A Stochastic Attribute Grammar for Robust Cross-View Human Tracking

Xiaobai Liu, Yuanlu Xu, Lei Zhu, and Yadong Mu

Abstract—**In computer vision, tracking humans across camera views remain challenging, especially for complex scenarios with frequent occlusions, significant lighting changes, and other difficulties. Under such conditions, most existing appearance and geometric cues are not reliable enough to distinguish humans across camera views. To address these challenges, this paper presents a stochastic attribute grammar model for leveraging complementary and discriminative human attributes for enhancing cross-view tracking. The key idea of our method is to introduce a hierarchical representation, *parse graph*, to describe a subject and its movement trajectory in both space and time domains. These results in a hierarchical compositional representation, comprising trajectory entities of varying level, including human boxes, 3D human boxes, tracklets, and trajectories. We use a set of grammar rules to decompose a graph node (e.g., tracklet) into a set of children nodes (e.g., 3D human boxes), and augment each node with a set of attributes, including geometry (e.g., moving speed and direction), accessories (e.g., bags), and/or activities (e.g., walking and running). These attributes serve as valuable cues, in addition to appearance features (e.g., colors), in determining the associations of human detection boxes across cameras. In particular, the attributes of a parent node are inherited by its children nodes, resulting in consistency constraints over the feasible parse graph. Thus, we cast cross-view human tracking as finding the most discriminative parse graph for each subject in videos. We develop a learning method to train this attribute grammar model from weakly supervised training data. To infer the optimal parse graph and its attributes, we develop an alternative parsing method that employs both top-down and bottom-up computations to search the optimal solution. We also explicitly reason the occlusion status of each entity in order to deal with significant changes of camera viewpoints. We evaluate the proposed method over public video benchmarks, and demonstrate with extensive experiments that our method clearly outperforms the state-of-the-art tracking methods.**

Index Terms—**Object tracking, multi-view, video grammar.**

I. Introduction

A. Background and Motivations

TRACKING humans across multiple cameras while observing them moving in the scene has been playing a critical role in most high-level video understanding tasks, e.g., activity recognition, and thus has attracted many attentions in the past decade [50]. Most existing methods employ appearance cues [20] to train discriminative [53] or generative models [44] with shallow [7], [19] or deep [44] representations. The recent technical breakthroughs in deep learning techniques [17], [38] achieved remarkable improvements in multiple recognition problems, e.g., face, audio, etc. These techniques, however, have two limitations which restrict their applications in surveillance systems. (I) Deep learning based methods are mostly driven by the availability of large-scale labeled data and the effective end-to-end training on powerful computing devices, which are difficult to collect for human tracking tasks. (II) Appearances of the same person might be significantly different while observing he/she moving in the scene because of the varying imaging conditions (illuminations, occlusions, etc.). To address these limitations, in this work, we will develop a unified multi-view human tracking framework to leverage the advantages of deep representations with minimal efforts of data preparation.

The proposed tracking solution is motivated with the fact that a human observer can robustly identify persons who appear in multiple surveillance areas of complex scenarios. Such a correspondence problem is the core of cross-view human tracking task. While the intrinsic working schema of human brain remains unclear, it is well accepted that we human being can immediately perceive object's attributes and use them to guide the matching process [15]. A few typical examples are shown in Figure 1 which includes four camera views of the same scene at a certain time. The three persons in subfigure (a) are performing different activities, i.e. playing baseball, walking, and standing, respectively. The re-identification of these three persons in other three camera views becomes relatively straightforward if we can recognize the activity labels of the detected human boxes. The other possible cues for boosting cross-view identification task include (I) accessories, e.g., wearing hats or t-shirt, holding baseball bat; and (II) geometry information, i.e., facing into or walking toward a landmark (e.g. the building). These attributes directly confine the search space of cross-view human re-identification as well as cross-view human tracking. Moreover, in the

Manuscript received January 25, 2017; revised August 13, 2017; accepted October 6, 2017. Date of publication December 11, 2017; date of current version October 24, 2018. The work of X. Liu was supported by the NSF Program under Grant 1657600. The work of Y. Mu was supported by the NSF China Program under Grant 61772037. This paper was recommended by Associate Editor W. Zuo. *(Corresponding author: Xiaobai Liu.)*

X. Liu is with the Department of Computer Science, San Diego State University, San Diego, CA 92125 USA (e-mail: xbliu.lhi@gmail.com).

Y. Xu is with the Department of Computer Science, University of California at Los Angeles, Los Angeles, CA 92104 USA.

L. Zhu is with the School of Information Science and Engineering, Shandong Normal University, Jinan 250358, China.

Y. Mu is with the Institute of Computer Science and Technology, Peking University, Beijing 100000, China.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCSVT.2017.2781738

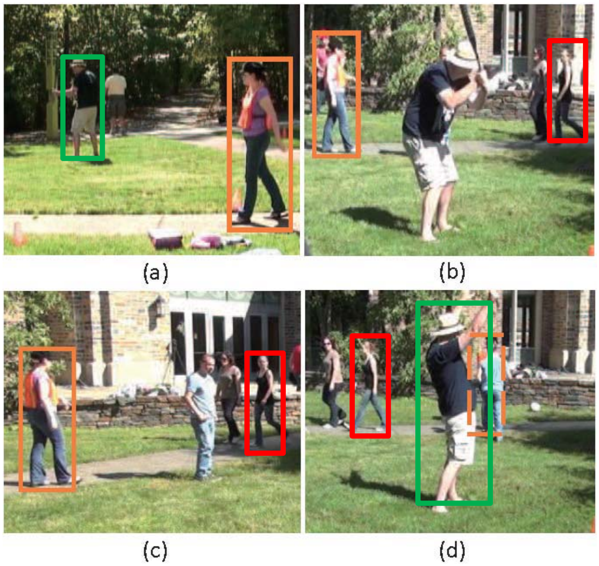


Fig. 1. Human tracking across camera views. (a)-(d): four camera views. The subject in orange has significantly different appearance while being observed in different camera views, and is occluded by other subjects in the view (d). This subject can be readily tracked if we can identify its attributes, e.g., activities, accessories, moving direction etc.

past decade most such recognitions have reached a level of accuracy, even under the various challenges (e.g., illumination changes, occlusions) [12].

There is thus a demand of leveraging various human attributes, either semantic or geometric, static or dynamic, for robust multi-view human tracking in videos. Though promising, a critical problem of this methodology is *how to deal with the potential errors made in the recognitions of human attributes*. In this work, we will introduce a stochastic grammar model to exploit human attributes as extra cues, and develop an unified energy minimization formula to avoid premature decisions during inference. In comparisons to previous efforts, our method will contribute in hierarchical representation of human trajectories, bottom-up and top-down inference, and effective learning (see the next subsection). Figure 1 shows a person in subfigure (a) and the tracked boxes of the same subject in other three camera views.

B. Overview

The objective of this work is aimed at developing a robust solution to cross-view human tracking in complex scenarios, which might include various challenges, e.g., low resolution, frequent occlusions, significant illumination changes, etc. Under such conditions, appearance information are not reliable enough to identify the same subject across camera views and thus lead to errors (e.g. ID switches) in human tracking.

We propose an attribute grammar model for robust human tracking. Our model embraces two principles. (I) *Composition*. We describe the tracking of a human as a composition process, which decomposes a human trajectory into tracklets, 3D human boxes, and detection boxes, resulting in a hierarchical graphical structure, called *parse graph*. Our grammar model comprises of a few grammar rules, each of which is used to generate graph nodes of parse graphs. These grammar rules explicitly define the composition process, e.g., associating multiple boxes in different camera views to be a 3D human box, grouping multiple sequent 3D boxes as a tracklet, etc. (II) *Attribution*. In parse graph, we augment each graph node with a set of attributes, including appearance (e.g., color, texture, gradients), motion (e.g., speed, direction), accessories (e.g., glasses, bags, purses), and activities (e.g., walking, running, turning). Attributes of a node will be inherited by its children nodes, resulting in consistency constraints between sibling nodes. Thus, given a video sequence, our goal is to retrieve the most probable parse graph subjecting to various attribute constraints.

We formulate the construction of parse graph from videos as an energy minimization problem [34]. The energy function of a candidate parse graph is a linear combination of energy terms over individual graph nodes, of which describe the dissimilarities between sibling nodes regarding their various attributes. The weights of these terms are discriminatively learned from weakly supervised training samples. Thus, the energy function is used to measure how plausible a parse graph is as a valid trajectory representation.

We develop a bottom-up and top-down inference algorithm to retrieve the most probable parse graph for each subject in videos. With an initial parse graph, our algorithm follows the Metropolis-Hasting principle [34] to reconfigure the current parse graph using a set of dynamics so as to simulate a Marko Chain in the joint solution space. Those dynamics will select a subtree of graph nodes in either top-down or bottom-up fashions and assign new subject ID to the nodes of the selected subtree. We introduce a binary indicator variable for terminal nodes, i.e. human boxes in individual camera views, and explicitly infer its status. We will also propagate attributes through the tree-structure from parent nodes to children nodes, i.e. in a top-down fashion. In comparisons to previous sampling methods [19], the designs of our method will be able to make distant proposals, i.e. proposals those are far from the current solution but still have high probabilities to be accepted.

C. Relationships to Previous Works

This work is closely related to the following research streams in computer vision.

Multi-view object tracking is often formulated as a data association task. A key research question is: how to find cross-view correspondence at either pixel level [32] or region-level [2], [13] or object-level [43]. Typical data association methods are developed based on integer programming [11], network flow [4], [42], marked point process [35], multi-commodity network [30], and multi-view SVM [53]. Among these approaches, sampling techniques bear the advantages of solving intractable optimization and have been extensively studied in the past literature. For example, Khan and Shah [13] integrated Markov Chain Monte Carlo method with particle filer tracking framework. Yu et al. [51] utilized single site sampler for associating foreground blobs to trajectories. Liu et al. [19] introduced a spatial-temporal graph to jointly solve region labeling and object tracking by Swendsen-Wang Cut method [3]. While promising, all these algorithms use shallow representations which are sensitive to various challenges (e.g., illuminations changes). In this work, we propose to integrate sampling techniques with deep representation of human trajectories and design a set of reversible dynamics that can efficiently search the joint solution space.

Tracking under wild conditions Tracking subjects of interests across multiple camera views with wide-baselines is essentially an identification problem and the most popular features are extracted based on appearance information (e.g., color, gradient). However, in these scenes with significant illumination changes or frequent occlusions, appearance information are not reliable, as shown in Figure 1. In a particular camera view, a subject might be occluded by other objects or is not visible. To address these fundamental challenges, a natural solution is to integrate high-level recognition outcomes with human tracking [5]. Moreover, Yang et al. [47] explicitly addressed occlusions in a probabilistic framework for multi-target tracking. Zhang et al. [52], Henriques et al. [9] and Pirsiavash et al. [26] introduced global optimization frameworks to track objects over long-range, which are helpful to recovering trajectories from occlusions. Milan et al. [22] addressed multiple object tracking by defining bi-level exclusions. Wang et al. [37] proposed to infer tracklets, i.e. short trajectories, and further solved data association problem. Possegger et al. [28] relied on geometric information to efficiently overcome detection failures when objects are significantly occluded. These algorithms achieved promising results but are restricted to shallow data representations and lacks of formal modeling of human attributes. In this work, we develop an attribute grammar to fill in this gap and demonstrate its superiorities over alternative tracking methods.

Joint video parsing with multiple objectives has been approved to be an effective way for boosting the performance of individual objectives. For example, Wei et al. [39] introduced a probabilistic framework for joint event, recognition, and object localization. Shu et al. [31] proposed to jointly infer groups, events, and human roles in aerial videos. Nie et al. [24] employed human poses to improve action recognition. Park and Zhu [25] introduced an stochastic grammar to jointly estimate human attributes, parts and poses. Weng and Fu [40] utilized trajectories and key pose recognitions to improve human action recognition. Yao et al. [49] investigated how to use pose estimation to enhance human action recognition. Kuo and Nevatia [16] studied how person identity recognition can help multi-person tracking. Xu et al. [45] developed a spatial-temporal reasoning framework for jointly exploiting appearance, gestures, and actions of humans for robust tracking. In this work, we follow the same methodology and present a stochastic attribute grammar, as a formal language, for joint video parsing. Our parsing framework can leverage various semantic human attributes, including orientations, poses, and actions, to narrow the search space in cross-view tracking task, and significantly improve tracking robustness and accuracies.

D. Contributions and Organizations

The three contributions of this work include (i) a stochastic attribute grammar model capable of integrating a diverse set of human attributes for robust cross-view human tracking in complex scenarios; (ii) an effective computational framework that can learn grammar model from weakly supervised training data and infer the most probable parse graph for each subject in videos; and (iii) state-of-the-art performance on both public video datasets and newly collected videos.

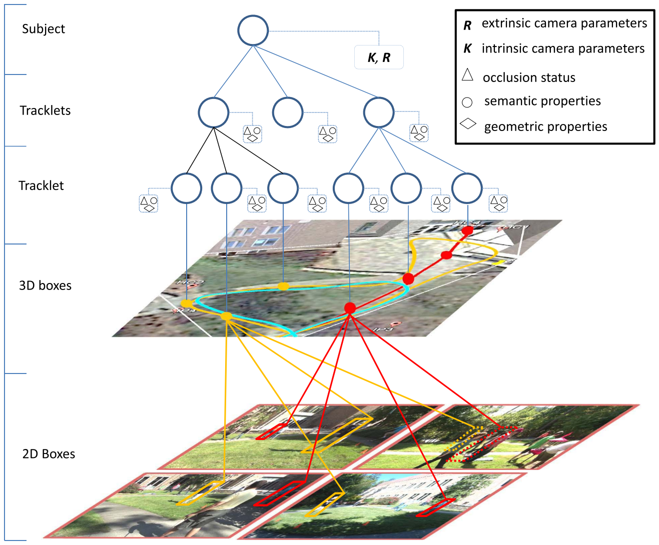


Fig. 2. Parse graph for a human trajectory. A video might include more than one trajectories and thus multiple parse graphs.

The rest of this paper is organized as follows. In section II, we introduce the proposed stochastic grammar for human tracking problem. In section III, we present how to efficiently learn the grammar model from training data and perform effective inference in videos. In section IV, we report evaluation results of the proposed models and alternative methods on public video datasets. In section V, we conclude this work and remark the future research directions.

II. Stochastic Attribute Grammar for Cross-view Human Tracking

This section presents a stochastic attribute grammar model for cross-view human tracking.

A. Compositional Human Representation

We develop a compositional representation to describe the moving trajectory of a human in videos. Figure 2 illustrates the proposed graph representations, which embodies two principles.

(I) **Composition**. As illustrated, a human trajectory (top row) can decompose into multiple tracklets, a tracklet comprises of multiple 3D human boxes, and a 3D box corresponds to multiple 2D human boxes in individual cameras. This hierarchical decomposition results in a tree-like structure, i.e. Parse Graph, which includes both terminal nodes (i.e. human boxes) and intermediate nodes (tracklets). For each node, there are often more than one ways of compositions, and thus the compositional process must determine the most probable structure in the compositional space.

TABLE I
LIST OF NINE HUMAN ATTRIBUTES

Category	Property	Exemplar Values
Geometry	Direction	vector
	Speed	scalar
Accessories	Glasses	‘Yes’, ‘No’
	Bags	‘Yes’, ‘No’
	Clothes	‘T-shirt’, ‘Coat’, ‘Suit’
	Hats	‘Yes’, ‘No’
Semantics	Activities	‘walking’, ‘running’, ‘riding bike ’
	Gesture	’standing’, ‘sitting’, ’bending’
	Gender	‘male’, ‘female’

(II) **Attribution**. Every graph node in the hierarchy represents a trajectory, a tracklet, a 3D human box, or a 2D human box, and is associated with a set of geometric and semantic properties. For example, geometric attributes include moving directions, locations, poses, occlusion statuses, and moving speed. Semantic attributes include accessories (e.g. bags, clothes) and activities (e.g., walking, running). We cluster these attributes into three groups, and summarize them in Table I. These attributes are used as complementary discriminative information for identifying humans across camera views that have significant viewpoint changes or illumination variances. In such scenes, the conventional appearance- or motion- based measures are not reliable and usually result in failures of tracking. In contrast, object attributes (e.g., gender, activities), once recognized, are intrinsically invariant against illuminations or viewpoint changes and are thus fairly reliable in complex scenarios.

The proposed attributed parse graph serves as a redundant and informative deep representation for human trajectories in videos. In comparisons to the previous shallow representation [19], [50], attributed parse graph has the following two advantages. *First*, our method allows different trajectory entities of the same subject to be grouped based on different cues. For example, our method might group tracklets A and B together since they have similar moving speed, and group tracklets B and C together since they both wear a red hat. *Second*, our representation model can adaptively exploit both low-level attributes (e.g., speed) and high-level attributes (e.g., activities), which is critical to the success of human tracking in complex scenario. *Third*, as a redundant representation, for a given video sequence the optimal parse graph for a subject is not unique which means that even local minimal solution can still convey plausible parse graph for tracking purpose.

B. Stochastic Grammar

We introduce a stochastic grammar model to guide the construction of attributed parse graphs.

Definition 1: The movement trajectory of a subject in videos is described using a stochastic grammar, specified by a five-tuple $G = (\mathcal{T}, \mathcal{N}, P, S, R)$.

In the representation model G, T denotes terminal nodes, N denotes non-terminal nodes, i.e. tracklets, P the probabilistic models, S the root node standing for a subject, and $R = \{r_1, r_2, \ldots\}$ a set of grammar rules. Note that a parse graph G is used to describe a single subject, and a video sequence might include multiple subjects.

Our grammar model comprises of a set of grammar rules $r : A \rightarrow (A_1, A_2, \ldots,)$, each of which defines a type of generation relationships between a parent node $A \in \mathcal{N}$ and its children nodes $A_i \in \mathcal{N} \cup T$. A children node can further decompose into a set of children nodes. These production rules will be applied recursively to generate a parse graph, representing a subject's trajectory in videos. In this work, we define four grammar rules.

- $r_1 : S \rightarrow (A_1, A_2, \ldots,)$, where the parent node S represents a trajectory and the children nodes are tracklets.
- $r_2 : A \rightarrow (A_1, A_2, \ldots,)$ where the parent node A denotes a tracklet and the children nodes are either tracklets or 3D human boxes. This grammar rule is used to recursively decompose a tracklet into finer-level tracklets.
- $r_3 : A \rightarrow (A_1, A_2, \ldots,)$ where A and A_i denote 3D human boxes and 2D human boxes, respectively.
- $r_4 : A \rightarrow a$ which instantiates a non-terminal node A to be a terminal node a. Herein, A represents a 2D human box, and a represents the visual observation of A in images.

Among the above grammar rules, the rule r_2 can be recursively applied, and the rest rules will appear multiple times at certain levels of a parse graph. In contrast to the recursive grammar model, e.g., Liu et al. [21], our grammar model is non-recursive and generates much less plausible parse graphs.

Definition 2: A parsing graph G is a tree structure expanded from a root node by a sequence of grammar rules while respecting the various attribute constraints.

We can expand a nonterminal node to a collection of nonterminal or terminal nodes by applying grammar rules sequentially. Each expansion generates a subtree. A terminal node represents a human box detected in videos. A nonterminal node $A \in \mathcal{N}$ represents a sequence of human boxes over a certain period of time (i.e. tracklets), or a cluster of 2D human boxes across multiple camera views at the same time-point.

C. Attributes and Constraints

We augment every graph node with a set of attributes to describe subject's states in space-time domain. The attributes of a non-terminal node $A \in \mathcal{N}$ are defined as follows:

$$X(A) = (\mathbf{v}, l, t) \tag{1}$$

where $\mathbf{v}$ are a set of attribute values,l the center location of node A in the scene, and t the time-stamp. Location coordinates are defined on a reference camera view which can be projected into other camera views or a world reference coordinate [18]. Similarly, the attributes of a terminal node $a \in \mathcal{T}$ are defined as follows

$$X(a) = (o, l, t) \tag{2}$$

where the binary variable $o \in 1, 0$ indicates the visibility of the terminal node a. We have $o = 1$ if a subject is visible in the current camera view; otherwise, $o = 0$.

In a parse graph, the attributes $X(A)$ of a parent node A will be inherited by its offspring nodes, which imposes a set of constraint equations. For a graph node $A \rightarrow (A_1, A_2)$, the associated equations are defined over the attributes of A and A_1, A_2:

$$g_i[X(A)] = f_i[X(A_1), X(A_2)], \quad i = 1, 2, 3, \ldots \tag{3}$$

where $g_i(), f_i()$ are projection functions of the attribute vectors. For instance, let $X(A) = (X_1, X_2)$ and $X(B) = (X_1, X_2, X_3)$, $B \in A.Child$. Table I summarizes three groups of attributes used in this work. Then an equation could be simply an equivalence constraint (or assignment) for passing the information between nodes A and A_1 in either directions,

$$A.X_1 = B.X_1 \tag{4}$$

The above equation is also used to define two parsing procedures. (I) *bottom-up* message parsing, which passes the attributes of a child node (i.e. $B.X_1$) to its parent node $A.X_1$. (II) *top-down* message parsing, which passes the attributes of a parent node (i.e. $A.X_1$) to its children nodes (i.e. $B.X_1$). We will develop an inference algorithm that alternates these two procedures for effective computing.

Proposition 1: Human detections of the identical person are terminal nodes of the same parse graph.

The above proposition holds by the definition of parse graph. Thus, the cross-view human tracking problem can be cast as finding the most probable parse graph from videos. Specifically, given human boxes detected in videos, we will apply the grammar rules to group these 2D detections across camera views to form 3D human boxes, associate 3D human boxes to get tracklets and cluster tracklets to obtain human trajectories. Among the composition process, the attributes of levels of nodes should be consistently assigned so as to ensure all attribute constraints are satisfied. It is noteworthy that we will create multiple parse graphs from the input videos, each corresponding to one of the subjects in the scene.

D. Energy Function

We formulate the construction of parse graphs as an energy minimization problem (or maximizing a posterior probability). We define the energy of a parse graph G to be the sum of the energies of non-terminal graph nodes in G plus the energy of placing terminal nodes in video frames.

$$E(G) = \sum_{A \in \mathcal{N}} E^{attr}(A) + E^{term}(\mathcal{T}) \tag{5}$$

For each non-terminal node A, its energy is defined over the attributes,

$$E^{attr}(A) = \sum_{B \in A.Child} \sum_{k} \mathbf{1}(f_k(A) \neq g_k(B)) \tag{6}$$

where k is the index of all valid attribute constraints between non-terminal nodes A and B.

We define the energy term $E^{term}(T)$ to be the sum of appearance in-consistency energies between visible terminal nodes plus the constant penalties of placing nodes as invisible.

$$E^{term}(\mathcal{T}) = \sum_{a_i \neq a_j} o_i \cdot o_j \cdot E^{term}(a_i, a_j) + \sum_{a_i \in \mathcal{T}} \beta \cdot \mathbf{1}(o_i = 0) \tag{7}$$

where $o_i \in \{1, 0\}$ indicates the visibility of a node a_i, and β is a constant penalty. We define the appearance models for terminal nodes by associating a filter $\mathbf{H}$. Then $E^{term}(a_1, a_2) = \mathbf{H} \cdot (\phi(a_1) - \phi(a_2))$ is the dot product between the filter coefficients and the feature vectors extracted from the two human detections. We will introduce the extraction of feature vectors $\phi()$ in Section IV.

The energy function of Eq. (5) directly encodes attribute constraints over non-terminal nodes, appearance consistencies between terminal nodes, and visibilities of individual graph nodes. The model parameters will be automatically learned from weakly supervised data as introduced in the next section.

III. Inference and Learning

In this section we introduce the developed inference and learning algorithms for cross-view human tracking.

A. Inference for Cross-View Human Tracking

The goals of our inference algorithm are two-fold: (i)*model selection*, to determine the number of subjects in videos; (ii) *state estimation*, to construct the optimal parse graph for every subject, i.e., model estimation.

We solve the above two goals jointly in a Bayesian framework by maximizing a posterior probability $P(\mathbf{G}|\mathbf{I})$ where $\mathbf{G}$ pool all the parse graphs desired in the input videos $\mathbf{I}$. According to Bayes rules, we have:

$$P(\mathbf{G}|\mathbf{I}) \propto \prod_{m} P(\mathbf{I}|G_m) P(\mathbf{G}) \tag{8}$$

We define the likelihood model using the energy functions (5), i.e. $P(\mathbf{I}|G_m) = \exp\{-E(G_m)/\mathcal{K}\}$ where $\mathcal{K}$ is a constant. We define the prior model to encourage small number of parse graph $P(\mathbf{G}) = \exp\{-|\mathbf{G}|\}$ where $|\mathbf{G}|$ is the number of subjects.

We develop an efficient cluster sampling algorithm following the data-driven Markov Chain Monte Carlo (MCMC) schema [34]. Traditional sampling techniques, e.g. Gibbs sampler [14], often suffer from efficiency issues. In contrast, clustering sampling methods [3], [27] will group variables into clusters and re-label a cluster of nodes together. In our method, with an initial parse graph $\mathbf{G}$, we use a set of dynamics to reconfigure $\mathbf{G}$ and accept the the new state $\mathbf{G}'$ with a probability. The acceptance probability is defined following the Metropolis-Hasting strategy [34]:

$$\min \left[1, \frac{P(\mathbf{G}'|I) Q(\mathbf{G} \to \mathbf{G}')}{P(\mathbf{G}|I) Q(\mathbf{G}' \to \mathbf{G})} \right] \tag{9}$$

where $Q(\mathbf{G}' \to \mathbf{G}))$ is the proposal probability.

Algorithm 1 summarizes the proposed inference algorithm. We use five dynamics that specify either jump or diffusion moves between solution states. Jump dynamics are paired with each other to preserve detail-balancing in random walk. In the rest of this subsection, we first introduce the initializations of $\mathbf{G}$ and then introduce the designs of five dynamics.

Initializations Our inference algorithm comprises of the following three types of pre-processing.

- *C*ross-view camera calibration. To obtain the projection matrix between two camera views, we follow the conventional structure-from-motion pipeline [33]. It comprises of detection of interests point, finding corresponding using RANSAC method, and performing bundle adjustment method to obtain the camera motions.

- *H*uman detections. We employ the popular Faster Region-based Convolution Neural Network method [29] to detect humans in videos. We fine-tune the pre-trained network models over our training videos.
- *R*ecognitions of human attributes. Given a terminal node or non-terminal nodes (e.g. tracklets), we can directly estimate its speed and moving direction from visual inputs. The recognitions of other attributes, e.g., accessories, activities, gestures and genders, will need off-line training of machine learning models. We will introduce the training of human recognition modules in Section IV.

Dynamic I and II: addition/deletion of parse graphs. This pair of dynamics are used to add a new parse graph (or a subject) or remove one of the parse graphs in **G** at each iteration. As To add a new parse graph, we first collect all detected human boxes not assigned to any IDs, extract their appearance features (see Section IV), and run K-means method to get clusters of nodes. Each cluster is considered to be candidate parse graph. For each candidate, we use their average pair-wise similarities to define proposal probability $Q()$. We greedily apply the Dynamic III over the selected cluster of nodes to create a parse graph. For the dynamic II, we will randomly select one of the parse graphs in **G** and assign its terminal nodes to be background. The proposal probability is set to be a constant.

Dynamic III and IV: addition/deletion of non-terminal nodes in a parse graph. This pair of dynamics are used to reconfigure a parse graph through adding new graph nodes or deleting existing graph nodes. To add a new graph node (Dynamic III), we will randomly select one of the existing parse graphs in **G**, and create a list of candidate nodes, which have not been assigned to any ID. The proposal probability of selecting a candidate node is defined to be proportional to its average similarities with the terminal nodes in the selected parse graph. To delete a nonterminal node, we create a list of candidates involving all nonterminal nodes, and specify a proposal probability for each candidate according to its energy (i.e. (5)). Once selected a node, we will delete it and its offspring nodes together.

Dynamic V: switching nodes between parse graphs. This dynamic is used to split a trajectory entity (e.g., tracklet, 3D human boxes) from one parse graph and add it to another parse graph. To do so, we use the same strategy used for Dynamic IV to generate candidate nodes in a randomly selected parse graph. The selected node along its offspring are added to the corresponding layers (i.e., tracklets, 3D human boxes, or 2D human boxes of another parse graph selected.

Dynamic VI: changing attributes of graph nodes. The attributes of a graph nodes are mostly provided with confidences, and there is thus a demand to exploit the alternative recognition results. To do so, we will randomly select a node in the hierarchy and change one of its attributes to be alternative values with a probability. The proposal probability of a designed value is defined to be proportional to its recognition confidence. Once changed, we will propagate this new attribute to its offspring nodes.

Algorithm 1 Inference

```
Input: multiple-view video sequences
Initializations of cross-view calibrations, human detections
  and attribute recognitions.
Construct initial graphs G;
Iterate until convergence,
    - Randomly select one of five dynamics
    - Make proposals accordingly to change solution state
    - Accept the change with a probability
```

Among the above dynamics, the Dynamics I through V result in jump moves in the solution space through bottom-up computations, and the dynamic VI results in diffusion changes through top-down propagation. It is noteworthy that the proposed model is computational efficient due to the structured solution spaces, defined by the hierarchical parse graph. In particular, our method can adaptively determine the best scale to optimize, from low-level graph elements, e.g., tracklets, to high-level graph elements, e.g, long trajectories. In this way, the sampling method is able to switch the labels (trajectory IDs) of a big chuck of elements, and thus accelerate the mixing process.

B. Learning of Grammar Model

We utilize an empirical study over training samples to estimate the optimal parameters of the energy function $E(G)$, including filter parameters **H**, kernel widths used for the exponential functions and other hyper-parameters. We use weakly supervised training data, each of which is only provided with human trajectories, without parse graphs. Our goal is to select the optimal value for each parameter, i.e., the optimal parameter configuration. To do so, for each of these parameters we empirically quantize its possible values, e.g. $0.1, 0.3, \dots, 1$ for a constant. With each possible parameter configuration, we need to simulate a parse graph for every image from the trajectory annotations.

In parameter learning, we revise Algorithm 1 as follows: i) skip the step of initializations, e.g., detection, tracklets generations, since we have access to the annotated human trajectories; ii) only use the dynamics III and IV (birth/death of non-terminal nodes) during MCMC sampling. This revised inference usually converges within a hundred of iterations (with dozens of graph nodes). After convergence, we calculate the energy $E(G)$. Thus, we select the parameter configuration that achieves the minimal energy. Similar simulation based learning method has been used in previous works [34] [21].

IV. Experiments

We apply the proposed grammar model over multi-view videos to track humans in the scene and compare to other popular tracking methods on the same video dataset.

A. Datasets

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

(1) *DARPA dataset*. This is a video dataset collected for the DARPA MSEE program and was used by Liu et al. for multi-view human tracking [18]. The videos were captured in three scenes: parking lot, garden, and office areas. There are 8, 6, and 10 cameras mounted on top of building or wall, respectively. For each scene, there are two groups of cameras and each group has overlapping camera views. For each camera view, there is one video sequence of 8-10 minutes long.

(2) *PPL-DA dataset*. We collect a new dataset aiming to cover people's daily activities. This dataset consists of 3 public facilities: foot court, office reception, and 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. This dataset was used in our previous work [45].

(3) *EPFL dataset* . This dataset is collected by Berclaz et al. [4], including five scenes. For each scene, there are 3-5 cameras and each video is about 3-5 minutes long.

For each of the above three datasets, we incorporate 10% of the videos as augmented training set and the rest as testing set. The training data are used to learn model parameters and train classifiers for recognizing human attributes (as introduced later). The learning process is only done once and applied to all datasets. All parameters are fixed in the experiment.

We also annotate object attributes for all the videos of the DARPA dataset, and used the ground-truth annotations for ablation experiments. We only annotated the high-level attributes, i.e., accessories and activities. We use the labeling tool VATIC [36] to reduce the labeling efforts. In particular, we manually provide attribute labels for each object at a video frame and use VATIC to propagate these object attributes to the following video frames.

B. Implementation of the Proposed Method

We implement the Algorithm 1 as follows. To obtain feature vectors of terminal nodes, i.e. $\phi()$, we will employ the powerful deep convolution neural network [10]. In particular, we fine-tune the CaffeNet using people image samples with identity labels. The network consists of 5 convolutional layers, 2 max-pooling layers, 3 fully-connected layers and a 1000-dimensional layer connected by the classification loss. Similar to bag-of-words (BoW), such a network plays the role of a codebook, which describes a person image with common people appearance templates. For each image, we run the forward pass through the trained network to get the 1000-dimensional output layer as its feature vector.

In order to quantize the contributions of various human attributes (summarized in Table I, we implemented five variants of the proposed method. (a) Ours-I, that does not utilize any human attributes; (b) Ours-II, that only utilizes the geometry attributes, i.w. direction and speed, as shown in Table I; (c) Ours-III: that only uses the attributes of Accessories; (d) Ours-IV: that only uses the attributes of Semantics (i.e., activities, gesture and gender); (d) Ours-V, that uses all the human attributes. We apply these variants over the same testing videos for ablation analysis.

We extract human attributes (as listed in Table I) as follows. First, we compute average speed of each tracklet, and project its movement direction in images to the reference camera view. Two moving directions are considered to be same if their relative angle is less than 15 degrees. Second, we train a neural network to recognize accessories, including glasses, bags, clothes, and hats and genders. For each attribute, e.g., hats, we annotate attribute label (Yes or No) for each subject in training videos. Each of these attributes labels are related to two output units and share the same CNN network. These deep models were trained once, and are fixed through the experiments over various datasets. Third, to recognize gestures or activities of an individual, we train a deep neural network to categorize the classical human pose/action variations. We use the PASCAL VOC 2012 action dataset, augmented by our own collected images. We use four activities: 'walking', 'running', 'riding bike', 'skate boarding', and three gestures: 'sitting', 'standing' ,'bending', which cover people's common types of gestures/activities. With about 5000 training images, we fine-tune a 7 layer CaffeNet, with 5 convolutional layers, 2 max-pooling layers, 3 fully-connected layers. We consider each gesture or activity as a binary class and thus the final output of the CaffNet has 14 output units.

We employ the Faster Region-based Convolution Neural Network method [29] to detect human boxes in videos. We use the pre-trained model and fine-tune it over the training videos. In testing, the pruning threshold is set to be 0.3. We apply Sequential Shortest Path (SSP) [26] to initialize tracklets. The sampling is set to finish after 1000 iterations, which achieves decent results. In initializations, we assign two boxes in different camera views to the same subject if their projection boxes overlap with each other.

To handle streaming videos, we run Algorithm 1 over a window of 200 frames and slide it forward at the step of 20 frames. For each window, we utilize the results from the previous window as initial solution. Algorithm 1 usually converges within 1000 iterations. On an DELL workstation (with 64GB memory, $i7$ CPU @2.80GHz, and NVIDIA Tesla K40 GPU), our algorithm can process on average 10 frames per second.

C. Metrics

We evaluate the various tracking methods using the following metrics [46], including:

- **TA**, Multiple Object Tracking Accuracy, number of correctly matched detections over total number of ground-truth detections;
- **TP**, *Multi Object Tracking Precision*, the average ratio of the spatial intersection divided by the union of an estimated object bounding box and the ground-truth bounding box.
- **FRG**$^{\downarrow}$, number of trajectory fragments;
- **MT**, *mostly tracked*, percentage of ground truth trajectories which are covered by tracker output for more than 80% in length;
- **ML**$^{\downarrow}$, *mostly lost*, percentage of ground-truth trajectories which are covered by tracker output for less than 20% in length;

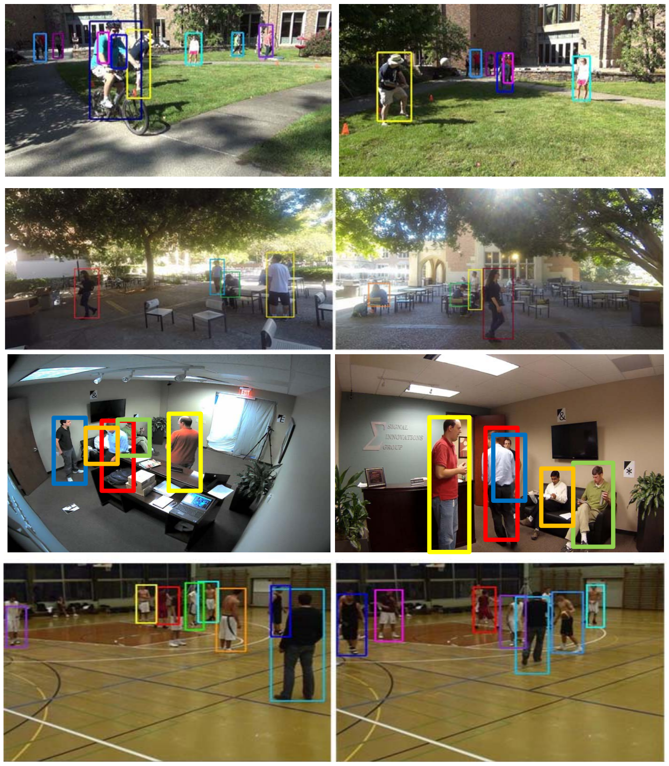


Fig. 3. Sampled qualitative results of our proposed method on DARPA (Row 1) and PPL-DA datasets (Rows 2 and 3), and EPFL (Row 4). Each row shows two camera views at the same time, and the tracked subjects are identified with colors.

- **IDSW**$^{\downarrow}$, *ID Switch*, the number of times that an object trajectory changes its matched id.

Herein, $^{\downarrow}$ indicates that a metric is better if smaller.

D. Qualitative Results

Fig. 3 shows exemplar results of the proposed method on three datasets, including DARPA (the first row), PPL-DA (the second and third rows), and EPFL (the fourth row). For each scene, we show two camera views which are overlaid with the tracked subjects. Every subject is identified with a unique color. These videos pose great challenges to cross-view tracking in many aspects, including severe occlusions (rows 3 and 4), significant lighting changes (Rows 1 and 2), and large pose changes (rows 3 and 4), etc. Under such complex conditions, the proposed method can still achieve robust tracking with the informative attribute grammar.

E. Ablation Experiments on the DARPA Dataset

We apply the proposed method over the DARPA dataset and perform ablation experiments to analyze the contributions of human attributes. Table II reports the precision rate and recall rate of the human attribute recognition method used in this work. The average recall rate is %86.5and the average precision is %85.7 While these results are moderately acceptable, there are still considerate amount of errors or false alarms made by the recognition algorithms. Therefore, it is critical to evaluate how the proposed method performs while human attributes are incorrectly recognized.

TABLE II

RESULTS OF ATTRIBUTE RECOGNITIONS ON THE TESTING SUBSET OF THE DARPA DATASET

Category	Attribute	Classes	Precision	Recall Rate	F1
Accessories	Glasses	Positive	73.5	77.3	75.3
		Negative	83.2	85.6	84.3
	Bags	Positive	91.2	93.5	93.3
		Negative	89.4	82.3	85.7
	Clothes	T-shirt	88.1	79.3	83.4
		Coat	85.0	88.4	86.6
		Suit	81.2	79.3	80.2
	Hats	Positive	90.2	89.5	89.8
		Negative	91.3	93.7	92.4
Semantics	Gender	Male	95.1	96.3	95.6
		female	96.3	94.7	95.5
	Gestures	standing	87.3	85.5	86.4
		sitting	75.3	74.6	74.9
		bending	85.2	83.1	84.1
	Activities	walking	78.5	81.3	79.8
		running	88.3	79.1	83.4
		biking	93.9	90.2	92.0
		skating	84.1	8.1	86.5
Average			86.5	85.7	86.1

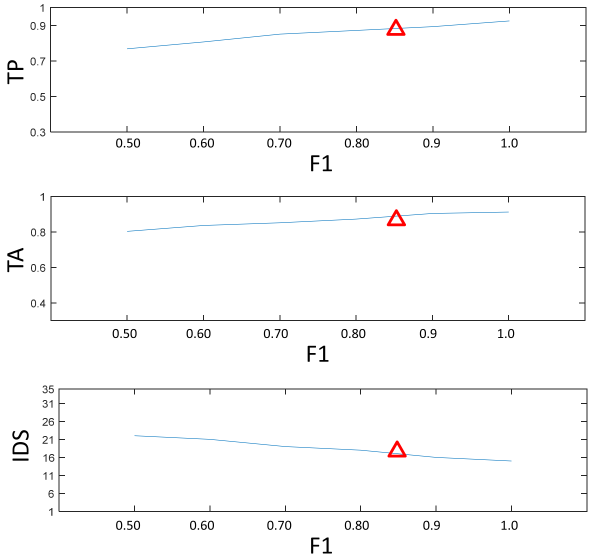


Fig. 4. Tracking Performance v.s. accuracies of human attribute recognitions. horizontal-axis: average F1 scores; vertical-axis: TP (top), TA (middle) and IDS (bottom). The triangle in red indicates the accuracy of the attribute recognition method used in this work.

To do so, we progressively add errors to the ground-truth attribute annotations as follows: randomly select a human instance and set one of its attributes (e.g., 'gender') to be a wrong label (e.g., 'female'). We repeat the above process to add more errors. With these flawed human attributes, we apply Ours-V to get cross-view object trajectories, and calculate the various tracking metrics. Figure 4 reports the accuracies of Ours-V while using various qualities of human attribute recognitions. In particular, the x-direction represents the accuracies of attribute recognition in terms of Average F1 Score (i.e., $2*\frac{Precision*Recall}{Precision+Recall}$), and the y-direction represents the tracking performance in terms of IDS, TA, or TP, respectively. These comparisons are used to analyze the impacts of erroneous attribute recognitions, which is inevitable even in state-of-the-art recognition methods, over the proposed tracking system. From the figures, we can observe that our method is relatively robust and consistent even while the attributes are not properly recognized. Note that the last column of each sub-figure represents the method using ground-truth attributes (F1: 1.0) and the red triangle represents the outcome of by the attribute recognition method used in this work.

F. Results on the DARPA and EPFL Datasets

We further apply the proposed method over the DARPA and EPFL dataset, and compare it to the other popular trackers. We use two recent multi-view trackers: (i) the K-shortest Path (KSP) method by Berclaz et al. [4]; (ii) the multi-view SVM method (mvSVM) by Zhang et al. [53]. We also implemented several single-view based human trackers for comparisons, including: (iii) The local sensitive histogram based tracker (LSH) [8]; (iv) The discrete-continuous tracking (DCT) method proposed by Andriyenko and Schindler [1]; (v) The occlusion geodesic (Geodesic) based tracker [28]. We use the default parameter configuration in their source codes. We also include the tracking results of SSP method [26], which are used to initialize the proposed methods. In addition to the above methods, we employed a recent neural network based method, MDNet [23], that employs a Multi-Domain Convolutional Neural Network for visual tracking, where each object of interest(or domain) is represented as a separate CNN network. MDNet achieved state-of-the-art tracking performance in multiple visual tracking benchmarks [23]. As most other deep learning trackers, MDnet employs extra training images and ground-truth trajectories to train the networks as a deep representation of the objects of interest. We pretrained the MDnet on the OTB dataset [41], as discussed in the original paper, and fine-tuned it using the training videos of the DARPA and EPFL datasets. We use the recommended parameters (e.g. layers, activation functions) in the original work [23]. It is noteworthy that MDNet is developed for single-view tracking and we apply MDNet over individual video sequences.

TABLE III

QUANTITATIVE TRACKING RESULTS ON THE DARPA DATASET. THERE ARE FIVE IMPLEMENTATIONS OF THE PROPOSED METHODS: OURS-I, THAT DOES NOT EXPLORE HUMAN ATTRIBUTES; OURS-II, THAT EXPLORES ONLY GEOMETRY ATTRIBUTES; OURS-III, THAT EXPLORES ONLY THE ATTRIBUTES OF ACCESSORIES; OURS-IV, THAT EXPLORES ONLY THE SEMANTIC ATTRIBUTES; OURS-V, THAT EXPLORES ALL HUMAN ATTRIBUTES

Metrics	TA(%)	TP(%)	MT(%)	ML(%)↓	FRG↓	IDS↓
MDNet [24]	86.7	88.9	92.3	8.6	65	14
Ours-V	**85.2**	**87.1**	**84.5**	**9.5**	**68**	**19**
Ours-IV	81.3	86.1	84.5	10.1	75	25
Ours-III	78.3	83.2	83.3	11.7	81	27
Ours-II	75.9	80.8	82.6	12.0	89	31
Ours-I	74.5	78.9	80.7	12.5	97	43
SSP [27]	72.3	74.5	75.0	14.6	102	59
mvSVM [54]	68.5	71.8	72.7	15.9	124	82
KSP [5]	71.6	73.4	74.3	14.1	244	59
DCT [1]	52.4	54.3	69.4	18.8	243	85
AVT [49]	63.5	64.1	78.8	17.2	198	71
LSHT [9]	62.1	60.7	70.6	15.3	173	79
Geodesic[29]	64.2	66.1	74.2	14.5	340	73

TABLE IV

QUANTITATIVE TRACKING RESULTS ON THE EPFL [4]. THERE ARE FIVE IMPLEMENTATIONS OF THE PROPOSED METHODS: OURS-I, THAT DOES NOT EXPLORE HUMAN ATTRIBUTES; OURS-II, THAT EXPLORES ONLY GEOMETRY ATTRIBUTES; OURS-III, THAT EXPLORES ONLY THE ATTRIBUTES OF ACCESSORIES; OURS-IV, THAT EXPLORES ONLY THE SEMANTIC ATTRIBUTES; OURS-V, THAT EXPLORES ALL HUMAN ATTRIBUTES

Metrics	TA(%)	TP(%)	MT(%)	ML(%)↓	FRG↓	IDS↓
MDNet	88.9	91.2	91.9	7.1	12	18
Ours-V	**85.7**	**88.9**	**89.6**	**7.6**	**14**	**16**
Ours-IV	83.9	87.3	86.3	9.1	21	28
Ours-III	81.5	85.9	83.7	10.9	34	47
Ours-II	80.2	85.2	80.6	11.0	47	54
Ours-I	80.1	84.5	79.5	11.3	53	63
SSP [27]	78.1	76.7	74.3	18.9	89	67
mvSVM [54]	79.5	75.3	76.3	12.9	112	34
KSP [5]	78.6	76.1	75.3	14.3	189	25
DCT [1]	62.4	69.6	68.1	16.2	214	61
AVT [49]	73.3	72.8	70.4	14.1	145	53
LSHT [9]	69.4	67.2	68.3	15.4	214	48
Geodesic[29]	73.2	72.1	69.2	15.2	114	41

Tables III and IV report quantitative results of various methods on the DARPA dataset and the EPFL dataset, respectively. Among these baselines, the mvSVM [53], KSP [4] and the proposed methods are multi-view trackers,whereas the other methods work on individual video sequences. Note that mvSVM and KSP are two widely used methods for multi-view tracking, and the MDNet is the most recent state-of-the-art tracker. From the results, we have the following observations. (I) The proposed method Ours-V outperforms the baseline methods mvSVM and KSP, as well as the single-view tracking methods, DCT, AVT, LSHT, and Geodesic. In particular, our method generated much less false alarms than other methods. For example, on the DARPA dataset, our method achieves IDS of 19, while the best score among the baselines is 59 (KSP). These methods, however, are inferior to the learning based method MDNet which was trained using extra training samples with annotations. It is also noteworthy that the comparisons between our methods and MDNet are not fair since the later can only track objects in individual camera views. (II) The method Ours-V clearly outperforms its variants Ours-I that does not explore any human attributes and Ours-II that uses only low-level attributes of geometry. The comparisons between Ours-I, Ours-II, Ours-III and Ours-IV show that system accuracies can be further improved through additionally using the attributes of accessories (Ours-III) and Semantic attributes (Ours-IV). These ablation analysis clearly demonstrate the advantages of leveraging human attributes for visual tracking task.

G. Quantitative Results on PPL-DA Dataset

We further apply the proposed methods on the PPL-DA dataset [45] and compare to two state-of-the-arts methods:

TABLE V

QUANTITATIVE RESULTS AND COMPARISONS ON PPL-DA DATASET. THERE ARE FIVE IMPLEMENTATIONS OF THE PROPOSED METHODS: OURS-I, THAT DOES NOT EXPLORE HUMAN ATTRIBUTES; OURS-II, THAT EXPLORES ONLY GEOMETRY ATTRIBUTES; OURS-III, THAT EXPLORES ONLY THE ATTRIBUTES OF ACCESSORIES; OURS-IV, THAT EXPLORES ONLY THE SEMANTIC ATTRIBUTES; OURS-V, THAT EXPLORES ALL HUMAN ATTRIBUTES

Seq-Court	TA(%)	TP(%)	MT(%)	ML(%) ↓	IDSW ↓	FRG ↓
MDNet [24]	53.1	82.1	32.2	21.3	53	42
Our-V	34.5	72.4	18.5	25.9	79	55
Our-IV	30.1	71.9	17.2	28.6	92	69
Our-III	28.3	71.6	15.2	31.7	108	75
Our-II	26.9	70.3	12.1	32.9	113	82
Our-I	26.8	70.2	11.1	33.3	114	90
HTC [45]	29.5	71.9	14.8	25.9	91	77
KSP [5]	24.7	64.4	0.00	44.4	318	291
POM [7]	22.3	65.4	0.00	51.9	296	269
Seq-Office	TA(%)	TP(%)	MT(%)	ML(%) ↓	IDSW ↓	FRG ↓
MDNet [24]	60.3	87.1	54.1	0.00	33	28
Out-V	47.4	73.7	42.9	0.00	45	31
Out-IV	44.5	69.5	33.5	0.00	57	44
Out-III	43.5	57.1	29.1	0.00	68	59
Out-II	41.2	56.3	28.1	0.00	69	62
Out-I	39.8	9.0	28.6	0.00	72	64
HTC [45]	41.2	70.7	28.6	0.00	66	59
KSP [5]	39.6	58.0	28.6	0.00	83	76
POM [7]	36.9	58.8	28.6	0.00	89	82
Seq-Plaza	TA(%)	TP(%)	MT(%)	ML(%) ↓	IDSW ↓	FRG ↓
MDNet [24]	27.4	68.9	18.5	12.7	112	98
Our-V	25.2	67.1	16.3	11.6	165	133
Our-IV	24.2	66.3	15.0	12.2	177	154
Our-III	22.4	65.1	14.2	14.2	195	172
Our-II	21.4	65.1	14.2	18.6	210	180
Our-I	20.6	65.1	11.6	18.6	244	199
HTC [45]	23.1	66.2	11.6	18.6	202	178
KSP [5]	17.3	57.5	7.0	27.9	356	311
POM [7]	16.7	57.9	4.6	32.6	339	295

Probabilistic Occupancy Map (POM) [6], K-Shortest Path (KSP) [4]. We use the publicly available softwares of POM and KSP. We also use the MDNet [23] as a baseline. We pre-trained the MDnet on the OTB dataset [41], and fine-tuned it using the training videos of PPL-DA dataset. In addition, we include our recent work, Hierarchical Trajectory Composition (HTC) [44] for comparisons.

Table V reports the quantitative results of various methods, including the five variants of the proposed methods, on the PPL-DA dataset. From the table, we can obtain similar observations as those on the DARPA and EPFL datasets. In particular, the proposed method Our-V clearly outperforms the three popular baselines on all three scenarios while using all Six metrics.Notably, our method can significantly reduce the number of ID switches (IDSW) on all scenarios, which is a critical indicator of the superiority of our method. Our-V also outperforms the other four variants on all testing settings, which directly justifies the key idea of this work, i.e. that integrating human attributes is capable of boosting system robustness while identifying subjects across camera views in complex scenarios. MDNet achieved the best performances on all video sequences mostly because it is directly trained for individual camera views. Like other single-view trackers, MDnet, however, is not be able to discover the cross-view correspondences, which is the main focus of this work.

V. CONCLUSIONS

This work presents a stochastic grammar model for leveraging various human attributes in cross-view human tracking. Our model can robustly track multiple persons while observing them moving in the scene through camera views, even in complex scenarios. To do so, we proposed a deep compositional representation, i.e. parse graph, and introduced an attribute grammar to guide the construction of parse graph from videos. We formulated such a challenging task in the Bayesian framework, and developed an alternative sampling algorithm to solve model selection and state estimation simultaneously. Exhaustive experiments over multiple video datasets clearly demonstrated the advantages of the proposed grammar model, as an effective way to leveraging various human attributes.

REFERENCES

[1] A. Andriyenko and K. Schindler, "Multi-target tracking by continuous energy minimization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 1265–1272.

[2] M. Ayazoglu, B. Li, C. Dicle, M. Sznaier, and O. I. Camps, "Dynamic subspace-based coordinated multicamera tracking," in *Proc. IEEE Int. Conf. Comput. Vis.*, Nov. 2011, pp. 2462–2469.

[3] A. Barbu and S.-C. Zhu, "Generalizing Swendsen–Wang to sampling arbitrary posterior probabilities," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp. 1239–1253, Aug. 2005.

[4] J. Berclaz, F. Fleuret, E. Türetken, and P. Fua, "Multiple object tracking using K-shortest paths optimization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 9, pp. 1806–1819, Sep. 2011.

[5] J. Fan, X. Shen, and Y. Wu, "What are we tracking: A unified approach of tracking and recognition," *IEEE Trans. Image Process.*, vol. 22, no. 2, pp. 549–560, Feb. 2013.

[6] F. Fleuret, J. Berclaz, R. Lengagne, and P. Fua, "Multi-camera people tracking with a probabilistic occupancy map," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 2, pp. 267–282, Feb. 2008.

[7] R. B. Girshick, P. F. Felzenszwalb, and D. A. McAllester, "Object detection with grammar models," in *Proc. Adv. Neural Inf. Process. Syst.*, 2011, pp. 442–450.

[8] S. He, Q. Yang, R. W. H. Lau, J. Wang, and M.-H. Yang, "Visual tracking via locality sensitive histograms," in *Proc. CVPR*, 2013, pp. 2427–2434.

[9] J. F. Henriques, R. Caseiro, and J. Batista, "Globally optimal solution to multi-object tracking with merged measurements," in *Proc. CVPR*, 2011, pp. 2470–2477.

[10] Y. Jia *et al.*, "Caffe: Convolutional architecture for fast feature embedding," in *Proc. 22nd ACM Int. Conf. Multimedia*, 2014, pp. 675–678.

[11] H. Jiang, S. Fels, and J. J. Little, "A linear programming approach for multiple object tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2007, pp. 1–8.

[12] J. Joo, S. Wang, and S.-C. Zhu, "Human attribute recognition by rich appearance dictionary," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 721–728.

[13] S. M. Khan and M. Shah, "A multiview approach to tracking people in crowded scenes using a planar homography constraint," in *Proc. Eur. Conf. Comput. Vis.*, 2006, pp. 133–146.

[14] C.-J. Kim and C. R. Nelson, *StateâŁ"Space Models With Regime Switching—Classical & GibbsâŁ"Sampling Approaches With Applications*, vol. 1. Cambridge, MA, USA: MIT Press, 1999.

[15] D. Kimura, "Dual functional asymmetry of the brain in visual perception," *Neuropsychologia*, vol. 4, no. 3, pp. 275–285, 1966.

[16] C.-H. Kuo and R. Nevatia, "How does person identity recognition help multi-person tracking?" in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 1217–1224.

[17] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, May 2015.

[18] X. Liu, "Multi-view 3D human tracking in crowded scenes," in *Proc. AAAI Conf. Artif. Intell.*, 2016, pp. 3553–3558.

[19] X. Liu, L. Lin, and H. Jin, "Contextualized trajectory parsing with spatio-temporal graph," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 12, pp. 3010–3024, Dec. 2013.

[20] X. Liu, L. Lin, S. Yan, H. Jin, and W. Jiang, "Adaptive object tracking by learning hybrid template online," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 21, no. 11, pp. 1588–1599, Nov. 2011.

[21] X. Liu, Y. Zhao, and S.-C. Zhu, "Single-view 3D scene parsing by attributed grammar," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 684–691.

[22] A. Milan, K. Schindler, and S. Roth, "Detection- and trajectory-level exclusion in multiple object tracking," in *Proc. CVPR*, 2013, pp. 3682–3689.

[23] H. Nam and B. Han, "Learning multi-domain convolutional neural networks for visual tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 4293–4302.

[24] B. X. Nie, C. Xiong, and S.-C. Zhu, "Joint action recognition and pose estimation from video," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1293–1301.

[25] S. Park and S.-C. Zhu, "Attributed grammars for joint estimation of human attributes, part and pose," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 2372–2380.

[26] H. Pirsiavash, D. Ramanan, and C. C. Fowlkes, "Globally-optimal greedy algorithms for tracking a variable number of objects," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 1201–1208.

[27] J. Porway and S.-C. Zhu, "C^4: Exploring multiple solutions in graphical models by cluster sampling," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 9, pp. 1713–1727, Sep. 2011.

[28] H. Possegger, T. Mauthner, P. M. Roth, and H. Bischof, "Occlusion geodesics for online multi-object tracking," in *Proc. CVPR*, 2014, pp. 1306–1313.

[29] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.

[30] H. B. Shitrit, J. Berclaz, F. Fleuret, and P. Fua, "Multi-commodity network flow for tracking multiple people," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 8, pp. 1614–1627, Aug. 2014.

[31] T. Shu, D. Xie, B. Rothrock, S. Todorovic, and S.-C. Zhu, "Joint inference of groups, events and human roles in aerial videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 4576–4584.

[32] J. Sun, N.-N. Zheng, and H.-Y. Shum, "Stereo matching using belief propagation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 7, pp. 787–800, Jul. 2003.

[33] P. H. S. Torr and A. Zisserman, "Feature based methods for structure and motion estimation," in *Proc. Int. Workshop Vis. Algorithms*, 1999, pp. 278–294.

[34] Z. Tu and S.-C. Zhu, "Image segmentation by data-driven Markov chain Monte Carlo," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 5, pp. 657–673, May 2002.

[35] A. Utasi and C. Benedek, "A 3-D marked point process model for multi-view people detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 3385–3392.

[36] C. Vondrick, D. Patterson, and D. Ramanan, "Efficiently scaling up crowdsourced video annotation," *Int. J. Comput. Vis.*, vol. 101, no. 1, pp. 184–204, Jan. 2013.

[37] B. Wang, G. Wang, K. L. Chan, and L. Wang, "Tracklet association with online target-specific metric learning," in *Proc. CVPR*, 2014, pp. 1234–1241.

[38] N. Wang and D.-Y. Yeung, "Learning a deep compact image representation for visual tracking," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 809–817.

[39] P. Wei, Y. Zhao, N. Zheng, and S.-C. Zhu, "Modeling 4D human-object interactions for joint event segmentation, recognition, and object localization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1165–1179, Jun. 2017.

[40] E.-J. Weng and L.-C. Fu, "On-line human action recognition by combining joint tracking and key pose recognition," in *Proc. IEEE/RSJ Conf. Intell. Robots Syst.*, Oct. 2011, pp. 4112–4117.

[41] Y. Wu, J. Lim, and M.-H. Yang, "Online object tracking: A benchmark," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 2411–2418.

[42] Z. Wu, N. I. Hristov, T. L. Hedrick, T. H. Kunz, and M. Betke, "Tracking a large number of objects from multiple views," in *Proc. IEEE Int. Conf. Comput. Vis.*, Sep./Oct. 2009, pp. 1546–1553.

[43] Y. Xu, L. Lin, W.-S. Zheng, and X. Liu, "Human re-identification by matching compositional template with cluster sampling," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 3152–3159.

[44] Y. Xu, X. Liu, Y. Liu, and S.-C. Zhu, "Multi-view people tracking via hierarchical trajectory composition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 4256–4265.

[45] Y. Xu, X. Liu, L. Qin, and S.-C. Zhu, "Cross-view people tracking by scene-centered spatio-temporal parsing," in *Proc. AAAI Conf. Artif. Intell.*, 2017, pp. 4299–4305.

[46] B. Yang and R. Nevatia, "An online learned CRF model for multi-target tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2034–2041.

[47] M. Yang, Y. Liu, L. Wen, Z. You, and S. Z. Li, "A probabilistic framework for multitarget tracking with mutual occlusions," in *Proc. CVPR*, 2014, pp. 1298–1305.

[48] M. Yang, J. Yuan, and Y. Wu, "Spatial selection for attentional visual tracking," in *Proc. CVPR*, vol. 1. 2007, pp. 1–8.

[49] A. Yao, J. Gall, G. Fanelli, and L. Van Gool, "Does human action recognition benefit from pose estimation?" in *Proc. Brit. Mach. Vis. Conf.*, 2011, pp. 67.1–67.11.

[50] A. Yilmaz, O. Javed, and M. Shah, "Object tracking: A survey," *ACM Comput. Surv.*, vol. 38, no. 4, 2006, Art. no. 13.

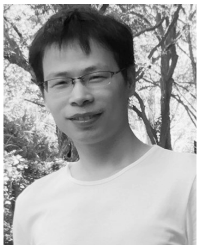
[51] Q. Yu, G. Medioni, and I. Cohen, "Multiple target tracking using spatio-temporal Markov chain Monte Carlo data association," in *Proc. CVPR*, 2007, pp. 1–8.

[52] L. Zhang, Y. Li, and R. Nevatia, "Global data association for multi-object tracking using network flows," in *Proc. CVPR*, 2008, pp. 1–8.

[53] S. Zhang, X. Yu, Y. Sui, S. Zhao, and L. Zhang, "Object tracking with multi-view support vector machines," *IEEE Trans. Multimedia*, vol. 17, no. 3, pp. 265–278, Mar. 2015.

Xiaobai Liu received the Ph.D. degree from the Huazhong University of Science and Technology, China. He is currently an Assistant Professor of computer science with San Diego State University, San Diego, CA, USA. He has published 38 peer-reviewed articles in top-tier conferences (e.g., ICCV, CVPR, and so on) and leading journals (e.g., TPAMI, TIP, and so on). His research interests focus on the development of theories, algorithms, and models for the core computer vision problems. He received a number of awards for his academic contributions, including the 2013 Outstanding Thesis Award by the China Computer Federation.

Yuanlu Xu received the B.E. degree (Hons.) from the School of Software, Sun Yat-sen University, Guangzhou, China, and the master's degree from the School of Information Science and Technology, Sun Yat-sen University. He is currently pursuing the Ph.D. degree with the University of California at Los Angeles, Los Angeles, CA, USA. His current advisor is Prof. S.-C. Zhu, and they have cooperated on publishing a couple of papers on computer vision. His research interests are in video surveillance, statistical modeling, and sptio-temporal inference.



Lei Zhu received the B.S. degree from the Wuhan University of Technology in 2009 and the Ph.D. degree from the Huazhong University of Science and Technology in 2015. He was a Post-Doctoral Research Fellow at the Data and Knowledge Engineering Research Group, The University of Queensland from 2016 to 2017, and Singapore Management University from 2015 to 2016. He is currently a Professor with the School of Information Science and Engineering, Shandong Normal University. His research interests are in the areas of multimedia analysis and search.

Yadong Mu received the Ph.D. degree from Peking University in 2009. He was with the National University of Singapore, Columbia University, Huawei Noah's Ark Lab, and AT&T Labs. He is currently an Assistant Professor with Peking University and also leading the Machine Intelligence Lab, Institute of Computer Science and Technology. His research interest is in large-scale machine learning, video analysis, and computer vision.