Cross-view People Tracking by Scene-centered Spatio-temporal Parsing*

Yuanlu Xu[1], Xiaobai Liu[2], Lei Qin[1,3] and Song-Chun Zhu[1]
[1]Dept. Computer Science and Statistics, University of California, Los Angeles (UCLA)
[2]Dept. Computer Science, San Diego State University (SDSU)
[3]Inst. Computing Technology, Chinese Academy of Sciences
yuanluxu@cs.ucla.edu, xiaobai.liu@mail.sdsu.edu, qinlei@ict.ac.cn, sczhu@stat.ucla.edu

Abstract

In this paper, we propose a Spatio-temporal Attributed Parse Graph (ST-APG) to integrate semantic attributes with trajectories for cross-view people tracking. Given videos from multiple cameras with overlapping field of view (FOV), our goal is to parse the videos and organize the trajectories of all targets into a scene-centered representation. We leverage rich semantic attributes of human, e.g., facing directions, postures and actions, to enhance cross-view tracklet associations, besides frequently used appearance and geometry features in the literature. In particular, the facing direction of a human in 3D, once detected, often coincides with his/her moving direction or trajectory. Similarly, the actions of humans, once recognized, provide strong cues for distinguishing one subject from the others. The inference is solved by iteratively grouping tracklets with cluster sampling and estimating people semantic attributes by dynamic programming. In experiments, we validate our method on one public dataset and create another new dataset that records people's daily life in public, e.g., food court, office reception and plaza, each of which includes 3-4 cameras. We evaluate the proposed method on these challenging videos and achieve promising multi-view tracking results.

Introduction

In this paper, we study a novel cross-view tracklet association algorithm for multi-view person tracking. We consider surveillance scenarios where there are 3-4 cameras looking at a target area (e.g., parking-lot, garden) from different viewpoints. The task is to compute the scene-centered overall trajectory of all the people within the scene. In comparison with the single-view setting (Liu, Lin, and Jin 2013; Andriyenko and Schindler 2011; Wu, Betke, and Kunz 2011; Dehghan, Assari, and Shah 2015), it remains unclear how to associate people trajectories across views, especially when the cameras have wide baselines or large view changes.

- Large appearance variations. A person is assumed to have similar appearance across space and time. Nevertheless, large camera view and scale changes compromise such assumption. For example, Fig. 1 shows a garden covered by four cameras. From these camera view snapshots, the person in navy blue looks different in front and back view.
- Inaccurate geo-localization. A common way for solving the task is to calibrate camera parameters and utilize cross-camera ground homographs, with which a person detected in one viewpoint can be registered in another view. However, the registration results are often not accurate enough to separate humans in the proximity because of the calibration errors or the inaccuracy of footprint estimation. For example, in Fig. 1 (c-d), people's feet are occluded by the wall and so it is difficult to register the detected human feet positions in other views.

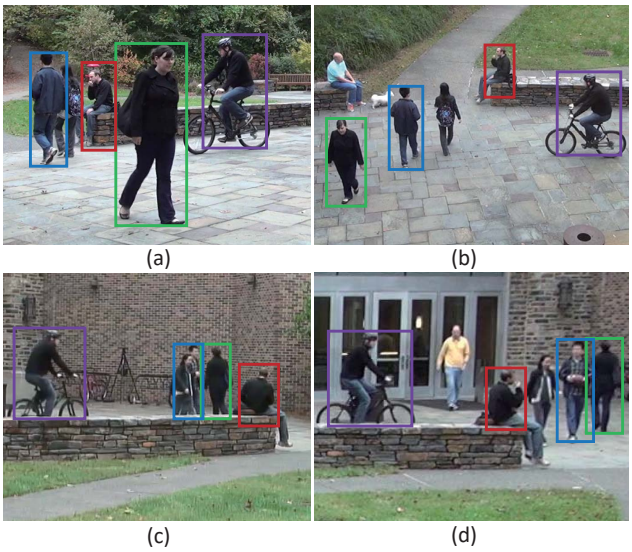

Figure 1: An example of cross-view data association for target tracking. (a)-(d) represents four different camera views of the same scene. Each color of the bounding box represents a unique person.

The main idea of our approach is to leverage semantic attributes, e.g., facing orientations, poses and actions (standing, running, etc.), for cross-view tracklet association. Taking Fig. 1 for example, attributes of person can help prune the ambiguities in cross-view data association. Specifically,

*This work is supported by DARPA SIMPLEX Award N66001-15-C-4035, ONR MURI project N00014-16-1-2007, and NSF IIS 1423305. Lei Qin is a visiting scholar at UCLA and the correspondence author is Xiaobai Liu.
Copyright © 2017, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

if the orientation of every human box can be correctly identified, we can associate the green box across views because there is only one person facing the building. In addition, since there is only one person sitting (red boxes) and one person on the bike (purple boxes), the pose and action recognition can be used to narrow down the association space. With the recent advances in computer vision and machine learning, these semantic attributes can be readily detected with a level of accuracy from a single view, serving as powerful cues for associating human boxes or trajectories across cameras.

We use Spatio-temporal Attributed Parse Graph (ST-APG) to integrate the semantic attributes with the people trajectories, and pose multi-view people tracking as spatio-temporal parsing problem. As illustrated in Fig. 2, the scene is decomposed into people trajectories and trajectories consists of tracklets with the same identity. A tracklet is a series of human boxes grouped by spatial coherency and perceptual similarity. The parse graph is enriched with attributes across different levels. The scene is incorporated with the camera information while tracklets with four types of attributes: i) appearance; ii) geometry, e.g., footprints; iii) motion, e.g., facing direction and speed; iv) pose/action, e.g., standing, sitting, walking, running, biking. These attributes can be recognized with a single image or a monocular video. We use these attributes to impose consistency constraints for cross-view tracklet associations. The constraints are used as additional energy term in the probabilistic formula, instead of hard constraints, to reduce errors made in bottom-up predictions.

To infer the ST-APG, we propose an efficient algorithm dealing with two sub-problems. I) We first employ a stochastic clustering algorithm (Barbu and Zhu 2005) to group the tracklets, which can efficiently traverse the combinatorial solution space. We explore two types of relationships among tracklets: i) being cooperative, i.e., tracklets from different view are allowed to be grouped together according to their appearance and semantic attributes; ii) being conflicting, e.g., tracklets with temporal overlaps in the same view, are conflicted to be grouped together. The conflicting relationships explicitly express the structure of the solution space. II) We use Dynamic Programming (DP) to estimate semantic attributes of the grouped tracklets. The trajectory is represented as a Markov Chain and DP are guaranteed to find the optimal solution. These two algorithms run iteratively until convergence.

We evaluate our approach on one public multi-view tracking dataset and collect a new multi-view dataset to cover daily activities (e.g., touring, dining, working). We use 4 GOPRO cameras to capture synchronized videos for 3 scenarios, including food court, office reception and plaza, which provides rich actions and activities. Results and comparisons with popular trackers show that our method obtains impressive results and sets up a new state-of-the-art for multi-view tracking.

Related Work

The proposed work is closely related to the following research streams in computer vision and artificial intelligence.

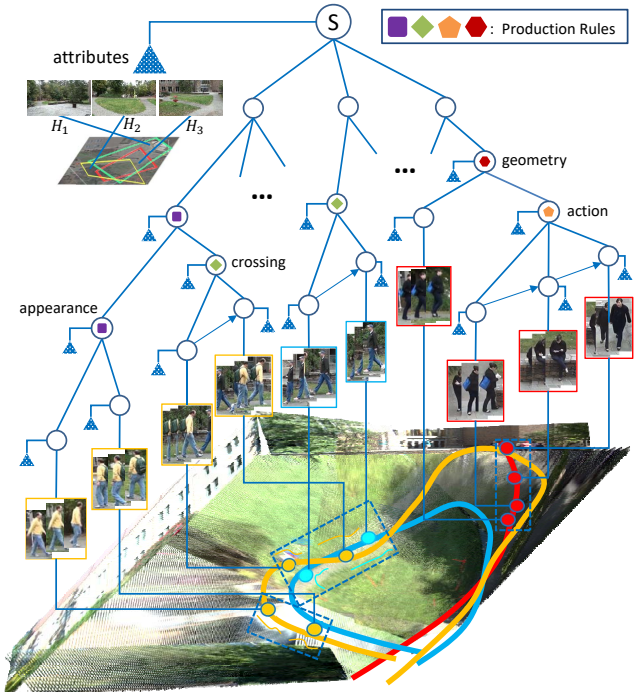

Figure 2: A Spatio-temporal Attributed Parse Graph (ST-APG). The scene S is generated by 3D reconstruction and associated with certain global attributes (e.g., homograph $H_1, \ldots, H_n$), and can be decomposed into tracklets belonging to different persons. Each tracklet is also leveraged with four types of semantic attributes (i.e., the blue triangles connected to nodes) and hierarchically organized in a parse graph.

Multi-view object tracking, like single-view tracking, is often formulated as a data association problem across cameras. A major question is to find cross-view correspondence at either pixel level (Sun, Zheng, and Shum 2003) or region-level (Khan and Shah 2006; Ayazoglu et al. 2011) or object-level (Xu et al. 2013; 2014). Typical data association methods are developed based on integer programming (Jiang, Fels, and Little 2007), network flow (Wu et al. 2009; Berclaz et al. 2011), marked point process (Utasi and Benedek 2011), multi-commodity network (Shitrit et al. 2013), and multi-view SVM (Zhang et al. 2015). Notably, Porway and Zhu (Porway and Zhu 2011) first introduced a cluster sampling method to explore both positive and negative relationships between samples, and Liu et al. (Liu, Lin, and Jin 2013) integrated a similar idea with motion information to construct a spatial-temporal graph for single-view tracking. In this work, we extend these two methods to further explore appearance, geometry, motion and pose/action relations between people tracklets in multi-view tracking.

Joint video parsing for solving multiple tasks simultaneously has been approved to be an effective way for boosting the performance of individual objectives. Wei et al. (Wei et al. 2016) presented a probabilistic framework for joint event, recognition, and object localization. Shu et al. (Shu

et al. 2015) proposed to jointly infer groups, events, and human roles in aerial videos. Nie et al. (Nie, Xiong, and Zhu 2015) used human poses to improve action recognition. Park and Zhu used an stochastic grammar to jointly estimate human attributes, parts and poses (Park and Zhu 2015). Weng and Fu (Weng and Fu 2011) utilized trajectories and key pose recognition to improve human action recognition. Yao et al. (Yao et al. 2011) employed pose estimation to enhance human action recognition. Kuo and Nevatia (Kuo and Nevatia 2010) studied how person identity recognition can help multi-person tracking. In this paper, we follow the same methodology to leverage semantic human attributes, including orientations, poses, and actions, to narrow the search space in cross-view data association.

Contributions. In comparison with previous methods, the contributions of this work is three-fold: i) a unified probabilistic framework of cross-view people tracking that can leverage multiple semantic attributes; ii) an efficient stochastic inference algorithm that can explore both positive and negative constraints between tracklets; iii) a comprehensive video benchmark regarding people's daily life which fosters research in this direction.

Spatio-temporal Attributed Parse Graph

In a common multi-view setting, activities in a scene S are captured by multiple cameras $\{C_1, C_2, \ldots, C_n\}$ with overlapping field of view (FOV). Videos from these cameras are synchronized in time. Given such data, our goal is to discover the trajectories Γ of every person within the scene, that is,

$$\Gamma = \{\Gamma_i : i = 1, ..., K\}, \tag{1}$$

where K indicates the total number of people appearing in the scene over a time period.

We use tracklets (i.e., trajectory fragments) as the basic unit. Tracklet is regarded as a mid-level representation to reduce the computation complexity, similar to superpixels/voxels in segmentation. A tracklet τ consists of a short sequence of object bounding boxes, which can be denoted as

$$\tau = \{(b_k, t_k) : \ k = 1, 2, \ldots, |\tau|\}, \tag{2}$$

where b_k indicates the bounding box and t_k the corresponding frame number. Normally, the duration of the tracklet is short (less than 300 frames, usually 50-200 frames) and the person identity and motion within the tracklet is consistent.

Given a tracklet set $\Gamma = \{\tau_j, j = 1, 2, \ldots, N\}$, we can re-write the scene-center trajectory of a person Γ_i as

$$\Gamma_i = \{\, \tau_j : \, l(\tau_j) = l_i, \, j = 1, 2, \ldots, N \,\}, \tag{3}$$

where K indicates the total number of existing people in the scene. Each tracklet τ_j will be assigned with a label $l_i \in \{0, 1, \ldots, K\}$, which can be regarded as the person ID which it belongs to. We also add $l_i = 0$ to denote this tracklet belongs to background.

Therefore, the problem of multi-view tracking can be formulated as a tracklet grouping problem, i.e. clustering tracklets of the same person into scene-centered trajectories. We further associate these tracklets with attributes and represent the scene as a Spatio-temporal Attributed Parse Graph (ST-APG) M, as illustrated in Fig. 2. A ST-APG consists of four components:

$$M = (\, S, \, X(S), \, \Gamma, \, X(\Gamma) \,), \tag{4}$$

where $X(S)$ denotes the global attributes (i.e., homographs $\{H_1, H_2, \ldots, H_n\}$ for each camera $\{C_1, C_2, \ldots, C_n\}$, $X(\Gamma)$ denotes the semantic attributes for tracklets. Therefore, solving multi-view people tracking is equivalent to finding the optimal ST-APG.

Semantic Attributes

Besides the identity label $l(\cdot)$, a tracklet τ_i is enriched with four kinds of attributes:

$$x(\tau_i) = (l(\tau_i), f(\tau_i), \, h(\tau_i), \, \vec{v}_i, \, \{a_{i,k}\}_{k=1}^{|\tau_i|}), \tag{5}$$

where $f(\tau_i)$ denotes the appearance attribute, $h(\tau_i)$ denotes the geometry attribute, $\vec{v}_i$ denotes the motion attribute of tracklet τ_i and $a_{i,k}$ the pose/action attribute at time $t_{i,k}$, i.e., the k-th frame of tracklet τ_i.

Similar to the literature, we define the appearance attribute $f(\tau_i)$ as a feature descriptor, which implicitly models the visual evidence, e.g., clothing, face, fair of a person. We also define the geometry attribute $h(\tau_i)$ as the 2D object bounding boxes and projected footprints on the 3D ground plane. Besides appearance and geometry attributes, we further leverage two kinds of human semantic attributes to specifically handle the task of people tracking.

Motion Attributes. We assume the facing direction of a person is same as his/her motion direction. The average speed $\vec{v}_i$ is computed for each tracklet τ_i. However, 2D view-based motion not only suffers from the scale problem, but also is useless for cross-view comparisons. We thus transform the 2D view-based motion into the 3D real motion. Given the camera calibration, the foot point of each 2D bounding box is calculated and projected back onto the 3D ground. The speed and facing direction are thus computed and regarded as the motion attributes.

Pose/Action Attributes. To describe the actions and poses a_i of an individual, we apply a DCNN to categorize the classical human pose/action variations. We use the PASCAL VOC 2012 action dataset, augmented by our own collected images. The training set has 7 categories, including standing, sitting, bending, walking, running, riding bike, skateboarding, which covers people's common type of actions/poses in daily activities. The collected training set consists 5000 images. We thus fine tune a 7 layer CaffeNet, with 5 convolutional layers, 2 max-pooling layers, 3 fully-connected layers. The final output give us a 7d human pose/action confidence score and can be regarded as the local attribute probability $p(a_i)$.

Besides the unary pose/action confidence, we further learn a binary temporal consistency table $T(a_i, a_j)$ to describe the possible transitions between two successive pose/action attributes. The consistency table is learned from our newly collected multi-view dataset and apply in all experiments. There are around 1000 training samples in total. In learning, we initialize impossible transitions (e.g., bending→running, sitting→riding) as 0 and else as 0.05.

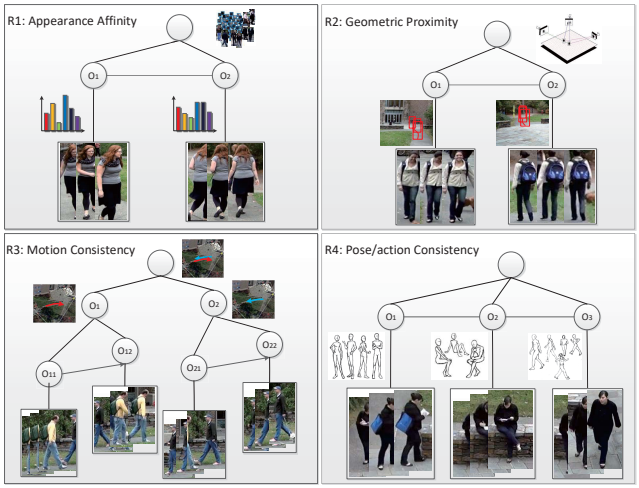

Figure 3: An illustration of four kinds of relations we utilize in this paper.

Bayesian Formulation

According to Bayes rule, M can be solved by maximizing a posterior (MAP), that is,

$$\begin{aligned} M^* &= \arg\max_M p(M|\Gamma;\theta) \\ &= \arg\max_M \frac{1}{Z}\exp\{-\mathcal{E}(\Gamma|M;\theta)-\mathcal{E}(M;\theta)\}, \end{aligned} \tag{6}$$

where θ indicates the model parameters.

Likelihood term $\mathcal{E}(\Gamma|M;\theta)$ measures how well the observed data (video bundle) satisfies a certain object trajectory. Assuming the likelihood of each bundle is calculated independently given the partition, then $\mathcal{E}(\Gamma|M;\theta)$ can be written as

$$\mathcal{E}(\Gamma|M) = \sum_{\tau_i\in\Gamma}\mathcal{E}(\tau_i|M;\theta). \tag{7}$$

Each term $\mathcal{E}(\tau_i|M;\theta)$ measures how the tracklet τ_i discriminates from the background. Therefore we treat this term as the constraint of itself being consistent with a foreground trajectory of a certain person. We estimate $\mathcal{E}(\tau_i|M;\theta)$ as a Markov chain structure, where the unary term $\mathcal{E}(a_i)$ is the attributes confidence probability, and the pairwise term $\mathcal{E}(a_i,a_j)$ is the attribute consistency in two successive frames, that is

$$\mathcal{E}(\tau_i|M;\theta) = \sum_{k=1}^{|\tau_i|}\mathcal{E}(a_{i,k}) + \sum_{k=1}^{|\tau_i|-1}\mathcal{E}(a_{i,k},\, a_{i,k+1}). \tag{8}$$

Note the motion information is trivial for successive frames and we thus ignore this part.

In this paper, we utilize **prior term** $\mathcal{E}(M;\theta)$ imposes constraints on people trajectories and their interactions. To do so, we develop four types of relations between two tracklets, as illustrated in Fig. 3. Given two tracklets, we consider both traditional visual relations (i.e., appearance and geometry) and leveraged semantic attribute relations (i.e., motion and pose/action).

Appearance similarity. This constraint assumes that the same person should share similar appearance across time and cameras. We adopt the appearance measurement proposed in (Xu et al. 2016), which basically uses a DCNN as codebook and encodes human body appearance as a 1000d feature vector. We measure the appearance similarity rule by

$$\mathcal{E}_e^{app}(\tau_i,\tau_j) = \sum_1^{|\tau_i|}\sum_1^{|\tau_j|}\frac{\|f(\tau_i)-f(\tau_j)\|_2}{|\tau_i|\cdot|\tau_j|}, \tag{9}$$

where $f(\tau_i)$ denotes the encoded feature vector of τ_i.

Geometric proximity measures how far two tracklets are located. We project the foot points of two tracklets onto the scene 3D ground plane using the given 2D to 3D homograph, and then compute the proximity of two tracklets as

$$\mathcal{E}_e^{geo}(\tau_i,\tau_j) = D(h(\tau_i),\, h(\tau_j)). \tag{10}$$

$D(\cdot,\cdot)$ denotes the averaged Euclidean distance between foot points of τ_i and τ_j over all overlapped frames.

Motion consistency. Given two proximate tracklets, the motion direction actually provides a solid evidence to show whether these tracklets belong to a same person or two persons crossing each other. Therefore, we can compute the angle between two motion directions. that is,

$$\mathcal{E}_e^{mov}(\tau_i,\tau_j) = \arccos\frac{\vec{v}_i\cdot\vec{v}_j}{|\vec{v}_i||\vec{v}_i|}. \tag{11}$$

If the angle is large, this probably indicates that two persons are moving in different directions.

Pose/action consistency. Noticing the pose/action of a same person across different views should also be consistent, we thus use the learnt temporal consistency table $p(a_i,a_j)$ to describe the consistency between two actions/poses. The rule is computed as

$$\mathcal{E}_e^{act}(\tau_i,\tau_j) = \sum_{t_m\in\{t_{i,k}\}\cap\{t_{j,k}\}}\mathcal{E}(a_{i,m},\, a_{j,m}). \tag{12}$$

Note that we only consider such relation among overlapped frames of two tracklets τ_i and τ_j.

We further introduce an adjacency graph $G=<\Gamma,E>$ to describe connections among tracklets. Each tracklet $\tau_i\in\Gamma$ is treated as a graph vertex and each edge $e_{ij}=<\tau_i,\tau_j>\in E$ describes the relation between two adjacent (neighboring) tracklets τ_i and τ_j. In this paper, two tracklets τ_i and τ_j are regarded as neighbors $\tau_i\in nbr(\tau_j)$ if only their temporal difference is no more than $\Delta_t=30$ frames and no far than $\Delta_d=5m$.

We regard edges generated by four types of constraints as cooperative edges E^+. The edge set E is further extended with conflicting edges E^-, that is, $E=E^-\cup E^+$. We enforce hard constraints to guarantee that i) two tracklets from the same view with temporal overlap will never be grouped together; ii) two adjacent tracklets with same identities will never have impossible pose/action transitions defined in temporal consistency table $T(a_i,a_j)$. Both types of relationships are utilized to help us group tracklets with similar characteristics together and with conflicting characteristics being dispelled.

Algorithm 1: Sketch of our inference algorithm

Input: Tracklet set Γ, global attributes $X(S)$
Output: Spatio-temporal Attributed Parse Graph M
Assign semantic attributes for each tracklet τ_i by DP ;
Construct adjacency graph G by computing cooperative and conflicting relations among Γ ;
Initialize $K = |\Gamma|, l_i = i$;
repeat
  Generate a cluster V_{cc};
  Randomly relabel cluster V_{cc} and obtain a new state M' ;
  Accept the new state with acceptance rate $\alpha(M \to M')$;
  Re-run DP on each new trajectory to update semantic attributes ;
until *convergence*;

Therefore, we can decompose the **prior term** $\mathcal{E}(M;\theta)$ into pairwise potentials between every two adjacent tracklets within G, that is,

$$\mathcal{E}(M;\theta) = \sum_{l_i=l_j, e_{ij}\in E^+} \mathcal{E}_e^+(\tau_i,\tau_j) + \sum_{l_i=l_j, e_{ij}\in E^-} \mathcal{E}_e^-(\tau_i,\tau_j), \tag{13}$$

where p_e^+ and p_e^- are the corresponding cooperative and conflicting edge probability defined above.

Inference

Given a scenario, finding the optimal ST-APG includes two sub-tasks: (1) partitioning tracklet set Γ into trajectories belonging to different people Γ_i, (2) inferring the semantic human attributes for each person. Noticing that sub-task (1) is a combinatorial optimization problem and jointly solving these two sub-tasks is infeasible, we therefore propose an inference algorithm to optimize these two sub-tasks iteratively. For sub-task (1), we apply a stochastic clustering algorithm, i.e., Swensden-Wang Cuts (Barbu and Zhu 2005), which could efficiently and effectively traverses through the grouping solution space. For sub-task (2), given grouped tracklets, we can use Dynamic Programming to update the semantic attributes of tracklets within every group (i.e., person trajectory). These two algorithms are are iterated one after another until convergence.

Associating Tracklets by Stochastic Clustering

Traditional sampling algorithms usually suffer from the efficiency issues. On the contrary, cluster sampling algorithm overcomes this issue by randomly grouping clusters and resampling cluster as a whole. The algorithm consists of two steps:

(I) **Generating cluster set**. Given an adjacency graph $G = <\Gamma, E>$ and the current state M, we regard every edge e_{ij} in this graph as a switch. We turn on every edge e_{ij} probabilistically with its edge probability p_e. Afterwards, we regard candidates connected by "on" positive edges as a cluster V_{cc} and collect separate clusters to produce the cluster set.

(II) **Relabeling cluster set**. We randomly choose a cluster V_{cc} from the produced cluster set and randomly change the label of the selected cluster, which generates a new state M'. This is essentially changing the ID of a group of tracklets. This group of tracklets can either be merged into another trajectory, or set to background noises. Following the Markov chain Monte Carlo principal, we accept the transition from state M to new state M' with a rate $\alpha(\cdot)$ defined by the Metropolis-Hastings method (Metropolis et al. 1953):

$$\alpha(M \to M') = \min(1, \frac{p(M' \to M) \cdot p(M'|\Gamma)}{p(M \to M') \cdot p(M|\Gamma)}), \tag{14}$$

where $p(M' \to M)$ and $p(M \to M')$ are the state transition probability, $p(M'|\Gamma)$ and $p(M|\Gamma)$ the posterior defined in Equation.(6). This guarantees the stochastic algorithm can find better states and obtains reversible jumps between any two states.

Following instructions in (Barbu and Zhu 2005), the transition probability ratio can be calculated as

$$\frac{p(M' \to M)}{p(M \to M')} \propto \frac{p(V_{cc}|M')}{p(V_{cc}|M)} \propto \frac{\prod_{e\in E^\star_{M'}}(1-p_e)}{\prod_{e\in E^\star_M}(1-p_e)}, \tag{15}$$

where $E^\star$ denotes the sets of edges being turned off around V_{cc}, that is,

$$E^\star = \{e \in E : \tau_i \in V_{cc}, \tau_j \notin V_{cc}, l(\tau_i) = l(\tau_j)\}. \tag{16}$$

Assigning Semantic Attributes by DP

Given a trajectory, we first find trajectory gaps (i.e., no bounding box presented) below 60 frames, we then apply a linear interpolation to fill-in the missing bounding boxes.

After that, assigning the semantic attribute is similar to estimating the likelihood term $p(\tau_i|M)$. The whole trajectory is also treated as a Markov chain structure. We therefore apply the standard factor graph belief propagation (sum-product) algorithm to infer the semantic human attributes of a trajectory.

A short summary of our proposed inference algorithm is shown in Algorithm 1.

Experiment

To evaluate the proposed method, we compare with other state-of-the-arts using two datasets:

(1) **CAMPUS dataset** (Xu et al. 2016). This is a newly published dataset targeting multi-view tracking. There are four sequences, i.e., two gardens, parking lot, auditorium, each of which is shot by 3-4 1080P cameras. The recorded videos are 3-4 minutes long and with 30fps. This dataset contains people with huge pose variations and lots of actions (e.g., running, riding bikes, sitting), providing richer semantic human attributes.

(2) **PPL-DA dataset**. We collect a new dataset aiming to cover people's daily activities. The new dataset consists of 3 public facilities: foot court, office reception, plaza. The scenes are recorded with 4 GoPro cameras, mounted on around 1.5 meters high tripods. The produced videos are also around 4 minutes long and in 1080P high quality. We further annotate the trajectories of every person inside the scene with cross-view consistent ID.

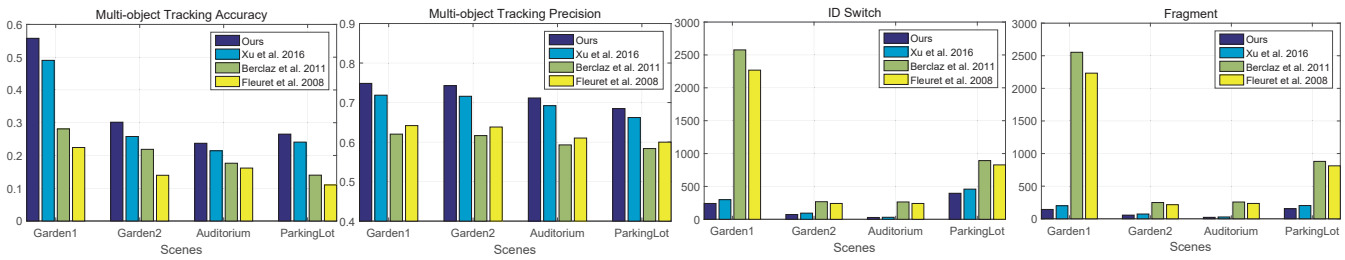

Figure 4: Comparison charts of major metrics on CAMPUS datasets.

Seq-Court	TA(%)	TP(%)	MT(%)	ML(%)	IDSW	FRG
Our-full	**34.47**	**72.38**	**18.52**	**25.93**	**79**	**55**
Our-1	26.82	70.23	11.11	33.33	114	90
HTC	29.51	71.87	14.81	25.93	91	77
KSP	24.72	64.40	0.00	44.44	318	291
POM	22.26	65.39	0.00	51.85	296	269
Seq-Office	TA(%)	TP(%)	MT(%)	ML(%)	IDSW	FRG
Our-full	**47.38**	**73.70**	**42.86**	**0.00**	**45**	**31**
Our-1	39.79	68.99	28.57	0.00	71	64
HTC	41.17	70.65	28.57	0.00	66	59
KSP	39.62	58.01	28.57	0.00	83	76
POM	36.86	58.77	28.57	0.00	89	82
Seq-Plaza	TA(%)	TP(%)	MT(%)	ML(%)	IDSW	FRG
Our-full	**25.18**	**67.10**	**16.28**	**11.63**	**165**	**133**
Our-1	20.59	65.15	11.63	18.60	244	199
HTC	23.11	66.24	11.63	18.60	202	178
KSP	17.30	57.49	6.98	27.91	356	311
POM	16.71	57.87	4.65	32.56	339	295

Table 1: Quantitative results and comparisons on PPL-DA dataset. Our-1 and Our-full are two variants of the proposed framework. See text for detailed explanations.

For both datasets, we incorporate 10% of the videos as augmented training set and the rest as testing set. The augmented data, together external dataset described in previous section, helps us learn the action labels and transitions. The learning process is only done once and applied to both datasets. All parameters are fixed in the experiment. We use fast r-cnn (Girshick 2015) to generate people's bounding boxes. The pruning threshold is set to 0.3. We apply Sequential Shortest Path (SSP) (Pirsiavash, Ramanan, and Fowlkes 2011) to initialize tracklets. The sampling is set to finish after 1000 iterations, which achieves decent results.

The proposed approach is compared with 3 state-of-the-arts methods: Probabilistic Occupancy Map (POM) (Fleuret et al. 2008), K-Shortest Path (KSP) (Berclaz et al. 2011) and Hierarchical Trajectory Composition (HTC) (Xu et al. 2016). The public implementations of POM and KSP are adopted. We further implement HTC on our own, using the default parameters mentioned in their paper. For quantitative results, we apply multi-object tracking accuracy (TA), multi-object tracking precision (TP), mostly tracked/lost trajectories (MT/ML), identity switches (IDSW) and trajectory fragments (FRG). DA, DP, TA and TP mainly measure the

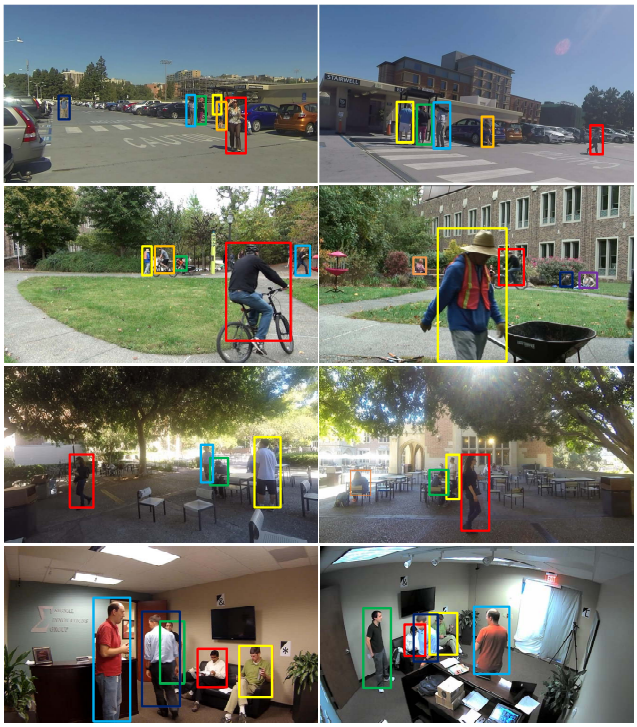

Figure 5: Sampled qualitative results of our proposed method on CAMPUS and PPL-DA datasets.

percentage of true positives while MT/ML, IDSW and FRG mainly measure the completeness and identity consistency of the result trajectories. A higher value means better for TA, TP and MT while a lower value means better for ML, IDSW and FRG.

We report quantitative results on CAMPUS datasets in Fig. 4 and on PPL-DA dataset in Table 1. From the results, the proposed method obtains a significant improvement over the competing methods. An interesting observation is that tracking by associating bounding boxes (i.e., KSP, POM) yields much worse results than tracking by associating tracklets (i.e., Ours, HTC).

We set up a baseline **Our-1** to further analyze the effectiveness of leveraged semantic attributes. **Our-1** only uses appearance and geometry information for multi-view tracking. From the results we can observe that when people with

various actions present, the proposed method is able to exploit this visual information and significantly improves the tracking results. However, when lack of such variations (e.g., Auditorium, ParkingLot, Plaza), the proposed method can only utilize people motion information and obtains slightly better results. some qualitative results are visualized in Fig. 5.

We implement the proposed method with MATLAB and test it on a workstation with I7 3.0GHz CPU, 32GB memory and GTX1080 GPU. For a scene shot by 4 cameras and lasting for around 4 minutes, our algorithm obtains 5 frames per second on average. With further code optimization and batch-based data parallelization, our proposed method can run in realtime.

Conclusion

In this paper, we propose a novel multi-view multi-object tracking approach. Tracking people is leveraged with rich semantic attributes and therefore the association of tracklets are further constrained. By incorporating the motion attributes, pose attributes and action attributes, our algorithm outperforms the competing methods only using appearance and geometry information. In the future, we will continue to explore more high-level information (e.g., people interactions, group information) among tracklets and more efficient inference algorithms.

References

Andriyenko, A., and Schindler, K. 2011. Multi-target tracking by continuous energy minimization. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Ayazoglu, M.; Li, B.; Dicle, C.; Sznaier, M.; and Camps, O. 2011. Dynamic subspace-based coordinated multicamera tracking. In *IEEE International Conference on Computer Vision*.

Barbu, A., and Zhu, S. 2005. Generalizing swendsen-wang to sampling arbitrary posterior probabilities. *IEEE Transaction on Pattern Analysis and Machine Intelligence* 27(8):1239–1253.

Berclaz, J.; Fleuret, F.; Turetken, E.; and Fua, P. 2011. multiple object tracking using k-shortest paths optimization. *IEEE Transaction on Pattern Analysis and Machine Intelligence* 33(9):1806–1819.

Dehghan, A.; Assari, S.; and Shah, M. 2015. Gmmcp-tracker:globally optimal generalized maximum multi clique problem for multiple object tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Fleuret, F.; Berclaz, J.; Lengagne, R.; and Fua, P. 2008. Multicamera people tracking with a probabilistic occupancy map. *IEEE Transaction on Pattern Analysis and Machine Intelligence* 30(2):267–282.

Girshick, R. 2015. Fast r-cnn. In *IEEE International Conference on Computer Vision*.

Jiang, H.; Fels, S.; and Little, J. 2007. A linear programming approach for multiple object tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Khan, S., and Shah, M. 2006. A multiview approach to tracking people in crowded scenes using a planar homography constraint. In *European Conference on Computer Vision*.

Kuo, C., and Nevatia, R. 2010. How does person identity recognition help multi-person tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Liu, X.; Lin, L.; and Jin, H. 2013. Contextualized trajectory parsing via spatio-temporal graph. *IEEE Transaction on Pattern Analysis and Machine Intelligence* 35(12):3010–3024.

Metropolis, N.; Rosenbluth, A.; Rosenbluth, M.; Teller, A.; and Teller, E. 1953. Equation of state calculations by fast computing machines. *Journal of Chemical Physics* 21(6):1087–1092.

Nie, B. X.; Xiong, C.; and Zhu, S. 2015. Joint action recognition and pose estimation from video. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Park, S., and Zhu, S. 2015. Attributed grammars for joint estimation of human attributes, parts and poses. In *IEEE International Conference on Computer Vision*.

Pirsiavash, H.; Ramanan, D.; and Fowlkes, C. 2011. Globally-optimal greedy algorithms for tracking a variable number of objects. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Porway, J., and Zhu, S. 2011. C4 : Computing multiple solutions in graphical models by cluster sampling. *IEEE Transaction on Pattern Analysis and Machine Intelligence* 33(9):1713–1727.

Shitrit, H.; Berclaz, J.; Fleuret, F.; and Fua, P. 2013. Multi-commodity network flow for tracking multiple people. *IEEE Transaction on Pattern Analysis and Machine Intelligence* 36(8):1614–1627.

Shu, T.; Xie, D.; Rothrock, B.; Todorovic, S.; and Zhu, S. 2015. Joint inference of groups, events and human roles in aerial videos. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Sun, J.; Zheng, N.; and Shum, H. 2003. Stereo matching using belief propagation. *IEEE Transaction on Pattern Analysis and Machine Intelligence* 25(7):787–800.

Utasi, A., and Benedek, C. 2011. A 3-d marked point process model for multi-view people detection. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Wei, P.; Zhao, Y.; Zheng, N.; and Zhu, S. 2016. Modeling 4d human-object interactions for joint event segmentation, recognition, and object localization. *IEEE Transaction on Pattern Analysis and Machine Intelligence*.

Weng, E., and Fu, L. 2011. On-line human action recognition by combining joint tracking and key pose recognition. In *IEEE/RSJ Conference on Intelligent Robots and Systems*.

Wu, Z.; Betke, M.; and Kunz, T. 2011. Efficient track linking methods for track graphs using network-flow and set-cover techniques. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Wu, Z.; Hristov, N.; Hedrick, T.; Kunz, T.; and Betke, M. 2009. Tracking a large number of objects from multiple views. In *IEEE International Conference on Computer Vision*.

Xu, Y.; Lin, L.; Zheng, W.; and Liu, X. 2013. Human re-identification by matching compositional template with cluster sampling. In *IEEE International Conference on Computer Vision*.

Xu, Y.; Ma, B.; Huang, R.; and Lin, L. 2014. Person search in a scene by jointly modeling people commonness and person uniqueness. In *ACM Multimedia*.

Xu, Y.; Liu, X.; Liu, Y.; and Zhu, S. 2016. Multi-view people tracking via hierarchical trajectory composition. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Yao, A.; Gall, J.; Fanelli, G.; and Gool, L. 2011. Does human action recognition benefit from pose estimation? In *British Machine Vision Conference*.

Zhang, S.; Yu, X.; Sui, Y.; Zhao, S.; and Zhang, L. 2015. object tracking with multi-view support vector machines. *IEEE Transaction on Multimedia* 17(3):265–278.