Causal Reasoning in Multi-Object Interaction on the Traffic Scene: Occlusion-Aware Prediction of Visibility Fluent

XUANPENG LI, (Member, IEEE), QIFAN XUE, JINGWEN ZHAO, AND DONG WANG

School of Instrument Science and Engineering, Southeast University, Nanjing 210096, China
Corresponding author: Xuanpeng Li (li_xuanpeng@seu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61906038, and in part by the Fundamental Research Funds for the Central Universities under Grant 2242019K40039.

ABSTRACT

Occlusion caused by multi-object interaction makes the traffic scene understanding intractable. In this paper, we focus on predicting the visibility status of vehicle in the framework of causality perception. The visibility fluent is employed to present the varying state of an object, involving visible and occluded. We introduce a probabilistic grammar model, named Hierarchical And-Or Graph (H-AOG), to construct the causal relations between fluents and actions. It consists of a Causal And-Or Graph (C-AOG) module and an Action And-Or Graph (A-AOG) module. An influence field is constructed by the polar coordinate transformation to model interactions in the A-AOG module. This method interprets the occurrence of occlusion due to multi-vehicle interaction. We evaluate our approach on both synthetic data and real data from the KITTI dataset. Compared to the state-of-the-art models like LSTM/GRU, it proves to achieve a promising accuracy in prediction of objects’ visibility states and have better generalization on real data.

INDEX TERMS

Visibility fluent, causal And-Or graph, influence field, traffic scene.

I. INTRODUCTION

Multi-object interaction on the traffic scene leads to change in objects’ states and responsive actions. Especially, the visibility of an object frequently varies over time, e.g., it changes from visible to invisible, which influences the performance of motion prediction. Once an object is fully occluded, most methods of object detection and tracking tend to be failed. It is hard to predict the motion of occluded object in most state-of-the-art trackers. For autonomous vehicle, if it has no “cognition” of the existence of occluded object, its planning would be full of uncertainty and risk. Thus, modeling interactions between multiple objects on the traffic scene is rather significant, helping to prejudge the occurrence of abnormal events and improve driving safety. Many recent work focuses on the interaction modeling by various ways [1]–[3], but it is still a challenging task until now.

Some filter-based methods as Kalman filter [7] and particle filter [8] can only track unobservable objects in limited time, which have not the capability of long-term prediction. Therefore, it is necessary to mine common sense such as causality in multi-object interaction, and construct an approximate dynamic system model based on causality for interactions among multiple agents, which can predict objects’ states in a long-term way.

Causal modeling involves a variety of approaches, typically using a Bayesian network based on the directed graph [9]–[11], or grammar model based on the And-Or graph [12]–[14]. Especially in recent years, the causal And-Or graph modeling method gradually gets the attention of researchers in the field of intelligent agents.

Bayesian network based on graph model provides an important example for solving the problems of incomplete data and sample bias in causal modeling. Mohan and Pearl [9], [10] have conducted long-term research on how to solve these problems, and have proved that the data missing problem can be encoded using the graph model with conditional independence hypothesis. Ayazoglu et al. [15] transformed the causal modeling problem of interaction between objects into a topology of the directed graph in video. By using the target motion as the graph node and the causal relationship as the inter-node connection, the graph optimization problem is formed with the sparse constraint. The structural rank minimization method is employed to solve the
singularity problem caused by data missing and false detection caused by occlusion. Xie et al. [16] also explored the hierarchical Bayesian modeling in order to construct causality for crowd behavior recognition and conduct an analysis on intelligent interaction in video.

On the other hand, for the causal modeling based on grammar model, Fire and Zhu [12] proposed a fluent-based causal perception method, using heuristic learning to build the causal relationship between behavior and states in video, and then predict the hidden state and behavior [13]. The fluent is firstly used by Newton to denote the varying states over time. In cognitive science, fluent characterizes the changes of objects’ states in sequence. It can be used to represent the state of vehicle and pedestrian for prediction of object’s behavior on the traffic scene.

For instance, Xu et al. [14] studied the visibility fluent of pedestrian in the surveillance video. Since changes in the visual state are caused by interactions with surrounding objects, the causal relations between object’s activities and fluents are established by the Causal And-Or Graph (C-AOG). The C-AOG is a bottom-up hierarchical model. At the bottom level, the interactions between pedestrians and the outside world are denoted as the Leaf nodes. Then, the And-Or nodes establish the causal connections between multiple behaviors and changes of visibility. By constructing a causal relationship between behavior and change of visibility through multi-level And-Or nodes, it can track multiple pedestrians in complex scenes. In addition, Li et al. [17] constructed a spatial-temporal And-Or graph framework to jointly recognize semantic car fluents, and the status of blurred, heavily occluded or deformed vehicle components (such as doors and trunks) in video. Wu et al. [18] adopted the And-Or graph and spatial-temporal dynamic programming algorithms inferring object bounding box in real time. It achieves synchronous tracking, learning and parsing of unknown objects in video sequence.

In this paper, we propose to reason the visibility using a Hierarchical And-Or Graph (H-AOG), which represents the causal relations between visibility fluents and multi-object interactions. As shown in Fig. 1, the visibility status of vehicle A changes first from “visible” to “partly occluded”, and then to “fully occluded”. This image sequence demonstrates the visibility fluents changing over time due to the object’s interaction with the surrounding vehicles.

In the following sections, the problem formulation and details of the proposed method are presented in Sec. II. We evaluate our method with the synthetic data and the real-world data from the KITTI dataset in Sec. III and finally make a conclusion in Sec. IV.

II. APPROACH

A. PROBLEM FORMULATION

In this work, 3 states are defined to describe the visibility fluent of an object on the traffic scene: visible (V), partly-occluded (P), and fully-occluded (F). We assign an individual fluent $F_i^t$ to the object $i$ at the time $t$ as follows:

$$F_i^t \in \{V, P, F\}. \quad (1)$$

The fluent of the object $i$ may vary over time, e.g., transition from V to P, and then to F, due to some actions, denoted as A. There are 4 kinds of fluent changes, denoted as $\Delta F$, as:

$$\Omega_{\Delta F} = \{V \rightarrow P, P \rightarrow F, F \rightarrow P, P \rightarrow V\}. \quad (2)$$

The fluent keeps unchanged if there is no occurrence of action, denoted as $\Delta F = 0$. For instance, a vehicle would keep the fully-occluded status for a while, until another vehicle completely leaves.

Given a data sequence $S$ with length $T$, regarded as $S = \{s_1, s_2, \ldots, s_T\}$. The scene $R$ from the data $s_t$ at the time $t$ can be formulated as

$$R_t = \{o_t^i|s_t : i = 1, 2, \ldots, N_t\}, \quad (3)$$

$$o_t^i = \{x_t^i, y_t^i, w_t^i, h_t^i, F_t^i, A_t^i\}, \quad (4)$$

where $o_t^i$ denotes the object $i$, including 6 features at the time $t$: object location $x_t^i, y_t^i$, object dimension $w_t^i, h_t^i$, visibility status $F_t^i$, and action space $A_t^i$ at the time $t$ and $N_t$ is the number of all objects.

Considering the rationality and consistency of actual transitions, we make two hypotheses for the problem formulation. 

Hypothesis 1: The fluent changes due to a limited number of actions, denoted as

$$A \rightarrow \Delta F, \quad \Delta F \in \Omega_{\Delta F}. \quad (5)$$

This hypothesis indicates that there is a causal relation between changes in object’s fluent and object’s actions/sub-events. In addition, the visibility status is only changed when the predefined causal actions occurs and it maintains unchanged unless an action is detected. For example, a vehicle is occluded by the surrounding vehicles, without considering certain case like camera’s lens is stained. Given a sequence of multiple vehicles activities, their fluents can be predicted over time even if observations are intractable, e.g., tracking an occluded vehicle.

Hypothesis 2: Visibility fluent reasoning follows the Markov assumption as

$$P(F_t|F_{t-1}, F_{t-2}, \ldots, F_1, A_{t-1}, A_{t-2}, \ldots, A_1) = P(F_t|F_{t-1}, A_{t-1}). \quad (6)$$
The Markov process is a stochastic model describing a sequence of possible events in which the probability of each event only depends on the state attained in the previous event [19]. The visibility fluent modeling is Markovian since the current state of an object only relies on the previous state. Markov property helps to reduce the complexity of causal-effect modeling, and we only require the detected actions and fluents at the last step.

In addition, there are two kinds of timestamps in a sequence: key time \( t_{\text{key}} \) and non-key time \( t_{\text{nonkey}} \). According to the hypothesis 1, we assume that action and fluent change only happen at the key time. Beside, at the non-key time, we suppose that no interactive action occurs, and the visibility fluent remains unchanged.

\[
s_t = \begin{cases} 
\{F_1, F_2, \ldots, F_i, A_1, A_2, \ldots, A_j\}, & \text{for } t \in t_{\text{key}} \\
\{F_1, F_2, \ldots, F_i\}, & \text{for } t \in t_{\text{nonkey}}.
\end{cases}
\]  

(7)

Overall, if action at the key time can be clearly observed, it leads to fluent change. Otherwise, the visibility fluent remains unchanged at the non-key time and their value are determined by the previous fluent at the last key time.

## B. FRAMEWORK

In this paper, the fluent can be estimated by jointly utilizing the detected action, the previous status of object, the causal relation and the prior knowledge. We employ a grammar model based on the H-AOG in a bottom-up way to represent the causal relations between fluents and actions, as shown in Fig. 2. Given a data sequence, such as image sequence and LiDAR point clouds, our aim is to create the optimal H-AOG by learning and get the optimal parse graph by inference. A parse graph consists of a set of selected Or-nodes at different levels of AOG. It provides a causal interpretation for data sequence.

The framework of the H-AOG involves two steps, regarded as the causal learning step and the causal inference step, and contains 3 main components as follows.

1) **INFLUENCE FIELDS**

We convert the object’s state space to the influence field by using the polar coordinate transformation. The influence field is the basic unit to describe interactions among multiple objects.

2) **ACTION AND-OR GRAPH (A-AOG)**

This module decomposes object’s interaction into pairs of influence fields, which quantitatively represent objects’ interactions for constructing the causal relations.

3) **CAUSAL AND-OR GRAPH (C-AOG)**

This module utilizes the weights obtained at the causal learning step, observed fluent, detected action and consistence to predict the visibility fluent.

### C. INFLUENCE FIELDS

The occlusion issue is related to the viewpoint. Considering the perception ranges of sensors like cameras and LiDARs are normally fan-shaped, we employ the polar coordinate conversion to present multi-vehicle interaction in the bird-eye view. The field of view from the ego-vehicle is divided into a numbers of blocks according to various angles and radius as shown in Fig. 3.

The blocks are labeled and scored according to various influence fields as shown in Fig. 4. There are 3 kinds of influence fields for a vehicle: side field, heading field and rear field, whose intensity ranges from 0% to 100%. The side fields correspond to the blocks occupied by both sides of the vehicle, and the intensity of the side fields maintains 100%. The heading fields are related to the blocks in front of the vehicle, while the rear fields involve the blocks behind the vehicle. The intensities of the heading and rear fields decrease along the distance away from the vehicle. The descent rate

VOLUME 8, 2020
depends on the position and orientation of the vehicle. Especially, the descent of the rear field is faster than the descent of the heading field, because the rear part is less likely to affect others.

Moreover, if the field of a vehicle overlays the field of another vehicle at certain blocks, the overlaps are regarded as the interaction blocks marked in dark shaded gray as shown in Fig. 3. Actions depend on the configurations of the interaction blocks. Besides, fluent score depends on the types of the occupied blocks.

There are 9 pairs of the interaction fields between various influence fields as illustrated in Fig. 5, consisting of the influence nodes in the A-AOG. The left column indicates 3 influence fields of certain vehicle. The top row denotes 3 influence fields of another vehicle. In our method, the scores of the influence nodes in the A-AOG are computed by the overlap of influence fields.

In the framework of H-AOG, a parse graph is a couple of linked nodes. A parse graph is illustrated as the red linking path in Fig. 6. Totally, there are \(2^{N_{or}}\) (\(N_{or}\) is the number of the Or-nodes) possible parse graphs in an AOG. At the time \(t\), the parse graph \(pg\) involves the layouts of various functions as follows:

1) FLUENT LAYER
It presents the visibility fluent of the object \(i\) at the time \(t\).

2) STATE LAYER
There are three visibility states: visible, partly-occluded, and fully-occluded. At the time \(t\), the parent fluent layers choose one of these states.

3) TRANSITION LAYER
There are totally 9 transition pairs of the states: initially visible, staying visible, partly-occluded to visible, initially partly-occluded, visible to partly-occluded, staying partly-occluded, fully-occluded to partly-occluded, partly-occluded to fully-occluded, and staying fully-occluded. If the state changes, the transition belongs to the 11 pairs. Note that the state cannot jump from visible to fully-occluded, because it normally changes from visible to partly-occluded before being fully-occluded.

4) ACTION LAYER
This layer is shared by the C-AOG and A-AOG. An action is the basic reason why the fluent changes. However, vehicle action is hard to be directly described (e.g., the overtaking is a process lasting several seconds, which involves approaching, leaving, and other sub-actions). Hence, we decompose the action into two following layers.

5) POSITION LAYER
This layer denotes the relative position of the objects. For example, being in the front, being side by side and following are different positions of objects.

6) INFLUENCE LAYER
We propose the influence field to score the influence of interactive objects, which is presented in Section II-C.

D. HIERARCHICAL AND-OR GRAPH
The H-AOG consists of the A-AOG module and the C-AOG module as shown in Fig. 6. There are 3 types of nodes in the A-AOG and C-AOG: And-node, Or-node and Leaf-node. The And-node represents the decomposition of the top-level entities. The Or-node represents variations or choices. The Leaf-node is the basic bottom unit, e.g., object’s actions. Actions are decomposed by the And-nodes into a couple of leaf events. The leaf events can be directly detected from the data sequence.

Some state-of-the-art detection algorithms based on the deep convolution networks can be explored. Besides the nodes, there exist the edges linking the nodes, which represent the causal relations.

There are 6 layers in the framework of the A-AOG and the C-AOG from the top level to the bottom level: a fluent layer, a state layer, a transition layer, an action layer, a position layer and an influence layer. The C-AOG involves the top 4 layers which represent the causal relations between actions and fluents. The A-AOG contains the bottom 3 layers, which decompose the actions into single interactions of objects that can be directly detected from the sequence. All edges in the A-AOG have the same weight.

In the framework of H-AOG, a parse graph is a couple of linked nodes. A parse graph is illustrated as the red linking path in Fig. 6. Totally, there are \(2^{N_{or}}\) (\(N_{or}\) is the number of the Or-nodes) possible parse graphs in an AOG. At the time \(t\), the parse graph \(pg\) involves the layouts of various functions as follows:

1) FLUENT LAYER
It presents the visibility fluent of the object \(i\) at the time \(t\).

2) STATE LAYER
There are three visibility states: visible, partly-occluded, and fully-occluded. At the time \(t\), the parent fluent layers choose one of these states.

3) TRANSITION LAYER
There are totally 9 transition pairs of the states: initially visible, staying visible, partly-occluded to visible, initially partly-occluded, visible to partly-occluded, staying partly-occluded, fully-occluded to partly-occluded, partly-occluded to fully-occluded, and staying fully-occluded. If the state changes, the transition belongs to the 11 pairs. Note that the state cannot jump from visible to fully-occluded, because it normally changes from visible to partly-occluded before being fully-occluded.

4) ACTION LAYER
This layer is shared by the C-AOG and A-AOG. An action is the basic reason why the fluent changes. However, vehicle action is hard to be directly described (e.g., the overtaking is a process lasting several seconds, which involves approaching, leaving, and other sub-actions). Hence, we decompose the action into two following layers.

5) POSITION LAYER
This layer denotes the relative position of the objects. For example, being in the front, being side by side and following are different positions of objects.

6) INFLUENCE LAYER
We propose the influence field to score the influence of interactive objects, which is presented in Section II-C.
E. PROBABILISTIC REPRESENTATION

For selecting the best parse graph at the time \( t \), we raise a Bayesian probability distribution updating scheme as

\[
P(p_{g_i}^t | S_{1:t}) = \frac{P(p_{g_i}, S_{1:t})}{P(S_{1:t})} = \frac{P(p_{g_i}, S_{1:t-1}, S_t)}{P(S_{1:t})} = \frac{P(S_t | p_{g_i}, S_{1:t-1})P(p_{g_i}, S_{1:t-1})}{P(S_{1:t})} \propto P(S_t | p_{g_i}; \theta)P(p_{g_i} | S_{1:t-1}; \theta),
\]

where \( p_{g_i} \) represents the best parse graph for the vehicle \( i \) at the time \( t \), \( \theta \) denotes the parameters learned in our method or based on the prior. \( \psi(s_t | p_{g_i}; \theta) \) is the likelihood energy, which measures how well the parse graph explains the data. It consists of several energy parts as:

\[
\psi(s_t | p_{g_i}; \theta) = \sum_{d_m \in D_{\text{and}}(p_{g_i})} \psi(d_m^i) + \sum_{d_m \in D_{\text{or}}(p_{g_i})} \psi(d_m^i) + \sum_{c_r_m \in \text{CR}(p_{g_i})} \psi(c_r_m),
\]

where \( \psi(F_t^i) \) represents \( F_t^i \), \( D_{\text{and}}(p_{g_i}) \) and \( D_{\text{or}}(p_{g_i}) \) represent the active And-nodes and Or-nodes in the parse graph \( p_{g_i} \). \( \psi(d_m^i) \) represents the energy of the And-nodes and Or-nodes. \( \psi(c_r_m^i) \) represents the causal relation in parse graph \( p_{g_i} \). \( \psi(d_m^i) \) represents the energy of the causal relation \( c_r_m^i \). Causal relations in the A-AOG have no extra energy. The energy of And-nodes and Or-nodes is calculated according to their child nodes. \( \psi(d_m^i) = \sum_{d_c \in \text{child}(d_m^i)} \psi(d_c^i), \psi(d_m^i) \in \{ \psi(\text{child}(d_m^i)) \} \), \( \text{child}(d_m^i) \) represents the child nodes of the node \( d_m^i \). Hence, the energy of And-nodes and Or-nodes can be hierarchically represented as the sum of the leaf-nodes.

The likelihood energy can be rewritten as:

\[
\sum_{d_m \in D_{\text{and}}(p_{g_i})} \psi(d_m^i) + \sum_{d_m \in D_{\text{or}}(p_{g_i})} \psi(d_m^i) = \sum_{d_m \in D_{\text{leaf}}(p_{g_i})} \psi(d_m^i | A_t^{1:j}),
\]

where \( A_t^{1:j} \) denotes all the actions at the time \( t \). \( D_{\text{leaf}}(p_{g_i}) \) represents the chosen leaf-nodes in the parse graph \( p_{g_i} \).

In addition, \( \psi(p_{g_i} | S_{1:t-1}; \theta) \) is the prior energy. It can be formulated as:

\[
\psi(p_{g_i} | S_{1:t-1}; \theta) = \psi(p_{g_0^i}; \theta) + \sum_{t=1}^{N} \psi(p_{g_{t-1}^i} | p_{g_{t-1}^i}),
\]

where \( \psi(p_{g_0^i}; \theta) \) is the initial energy based on the learning knowledge. \( \psi(p_{g_{t-1}^i} | p_{g_{t-1}^i}) \) is the transition energy which depends only on the previous parse graph \( p_{g_{t-1}^i} \) according to the hypothesis 2, which makes causation energy of the leaf nodes.

The likelihood energy can be rewritten as:

\[
\psi(d_m^i | A_t^{1:j}) = \begin{cases} 0, & \text{if } p_{g_i} \text{ consists with } p_{g_i-1} \\ \lambda, & \text{else} \end{cases}
\]

where \( \lambda \gg 1 \), which means that heavy punishment should be given if there exists inconsistency.

For prediction of visibility fluent, maximum a posterior probability (MAP) is employed to estimate the parse graph. The function of Eq. 8 can be rewritten as:

\[
p_{g_i}^* = \arg \max P(p_{g_i}^* | S_{1:t}) \propto \arg \max P(s_t | p_{g_i}^*; \theta)P(p_{g_i}^* | S_{1:t-1}; \theta)
\]

\[
= \arg \max \frac{1}{Z} \exp \left\{ -\psi(p_{g_0^i}; \theta) - \sum_{t=1}^{N} \psi(p_{g_{t-1}^i} | p_{g_{t-1}^i}) - \sum_{d_m \in D_{\text{leaf}}(p_{g_i})} \psi(d_m^i | A_t^{1:j}) - \sum_{c_r_m \in \text{CR}(p_{g_i})} \psi(c_r_m^i) \right\},
\]

where \( \psi(d_m^i | A_t^{1:j}) \) denotes the detection score of actions. \( \psi(c_r_m^i) \) denotes the score of causal relation.
In addition, we use a Viterbi algorithm to make dynamic programming for inference. A best parse graph chain is a sequence of the previous parse graphs, which determine the object’s visibility state at the key time.

F. CAUSAL LEARNING

Moreover, we use the method [12] to learn weights of causal relations in the C-AOG. The training process is formulated as an incremental construction of causal structure:

\[ P_0(S) \rightarrow P_1(S) \rightarrow \cdots \rightarrow P(S) \rightarrow P_+(S) \rightarrow P_k(S) \approx \hat{P}(S), \]

(14)

where each new model incorporates a new causal relation. KL-divergence is employed to pursue the causal relations. In a single iteration, the current model \( P(S) \) is augmented to a new model \( P_+(S) \). By using the minimization-maximization of information gain, the learning process consists of two steps. In the first step, the best causal relation \( cr_+ \) is computed through a greedy pursuit which leads to the maximum reduction of the KL-divergence between \( P_+ \) and \( P \):

\[ cr_+ = \arg \max_{cr} KL(P_+ || P). \]

(15)

In the second step, we minimize the KL-divergence between \( P_+ \) and \( P \), subject to matching the causal relation to the observed data. Given a selected relation, step two requires that the new model matches the observed data on the new causal relation. Finally, \( \hat{P}(S) \) denotes the learned probability of causal relations.

III. EXPERIMENTS

In the experiments, we evaluate our method with LSTM [20] and GRU [21], which have capability to predict the sequence with a long-term memory. Note that our method is based on the Markov assumption that the current prediction only relies on the last step, not considering more history data, while the methods based on LSTM/GRU take the last \( n \) samples into account. The experiment involves two parts: using the synthetic data and using the public real-world data from the KITTI dataset. All experiments run on a laptop with an Intel 2.6GHz i7-9750H CPU and 16GB memory.

Here, we employ two metrics to evaluate our method: a classification metric and a bias statistics metric. The former adopts the classification evaluation method to present the performance of inference on the visibility fluent. The latter employs a sequence evaluation method through the definition of a bias \( \delta_t \) as:

\[ \delta_t = |t_{gt} - t_{pred}|, \]

(16)

where \( t_{gt} \) denotes the time when a certain state changes to another state, and \( t_{pred} \) represents the predicted time of the corresponding state transition.

A. SYNTHETIC DATA

The synthetic data involves three cases on an area of 48m × 12m with 3.5m-width lanes and vehicles as shown in Fig. 7.

The data is generated randomly based on three overtaking actions, involving two observed vehicles A and B, which have various scales, positions, velocities and accelerations at the initialization. An ego-vehicle is located at the side of vehicles A and B. In the case I, the vehicle A accelerates in its own lane and then overtakes the vehicle B. In the case II, A follows B at first, then A changes its lane, and finally A overtakes B. Lastly, in the case III, A overtakes B and cuts into the lane where B is located.

We generate the independently identically distributed data: 200 × 3 sequences for training and 200 × 3 sequences for testing. The length of data sequence depends on the velocity and position of the vehicle B. The velocity of the vehicle B is randomly selected from \( 5 \, \text{m/s} \) to \( 15 \, \text{m/s} \), and it is randomly located at certain position of its lane between \( 0 \, \text{m} \) and \( 24 \, \text{m} \). The vehicle A runs faster than the vehicle B and its velocity and position are also randomly generated. The sampling frequency is set to 10 Hz. The synthetic data is self-annotated with the ground-truth states. We pay attention on the visibility status of the vehicle B, because its visibility fluent changes due to the motion of vehicle A.

Firstly, we evaluate the models based on a multi-layer LSTM/GRU with last \( N \{1, 2, 5, 10\} \) samples to predict the visibility fluent (V, P, F) at the next time. In addition, we test the different number of LSTM/GRU layers, including single layer, two layers and three layers. And we find the model with single layer could achieve best results, considering the complexity of data and the generalization of models. The results are shown in the Table 1. It is interesting that when the last samples \( N \) is set to 1, both models could get the best results, and the model based on GRU is a little better on the model based on LSTM. One reason may be that random noise on the data influences the LSTM/GRU model with the input of long sequence. For all models, most prediction of the visible status and fully-occluded status are much better than the prediction of partly-occluded status.

Then, we compare the H-AOG with the best models of LSTM and GRU as shown in Table 2. It is clear that the
TABLE 1. Prediction of visibility fluent by using LSTM and GRU models on the synthetic data.

| Input | Status | Precision | Recall | Accuracy | Precision | Recall | Accuracy |
|-------|--------|-----------|--------|----------|-----------|--------|----------|
| N=1   | V      | 98.48%    | 94.80% | 93.83%   | 99.59%    | 87.15% | 89.71%   |
|       | P      | 70.08%    | 90.77% | 94.95%   | 56.25%    | 89.94% | 94.95%   |
|       | F      | 98.89%    | 92.57% | 89.67%   | 98.78%    | 86.59% | 88.99%   |
| N=2   | V      | 97.74%    | 89.17% | 89.78%   | 98.78%    | 86.59% | 88.99%   |
|       | P      | 56.78%    | 84.06% | 90.99%   | 63.84%    | 88.12% | 97.35%   |
|       | F      | 98.08%    | 94.86% | 89.67%   | 83.77%    | 97.35% | 88.99%   |
| N=5   | V      | 98.15%    | 89.97% | 97.19%   | 98.73%    | 84.64% | 87.82%   |
|       | P      | 61.08%    | 86.30% | 90.99%   | 52.11%    | 90.03% | 96.57%   |
|       | F      | 96.73%    | 97.19% | 90.99%   | 97.10%    | 90.03% | 96.57%   |
| N=10  | V      | 98.37%    | 84.30% | 97.19%   | 97.19%    | 87.97% | 89.26%   |
|       | P      | 52.26%    | 87.84% | 97.19%   | 61.61%    | 85.00% | 89.26%   |
|       | F      | 96.11%    | 97.39% | 97.19%   | 91.29%    | 95.93% | 89.26%   |

TABLE 2. Prediction of visibility fluent by using various models on the synthetic data. The H-AOG model is compared with the best models based on the LSTM and GRU.

| Status | LSTM | GRU | H-AOG |
|--------|------|-----|-------|
| V      | 94.88% | 99.59% | 94.43% |
| P      | 70.08% | 94.95% | 94.15% |
| F      | 98.89% | 90.99% | 74.62% |

TABLE 3. Prediction of visibility fluent by using H-AOG models on each case.

| Case I | Case II | Case III |
|--------|---------|----------|
| Status | Precision | Recall | Accuracy | Precision | Recall | Accuracy | Precision | Recall | Accuracy |
| V      | 91.88% | 87.98% | 83.78% | 95.37% | 76.21% | 91.96% | 95.36% | 73.30% | 90.41% |
| P      | 65.12% | 74.23% | 80.06% | 76.38% | 80.83% | 90.31% | 66.71% | 73.66% | 90.41% |
| F      | 82.27% | 80.06% | 80.06% | 90.31% | 80.83% | 90.31% | 91.50% | 73.66% | 90.41% |

TABLE 4. The bias of prediction using LSTM, GRU and H-AOG in various cases.

| Visibility | 10% | 50% | 90% | 10% | 50% | 90% | 10% | 50% | 90% |
|------------|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| Case I     | V-P | 1   | 4   | 12  | 2   | 6   | 10  | 4   | 9   | 12  |
|            | P-P | 1   | 4   | 8   | 3   | 4   | 10  | 4   | 9   | 12  |
|            | P-P | 0   | 1   | 2   | 1   | 3   | 10  | 4   | 9   | 12  |
|            | V-V | 1   | 2   | 8   | 3   | 5   | 10  | 4   | 9   | 12  |
| Case II    | V-P | 0   | 2   | 6   | 0   | 3   | 7   | 5   | 7   | 8   |
|            | P-P | 0   | 2   | 4   | 0   | 3   | 7   | 5   | 7   | 8   |
|            | P-P | 4   | 7   | 11  | 2   | 5   | 10  | 7   | 8   | 9   |
|            | V-V | 5   | 9   | 13  | 7   | 13  | 14  | 2   | 3   | 5   |
| Case III   | V-P | 3   | 6   | 9   | 0   | 3   | 5   | 5   | 6   | 10  |
|            | P-P | 6   | 9   | 12  | 0   | 2   | 4   | 4   | 6   | 10  |
|            | P-P | 0   | 1   | 2   | 0   | 2   | 4   | 2   | 3   | 6   |
|            | V-V | 4   | 5   | 6   | 0   | 2   | 4   | 2   | 3   | 3   |

The prediction of visible state (V) and fully-occluded state (F) is more accurate than the prediction of the partly-occluded state (P). The reason may be that the partly-occluded state is a transition step, which is not salient as the other two states. We find that the H-AOG model could achieve similar accuracy as the best models of LSTM and GRU.

For the 3 kinds of overtaking cases, their respective results of prediction based on the H-AOG are shown in Table 3. The prediction on the case II obtains the best accuracy.

Secondly, we use the bias metric to evaluate various models. The biases in all cases are increasingly sorted and then presented by 10 percentile, 50 percentile, 90 percentile of the entire biases below the certain sampling intervals. It focuses on prediction on the probable state transitions. The bias of each state transition in the test set are increasingly sorted as shown in Table 4.

In the case I, the two vehicles drive straightly. H-AOG model takes the median 7-8 sampling intervals to predict the probable transitions. Most of the prediction are correctly made within 2 to 12 sampling intervals. The results of prediction on F-P and V-P are more accurate than the others. However, it should be noticed that over 20% prediction of V-P and P-V takes over 10 sampling intervals, which is not good enough. The H-AOG model obtains similar results as the GRU model, while the LSTM model is the best on in this case. Then, in the case II, all prediction improves compared
to the case I. The reason may be that the action of vehicle B influences the overtaking of vehicle A. The interaction field could be informative when there are interactions between multiple vehicles. Correspondingly, in the case III, the prediction shows the similar results as the case II. It would be a reasonable interpretation when these two vehicles (A & B) interact between each other.

B. REAL-WORLD DATA

The KITTI dataset provides an amount of image sequences and LiDAR point clouds about multi-vehicle interaction on the traffic scene. We select 3 data fragments from the multi-object tracking dataset in which there are multi-vehicle interaction, e.g., overtaking, as shown in Fig. 8. The data is labeled with frame, track id, type, truncated, occluded, observation angle, 2d bbox, 3d dimensions, 3d location, rotation_y, score. We directly adopt the models trained on the synthetic data to evaluate the generalization and accuracy of our method. The results are shown in the Table 5.

The H-AOG model distinctly outperforms the other two models in all cases. Considering there is no training step, the generalization of the H-AOG model is better than the others. Note that in the case 3 of Table 5, all models work not well. One reason is that the length of the sequence 0010 may be too short to make a reasonable prediction.

As shown in Table 2, the LSTM model and the GRU model could make higher accuracy on simple dataset after training because they could easily capture the latent distribution of the sequence, due to their powerful capacities of memory. However, they could be hard to directly work on new data with different distribution. While, the training of H-AOG is aimed to learn the link weights based on an interpretable graph like some ‘common’ knowledge, which would exist on data of various distributions. Although the accuracy of the H-AOG model on the synthetic data is a bit lower than that of the LSTM model and the GRU model, it would be a promising approach of inference on occlusion issue considering its performance on the experiments of real data with various distributions.

IV. CONCLUSION

In this paper, we propose a hierarchical And-Or graph model to jointly predict the vehicle’s visibility state, taking the occlusion into account. We employ the causal And-Or graph (C-AOG) to represent the causal relations between fluents and actions. The action And-Or graph (A-AOG) are proposed to decompose the comprehensive actions into the simple overlay of influence fields between multiple vehicles. In addition, we adopt a polar coordinate transformation to describe the interactions between observed vehicles. It has proven to have a reasonable performance on the synthetic data and a couple of real data from the KITTI dataset. It can be compatible to various sensor modals, e.g., image sequences and point clouds of LiDAR.

It should be mentioned that the hierarchical AOG model could explain the visibility variations between multi-object interactions in the traffic scenes of our synthetic data and KITTI data, but it could not cover all traffic scenes due to the specific definition of A-AOG. The opinion in this work is to mine and employ common sense for modeling multi-object interaction, and we expect it could improve the capability of scene understanding for intelligent vehicles. In the next step, the occlusion-aware visibility information would be integrated into multi-object tracking and trajectory prediction.

REFERENCES

[1] X. Wang, E. Turetken, F. Fleuret, and P. Fua, “Tracking interacting objects optimally using integer programming,” in Proc. Eur. Conf. Comput. Vis., Zurich, Switzerland, 2014, pp. 17–32.
[2] X. Wang, E. Turetken, F. Fleuret, and P. Fua, “Tracking interacting objects using intertwined flows,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 38, no. 11, pp. 2312–2326, Nov. 2016, doi: 10.1109/TPAMI.2015.2513406.
[3] C. Park, T. J. Woelfi, J. E. Evans, and N. D. Browning, “Minimum cost multi-way data association for optimizing multitarget tracking of interacting objects,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 37, no. 3, pp. 611–624, Mar. 2015, doi: 10.1109/TPAMI.2014.2346202.
[4] F. Chabot, M. Chaouch, J. Rarabarisoa, C. Teuliere, and T. Chatteau, “Deep MANTA: A coarse-to-fine many-task network for joint 2D and 3D vehicle analysis from monocular image,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Singapore, Jul. 2017, pp. 1827–1836.
[5] Y. Zhou, L. Liu, L. Shao, and M. Mellor, “DAVE: A unified framework for fast vehicle detection and annotation,” in Proc. Eur. Conf. Comput. Vis., Amsterdam, The Netherlands, 2016, pp. 278–293.
[6] R. Deng, B. Di, and L. Song, “Cooperative collision avoidance for overtaking maneuvers in cellular V2X-based autonomous driving,” IEEE Trans. Veh. Technol., vol. 68, no. 5, pp. 4434–4446, May 2019, doi: 10.1109/TVT.2019.2906509.
[7] F. Yang, H. Chen, J. Li, F. Li, L. Wang, and X. Yan, “Single shot multibox detector with Kalman filter for online pedestrian detection in video,” IEEE Access, vol. 7, pp. 15478–15488, 2019, doi: 10.1109/ACCESS.2019.2895376.
[8] Y. Du, W. Shangguan, and L. Chai, “Particle filter based object tracking of 3D sparse point clouds for autopilot,” in Proc. Chin. Autom. Cong. (CAC), Xi’an, China, Nov. 2018, pp. 1102–1107.
[9] F. Thoemmes and K. Mohan, “Graphical representation of missing data problems,” Struct. Equation Modeling, A Multidisciplinary J., vol. 22, no. 4, pp. 631–642, Oct. 2015, doi: 10.1177/10705511.2014.557378.
[10] K. Mohan and J. Pearl, “Graphical models for processing missing data,” 2018, arXiv:1801.03583. [Online]. Available: http://arxiv.org/abs/1801.03583

[11] E. Bareinboim and J. Pearl, “Causal inference and the data-fusion problem,” Proc. Nat. Acad. Sci. USA, vol. 113, no. 27, pp. 7345–7352, Jul. 2016, doi: 10.1073/pnas.1510507113.

[12] A. Fire and S.-C. Zhu, “Learning perceptual causality from video,” ACM Trans. Intell. Syst. Technol., vol. 7, no. 2, pp. 1–22, Jan. 2016, doi: 10.1145/2809782.

[13] A. Fire and S.-C. Zhu, “Inferring hidden statuses and actions in video by causal reasoning,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), Singapore, Jul. 2017, pp. 48–56.

[14] Y. Xu, L. Qin, X. Liu, “A causal and-or graph model for visibility fluent reasoning in tracking interacting objects,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Salt Lake City, UT, USA, Jun. 2018, pp. 2178–2187.

[15] M. Ayazoglu, B. Yilmaz, M. Sznaier, and O. Camps, “Finding causal interactions in video sequences,” in Proc. IEEE Int. Conf. Comput. Vis., Washington DC, USA, Dec. 2013, pp. 3575–3582.

[16] Z. Xie, T. Wu, X. Yang, L. Zhang, and K. Wu, “Jointly social grouping and identification in visual dynamics with causality-induced hierarchical Bayesian model,” J. Vis. Commun. Image Represent., vol. 59, pp. 62–75, Feb. 2019, doi: 10.1016/j.jvcir.2019.01.006.

[17] B. Li, T. Wu, C. Xiong, and S.-C. Zhu, “Recognizing car fluents from video,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Las Vegas, NV, USA, Jun. 2016, pp. 3803–3812.

[18] T. Wu, Y. Lu, and S.-C. Zhu, “Online object tracking, learning and parsing with and-or graphs,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 12, pp. 2465–2480, Dec. 2017, doi: 10.1109/TPAMI.2016.2644963.

[19] P. A. Gagniuc, Markov Chains: From Theory to Implementation and Experimentation, Hoboken, NJ, USA: Wiley, 2017.

[20] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” Neural Comput., vol. 9, no. 8, pp. 1735–1780, Nov. 1997, doi: 10.1162/neco.1997.9.8.1735.

[21] B. Luo, H. Wang, X. Jin, and H. Liu, “Empirical evaluation of gated recurrent neural networks on sequence modeling,” in Proc. Adv. Neural Inf. Process. Syst. Workshop Deep Learn., Montreal, QC, Canada, 2014.

XUANPENG LI (Member, IEEE) was born in Nanjing, China, in 1985. He received the B.S. degree in measurement and control technology and instrument and the M.S. degree in software engineering from Southeast University, Nanjing, in 2007 and 2010, respectively, and the Ph.D. degree in information technology from the Université de Technologie de Compiègne, Compiègne, France, in 2014.

From 2014 to 2015, he was a Postdoctoral Researcher at LIVIC, IFSTTAR, and VEDECOM, France. Since 2015, he has been an Assistant Professor with the School of Instrument Science and Engineering, Southeast University. His research interests include causal perception and scene understanding, driving behavior analysis, and risk analysis on the traffic scene in the field of intelligent transportation systems.

QIFAN XUE was born in Nantong, China, in 1995. He received the B.S. degree in measurement and control technology and instrument from Southeast University, China, where he is currently pursuing the Ph.D. degree in measurement and control technology and instrument. His research interests include causal perception and motion prediction on the traffic scene.

JINGWEN ZHAO was born in Qiqihar, China, in 1997. She received the B.S. degree in measurement and control technology and instrument from Southeast University, China, where she is currently pursuing the M.S. degree in measurement and control technology and instrument. Her research interest includes motion prediction in the field of intelligent transportation systems.

DONG WANG received the B.S. degree from theSchool of Instrument Science and Engineering, Southeast University, Nanjing, China, in 2011, and the Ph.D. degree from Southeast University, in 2016.

He is currently an Assistant Professor with the School of Instrument Science and Engineering, Southeast University. His research interests include intelligence transportation, sensor design, data processing, and inertial navigation.

* * *