage into long-range tracks with a greedy heuristic approach. In our approach, frame-to-frame and long-range similarity is incorporated into the objective function in a unified manner.

2. Model

We now turn to our mathematical abstraction of multiple people tracking as a minimum cost lifted multicut problem (LMP). The LMP is an optimization problem whose feasible solutions can be identified with decomposition of a graph. The minimum cost multicut problem (MP) [28] is defined w.r.t. a graph whose edges define possibilities of joining nodes directly into the same track. The LMP is defined, in addition, w.r.t. additional *lifted* edges that do not define possibilities of directly joining nodes.

Our motivation for modeling the *lifted* edges comes from the simple fact that persons of similar appearance are not necessarily identical. Given two detections that are far apart in time and similar in appearance, it is more likely that they



(a) MP

(b) LMP

(c) MP

(d) LMP

Figure 2. Comparison between Multicut Problem (MP) and Lifted Multicut Problem (LMP). Ground truth track of each person is depicted in gray. Regular edges are depicted in black, lifted edges are in green. Solid lines indicate joints, dotted lines indicate cuts. Costs of cutting edges are indicated by the numbers on the corresponding edges. (Best view in color)

represent the same person. At the same time, this decision has to be certified a posteriori by a track connecting the two. We achieve precisely this by introducing two classes of edges: *regular* edges and *lifted* edges. In order to assign two detections that are far apart in time and similar in appearance to the same cluster (person), there must exist a path (track) along the *regular* edges, that certifies this decision.

Two intuitive examples are given in Fig. 2. In (**a**) and (**b**) there are three persons in the scene, v_1 is the detection on the first person, v_2 and v_3 are the detections on the second, v_4 is on the third. The costs on the edges v_1v_2 and v_3v_4 are -3, suggesting strong rewards towards cutting the edges, and this is correct. However, the cost on the edge v_1v_4 suggests that the first and the third person look similar and introduces a strong reward towards connecting them. As a result, the MP incorrectly connects v_1 and v_4 as the same person; the LMP does not connect v_1 and v_4, as such long-range join is not supported by the local edges. (**c**) and (**d**) is another example where all the detections are on the same person, namely, a track that connects all the nodes in the graph is desirable. Due to partial occlusion or inaccurate bounding box localization, the costs on the local edges v_1v_2 and v_3v_4 could be ambiguous, sometimes even reverse. The long-range edge v_1v_4 correctly re-identifies the person. The MP, however, produces two clusters for a single person because the long-range edge does not introduce additional constraints on the local connections. In contrast, the LMP allows us to influence the entire chain of connections between person hypotheses with a single confident long-range observation.

In the following, we discuss in detail first the parameters, then the feasible set, and finally the objective function.

Parameters. Given an image sequence, we consider an instance of the LMP with respect to the parameters defined below. The estimation of these parameters from the image sequence is discussed in the next section.

- A finite set V in which every element $v \in V$ represents a *detection* of one person in one image, i.e., a bounding box. For every detection $v \in V$, we also define its height $h_v \in \mathbb{R}^+$, the image coordinates $x_v, y_v \in \mathbb{R}^+$ of its center and its frame number $t_v \in \mathbb{N}$.
- For every pair $v, w \in V$: a conditional probability $p_{vw} \in (0,1)$ of v and w to represent distinct persons, given their height, coordinates and appearance.
- A graph $G = (V, E)$ whose edges are regular edges that connect detections v, w in the same image $t_v = t_w$ and also connect detections v, w in distinct images $t_v \neq t_w$ that are *close in time*, i.e., for some fixed upper bounds $\delta_t \in \mathbb{N} : |t_v - t_w| \leq \delta_t$.
- A graph $G' = (V, E')$ with $E \subseteq E'$ whose additional edges $\{v, w\} \in E' \setminus E$ are lifted edges which connect detections v, w that are *far apart in time* and *similar in appearance*, i.e., for some fixed $p_0 \in (0, \frac{1}{2})$: $|t_v - t_w| > \delta_t$ and $p_{vw} \leq p_0$.

The graph G defines the decomposition space, and the graph G' adds *lifted* edges $E' \setminus E$ on top of G and defines the structure of the cost function. The *lifted* edges are introduced for the detections that are *far apart in time* and *similar in appearance*, because such pair of detections potentially indicates the same person that reappears after long-term occlusion.

Feasible Set. The feasible solutions of the LMP can be identified with the decomposition (clusterings) of the graph G. Here, in the context of tracking, every component (cluster) of detections defines a track of one person. It is therefore reasonable to think of our approach as tracking by clustering.

Formally, any feasible solution of the LMP is a 01-vector $x \in \{0,1\}^{E'}$ in which $x_{vw} = 1$ indicates that the nodes v and w are in distinct components. In order to ensure that x well-defines a decomposition of G, it is further constrained to the set $X_{GG'} \subseteq \{0,1\}^{E'}$ of those $x \in \{0,1\}^{E'}$ that satisfy the system of linear inequalities written below.

$$\forall C \in \text{cycles}(G)\, \forall e \in C :$$

$$x_e \leq \sum_{e' \in C \setminus \{e\}} x_{e'} \tag{1}$$

$$\forall vw \in E' \setminus E\, \forall P \in vw\text{-paths}(G) :$$

$$x_{vw} \leq \sum_{e \in P} x_e \tag{2}$$

$$\forall vw \in E' \setminus E\, \forall C \in vw\text{-cuts}(G) :$$

$$1 - x_{vw} \leq \sum_{e \in C} (1 - x_e) \tag{3}$$

The constraints (1) are generalized transitivity constraints which mean: For any neighboring nodes v and w, if there exists a path from v to w in G along which all edges are labeled as 0, then the edge vw can only be labeled as 0. The constraints (2) and (3) guarantee, for every feasible solution and every lifted edge $vw \in E' \setminus E$, that the label x_{vw} of this edge is 0 (indicating that v and w belong to the same track) if (2) and only if (3) v and w are connected in the smaller graph G by a path of edges labeled 0. By assigning a cost or reward $c_{vw} \in \mathbb{R}$ to a lifted edge $vw \in E' \setminus E$, we can thus assign this cost or reward precisely to those feasible solutions for which v and w belong to distinct tracks, *without* introducing the additional possibility of joining v and w directly.

Objective function. We consider instances of the LMP of the form

$$\min_{x \in X_{GG'}} \sum_{e \in E'} c_e x_e \tag{4}$$

with the costs c_e defined as

$$c_e = \log \frac{1 - p_e}{p_e} \quad . \tag{5}$$

The objective function is chosen such that solutions are decompositions of G into tracks that maximize the probability of detections representing the same or distinct persons. More specifically, we define p_e as a logistic form:

$$p_e := \frac{1}{1 + \exp(-\langle \theta_\gamma, f^{(e)} \rangle)}. \tag{6}$$

Then the cost c_e has the form:

$$c_e := \log \frac{1 - p_e}{p_e} = -\langle \theta_\gamma, f^{(e)} \rangle \quad . \tag{7}$$

The model parameter θ_γ is estimated on the training set by means of logistic regression. γ is the length of temporal interval between pair of detections. We estimate a separate set of edge-cost parameters θ_γ for each temporal interval between the detections. The feature $f^{(e)}$ describes the similarity between detections. In this work, $f^{(e)}$ is defined as a combination of person re-identification confidence (Sec. 3), deep correspondence matching, and spatio-temporal relations, which is discussed in Sec. 4

Optimization. The minimum cost lifted multicut problem defined by (4) is APX-hard [8]. Given the size of instances of our tracking problems, solving to optimality or within tight bounds using branch and cut is beyond feasibility. In this work, we exploit a primal heuristic proposed by [12], where the bi-partitions of a subgraph are updated by a set of sequences of transformations. The update has the worst-case complexity of $O(|V||E|)$ which is almost never reached in practice. Detailed run time analysis can be found in [12].

3. Person Re-identification for Tracking

Traditionally, person re-identification is the task to associate observed pedestrians in non-overlapping camera views. In the context of multi-person tracking, linking the detected pedestrians across the whole video can be viewed as re-identification with special challenges: occlusions, cluttered background, large difference in image resolution and inaccurate bounding box localization. In this section, we investigate several CNN architectures for re-identification for the multi-person tracking task. Our basic CNN architecture is VGG-16 Net [26]. Particularly, we propose a novel person re-identification model that combines the body pose layout obtained with state-of-the-art pose estimation methods.

Data Collection. One of the key ingredients of deep CNNs is the availability of large amounts of training data. To apply re-identification to tracking, we collect images from the MOT15 benchmark [18] training set and 5 sequences of the MOT16 benchmark [20] training set. We also collect person identity examples from the CUHK03 [19], Market-1501 [37] datasets that are captured by 6 surveillance cameras. We use the MOT16-02 and MOT16-11 sequences from the MOT16 training set as test sets. Overall a total of 2511 identities is used for training and 123 identities for testing.

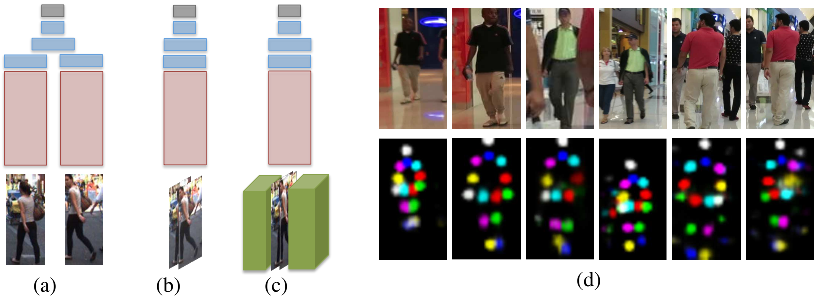
3.1. Architectures

In this work, we explore three architectures, namely ID-Net, SiameseNet, and StackNet.

ID-Net. We first learn a VGG net Φ to recognize $N = 2511$ unique identities from our data collection as a N-way classification problem. We re-size the training images to $112 \times 224 \times 3$. Each image $x_i, i = 1, ..., M$ associates to a ground truth identity label $y_i \in \{1, ..., N\}$. The VGG estimates the probability of each image being each label as $p_i = \Phi(x_i)$ by a forward pass. The network is trained by the softmax loss.

During testing, given an image from unseen identities, the final softmax layer is removed and the output of the fully-connected layer Φ_{f7} is used as the identity feature. Given a pair of images, the Euclidean distance between the two identity features can be used to decide whether the pair contains the same identity. In the experiments we observe that this identity feature already provides good accuracy. However, the performance is boosted by turning to a Siamese architecture and a StackNet, explained next.

SiameseNet. A Siamese architecture means the network contains two symmetry CNNs which share the parameters. We start with a commonly used Siamese architecture as shown in Fig. 3(a). To model the similarity we use fully connected layers on top of the twin CNNs. More specifically, the features $FC_6(x_i)$ and $FC_6(x_j)$ from a pair of images are extracted from the first fully-connected layer of the VGG-based Siamese network that shares the weights. Then the features


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

Model	Acc. (%)
ID-Net	80.4
SiameseNet	84.7
StackNet	86.9
StackNetPose	90.0

(e)

Figure 3. (a) SiameseNet. (b) StackNet. (c) StackNetPose. Red rectangles indicate the convolutional, relu and pooling layers of VGG16. Blue rectangles indicate the fully-connected layers. Grey rectangles on the top of each network are the loss layers. Green boxes are the stacked body part score maps. (d) Example results from StackNetPose. (e) Comparison of the person re-identification models.

are concatenated and transformed by two fully-connected layers (FC_7, FC_8), where FC_7 are followed by a ReLU non-linearity. FC_8 uses a softmax function to produce a probability estimation over a binary decision, namely the same identity or different identities.

StackNet. The most effective architecture we explored is the StackNet, where we stack a pair of images together along the RGB channel. The input to the network becomes $112 \times 224 \times 6$. Then the filter size of the first convolutional layer is changed from $3 \times 3 \times 3$ to $3 \times 3 \times 6$, and for the rest of the network we follow the VGG architecture. The last fully-connected layer models a 2-way classification problem, namely the same identity or different identities. During testing, given a pair of images, both SiameseNet and StackNet produce the probability of the pair being the same/different identities by a forward pass.

The StackNet allows a pair of images to communicate at the early stage of the network, but it is still limited by the lack of ability to incorporate body part correspondence between the images. Next, we propose a body part fusing method to explicitly allow modeling the semantic body part information within the network.

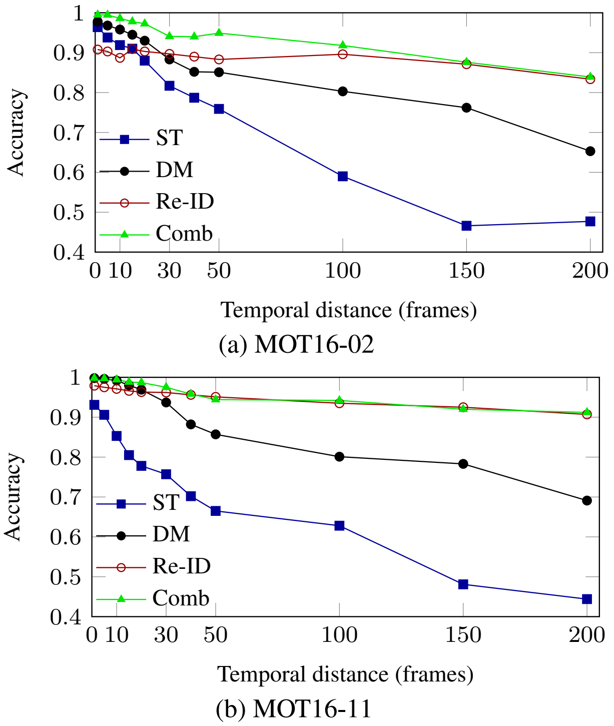
3.2. Fusing Body Part Information

A desirable property of the network is to localize the corresponding regions of the body parts, and to reason about the similarity of a pair of pedestrian images based on the localized regions and the full images. We implement such model by fusing body part detections into the CNN. More specifically, we utilize the body part detector [24] to produce individual score maps for 14 body parts, namely, head, shoulders, elbows, wrists, hips, knees, and ankles, each with left/right symmetry body parts except the head which is indicated by head top and head bottom. We combine the score maps from every two symmetry body parts which results in 7 scores maps; each has the same size as the input image. We stack the pair of images as well as the 14 score maps together to form a $112 \times 224 \times 20$ input volume. Now the filter size of the first convolutional layer is set as $3 \times 3 \times 20$, and the rest of the network follows the VGG16 architecture with a 2-way classification layer in the end. In Fig. 3(d) we show several examples of estimated body poses on our dataset. Note that augmenting the network with body layout information can be interpreted as an attention mechanism that allows us to focus on the relevant part on the input image. It can also be seen as a mechanism to highlight the foreground and to enable the network to establish corresponding regions between input images.

3.3. Experimental Analysis

Training. Our implementation is based on the Caffe deep learning framework [11]. To learn the ID-Net, our VGG model is pre-trained on the ImageNet Classification task. Following a common practice in face recognition/verfication literature [22], we use our ID-Net as initialization for learning the SiameseNet, StackNet and StackNetPose, which makes the training faster and produces better results.

Setup. We have 123 person identities as test examples which are collected from MOT16-02 and MOT16-11. More specifically, on these two sequences, detections that are considered as true positives for a certain identity are those whose intersection-over-union with the ground truth of the identity are larger than 0.5. Given the true positive detections for all the identities, we randomly select 1,000 positive pairs from the detections assigned to the same identity and 4000 negative pairs from the detections assigned to different identities as our test set. A larger ratio of negative pairs in the test set is to simulate the positive/negative distribution during the tracking. For every test pair, we estimate the probability of the pair of images containing the same person. For the positive (negative) pairs, if the estimated probabilities are larger (smaller) than 0.5, they are considered as correctly classified examples. The metric is the verification accuracy, the ratio of correctly classified pairs. For the ID-Net, the verification result of pairs of images is obtained by testing whether the distance between the extracted features is smaller than a threshold. The threshold is obtained on a separate validation data to maximize the verification accuracy.



(a) MOT16-02

(b) MOT16-11

Figure 4. Accuracy of pairwise affinity measures on the MOT16-02 **(a)** and the MOT16-11 **(b)** sequences.

Results. It can be seen from Fig. 3(e) that the l^2 distance of the Φ_{f7} features from the ID-Net already produces reasonable accuracy. The performance is improved by applying the SiameseNet, from 80.4% to 84.7%. The accuracy is further improved when using the StackNet, achieving 86.9% accuracy. Fusing the body part information (StackNetPose) outperforms all other models by a large margin, achieving 90.0% accuracy. For our tracking task, we use the StackNetPose model to generate person re-identification confidence. We show three pairs of detections that are correctly estimated by StackNetPose in Fig. 3(d). It can be seen that the body part maps enable the network to localize the person despite the inaccurate bounding boxes (the first/second pairs) and cluttered background (the third pair).

4. Pairwise Potentials

As discussed in Sec. 2, the cost c_e in the objective function (4) is defined as $c_e = -\langle\theta_\gamma, f^{(e)}\rangle$. In this section, we introduce the feature $f^{(e)}$, which is based on three information sources: spatio-temporal relations (ST), dense correspondence matching (DM) and person re-identification confidence (Re-ID) that is described in the previous section.

ST. The spatio-temporal relation based feature is commonly used in many multi-person tracking works [23, 34, 6], as it is a good affinity measure for pairs of detections that are in close proximity. Given two detections v and w, each has spatio-temporal locations (x, y, t) and height h. The ST feature is defined as $f_{st} = \frac{\sqrt{(x_v - x_w)^2 + (y_v - y_w)^2}}{\bar{h}}$, where $\bar{h} = \frac{(h_v + h_w)}{2}$. Intuitively, the ST features are able to provide useful information within a short temporal window. They model the geometric relations between bounding boxes but do not take image content into account.

DM. DeepMatching [32] is introduced as a powerful pairwise affinity for multi-person tracking by [28]. We apply it in this work as well. Given two detections v and w, each has a set of matched keypoints M. We define $MU = |M_v \cup M_w|$, and $MI = |M_v \cap M_w|$ between the set M_v and M_w. Then the pairwise feature between the two detections is defined as $f_{dm} = MI/MU$.

Re-ID. The DM feature is based on local image patch matching, which makes it robust to irregular camera motion and to partial occlusion in short temporal distance. As shown in [28] and in the experiment section of our work, the performance of the DM feature drops dramatically when increasing temporal distance. ReID is explicitly trained for the task of person re-identification. It is robust with respect to large temporal and spatial distance and allows long-range association. In this work, we utilize our deep re-identification model (StackNetPose) for modeling the long-range connections. Our final pairwise feature $f^{(e)}$ is defined as $(f_{st}, f_{dm}, f_{reID}, \xi_{min}, f_{st}^2, f_{st} \cdot f_{dm}, \ldots, \xi_{min}^2)$, where ξ_{min} is the lower detection confidence within the pair, and f_{reID} is the probability estimated by our StackNetPose. The quadratic terms introduce a non-linear mapping from the feature space to the cost space. In total the pairwise feature has 14 dimensions.

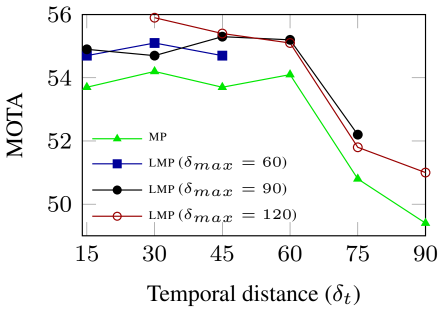
4.1. Experimental Analysis

In this section, we present an analysis of our pairwise features. We also choose MOT16-02 and MOT16-11 from the MOT16 training set for the analysis, as the imaging conditions and camera motion are largely different between these two sequences. The test example collection and the evaluation metric are the same as for evaluating the person re-identification networks, namely for every test pair, we estimate the probability of the pair of images containing the same person. For the positive (negative) pairs, if the estimated probabilities are larger (smaller) than 0.5, they are considered as correctly classified examples. Any bias toward cut or joint decreases the tracking performance. A higher accuracy leads to a better tracking performance. We conduct a comparison between features as a function of temporal distance. we demonstrate long temporal distance (200 frames), as our model is able to incorporate such information.

Results. It can be seen from Fig. 4 that the DM feature achieves good accuracy up to 10 frames, but its performance deteriorates for connections at longer time span. The performance of the ST feature drops quickly after 5 frames. This

MOT16-11					
δ_{max}	30	60	90	120	150
MP	54.2	54.1	49.4	43.9	32.1
LMP	54.5 (+0.3)	55.1 (+1)	55.3 (+5.9)	55.0 (+11.1)	51.1 (+19.1)
MOT16-02					
δ_{max}	30	60	90	120	150
MP	19.9	21.5	21.2	19.1	17.2
LMP	21.3 (+1.4)	22.4 (+0.9)	21.3 (+0.1)	22.3 (+3.2)	19 (+1.8)

(a) Varying δ_{max}



(b) Varying δ_t

Figure 5. Comparison of Multicut model (MP) and Lifted Multicut model (LMP) with different δ_{max} values (a) and different δ_t values (b).

is especially pronounced on the MOT16-11 sequence that has rapid camera motion. In contrast, the Re-ID feature is effective and maintains high accuracy over time. For example on the MOT16-11 sequence the Re-ID (red line) improves over DM (black line) by a notable margin for the temporal distances that are larger than 50 frames. When we combine the three features (Comb, green line in Fig. 4), we obtain the best accuracy at all the temporal distances. The reason is that, at different temporal distance, our combined feature is able to take advantage from different information sources. E.g., when the temporal distance is smaller than 30 frames (1 sec. for these two sequences), the DM and ReID features combine both low-level (local image patch matching) and high-level (person-specific appearance similarity) to produce high accuracy pairwise affinity measures. When the temporal distance increases gradually, the ReID feature becomes more and more informative. However, still adding the ST and DM feature improves the overall accuracy, because they act as a regularizer, that forbids physically impossible associations. Based on these results, we use the combined feature in our tracking experiments.

5. Tracking Experiments and Results

We perform our tracking experiments and compare to prior works on the MOT16 Benchmark [20]. The test set contains 7 sequences, where camera motion, camera angle, and imaging condition are largely different. For each test sequence, the benchmark also provides a training sequence that is captured in the similar setting. Therefore, we learn the model parameter θ_γ (defined in Eq. (7)) for the test sequences on the corresponding training sequences.

For analyzing our tracking models, we use MOT16-02 and MOT16-11 from the training set as the validation sequences, the same as previous sections. The model parameter θ_γ trained on MOT16-02 is used for MOT16-11 and vice versa. To obtain the final tracks from the clusters generated by MP or LMP, we estimate a smoothed trajectory from the detections that belongs to the same cluster, by using the code from [21]. When there are gaps in time due to occlusion or detection failures, we fill in the missing detections along the estimated trajectory. We do not consider any clusters whose size are less than 5 in all the experiments.

Evaluation Metric. We follow the standard CLEAR MOT metrics [4] for evaluating multi-person tracking performance. The metrics includes multiple object tracking accuracy (MOTA), which combines identity switches (IDs), false positives (FP), and false negatives (FN). Beside we also report multiple object tracking precision (MOTP), mostly tracked (MT), mostly lost (ML) and fragmentation (FM).

5.1. Lifted Edges versus Regular Edges

The graph for the lifted multicut (LMP) includes two types of edges: regular edges and lifted edges. The regular edges define the decomposition of the graph. The lifted edges introduce long-range information on which nodes should be joint/cut without modifying the set of feasible solutions. They penalize long-term false joint (e.g. similar looking people) by forcing valid paths in the feasible solution. As shown in Fig. 4, even beyond 50 frames, the accuracy of our pairwise affinity measure is still above 90%, Such good pairwise affinity should be leveraged into the tracking model. However, if we encode them by regular edges, we have 10% chances of making a false joint, such errors directly produce long false-positive tracks. If they are lifted edges, connecting those detections must be certified by the local regular edges. Two intuitive examples are shown in Fig. 2. In this section we perform experimental analysis on the two graph variants: Multicut (MP) and Lifted Multicut (LMP), to validate the effectiveness of the proposed methods. Note that we use the same pairwise feature (Comb. in Fig. 4) for the MP and LMP problems.

Given a tracking instance, intuitively, we would connect detections with regular edges up to a certain temporal distance to overcome potential missing detections due to occlusion. For the further distant detections, we would connect them with lifted edges to incorporate person re-identification information into the model to gain better tracking performance. Following the intuition, our MP is constructed in the

Method	MOTA	MOTP	FAF	MT	ML	FP	FN	ID Sw	Frag	Hz	Detector
CEM [21]	33.2	75.8	1.2	7.8%	54.4%	6837	114322	642	731	0.3	Public
TBD [10]	33.7	76.5	**1.0**	7.2%	54.2%	5804	112587	2418	2252	1.3	Public
LTTSC-CRF [16]	37.6	75.9	2.0	9.6%	55.2%	11,969	101,343	481	1,012	0.6	Public
OVBT [2]	38.4	75.4	1.9	7.5%	47.3%	11,517	99,463	1,321	2,140	0.3	Public
LINF1 [9]	41.0	74.8	1.3	11.6%	51.3%	7896	99224	430	963	**4.2**	Public
MHT [14]	42.9	76.6	**1.0**	13.6%	46.9%	**5668**	97919	499	659	0.8	Public
NOMT[6]	46.4	76.6	1.6	**18.3%**	41.4%	9753	87565	**359**	**504**	2.6	Public
Multicut [28]	46.3	75.7	1.1	15.5%	**39.7%**	6373	90914	657	1114	0.8	Public
Lifted Multicut (LMP)	**48.8**	**79.0**	1.1	18.2%	40.1%	6654	**86245**	481	595	0.5	Public

Table 1. Tracking Performance on the MOT16 test set. Best in bold, second best in blue.

way that besides having the regular edges between neighboring frames, we also introduce regular edges between all pairs of detections whose temporal distance are up to δ_{max}. The LMP has a combination of regular edges and lifted edges, we denote the temporal distance where we start to change the regular edges to the lifted edges as δ_t.

Varying δ_{max}. In our first analysis, we gradually change the value of δ_{max} from 1 to 150 frames. As shown in Fig. 5(a), on the MOT16-11 sequence, the MP achieves competitive MOTA (54.2%) when δ_{max} equals 30 frames, but the performance decreases significantly when δ_{max} is increased to 150 frames (5 sec on the MOT16-11). The reason is that the long-range regular edges change the feasible set of the MP. Although the accuracy of the pairwise affinity at 150 frames is near 90%, the model can still make catastrophic false joint, which introduces long-term false positive tracks. Similar results are obtained on the MOT16-02 sequence, MOTA drops to 17.2% when $\delta_{max} = 150$.

For the LMP, we also change δ_{max} from 1 to 150 frames and we set $\delta_t = \delta_{max}/2$. Comparing to the MP, the LMP obtains the best MOTA on the MOT16-11 sequence (55.3%) as well as on the MOT16-02 sequence (22.4%). Moreover, it presents a superior performance in all the settings. Particularly for the long-range connections, the margin between the MP and the LMP is more than 10% on the MOT16-11 sequence. Note that, these experiment results reveal a very desirable property of the LMP: stability with respect to the range of connections. Given a new tracking instance, due to unknown camera motion and imaging condition, it is not trivial to build a proper graph for the MP. As to the LMP, due to its robustness and stability, we are free to choose any sensible range of connections. In the next experiment, we further reveal the stability of the LMP by varying δ_t.

Varying δ_t. As shown in Fig. 5(b), we evaluate the influence of δ_t on LMP under 3 different δ_{max} settings, namely $\delta_{max} = 60, 90, 120$. As a baseline, the tracking performance of MP with $\delta_{max} = 15, 30, 45, 60, 75, 90$ is also shown in the Fig. 5(b), depicted as the green line. It can be seen that at all the temporal distances, adding lifted edges improves the tracking performance over MP, suggesting that long-range person re-identification information is useful for the tracking task. Furthermore, for the longer temporal distance (e.g. $\delta_{max} = 90$), MOTA of the MP drops significantly (49.4%); however, for the LMP with $\delta_{max} = 90$, MOTA maintains at higher levels for $\delta_t = 15, 30, 45, 60$ (black line), indicating that LMP is also robust to a large range of δ_t. Overall, the results show that our LMP is able to encode long-range information in a more rigorous manner, such that it produces much more stable and robust tracking results.

5.2. Results on the MOT16 Benchmark

Here we present our results on the MOT16 test set. We compare our method with the best published results on the benchmark, including NOMT[6], MHT-DAM [14], OVBT [2],LTTSC-CRF [16], CEM [21], TBD [10] and Multicut [28]. [28] is the most relevant approach comparing to our model, where the deep matching feature is employed and tracking is cast as the minimum cost multicut problem. It can be seen from Tab. 1 that our method establishes a new state-of-the-art performance in terms of MOTA, MOTP and false negative (FN). Comparing to the previous best result, we improve MOTA by 2.4% and MOTP by 3.1%. For FAF, MT, ML and FM, our method achieves the second best performance. The improvement over Multicut [28] demonstrates the advantage of incorporating the long-range person re-identification information with the lifted multicut formulation. The complete metrics and visualization are presented on the MOT16 benchmark website[2].

6. Conclusion

Incorporating long-range information for multi-person tracking is challenging. In this work, we propose to model such long-range information by pose aided deep neural networks. Given the fact that similar looking people are not necessarily identical, we propose a minimum cost lifted multicut formulation where the long-range person re-identification information is encoded in the way that it forces valid paths along the local edges. In the end, we show that the proposed tracking method outperforms previous works on the challenging MOT16 benchmark.

[2] https://motchallenge.net/results/MOT16/

References

[1] B. Andres. Lifting of multicuts. *CoRR*, abs/1503.03791, 2015. 2

[2] Y. Ban, S. Ba, X. Alameda-Pineda, and R. Horaud. Tracking Multiple Persons Based on a Variational Bayesian Model. In *ECCV Workshop on Benchmarking Mutliple Object Tracking*, 2016. 8

[3] J. Berclaz, F. Fleuret, E. Türetken, and P. Fua. Multiple object tracking using k-shortest paths optimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2011. 2

[4] K. Bernardin and R. Stiefelhagen. Evaluating multiple object tracking performance: The CLEAR MOT metrics. *Image and Video Processing*, 2008. 7

[5] W. Brendel, M. Amer, and S. Todorovic. Multiobject tracking as maximum weight independent set. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011. 2

[6] W. Choi. Near-online multi-target tracking with aggregated local flow descriptor. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015. 2, 6, 8

[7] A. Dehghan, Y. Tian, P. H. Torr, and M. Shah. Target identity-aware network flow for online multiple target tracking. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 2

[8] E. D. Demaine, D. Emanuel, A. Fiat, and N. Immorlica. Correlation clustering in general weighted graphs. *Theoretical Computer Science*, 2006. 4

[9] L. Fagot-Bouquet, R. Audigier, Y. Dhome, and F. Lerasle. Improving multi-frame data association with sparse representations for robust near-online multi-object tracking. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016. 8

[10] A. Geiger, M. Lauer, C. Wojek, C. Stiller, and R. Urtasun. 3d traffic scene understanding from movable platforms. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2014. 8

[11] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014. 5

[12] M. Keuper, E. Levinkov, N. Bonneel, G. Lavoue, T. Brox, and B. Andres. Efficient decomposition of image and mesh graphs by lifted multicuts. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015. 4

[13] M. Keuper, S. Tang, Z. Yu, B. Andres, T. Brox, and B. Schiele. A multi-cut formulation for joint segmentation and tracking of multiple objects. In *arXiv:1607.06317*, 2016. 2

[14] 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. 1, 2, 8

[15] R. Kumar, G. Charpiat, and M. Thonnat. Multiple object tracking by efficient graph partitioning. 2014. 2

[16] N. Le, A. Heili, and J.-M. Odobez. Long-term time-sensitive costs for crf-based tracking by detection. In *ECCV Workshop on Benchmarking Mutliple Object Tracking*, 2016. 8

[17] L. Leal-Taixé, C. Canton-Ferrer, and K. Schindler. Learning by tracking: Siamese CNN for robust target association. *arXiv:1604.07866*, 2016. 1, 2

[18] L. Leal-Taixé, A. Milan, I. Reid, S. Roth, and K. Schindler. MOTChallenge 2015: Towards a benchmark for multi-target tracking. *arXiv:1504.01942*, 2015. 1, 4

[19] W. Li, R. Zhao, T. Xiao, and X. Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014. 4

[20] A. Milan, L. Leal-Taixé, I. D. Reid, S. Roth, and K. Schindler. MOT16: A benchmark for multi-object tracking. *arXiv:1603.00831*, 2016. 1, 2, 4, 7

[21] A. Milan, S. Roth, and K. Schindler. Continuous energy minimization for multitarget tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2014. 2, 7, 8

[22] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2015. 5

[23] H. Pirsiavash, D. Ramanan, and C. C. Fowlkes. Globally-optimal greedy algorithms for tracking a variable number of objects. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011. 2, 6

[24] L. Pishchulin, E. Insafutdinov, S. Tang, B. Andres, M. Andriluka, P. Gehler, and B. Schiele. Deepcut: Joint subset partition and labeling for multi person pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 5

[25] E. Ristani and C. Tomasi. Tracking multiple people online and in real time. Springer, 2014. 2

[26] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014. 4

[27] S. Tang, B. Andres, M. Andriluka, and B. Schiele. Subgraph decomposition for multi-target tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 2

[28] S. Tang, B. Andres, M. Andriluka, and B. Schiele. Multi-person tracking by multicuts and deep matching. In *ECCV Workshop on Benchmarking Mutliple Object Tracking*, 2016. 1, 2, 6, 8

[29] S. Tang, M. Andriluka, A. Milan, K. Schindler, S. Roth, and B. Schiele. Learning people detectors for tracking in crowded scenes. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2013. 2

[30] S. Tang, M. Andriluka, and B. Schiele. Detection and tracking of occluded people. *International Journal of Computer Vision (IJCV)*, 2014. 2

[31] X. Wang, E. Turetken, F. Fleuret, and P. Fua. Tracking interacting objects using intertwined flows. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2016. 2

[32] 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. 6

[33] L. Wen, W. Li, J. Yan, Z. Lei, D. Yi, and S. Z. Li. Multiple target tracking based on undirected hierarchical relation hypergraph. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 2

[34] Y. Xiang, A. Alahi, and S. Savarese. Learning to track: Online multi-object tracking by decision making. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015. 2, 6

[35] A. R. Zamir, A. Dehghan, and M. Shah. GMCP-Tracker: Global multi-object tracking using generalized minimum clique graphs. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2012. 2

[36] 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. 2

[37] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 4