A Causal And-Or Graph Model for Visibility Fluent Reasoning in Human-Object Interactions

Lei Qin³*, Yuanlu Xu¹*, Xiaobai Liu², Song-Chun Zhu¹
¹Dept. Computer Science and Statistics, University of California, Los Angeles (UCLA)
²Dept. Computer Science, San Diego State University (SDSU)
³Inst. Computing Technology, Chinese Academy of Sciences
qinlei@ict.ac.cn, yuanluxu@cs.ucla.edu, xiaobai.liu@mail.sdsu.edu, sczhu@stat.ucla.edu

Abstract

Tracking humans that are interacting with the other subjects or environment remains unsolved in visual tracking, because the visibility of the human of interests in videos is unknown and might vary over times. In particular, it is still difficult for state-of-the-art human trackers to recover complete human trajectories in crowded scenes with frequent human interactions. In this work, we consider the visibility status of a subject as a fluent variable, whose changes are mostly attributed to the subject’s interactions with the surrounding, e.g., crossing behind another objects, entering a building, or getting into a vehicle, etc. We introduce a Causal And-Or Graph (C-AOG) to represent the causal-effect relations between an object’s visibility fluents and its activities, and develop a probabilistic graph model to jointly reason the visibility fluent change (e.g., from visible to invisible) and track humans in videos. We formulate the above joint task as an iterative search of feasible causal graph structure that enables fast search algorithm, e.g., dynamic programming method. We apply the proposed method on challenging video sequences to evaluate its capabilities of estimating visibility fluent changes of subjects and tracking subjects of interests over time. Results with comparisons demonstrated that our method clearly outperforms the alternative trackers and can recover complete trajectories of humans in complicated scenarios with frequent human interactions.

1 Introduction

Tracking objects of interest in videos is a fundamental computer vision task that has great potentials in many video based applications, e.g., security surveillance, disaster response, border patrol. In these applications, a critical problem is how to obtain the complete trajectory of the object of interest while observing it moving in the scene through camera views. This is a challenging problem since an object of interest might under frequent interactions with the surrounding, e.g., entering a vehicle or a building, or with the other objects, e.g., passing behind another subject. With these interactions, the visibility status of a subject will be varying over time, e.g., changes from invisible to visible and vice versa. In the literature, most state-of-the-art trackers employ appearance cues or motion cues to localize subjects in video sequences and are likely to fail to track the subjects whose visibility status change. To deal with the above challenges, in this work, we propose to explicitly reason subjects’ visibility status over time while tracking the subjects of interests in surveillance videos. The developed techniques, with slight modifications, can be generalized to other scenarios, e.g., hand-held cameras, driverless vehicles.

The key idea of our method is to introduce a fluent variable for each subject of interest to explicitly indicate his/her visibility status in videos. Fluent was firstly used by Newton to denote the time varying status of an object. It is also used to represent the varying object status in commonsense reasoning (Mueller 2014). In this paper, the visibility status of objects can be described as fluents varying over time. As illustrated in Fig. 1, the person3 and person2 are walking through the parking lot, while the person1 and person4 are entering a sedan. The visibility status of person1’s and person2’s changes first from "visible" to "occluded", and
then to "contained". This group of video frames directly demonstrates how object’s visibility fluent changes over time along with their interactions to the surrounding.

We introduce a graphical model to represent the causal relationships between object’s activities (actions/sub-events) and object’s visibility fluent changes.

Fig. 3 visualizes the action-fluent relationship using a Causal And-Or Graph (C-AOG). The occlusion status of an object might be caused by multiple actions, and we need to reason the actual causality from videos. These actions are alternative choices that lead to the same occlusion status, and form the Or-nodes. Each leaf node indicates an action or sub-event that can be described by and-nodes. Taking the videos shown in Fig. 1 for instance, the status of "occluded" can be caused by the following actions: i) walking behind a vehicle; ii) walking behind a person; or iii) inertial action that maintains the fluent unchanged.

The basic hypothesis of this model is that, for a particular scenario (e.g., parking-lot), there are only a limited number of actions that can cause the fluent to change. Given a video sequence, we need to create the optimal C-AOG and for each Or-node select the best choice in order to get the optimal causal parse graph, shown as red lines in Fig. 3.

We develop a probabilistic graph model to reason object’s visibility fluent changes using C-AOG representation. Our formula integrates object tracking purposes as well to enable joint solution of tracking and fluent change reasoning, which are mutually beneficial. In particular, for each subject of interest, our method uses two variables to represent i) subjects’ positions in videos; and ii) visibility status as well as the best causal parse graph. We utilize a Markov Chain Prior model to describe the transitions of these variables, i.e., the current state of a subject is only dependent on the previous state. Then we reformulate the problem into an Integer Linear Programming model, and utilize dynamic programming to search the optimal states over time. In evaluations, we apply the proposed method over a set of challenging sequences that include frequent human-vehicle or human-human interactions. Results showed that our method can readily predict the correct visibility statuses and correctly recover the complete trajectories. In contrast, most of the alternative trackers can only recover part of the trajectories due to the occlusion/containment.

**Contributions.** There are three major contributions of the proposed framework: i) a Causal And-Or Graph (C-AOG) model to represent object visibility fluents varying over time; ii) a joint probabilistic formulation for object tracking and fluent reasoning; and iii) a new occlusion reasoning dataset to cover objects with diverse fluent changes.

## 2 Related Work

The proposed research is closely related to the following three research streams in computer vision and AI.

**Multiple object tracking** has been extensively studied in the last decades. In past literature, tracking-by-detection has become the mainstream framework (Wen et al. 2014; Dehghan et al. 2015). Specifically, a general detector (Felzenszwalb et al. 2010; Ren et al. 2015) is first applied to obtain detection proposals, then data association techniques (Berczal et al. 2011; Dehghan, Assari, and Shah 2015; Yu et al. 2016) are employed to link detection proposals over times in order to get object trajectories. Our approach follows this pipeline as well but is focused on the reasoning of object visibility status.

**Occlusion handling** is perhaps the most fundamental problem in visual tracking, and has been extensively studied in the past literatures. These methods can be roughly divided into three categories: i) using depth information (Ess et al. 2009; Yilmaz, Xin, and Shah 2004), ii) modeling partial occlusion relations (Tang, Andriluka, and Schiele 2014), iii) modeling object appearing and disappearing globally (Wang et al. 2014; 2016; Maksai, Wang, and Fua 2016). These methods all have strong assumptions on the appearance or motion cues and mainly deal with short-term occlusions. In contrast, this paper presents a principled way to explicitly reason object visibility status during both short-term interactions, e.g., passing behind another object, or long-term interactions, e.g., entering a vehicle and moving together.

**Causal-effect reasoning** is a popular topic in AI but has not received much attention in the field of computer vision. It studies, for instances, the differences between co-occurrence and causality and aims to learn causal knowledge automatically from low-level observations, e.g., images or videos. There are two popular causality models: Bayesian Network (Griffiths and Tenenbaum 2005; Pearl 2009) and grammar models (Griffiths and Tenenbaum 2007; Liang et al. 2016). Notably, Fire and Zhu (Fire and Zhu 2016) introduced a causal grammar to infer the causal-effect relationships between object’s status, e.g., door open/close; and agent’s actions, e.g., pushing the door. They have studied this problem using manually designed rules and video sequences in lab settings. In this work, we extend the causal grammar models to infer objects’ visibility fluent and ground the task on challenging videos in surveillance systems.
Figure 3: (a) The proposed Causal And-Or Graph (C-AOG) model for the fluent of visibility. We use a C-AOG to represent the visibility status of an subject. Each OR node indicates a possible choice and an arrow shows how visibility fluent transits among states. (b) A series of atomic actions that could possibly cause visibility fluent change. Each atomic action describes interactions among people and interacting objects. "P", "D", "T", "B" denotes "person", "door", "trunk", "bag", respectively. The dash triangle denotes fluent. The corresponding fluent could be "visible", "occluded" or "contained" for a person; "open", "closed" or "occluded" for a vehicle door or truck. See text for more details.

3 Representation

In this paper, we define three states for visibility fluent reasoning: visible, (partially/fully) occluded, and contained. Most multiple object tracking methods are based on tracking-by-detection framework, which obtain good performance in visible and partially occluded situations. However, when full occlusions take place, these trackers usually regard the disappearing-and-reappearing objects as new objects. Although objects in fully occluded and contained states could be invisible, there are still evidences to infer the object’s location and fill-in the complete trajectory. We can distinguish object being fully occluded and object being contained from three empirical observations.

Firstly, motion independence. In fully occluded state, such as a person staying behind a pillar, motion of the person is independent of the pillar. While in contained state, such as person in vehicle, or bag in trunk, the position and motion of the person/bag would be the same as the vehicle. So it is important to infer the visibility fluent of the object, if we want to track objects accurately in a complex environment.

Secondly, coupling actions and object fluent changes. For example, as illustrated in Fig. 2, if a person gets into a vehicle, the related sequential atomic actions are: approaching a vehicle, opening the vehicle door, getting into the vehicle, and closing the vehicle door; the related object fluent changes are vehicle door closed → open → closed. The fluent change is a consequence of agent actions. If the fluent-changing actions do not happen, the object should maintain its current fluent. For example, a person that is contained in a vehicle will remain contained unless he/she opens the vehicle door and gets out of the vehicle.

Thirdly, visibility in the alternative camera views. In full occlusion state, such as a person occluded by a pillar, though the person could not be observed from the current viewpoint, he/she could be seen from the other viewpoints; while in contained state, such as a person in a vehicle, this person could not be seen from any viewpoints.

In this work, we mainly study the interactions of humans and the developed methods can be expanded to other objects (e.g., animal) as well.

3.1 Causal And-Or Graph

In this paper, we propose a Causal And-Or Graph (C-AOG) to represent the action-fluent relationship, as illustrated in Fig. 3(a). A C-AOG has two types of nodes: i) Or-nodes that represent the variations or choices, and ii) And-nodes that capture the decomposition of the top-level entity. The arrow represents the causal relation between action and fluent transition. For example, we can use a C-AOG to expressively model a series of action-fluent relations.

The C-AOG is capable of representing multiple alternatives for causes of occlusion and potential transitions. There are four levels in our C-AOG: visibility fluents, possible states, state transitions and agent actions. Or nodes represent alternative causes in visibility fluents and state levels; that is, one fluent can have multiple states and one state can have multiple transitions. An event can be decomposed into possible atomic actions and represented by And nodes, e.g., opening vehicle door, opening vehicle trunk, loading a baggage.

Given a video sequence I with length T, we represent the scene \( \mathcal{R} \) as

\[
\mathcal{R} = \{ O_t : t = 1, 2, ..., T \}, \quad O_t = \{ o^i_t : i = 1, 2, ..., N_t \},
\]

where \( O_t \) denotes all the object at time \( t \) and \( N_t \) is the size of \( O_t \), i.e. the number of objects at time \( t \). \( N_t \) is unknown and will be inferred from observations. Each object \( o^i_t \) is represented with its location (i.e., bounding boxes in the image), appearance features \( \phi^i_t \). To study the visibility fluent of a
subject, we further incorporate a state variable $s_i^t$ and an action label $a_i^t$, that is,

$$a_i^t = \left( l_i^t, \phi_i^t, s_i^t, a_i^t \right).$$

(2)

Thus, the state of a subject is defined as

$$s_i^t \in S = \{ \text{visible, occluded, contained} \}. \quad (3)$$

We define a series of atomic actions $\Omega = a_i : i = 1, \ldots, N_a$ that could possibly change the visibility status, e.g., walking, opening vehicle door. As illustrated in Fig. 3(b), a small set of actions $\Omega$ could cover most common interactions among people and vehicles.

Our task is to jointly find subject locations in video frames and estimate their visibility fluents $M$ from the video sequence $J$. Formally, we have,

$$M = \{ pg_t : t = 1, 2, \ldots, T \},$$

$$pg_t = \{ a_i^t = (l_i^t, \phi_i^t, s_i^t, a_i^t) \mid i = 1, 2, \ldots, N_t \}. \quad (4)$$

where $pg_t$ can be determined by the optimal causal parse graph at time $t$.

4 Problem Formulation

According to Bayes’ rule, we can solve our joint object tracking and fluent reasoning problem by maximizing a posterior (MAP),

$$M^* = \arg \max_M p(M \mid I; \theta) \propto \arg \max_M p(I \mid M; \theta) \cdot p(M; \theta)$$

$$= \arg \max_M \frac{1}{Z} \exp \{ -E(M; \theta) - E(I \mid M; \theta) \}.$$ 

(5)

The prior term $E(M; \theta)$ measures the temporal consistency between successive parse graphs. Assuming $G$ is a Markov Chain structure, we can decompose $E(M; \theta)$ as

$$E(M; \theta) = \sum_{t=1}^{T-1} E(pg_{t+1} \mid pg_t)$$

$$= \sum_{t=1}^{T-1} \sum_{N_t} \Phi(l_{t+1}^i, l_t^i, s_t^i) + \Psi(s_{t+1}^i, s_t^i, a_t^i). \quad (6)$$

The first term $\Phi(\cdot)$ measures the location displacement. It calculates the transition distance between two successive frames and is defined as:

$$\Phi(l_{t+1}^i, l_t^i, s_t^i) = \begin{cases} \delta(D_s(l_{t+1}^i, l_t^i) > \tau_s), & s_t^i = \text{Visible}, \\ 1, & s_t^i = \text{Occ, Cont}, \end{cases} \quad (7)$$

where $D_s(\cdot, \cdot)$ is the Euclidean distance between two locations, $\tau_s$ is the speed threshold and $\delta(\cdot)$ is an indicator function. The location displacement term measures the motion consistency of object in successive frames.

The second term $\Psi(\cdot)$ measures the state transition energy and is defined as:

$$\Psi(s_{t+1}^i, s_t^i, a_t^i) = -\log p(s_{t+1}^i \mid s_t^i, a_t^i). \quad (8)$$

Therefore, we utilize temporal information to generate candidate locations. We employ the SSP algorithm (Pirsiatrash,
Ramanan, and Fowlkes 2011) to generate trajectory fragments (i.e., tracklets). The candidate locations are identified as misses in complete object trajectories. The energy is thus defined as the cost of generating a virtual trajectory at this location. We compute this energy by computing the visual discrepancy between a neighboring tracklet $\varsigma_i^t$ before this moment and a neighboring tracklet $\varsigma_j^{t+1}$ after this moment. The appearance descriptor of a tracklet is computed as the average pooling of image descriptor over time. If the distance is below a threshold $\tau_c$, a virtual path is generated to connect these two tracklets using B-spline fitting.

The term $\Gamma(l_i^t, \phi_i^t, a_i^t)$ is defined over the object actions observed from data. In this work, we study the fluents of human and vehicles, that is,

$$\Gamma(l_i^t, \phi_i^t, a_i^t) = \sigma(D_h(l_i^t, \phi_i^t|a_i^t)) + \sigma(D_v(l_i^t, \phi_i^t|a_i^t)),$$  \hspace{1cm} (11)

where $\sigma(\cdot)$ is the sigmoid function. The definitions of the two data-likelihood terms $D_h$ and $D_v$ are introduced in the rest of this section.

A human is represented by his/her skeleton, which consists of multiple joints estimated by sequential prediction technology (Wei et al. 2015). The feature of each joint is defined as the relative distances of this joint to four saddle points (two shoulders, the center of the body, and the middle between the two hipbones). The relative distances are normalized by dividing the length of head to eliminate the influence of scale. A feature vector $\omega_i^h$ concatenating the features of all joints is extracted, which is assumed to follow a Gaussian distribution:

$$D_h(l_i^t, \phi_i^t|a_i^t) = - \log N(\omega_i^h; \mu_{a_i^t}, \Sigma_{a_i^t}),$$  \hspace{1cm} (12)

where $\mu_{a_i^t}$ and $\Sigma_{a_i^t}$ are the mean and the covariance of the action $a_i^t$ respectively, which are obtained from the training data.

A vehicle is described with its viewpoint, semantic vehicle parts, and vehicle part fluents. The vehicle fluent is represented by a Hierarchical And-Or Graph, as illustrated in Fig. 4. The feature vector of vehicle fluent $\omega_i^v$ is obtained by computing fluent scores on each vehicle part and concatenating them together. We compute the average pooling feature $\varpi_{a_i}$ for each action $a_i$ over the training data as the vehicle fluent template. Given vehicle fluent $\omega_i^v$ computed on image $I_t$, the distance $D_v(l_i^t, \phi_i^t|a_i^t)$ is defined as

$$D_v(l_i^t, \phi_i^t|a_i^t) = \| \omega_i^v - \varpi_{a_i} \|_2.$$  \hspace{1cm} (13)

### 5 Inference

We cast the intractable optimization of Eq. (5) as an Integer Linear Formulation (ILF) in order to derive a scalable and efficient inference algorithm. We denoted by $V$ the locations of vehicles, $E$ the edges between all possible pairs of nodes, whose time is consecutive and locations are close. The whole transition graph $G = (V, E)$ is shown as Fig. 5.

Then the energy function Equation (5) can be re-written as:

$$f^* = \arg \max_f \sum_{mn \in E_v} c_{mn} f_{mn},$$

$$c_{mn} = -\Phi(l_m, \phi_m, s_m, a_m) - \Psi(l_m, \phi_m, s_m) - \Gamma(l_m, \phi_m, a_m),$$

$$s.t. \quad f_{mn} \in \{0, 1\}, \quad \sum_m f_{mn} \leq 1, \quad \sum_m f_{mn} = \sum_k f_{nk},$$  \hspace{1cm} (14)

where $f_{mn}$ is the number of object moving from node $V_m$ to node $V_n$, $c_{mn}$ is the corresponding cost.

Since the subject of interest can only enter a nearby container (e.g., vehicle), to discover the optimal causal parse graph, we need to jointly track the container and the subject of interest. Similar to Equation (14), the energy function of container is as following.

$$g^* = \arg \max_g \sum_{mn \in E_v} d_{mn} g_{mn},$$

$$d_{mn} = h_v(l_m, \phi_m) - 1,$$

$$s.t. \quad g_{mn} \in \{0, 1\}, \quad \sum_m g_{mn} \leq 1, \quad \sum_m g_{mn} = \sum_k g_{nk},$$  \hspace{1cm} (15)

where $h_v(l_m, \phi_m)$ is the container detection score at location $l_m$. Then we add the contained constrains as:

$$\sum_{mn \in E_v} g_{mn} \geq \sum_{ij \in E_v} f_{ij},$$

$$s.t. \quad t_n = t_j, \quad \|l_n - l_j\|_2 < \tau_c,$$  \hspace{1cm} (16)

where $\tau_c$ is the distance threshold. Finally, we combine Equation (14-16) all together:

$$f^*, g^* = \max_{f,g} \sum_{mn \in E_v} c_{mn} f_{mn} + \sum_{ij \in E_v} d_{mn} g_{mn},$$

$$s.t. \quad f_{mn} \in \{0, 1\}, \quad \sum_m f_{mn} \leq 1, \quad \sum_m f_{mn} = \sum_k f_{nk},$$

$$g_{mn} \in \{0, 1\}, \quad \sum_m g_{mn} \leq 1, \quad \sum_m g_{mn} = \sum_k g_{nk},$$

$$t_n = t_j, \quad \|l_n - l_j\|_2 < \tau_c.$$  \hspace{1cm} (17)

This is the objective function we want to optimize.

As illustrated in the Fig. 5, the re-formulated graph structure is a directed acyclic graph (DAG). Thus we can adopt the Dynamic Programming technique to efficiently search the optimal solution for Eqn. (17).

### 6 Experiments

We apply the proposed method over two multiple object-tracking datasets and evaluate the improvement in visual tracking brought by the outcomes of visibility status reasoning.

#### 6.1 Implementation Details

For person and suitcase detection, we utilize the Faster R-CNN models (Ren et al. 2015) trained on the MS COCO dataset. The used network is the VGG-16 net, with score
threshold 0.4 and NMS threshold 0.3. The tracklets similarity threshold $\tau_s$ is set as 0.8. The contained distance threshold $\tau_c$ is set as the width of container. For appearance descriptors $\phi$, we employ the dense sampling ColorNames descriptor (Zheng et al. 2015), which applies square root operator (Arandjelovic and Zisserman 2012) and Bag-of-word encoding to the original ColorNames descriptors. For human skeleton estimation, we use the public implementation of (Wei et al. 2015). For vehicle detection and semantic part status estimation, we use the implementation provided by (Li et al. 2016) with default parameters mentioned in their paper.

We adopt the widely used CLEAR metrics (Kasturi et al. 2009) to measure the performances of tracking methods. It includes four metrics, i.e., Multiple Object Detection Accuracy (MODA), Detection Precision (MODP), Multiple Object Tracking Accuracy (MOTA) and Tracking Precision (MOTP), which take into account three kinds of tracking errors: false positives, false negatives and identity switches. We also report the number of false positives (FP), false negatives (FN), identity switches (IDS) and fragments (Frag). A higher value means better for TA and TP while a lower value means better for FP, FN, IDS and Frag. If the Intersection-over-Union (IoU) ratio of tracking results to ground truth is above 0.5, we accept the tracking result as a correct hit.

6.2 Datasets

People-Car dataset (Wang et al. 2014). This dataset consists of sequences on a parking lot with two synchronized bird-view cameras, with length 300 – 5100 frames. In this dataset, there are many instances of people getting in and out of cars. This dataset is challenging for the frequent interactions, light variation and low object resolution.

Tracking Interacting Objects (TIO) dataset. For current popular multiple object tracking datasets (e.g., PETS09 (Ferryman and Shahrokni 2009), KITTI dataset (Geiger, Lenz, and Urtasun 2012)), most tracked objects are pedestrian and no evident interaction visibility fluent changes. Thus we collect two new scenarios with typical human-object interactions: person, suitcase, and vehicle on several places.

Plaza. We capture 22 video sequences in a plaza that describe people walking around, getting in/out vehicles.

ParkingLot. We capture 3 video sequences in a parking lot that shows vehicles entering/exiting the parking lot, people getting in/out vehicles, people interacting with trunk/suitcase.

All the sequences are captured by a GoPro camera, with frame rate 30fps and resolution $1920 \times 1080$. The total number of frames of TIO dataset is more than 30K. There exist severe occlusions and large scale changes, making this dataset very challenging for traditional tracking methods.

Beside the above testing data, we collect another set of video clips for training. To avoid over-fitting, we set up different camera positions, different people and vehicles from the testing settings. The training data consists of 380 video clips covering 9 events: walking, opening vehicle door, entering vehicle, exiting vehicle, closing vehicle door, opening vehicle trunk, loading baggage, unloading baggage, closing vehicle trunk. Each action category contains 42 video clips on average.

Both the datasets and short clips are annotated with the bounding boxes for people, suitcase, vehicles, and visibility fluents of people and suitcase. The types of status are "visible", "occluded", and "contained". We utilize VATIC (Vondrick, Patterson, and Ramanan 2013) to annotate the videos.

6.3 Results and Comparisons

For People-Car dataset, we compare our proposed method with 5 state-of-the-arts methods: successive shortest path algorithm (SSP) (Pirsiavash, Ramanan, and Fowlkes 2011), K-Shortest Paths Algorithm (KSP) (Bercelaz et al. 2011), Probability Occupancy Map (POM) (Fleuret et al. 2008), Linear Programming (LP2D) (Leal-Taixé et al. 2014), Tracklet-Based Intertwined Flows (TIF-MIP) (Wang et al. 2016). The table below shows the quantitative results and comparisons of false positive (FP) rate, false negative (FN) rate and identity switches (IDS) rate on People-Car Dataset.

| Sequence | Method   | MODA | FP    | FN   | IDS |
|----------|----------|------|-------|------|-----|
| Seq.0    | Our-full | 0.69 | 0.18  | 0.08 | 0.05|
|          | Our-1    | 0.62 | 0.21  | 0.12 | 0.05|
|          | TIF-MIP  | 0.67 | 0.07  | 0.25 | 0.04|
|          | KSP      | 0.49 | 0.10  | 0.41 | 0.07|
|          | LP2D     | 0.47 | 0.05  | 0.48 | 0.06|
|          | POM      | 0.47 | 0.06  | 0.47 | N/A |
|          | SSP      | 0.20 | 0.04  | 0.76 | 0.04|
| Seq.1    | Our-full | 0.61 | 0.23  | 0.12 | 0.04|
|          | Our-1    | 0.53 | 0.29  | 0.14 | 0.04|
|          | TIF-MIP  | 0.58 | 0.17  | 0.25 | 0.04|
|          | KSP      | 0.04 | 0.71  | 0.25 | 0.12|
|          | LP2D     | 0.02 | 0.77  | 0.21 | 0.17|
|          | POM      | -0.21| 0.98  | 0.23 | N/A |
|          | SSP      | 0.00 | 0.75  | 0.25 | 0.12|

Available at cvlab.epfl.ch/research/surv/interacting-objects
For TIO dataset, we compare the proposed method with 6 state-of-the-arts: successive shortest path algorithm (SSP) (Pirsiavash, Ramanan, and Fowlkes 2011), multiple hypothesis tracking with distinctive appearance model (MHT_D) (Kim et al. 2015), Markov Decision Processes with Reinforcement Learning (MDP) (Xiang, Alahi, and Savarese 2015), Discrete-Continuous Energy Minimization (DCEM) (Milan, Schindler, and Roth 2016), Discrete-continuous optimization (DCO) (Andriyenko, Schindler, and Roth 2012) and Joint Probabilistic Data Association (JPDA_m) (Hamid Rezatofighi et al. 2015). We use the public implementations of these methods. We set up a baseline Our-1 to analyze the effectiveness of different components. Our-1 only utilizes human data-likelihood term while Our-full utilizes both human and vehicle data-likelihood terms. Note we also utilize the estimated human joint key-points to refine the detected people bounding boxes.

We report quantitative results and comparisons in Table 2 for TIO dataset. From the results, we can observe that our method obtains superior performance to the other methods on most metrics. It validates that the proposed method can not only track visible objects correctly, but also reason locations for occluded or contained objects. The alternative methods do not work well mainly due to lack of the ability to track objects under long-term occlusion or containment in other objects. Based on comparisons of Our-1 and Our-full, we can also conclude that each type of fluent plays its role in improving the final tracking results. Some qualitative results are visualized in Fig. 6.

**Table 2: Quantitative results and comparisons of false positive (FP), false negative (FN), identity switches (IDS), and fragments (Frag) on TIO dataset.**

| Method   | MOTA | MOTP | FP  | FN  | IDS | Frag |
|----------|------|------|-----|-----|-----|------|
| Our-full | 46.0%| 76.4%| 99  | 501 | 5   | 8    |
| Our-1    | 32.5%| 75.3%| 75  | 605 | 25  | 30   |
| MHT_D    | 34.3%| 73.8%| 56  | 661 | 15  | 18   |
| MDP      | 32.9%| 73.2%| 24  | 656 | 9   | 7    |
| DCEM     | 32.3%| 76.5%| 2   | 675 | 2   | 2    |
| SSP      | 31.7%| 72.1%| 19  | 678 | 21  | 25   |
| DCO      | 29.5%| 76.4%| 22  | 673 | 6   | 2    |
| JPDA_m   | 13.5%| 72.2%| 163 | 673 | 6   | 3    |

| Method   | MOTA | MOTP | FP  | FN  | IDS | Frag |
|----------|------|------|-----|-----|-----|------|
| Our-full | 38.6%| 78.6%| 418 | 1954| 6   | 5    |
| Our-1    | 28.9%| 78.4%| 544 | 2203| 14  | 16   |
| MDP      | 30.1%| 76.4%| 397 | 2296| 26  | 22   |
| DCEM     | 29.4%| 77.5%| 383 | 2346| 16  | 15   |
| SSP      | 28.9%| 75.0%| 416 | 2337| 12  | 14   |
| MHT_D    | 35.6%| 75.0%| 720 | 2170| 15  | 12   |
| DCO      | 24.3%| 78.1%| 536 | 2367| 38  | 10   |
| JPDA_m   | 12.3%| 74.2%| 1173| 2263| 28  | 17   |

Empirical Studies. We further report fluent estimation results on TIO-Plaza sequences and TIO-ParkingLot sequences in Fig. 8. From the results, we can see that our method can successfully reason the visibility status of subjects. Note that the precision of containment estimation is not high, since some people get in/out the vehicle from the opposite side towards the camera, as shown in Fig. 7. Under such situation, multi-view setting might be a better way to
reduce the ambiguities.

7 Conclusion
In this paper, we propose a Causal And-Or Graph model to represent the causal-effect relations between object visibility fluents and various human interactions. By jointly modeling short-term occlusions and long-term occlusions, our method can explicitly reason the visibility of subjects as well as their locations in videos. Our method clearly outperforms the alternative methods in complicated scenarios with frequent human interactions. In this work, we focus on the human-interactions as a running-case of the proposed technique, which can be easily extended for the other types of objects (e.g., animal, drones).

References
Andriyenko, A.; Schindler, K.; and Roth, S. 2012. Discrete-continuous optimization for multi-target tracking. In IEEE Conference on Computer Vision and Pattern Recognition.
Arandjelovic, R., and Zisserman, A. 2012. Three things everyone should know to improve object retrieval. In IEEE Conference on Computer Vision and Pattern Recognition.
Berczaj, J.; Fleuret, F.; Turetken, E.; and Fua, P. 2011. Multiple object tracking using k-shortest paths optimization. IEEE Transactions 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.
Dehghan, A.; Tian, Y.; Torr, P.; and Shah, M. 2015. Target identity-aware network flow for online multiple target tracking. In IEEE Conference on Computer Vision and Pattern Recognition.
Ess, A.; Schindler, K.; Leibe, B.; and Gool, L. V. 2009. Improved multi-person tracking with active occlusion handling. In IEEE ICRA Workshop.
Felzenszwalb, P.; Girshick, R. B.; McAllester, D.; and Ramanan, D. 2010. Object detection with discriminatively trained part based models. IEEE Transactions on Pattern Analysis and Machine Intelligence 32(9):1627–1645.
Ferryman, J., and Shahrokni, A. 2009. Pets2009: Dataset and challenge. In IEEE International Workshop on Performance Evaluation of Tracking and Surveillance.
Fire, A., and Zhu, S.-C. 2016. Learning perceptual causality from video. ACM Transactions on Intelligent Systems and Technology 7(2).
Fleuret, F.; Berczaj, J.; Lengagne, R.; and Fua, P. 2008. Multicamera people tracking with a probabilistic occupancy map. IEEE Transactions on Pattern Analysis and Machine Intelligence 30(2):267–282.
Geiger, A.; Lenz, P.; and Urtasun, R. 2012. Are we ready for autonomous driving? the kitti vision benchmark suite. In IEEE Conference on Computer Vision and Pattern Recognition.
Griffiths, T., and Tenenbaum, J. 2005. Structure and strength in causal induction. Cognitive Psychology 51(4):334–384.
Griffiths, T., and Tenenbaum, J. 2007. Two proposals for causal grammars. Causal learning: Psychology, philosophy, and computation 323–345.
Hamid Rezatofighi, S.; Milan, A.; Zhang, Z.; Shi, Q.; Dick, A.; and Reid, I. 2015. Joint probabilistic data association revisited. In IEEE International Conference on Computer Vision.
Kasturi, R.; Goldgof, D.; Soundararajan, P.; Manohar, V.; Garofolo, J.; Bowers, R.; Boonstra, M.; Korzhova, V.; and Zhang, J. 2009. Framework for performance evaluation of face, text, and vehicle detection and tracking in video: Data, metrics, and protocol. IEEE Transactions on Pattern Analysis and Machine Intelligence 31(2):319–336.
Kim, C.; Li, F.; Ciptadi, A.; and Rehg, J. M. 2015. Multiple hypothesis tracking revisited. In IEEE International Conference on Computer Vision.
Leal-Taixé, L.; Fenzl, M.; Kuznetsova, A.; Rosenhahn, B.; and Savarese, S. 2014. Learning an image-based motion context for multiple people tracking. In IEEE Conference on Computer Vision and Pattern Recognition.
Li, B.; Xu, T.; Xiong, C.; and Zhu, S.-C. 2016. Recognizing car fluents from video. In IEEE Conference on Computer Vision and Pattern Recognition.
Liang, W.; Zhao, Y.; Zhu, Y.; and Zhu, S.-C. 2016. What are where: Inferring containment relations from videos. In International Joint Conference on Artificial Intelligence.
Maksai, A.; Wang, X.; and Fua, P. 2016. What players do with the ball: A physically constrained interaction modeling. In IEEE Conference on Computer Vision and Pattern Recognition.
Milan, A.; Schindler, K.; and Roth, S. 2016. Multi-target tracking by discrete-continuous energy minimization. IEEE Transactions on Pattern Analysis and Machine Intelligence 38(10):2054–2068.
Mueller, E. T. 2014. Commonsense reasoning: An event calculus based approach. Morgan Kaufmann.
Pearl, J. 2009. Causality: Models, reasoning and inference. Cambridge University Press.
Pirsiavash, H.; Ramanan, D.; and Fowlkes, C. C. 2011. Globally-optimal greedy algorithms for tracking a variable number of objects. In IEEE Conference on Computer Vision and Pattern Recognition.
Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In Conference on Neural Information Processing Systems.
Tang, S.; Andriluka, M.; and Schiele, B. 2014. Detection and tracking of occluded people. International Journal of Computer Vision 110(1):5869.
Vondrick, C.; Patterson, D.; and Ramanan, D. 2013. Efficiently scaling up crowdsourced video annotation. International Journal of Computer Vision 101(1):184–204.
Wang, X.; Turetken, E.; Fleuret, F.; and Fua, P. 2014. Tracking interacting objects optimally using integer programming. In European Conference on Computer Vision.
Wang, X.; Turetken, E.; Fleuret, F.; and Fua, P. 2016. Tracking interacting objects using intertwined flows. IEEE Transactions on Pattern Analysis and Machine Intelligence 38(11):2312–2326.
Wei, S.-E.; Ramakrishna, V.; Kanade, T.; and Sheikh, Y. 2015. Convolutional pose machines. In IEEE Conference on Computer Vision and Pattern Recognition.
Wen, L.; Li, W.; Yan, J.; and Lei, Z. 2014. Multiple target tracking based on undirected hierarchical relation hypergraph. In IEEE Conference on Computer Vision and Pattern Recognition.
Xiang, Y.; Alahi, A.; and Savarese, S. 2015. Learning to track: Online multi-object tracking by decision making. In IEEE International Conference on Computer Vision.
Yilmaz, A.; Xin, L.; and Shah, M. 2004. Contour based object tracking with occlusion handling in video acquired using mobile cameras. *IEEE Transaction on Pattern Analysis and Machine Intelligence* 26(11):1532–1536.

Yu, S.-I.; Meng, D.; Zuo, W.; and Hauptmann, A. 2016. The solution path algorithm for identity-aware multi-object tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Zheng, L.; Shen, L.; Tian, L.; Wang, S.; Wang, J.; and Tian, Q. 2015. Scalable person re-identification: A benchmark. In *IEEE International Conference on Computer Vision*. 