A Graph-based MAP Solution for Multi-person Tracking using Multi-camera Systems

Xiaoyan Jiang, Marco Körner, Daniel Haase and Joachim Denzler
Computer Vision Group, Friedrich Schiller University of Jena, Jena, Germany

Keywords: Multi-person tracking, Multi-camera, Min-cost, MAP.

Abstract: Accurate multi-person tracking under complex conditions is an important topic in computer vision with various application scenarios such as visual surveillance. Taking into account the difficulties caused by 2D occlusions, missing detections, and false positives, we propose a two-stage graph-based object tracking-by-detection approach using multiple calibrated cameras. Firstly, data association is formulated into a maximum a posteriori (MAP) problem. After transformation, we show that this single MAP problem is equivalent of finding min-cost paths in a two-stage directed acyclic graph. The first graph aims to extract an optimal set of tracklets based on the hypotheses on the ground plane by using both 2D appearance feature and 3D spatial distances. Subsequently, the tracklets are linked into complete tracks in the second graph utilizing spatial and temporal distances. This results in a global optimization over all the 2D detections obtained from multiple cameras. Finally, the experimental results on three difficult sequences of the PETS'09 dataset with comparison to the state-of-the-art methods show the precision and consistency of our approach.

SCITEPRESS
SCIENCE AND TECHNOLOGY PUBLICATIONS

1 INTRODUCTION

Automatic initialization and tracking of multiple, potentially changing number of persons in real situations are a classic but challenging topic in computer vision. Along with the development of object detection approaches, the tracking-by-detection framework is adopted widely for multi-object tracking scenarios. Given discrete detections in separate time steps, the task afterwards is to assign the right detections to individual targets. Hence, data association is a key issue for multi-object tracking.

Massive works with regard to single-camera based multi-object tracking have shown the limitation of tracking performance (Breitenstein et al., 2011). This is mainly due to the large portion of false positive and missing detections caused by severe occlusions or bad lightness conditions. By contrast, for tracking using multiple cameras, one view can change information with others and compensate the data scarcity. However, data association from multiple cameras occurs an extra difficulty known as ”ghost effect” (Wu et al., 2012) caused by triangulation of objects in 3D space.

Accordingly, we propose a global optimization approach using two graphs for multi-person tracking in multiple calibrated camera systems. To simplify the calculation, we adopt hypotheses on the ground plane reconstructed from 2D detections from all available views. Afterwards, track fragments (Tracklets) are extracted by finding the min-cost paths in a so-called hypothesis graph. Finally, complete tracks are generated by linking those tracklets through a so-called tracklet graph.

1.1 Related Work

There are much effort made for efficient data association for multi-object tracking in previous years. In many works, tracking-by-detection was defined as a maximum a posteriori (MAP) problem (Zhang et al., 2008)(Berclaz et al., 2011)(Hofmann et al., 2013), which aims to find the optimal set of trajectories with maximum posteriori probabilities given all the observations from every video frame (Xing et al., 2009).

Step by step assignment such as particle filtering (Breitenstein et al., 2011)(Jiang et al., 2012), kalman filter (Satoh et al., 2004) propagated the object state vector according to a given motion model and performed data association between detections and tracks (Breitenstein et al., 2011) when a new frame came. However, during the period when the observation model that was normally defined as local features changed much due to occlusion or illumination, it was easy for the tracker to make wrong decisions or lose

Jiang X., Körner M., Haase D. and Denzler J..
A Graph-based MAP Solution for Multi-person Tracking using Multi-camera Systems.
DOI: 10.5220/0004690103430350
In *Proceedings of the 9th International Conference on Computer Vision Theory and Applications* (VISAPP-2014), pages 343-350
ISBN: 978-989-758-009-3

Copyright © 2014 SCITEPRESS (Science and Technology Publications, Lda.)

the target totally. Approaches such as Hungarian algorithm (Huang et al., 2008), bipartite graph matching (Bredereck et al., 2012), and energy minimization (Andriyenko and Schindler, 2011) tried to find local maxima or minima of matching, while global information was not considered.

Recently, many researches concerning global optimization schemes based on flow networks (Zhang et al., 2008)(Wu et al., 2011)(Hofmann et al., 2013) and graphs (Leal-Taixé et al., 2012)(Collins, 2012) have been widely presented in the literature. They converted the multi-object tracking problem to the searching of multiple min-cost paths in the network or the graph. Generally, each node of the network or the graph represents a single object's hypothesis of state with a specific time stamp. Trajectories of the targets are then obtained by traversing the found min-cost paths from the sink node to the source node (Jiang et al., 2013). In (Berclaz et al., 2011), 2D detections were firstly mapped into a probabilistic occupancy map on the ground plane. Afterwards, they built a flow network based on this probabilistic occupancy map (Fleuret et al., 2008). Tracking was ultimately formulated to a Integer Programming Problem with the basic restriction that the flows arriving at any position on the probabilistic occupancy map equal to the ones departing from this location. Authors in (Leal-Taixé et al., 2012) constructed local graphs for each view and then considered every pair of cameras as a possible unit to build a higher level graph. However, this was considered to be not intuitive (Hofmann et al., 2013). In (Wu et al., 2011), tracks were obtained in each view by track graphs. Subsequently, the set cover algorithm was implemented for linking among track segments from multiple views. Henriques et al., modeled the merge and split activities during tracking and removed the common restriction used in graph based approaches that one node in the graph belongs to one target (one-one match) at most (Henriques et al., 2011).

In this paper, we argue that the one-one match is necessary for multi-object tracking in multi-camera systems. Data association among multiple cameras is globally modeled by a two-stage graph. Additionally, we incorporate local features to the cost function in the first graph. Finally, we evaluate the proposed paradigm on 3 sequences with different difficulties of tracking from the PETS'09 dataset. The experimental results show the accuracy and consistency of the approach, as well as the cheap computational time it needs.

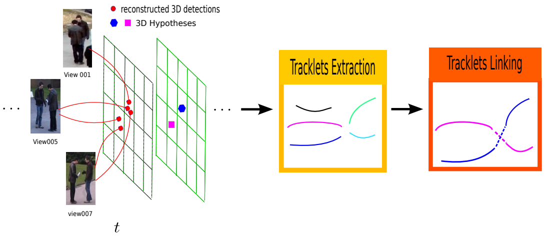

Figure 1: Diagram of our approach: hypothesis generation, tracklet extraction, and tracklet linking. The left part is the generation of $\mathcal{H}_t$: each $\mathbf{o}_i^{c_j}$ has a 3D detection on the ground plane (dashed grids). Those 3D detections whose back projections in different image views are nearest to the same detections in corresponding views are considered to be identical objects. $\mathcal{H}_t$ therefore consists of the 3D detections that have the minimum average back projection errors (solid grids).

1.2 Outline and Contributions

Our two-stage graph-based multi-person tracking using multi-camera systems approach is shown as Fig. 1 (detailed in Sec. 2). The approach contains three key components: hypothesis generation, tracklet extraction, and tracklet linking. By contrast with the indicated studies and the work in (Jiang et al., 2013) where a two-stage graph was used as well, our contributions are as follows:

1. We formulate the multi-object tracking problem into two MAP problems and solve each MAP problem by an individual graph. The graphs are conducted on the ground plane directly, which is more straightforward than the methods that construct local graphs for each view.

2. Local features such as appearance and size in image scale of the person are integrated to the assignment of costs for edges in the hypothesis graph.

The rest of the paper is structured as follows: formulation of two MAP problems is discussed in Sec. 2. Sec. 3 presents the details of mapping the two MAP problems into a two-stage graph. Subsequently, qualitative and quantitative results on the PETS'09 dataset with a comparison to the state-of-the-art algorithms are stated in Sec. 4. Finally, Sec. 5 summarizes the paper and gives an outlook of the future work.

2 TRACKING FORMULATION

After applying a detector to video frames from each camera view, 2D detections are obtained as input to tracking approaches. Normally, a 2D observation is formulated as $\mathbf{o}_i^{c_j} = \{x_i, s_i, t_i\}$ indicating the position x, size s and time index t of detection i in camera c_j (Felzenszwalb et al., 2010). Assume that the total number of cameras in the system is N, $c_j \in \{1, \cdots, N\}$, we define $\mathbf{O}_t^{c_j} = \{\mathbf{o}_i^{c_j}\}$ to be the set of observations from camera c_j at time t. $\mathbf{O}_{1:t}^{c_j} = \{\mathbf{O}_1^{c_j}, \cdots, \mathbf{O}_t^{c_j}\}$ is the

SCIENCE AND TECHNOLOGY PUBLICATIONS

set of observations until time t in camera c_j. Therefore, known the set of observations for all cameras $\mathbf{O}_{1:t} = \{\mathbf{O}^1_{1:t}, \cdots, \mathbf{O}^C_{1:t}\}$, the trajectories of the targets until time t, that is $\mathcal{T}_{1:t}$, are searched in this huge observation space. One step further, for multi-object tracking in multi-camera systems in this work, data association is a MAP problem based on 3D hypotheses $\mathcal{H}_{1:t}$:

$$\mathcal{T}^*_{1:t} = \arg\max_{\mathcal{T}_{1:t}} P(\mathcal{T}_{1:t}|\mathcal{H}_{1:t}) \ , \tag{1}$$

$$\mathcal{H}_{1:t} \subseteq \mathcal{R}_{1:t} = \mathcal{R}(\forall\, \mathbf{O}^{c_j}_{1:t} \in \mathbf{O}_{1:t}) \ , \tag{2}$$

with $\mathcal{T}^*_{1:t}$ is the set of optimal trajectories until time t. $\mathcal{H}_{1:t}$ is a subspace of all the possible reconstructed 3D detections $\mathcal{R}_{1:t}$ on ground plane accompanying local features in visible 2D images. $\mathcal{R}(\cdot)$ is the reconstruction function. Additionally, 3D detections belong to identical objects are integrated into single hypotheses. The process of generation of $\mathcal{H}_t$ is shown on the left part of Fig. 1.

Now $\mathcal{T}_{1:t}$ is a subset of $\mathcal{H}_{1:t}$. Recursive searching for every possible combination is impractical because of the computational complexity. We intend to use a global optimization strategy to find the possible associations that have the highest posteriori probabilities over the whole sequences with T frames.

Denote $\tau_k = \{\zeta_k, t_{\tau_k,0}, t_{\tau_k,1}\} \in \Upsilon$ is a tracklet from frame $t_{\tau_k,0}$ to frame $t_{\tau_k,1}$ with a set of connected 3D locations $\zeta_k = (p^{t_{\tau_k,0}}, \cdots, p^{t_{\tau_k,1}})$ and refer to (Xing et al., 2009), finding the optimal trajectories for multiple targets can be written as:

$$\mathcal{T}^*_{1:T} = \arg\max_{\mathcal{T}_{1:T}} P(\mathcal{T}_{1:T}|\Upsilon, \mathcal{H}_{1:T}) \tag{3}$$

$$= \arg\max_{\mathcal{T}_{1:T}} P(\mathcal{T}_{1:T}|\Upsilon) \cdot P(\Upsilon|\mathcal{H}_{1:T}) \tag{4}$$

$$= \arg\max_{\mathcal{T}_{1:T}} \prod_k P(\mathcal{T}_{1:T}|\tau_k) \prod_k P(\tau_k|\mathcal{H}_{1:T}) . \tag{5}$$

Equ. 5 is true because of the independence of individual tracklets.

We also follow the non-overlap constraint:

$$\tau_k \cap \tau_l = \emptyset, \ \forall k \neq l, \ \forall \tau_k, \tau_l \in \Upsilon \ . \tag{6}$$

From Equ. 5, we can see that the MAP problem is separated into two MAP sections: optimal tracklets extraction and optimal tracklets linking. The two MAP problems are equivalent of searching for the paths with minimum costs in individual graphs accordingly, which is discussed in the following Sec. 3.

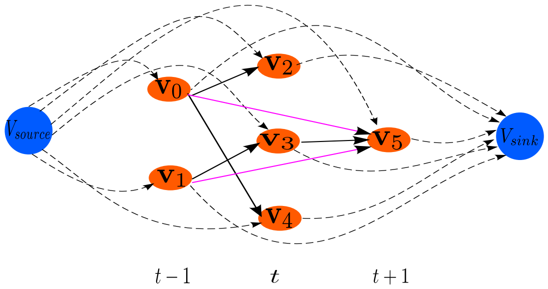

Figure 2: An exemplary hypothesis graph consists of 3 time steps. Dashed lines from and to virtual sink node and source node allow every possible entering and exiting position respectively. Here, we also consider missing detections by allowing edges composed by hypotheses with the time difference larger than one, which are shown by purple lines.

3 MAPPING TO A TWO-STAGE GRAPH

In the first-stage graph, all available tracklets are extracted without knowing the number of objects as a priori (Leal-Taixé et al., 2012) or assuming entrance and exit regions (Hofmann et al., 2013). In the second-stage graph, tracklets are linked using temporal and spatial distances to form complete tracks. Finally, the trajectories are refined to generate a unique trajectory per target.

3.1 Tracklet Extraction

We define a direct acyclic graph $\mathcal{G} = (\mathbf{V}, \mathbf{E}, \mathbf{c})$, which is called hypothesis graph, to extract tracklets from $\mathcal{H}_{1:T}$. A vertice $\mathbf{v} \in \mathbf{V}$ in $\mathcal{G}$ represents one hypothesis $\mathbf{h}^i_t = \{p^i_t, a_t^{i,1:n(c'_j)}, s_t^{i,1:n(c'_j)}\} \in \mathcal{H}_t, c'_j \in N$ that contains 3D reconstructed position on the ground plane p^i_t, appearance features and sizes in 2D images from number $n(c'_j)$ visible camera views with time stamp t. We introduce one virtual vertice of source v_{source} along with one virtual vertice of sink v_{sink} as shown in Fig. 2 to start and terminate paths separately. Since the number of objects varies time by time, each vertice $\mathbf{v} \in \mathbf{V}$ has probabilities incoming to v_{sink} and outgoing from v_{source} that are proportional to the corresponding frame index t:

$$P_{e=(v_{source}, \mathbf{v})} = 1 - t/T \ , \tag{7}$$

$$P_{e=(\mathbf{v}, v_{sink})} = t/T \ , \tag{8}$$

where T is the total number frames.

Denote $\mathbf{e}^{\Delta t}_{i,j} = \{\mathbf{v}^t_i, \mathbf{v}^{t+\Delta t}_j\} \in \mathbf{E}$, $\Delta t \in [1, t_{max}]$ is the number of frame differences. The transition probabil-

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS

ity $P^{\Delta t}_{i,j}$ assigned to $\mathbf{e}^{\Delta t}_{i,j}$ is defined as:

$$P^{\Delta t}_{i,j} = \begin{cases} P_{pena} \cdot P_{spat} \cdot P_s \cdot P_a \ , & d_{spat} < th_{vel} , \\ \infty \ , & \text{else,} \end{cases} \tag{9}$$

where $P_{pena} = \rho^{\Delta t}, \rho < 1$ is the penalty for skipping frames of missing detections. And

$$P_{spat} = 1 - d_{spat}/th_{vel} \tag{10}$$

computes the spatial affinity in 3D world coordinate system with a maximum defined motion th_{vel}. And $d_{spat} = \|p^i_t - p^j_{t+\Delta t}\|_2$ is the Euclidean distance.

The average probability for size affinity from all visible views is

$$P_s = (\sum_{c_j} \frac{\min(s_i, s_j)}{\max(s_i, s_j)})/n(c'_j) \ . \tag{11}$$

The average appearance similarity from all visible views is

$$P_a = (\sum_{c_j} sim(a_i, a_j))/n(c'_j) \ . \tag{12}$$

SCITEPRESS
SCIENCE AND TECHNOLOGY PUBLICATIONS

We use Bhattacharyya coefficient to evaluate the similarity of RGB histograms of the objects as appearance features.

After the configuration and refer to (Zhang et al., 2008), we can convert one of the MAP problem of extracting optimal tracklets Υ^* in Equ. 5 to k-shortest paths algorithm conducted on $\mathcal{G}$ through negative logarithm transformation:

$$\begin{aligned} \Upsilon^* &= \arg\max_{\Upsilon} \prod_{k} P(\tau_k|\mathcal{H}_{1:T}) && (13) \\ &= \arg\min_{\Upsilon} \sum_{k} -\log P(\tau_k|\mathcal{H}_{1:T}) && (14) \\ &= \arg\min_{\Upsilon} \sum_{i,j} -\log P^{\Delta t}_{i,j} && (15) \\ &+ (-\log P_{e=\{v_{source},\mathbf{v}\}}) + (-\log P_{e=\{\mathbf{v},v_{sink}\}}) && (16) \\ &= \arg\min_{\Upsilon} (\sum_{i,j} c_{i,j} + c_{en} + c_{ex}) \ . && (17) \end{aligned}$$

Thus, the costs are naturally defined as:

$$c_{i,j} = -\log P^{\Delta t}_{i,j} \tag{18}$$

$$c_{en} = -\log P_{e=(v_{source},\mathbf{v})} \tag{19}$$

$$c_{ex} = -\log P_{e=(\mathbf{v},v_{sink})} \ . \tag{20}$$

We iteratively employ Dijkstra's *shortest path* algorithm (Dijkstra, 1959) to find a number of relatively min-cost paths $\mathcal{P} = (v_{sink}, \mathbf{v}^{t_{\mathcal{P},0}}, \cdots, \mathbf{v}^{t_{\mathcal{P},1}}, v_{source})$. Depends on the non-overlap constraint between tracklets, costs of edges to and from vertices in found $\mathcal{P}$ are set to be infinite. Afterwards, tracklets are obtained by traversing each $\mathcal{P}$ from v_{sink} to v_{source} with exiting frame index and entering frame index.

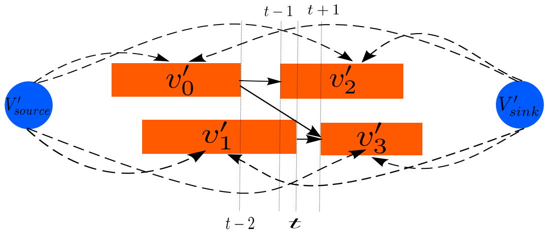

Figure 3: An exemplary tracklet graph consists of four vertices (Jiang et al., 2013). Dashed lines from and to virtual sink vertice and source vertice allow every tracklet to start and terminate a final track respectively. Edges only exist between nodes with temporally consistent order.

3.2 Tracklet Linking

Since tracklets are fragments of final trajectories, we define another directed acyclic tracklet graph $\mathcal{G}' = (\mathbf{V}', \mathbf{E}', \mathbf{c}')$ to globally choose the optimal combinations of tracklets. We again define a virtual source vertice v'_{source} and a virtual sink vertice v'_{source} to make all the paths found in $\mathcal{G}'$ begin and stop by them. Each $\mathbf{v}'_k \in \mathbf{V}'$ represents a tracklet $\tau_k = \{\zeta_k, t_{\tau_k,0}, t_{\tau_k,1}\}$ with starting and terminating frame indexes and 3D locations in between. Similarly, the entrance cost and exit cost for $\mathbf{v}'_k$ are:

$$c'_{en} = t_{\tau_k,0} \cdot c_{pe} \ , \tag{21}$$

$$c'_{ex} = (T - t_{\tau_k,1}) \cdot c_{pe} \ . \tag{22}$$

Here, c_{pe} is the penalty cost which is manually set. From this we can see that the tracklets start from the first frame and terminate in the last frame of the video have a lower entrance/exit cost.

The cost $c'(k,l)$ for $e'(k,l) = (\mathbf{v}'_k, \mathbf{v}'_l) \in \mathbf{E}'$ is defined as follows:

$$c'_{k,l} = \begin{cases} d^{k,l}_{spat} \cdot d^{k,l}_{temp} & 0 < t_{\tau_l,0} - t_{\tau_k,1} < th_{temp} \\ \infty & \text{else .} \end{cases} \tag{23}$$

The spatial distance

$$d^{k,l}_{spat} = \|\zeta_l^{t_{\tau_l,0}} - \zeta_k^{t_{\tau_k,1}}\|_2 \tag{24}$$

is the Euclidean distance between the corresponding terminating and starting points of τ_k, τ_l, and

$$d^{k,l}_{temp} = t_{\tau_l,0} - t_{\tau_k,1} \tag{25}$$

is the temporal distance between two tracklets. Therefore, only the pairs of tracklets who have no temporal overlap and the frame differences are smaller than a certain threshold th_{temp} have weighted edges. Otherwise, they are assigned by infinite costs to invalid the specific edges. The exemplary figure of $\mathcal{G}'$ is indicated in Fig. 3.

Similar to Subsect. 3.1, the second MAP problem of finding optimal trajectories from tracklets in Equ. 5

can be converted into searching for min-cost paths conducted on $\mathcal{G}'$:

$$\mathcal{T}^*_{1:T} = \arg\min_{\mathcal{T}_{1:T}} \prod_k P(\mathcal{T}_{1:T}|\tau_k) \tag{26}$$

$$= \arg\min_{\mathcal{T}_{1:T}} (\sum_{k,l} c'_{k,l} + c'_{en} + c'_{ex}) \ . \tag{27}$$

We again iteratively employ the Dijkstra's *shortest path* algorithm (Dijkstra, 1959) on $\mathcal{G}'$. The final trajectories are consequently obtained by traversing the linked tracklets.

3.3 Tracking Refinement

After both stages of tracklet extraction and tracklet linking, tracklets/tracks that belong to identical objects are recursively merged according to their beginning and finishing frame indexes. A pair of tracklets/tracks is judged to be identical objects when their spatial Euclidean distance in the same frame index is closer than a threshold th_{mer} for a minimum number of frames f_{mer}. And f_{mer} is proportional to the minimum length of the pair of tracklets/tracks considered.

The missed positions within tracklets in the first stage and the gaps between linked tracklets in the second stage are both linearly interpolated.

4 EXPERIMENTS

4.1 Dataset

We use the public available dataset of PETS'09 (Ferryman and Shahrokni, 2009) to evaluate the performance of our multi-object tracking using multi-camera systems approach. The dataset has three object-tracking sequences with different levels of people density: sparse (S2.L1), medium (S2.L2), and high (S2.L3). They were recorded by different number of cameras located in different positions. These videos are very challenging since they contain many different types of occlusion, for example inter-object occlusion, object-obstacle occlusion (people are occluded by a light pole with a big sign). The resolution in the first view is 768 × 576. The frame rate is 7 f/s, for which persons can move very fast between neighboring frames. Additionally, we report the detection results as well for comparison, since the tracking-by-detection paradigm depends much on the detection performance.

4.2 Implementation

We employ the deformable part models based object detector (Felzenszwalb et al., 2010) to get 2D detections in video frames of each camera. For PETS'09 S2.L1, we utilize 6 out of 7 recorded cameras as the fourth view suffers from frame rate instability (Ferryman and Shahrokni, 2009). The middle bottom points of the 2D detections are adopted for the reconstruction of 3D detections on the ground plane which are integrated into hypotheses detailed in Sec. 2.

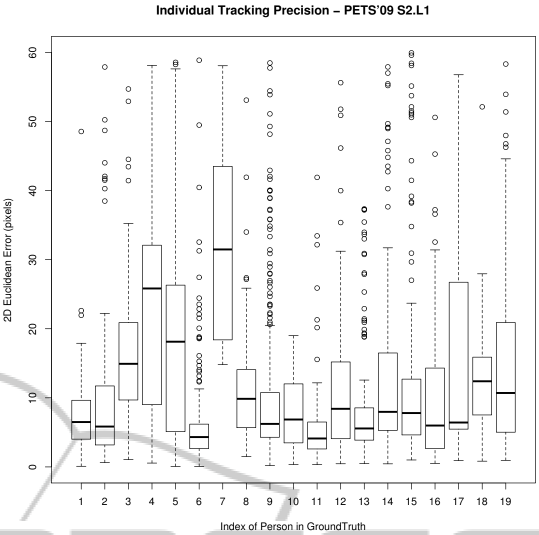

Figure 4: Boxplots of the pixel wise precisions for each person according to the ground truth data on PETS'09 S2.L1 sequence. Most of the targets are tracked and have a localization error around 10 pixels.

SCITEPRESS
SCIENCE AND TECHNOLOGY PUBLICATIONS

Parameters. The parameters configured in graph based approaches affect a lot on the final results. In our experiments, we set th_{mer} = 1000. Thus, positions that are spatially near than 1 m in 3D space are considered to be identical. The cost for entering or exiting a path is set to c_{pe} = 60000, which is larger than the transition cost to encourage the system to link all these vertices which are associated. Furthermore, we configured ρ = 0.95 to make the system penalize much more to the edges that link vertices across more number of frames. The fastest speed a person can walk is limited to be 3.5 m/s in the dataset, therefore th_{vel} = 500.

For tracking in sequence S2.L1, we extract 100 tracklets and set the maximum frame gap to $t_{max} = 4$, since a relatively higher or lower numbers reduce the performance. Eventually, 25 tracks are generated from these tracklets. Also, we select 1000 tracklets, set $t_{max} = 10$, 120 and 51 tracks are extracted on the video S2.L2 and S2.L3 respectively.

4.3 Evaluation

For evaluation, we employ the *multiple object tracking precision* (MOTP) and *multiple object tracking accuracy* (MOTA) (Bernardin and Stiefelhagen, 2008)

Table 1: Quantitative results on PETS'09 S2.L1, S2.L2 and S2.L3 dataset. We compared MOTP, MOTA, False Positive Rate, Miss Rate and Id switches with particle filter based tracking-by-detection (Jiang et al., 2012) (Breitenstein et al., 2011), Energy minimization (Andriyenko and Schindler, 2011), k-shortest paths (Berclaz et al., 2011), and Probabilistic tracking (J. Yang and Teizer, 2009).

Sequence	Method	MOTP	MOTA	False Pos. Rate	Miss Rate	Id switches
PETS'09 S2.L1	Jiang et al., (Jiang et al., 2012)	78.8%	60.8%	*n/a*	*n/a*	*n/a*
	Yang et al., (J. Yang and Teizer, 2009)	53.8%	75.9 %	*n/a*	*n/a*	*n/a*
	Breitenstein et al., (Breitenstein et al., 2011)	56.3%	79.7%	*n/a*	*n/a*	*n/a*
	Berclaz et al., (Berclaz et al., 2011)	60.0 %	66.0 %	*n/a*	*n/a*	*n/a*
	Andriyenko et al., (Andriyenko and Schindler, 2011)	76.1 %	**81.4%**	*n/a*	*n/a*	15
	Our Approach	**81.44%**	77.74%	7.83%	13.91%	24
PETS'09 S2.L2	Breitenstein et al., (Breitenstein et al., 2011)	51.3%	50.0%	*n/a*	*n/a*	*n/a*
	Our Approach	**60.14%**	**55.54%**	0.85%	40.83%	287
PETS'09 S2.L3	Breitenstein et al., (Breitenstein et al., 2011)	**52.1%**	67.5%	*n/a*	*n/a*	*n/a*
	Our Approach	50.08%	**67.71%**	0.0%	29.84%	107

metrics which have become *de facto* standard in the field of multi-object tracking. MOTP considers the average error of tracked positions over the whole sequence. False positives, misses, and mismatches compose MOTA that aims to estimate the tracker's ability of recognition and consistency.

The ground truth in the first view was provided by Anton Andriyenko (Andriyenko and Schindler, 2011). We back project our tracking results to this single view for measurement. The assignment between tracking and the ground truth in image coordinate system has a threshold of 60 pixels according to the average width of people appeared in the view. Because of the character of tracking-by-detection framework and in order to have a fair evaluation on tracking, we provide the false positive rate and the false negative rate for the detection by removing objects' Id numbers in the ground truth data. The detector we use has a distinct performance of different videos in the first view: the false positive rate is 0.05, 0.01, and 0.0, while the missing rate is 0.1, 0.6, and 0.5 for S2.L1, S2.L2, and S2.L3 respectively.

Tab. 1 shows the quantitative results of our approach compared to the state-of-the-art methods. From the results of S2.L1, it can be seen that we have the highest MOTP, which indicates the precision of target localization of our method. MOTA for S2.L1 is also comparable with others. Additionally, the pixel wise precisions for individual persons are shown as boxplots in Fig. 4. We can state that the average bias of the tracked objects' localization is approximately 10 pixels, which is probability the same error between different ground truth data labeled by diverse persons.

Besides, our approach has the highest MOTP and MOTA on S2.L2 sequence, which is partially because of the allowance of linking between hypotheses with number of frame gaps in the first graph. This benefit is also obvious when comparing detection to tracking with their missing rates. For S2.L3, our tracking performance is also comparable with the work that reported their results.

Therefore, from Tab. 1, we can see that our method has comparable or relatively better results for the three sequences on PETS'09. Although the numbers of Id switches are slightly higher, the tracker keeps tracking after switching. This happens when people are merging and splitting which cause wrong relatively low costs.

Qualitative results of PETS'09 S2.L1, S2.L2 and S2.L3 from different number of cameras are shown in Fig. 5. We visualize the trajectories from tracking results and obtain corresponding rectangles of each frame by finding the nearest 2D detections conducted on the frame. From the figure we can see that our approach is able to consistently track people who have already been occluded by others or the obstacle in the scene for a long time. In the third view of S2.L1 where the scenario is under bad lightness condition, we can still recognize and recover the trajectories using the data from other views, which indicate the benefit of utilization of multiple cameras.

4.4 Complexity and Runtime

The entire system was realized in c++. Our approach has a complexity of $O(k \cdot (n \log n + m))$, where n is the number of vertices, m is the number of edges, and k is the number of min-cost paths to be found in the respective graph. For 6 cameras with 795 frames each in S2.L1, it needs approximately 1.5 minutes for the extraction of 100 tracklets from a 1.5×10^4 node hypothesis graph and linking of final 25 tracks from the tracklet graph. Similar runtime is needed for the other two sequences we conducted. All measurements were conducted on a standard desktop computer with an Intel® Core$^{\mathrm{TM}}$ i5-760 CPU (2.80 GHz).

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS

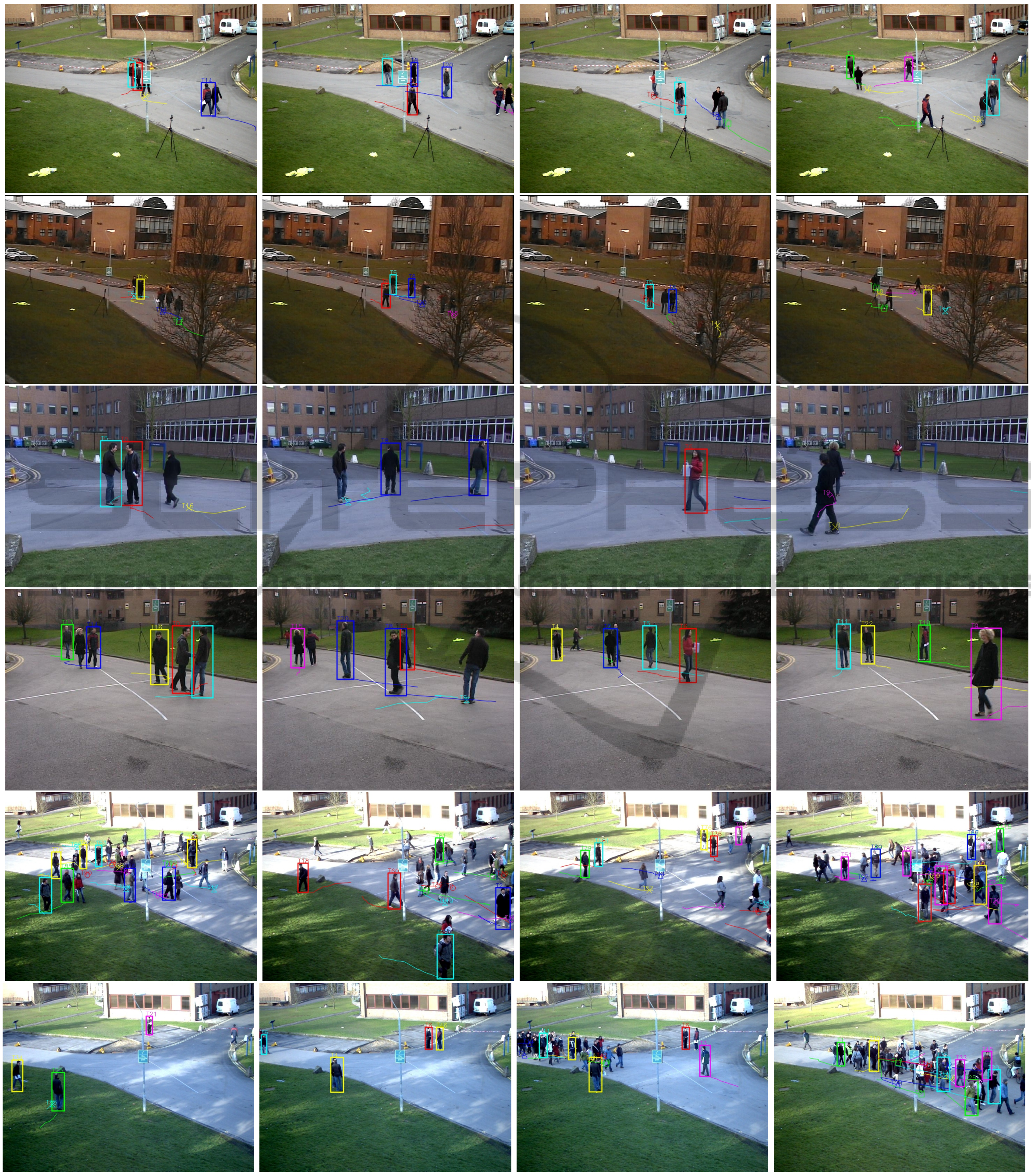

Figure 5: Qualitative results of our approach on PETS'09 S2.L1, S2.L2, and S2.L3 sequences. First four rows are for S2.L1 from four cameras separately. The fifth row shows the results from S2.L2 in the first view and results for the first view of S2.L3 are shown in the sixth row. Different colors and shapes indicate the different identities of the targets. The trajectories are shown by linking previous tracking results for up to 20 frames. Rectangles of the tracks are obtained by back projecting the tracking results to individual views and finding the nearest 2D detections. Therefore, the trajectories without rectangles denote the frames where existing missing detections while the tracker can still keep tracking.

5 CONCLUSIONS AND OUTLOOK

In this work, we firstly had a review on the recent studies of multi-object tracking using a single camera or multiple cameras and discussed the recent researches for global optimization based on flow networks or graphs. After the formulation of data association into two MAP problems. we proposed a two-stage graph-based multi-person multi-camera tracking approach. Firstly, a hypothesis graph was constructed to extract possible associated tracklets from reconstructed 3D detections. While most of the papers considered global features only, we incorporated local features such as appearance, size into the computation of costs for the edges in the hypothesis graph. Importantly, the hypotheses for tracking on the ground plane were arose from the reconstructions of 2D detections from each view at the same time step. Those reconstructed 3D detections who were regarded to be the same object were replaced by the one with the minimum back projection error. Consequently, the task of the second graph was to link tracklets into complete tracks. For this sake, the cost function accordingly took temporal and spatial distances into account. All in all, this framework is general for multi-object tracking in multi-camera systems.

From our experiments, we conclude that it is important to have the optimal outcome from the first step of hypothesis generation for multi-object tracking in multi-camera systems. Due to the impact of calibration and object detection errors, the precision of recognizing of identical objects can be improved by more restrict constraints. Hence, in the future, we would like to focus on the modeling of calibration and detection errors and incorporating them into the framework. Incremental learning might be able to refine the modeling as more and more frames are processed. Additionally, the tracklet linking stage could consider more information such as histogram of motion in the cost function to reduce the false positive rate and the number of Id switches.

REFERENCES

Andriyenko, A. and Schindler, K. (2011). Multi-target tracking by continuous energy minimization. In *CVPR*, pages 1265–1272.

Berclaz, J., Fleuret, F., Turetken, E., and Fua, P. (2011). Multiple object tracking using k-shortest paths optimization. *TPAMI*, 33:1806–1819.

Bernardin, K. and Stiefelhagen, R. (2008). Evaluating multiple object tracking performance: The clear mot metrics. *EJIVP*, 246309.

Bredereck, M., Jiang, X., Körner, M., and Denzler, J. (2012). Data association for multi-object tracking-by-detection in multi-camera networks. In *ICDSC*.

Breitenstein, M. D., Reichlin, F., Leibe, B., Koller-Meier, E., and Gool, L. V. (2011). Online muti-person tracking-by-detection from a single, uncalibrated camera. *PAMI*, 33:1820 – 1833.

Collins, R. T. (2012). Multitarget data association with higher-order motion models. In *CVPR*.

Dijkstra, E. W. (1959). A note on two problems in connexion with graphs. *NUMERISCHE MATHEMATIK*, 1:269–271.

Felzenszwalb, P., Girshick, R., and McAllester, D. (2010). Cascade object detection with deformable part models. In *CVPR*.

Ferryman, J. and Shahrokni, A. (2009). Pets2009: Dataset and challenge. In *2009 Twelfth IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS-Winter)*.

Fleuret, F., Berclaz, J., Lengagne, R., and Fua, P. (2008). Multi-camera people tracking with a probabilistic occupancy map. *TPAMI*, 30:267–282.

Henriques, J. F., Caseiro, R., and Batista, J. (2011). Globally optimal solution to multi-object tracking with merged measurements. In *ICCV*.

Hofmann, M., Wolf, D., and Rigoll, G. (2013). Hypergraphs for joint multi-view reconstruction and multi-object tracking. In *CVPR*.

Huang, C., Wu, B., and Nevatia, R. (2008). Robust object tracking by hierarchical association of detection responses. In *ECCV*, pages 788–801.

J. Yang, Z. Shi, P. V. and Teizer, J. (2009). Probabilistic multiple people tracking through complex situations. In *IEEE Workshop Performance Evaluation of Tracking and Surveillance*.

Jiang, X., Haase, D., Körner, M., Bothe, W., and Denzler, J. (2013). Accurate 3d multi-marker tracking in x-ray cardiac sequences using a two-stage graph modeling approach. In *the 15th Conference on Computer Analysis of Images and Patterns (CAIP)*.

Jiang, X., Rodner, E., and Denzler, J. (2012). Multi-person tracking-by-detection based on calibrated multi-camera systems. In *ICCVG*, pages 743–751.

Leal-Taixé, L., Pons-Moll, G., and Rosenhahn, B. (2012). Branch-and-price global optimization for multi-view multi-target tracking. In *CVPR*, pages 1987–1994.

Satoh, Y., Okatani, T., and Deguchi, K. (2004). A color-based tracking by kalman particle filter. In *ICPR*, pages 502–505.

Wu, Z., Kunz, T. H., and Betke, M. (2011). Efficient track linking methods for track graphs using network-flow and set-cover techniques. In *CVPR*, pages 1185–1192.

Wu, Z., Thangali, A., Sclaroff, S., and Betke, M. (2012). Coupling detection and data association for multiple object tracking. In *CVPR*.

Xing, J., Ai, H., and Lao, S. (2009). Multi-object tracking through occlusions by local tracklets filtering and global tracklets association with detection responses. In *CVPR*.

Zhang, L., Li, Y., and Nevatia, R. (2008). Global data association for multi-object tracking using network flows. In *CVPR*.