The Hierarchical And-Or Graph Based Visibility Reasoning on Road Scenes

Qifan Xue, Xuanpeng Li*, Qixu Zhang, Leixin Zheng, Weigong Zhang

School of Instrument Science and Engineering, Southeast University, Nanjing 210096, China
* Corresponding Author li_xuanpeng@seu.edu.cn

Abstract: Occlusion caused by interaction between multiple objects makes the road scene understanding intractable. In this paper, we focus on the prediction of the objects' states by causal reasoning. In this paper, the visibility fluent is used to present the varying state of objects involving *visible*, *occluded*, and *lost*. Then, a Causal And-Or Graph (C-AOG) is constructed to present the causal relations. Besides, an Action And-Or Graph (A-AOG) and the influence field are proposed to encode the interaction of multi-objects. Finally, a probabilistic grammar model is proposed to jointly make inference of visibility fluents. We evaluate our approach on the synthetic data. It proves to achieve a promising performance in the prediction of the objects' states.

Keywords: Road scene understanding; Causal reasoning; Visibility fluent; And-or graph

1 Introduction

In the field of scene understanding, the prediction of object's state helps autonomous robots to estimate collision risk and avoid accidents ahead of the time. However, due to the serious occlusion, it is hard to predict the motions of objects [1, 2]. For instance, as shown in Figure 1, the light gray area represents the field of view from the ego-car. The shaded gray part is the area outside the field of view. There are three cars *A*, *B* and *C* in front of the ego-car. *A* drives alongside the ego-car, changing from *lost* to *visible*. *B* and *C* drive towards the ego-car. The visual states of *B* and *C* become *occluded* since *A* blocks the sight of ego-car. Then, *C* suddenly crashes into the view. The states of three cars are drawn in the bottom of Figure 1.

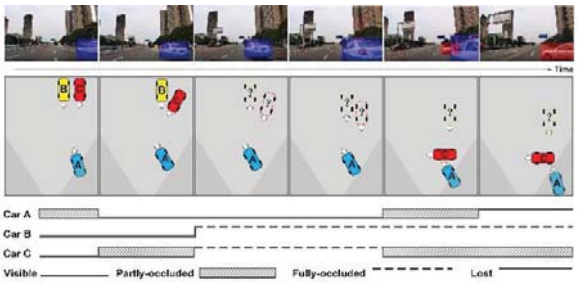

Figure 1 The road scene of multiple objects' interaction.

Modeling of multi-object interaction helps autonomous robots (e.g., Intelligent Vehicles) to prejudge the occurrence of abnormal events and improve driving safety [3]. Especially, in complex road scenes, the uncertainty of state estimation due to incomplete observation, such as partly/fully occlusion, multi-object interaction modeling is still a challenging task. Some filter-based methods as Kalman filter [4] and Particle filter [5] can only track unobserved objects in limited time, which have not the capability of long-term prediction. Therefore, it is necessary to mine the "common sense" such as causality in the environment, and construct an approximate dynamic system model for the interaction between multiple objects.

The paper is organized as follows: section 2 presents relevant research. In the section 3, the basic modeling formulation is presented with the hypotheses, the causal and-or graph, the influence field and the transformation based on polar coordinate. The experiment is carried out with the simulated data in the section 4. Finally, the conclusion is given in the last section.

2 Related work

Causal modeling involves a variety of approaches, typically using a Bayesian network based on the directed graph [6, 7, 8], or grammar model based on the and-or graph [9, 10, 11]. Especially in recent years, the causal and-or graph modeling method get the attention of researchers in visual applications.

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 [6, 7] conduct long-term research on how to solve these problems, and proves that the data missing problem can be encoded using the graph model with conditional independence hypothesis. Ayazoglu et al. [12] transforms the causal modeling problem of interaction between objects into a topology of the Directed Graph from the video. By using the object 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 the data missing and false detection caused by occlusion. Xie et al. [13] also explores the hierarchical Bayesian modeling in order to construct causality for crowd behavior recognition, and realize the analysis of intelligent interaction in video.

On the other hand, for the causal modeling based on grammar model, Fire et al. [9] proposes a fluent-based

978-1-7281-3863-3/19/$31.00 ©2019 IEEE

causal perception method, using heuristic learning to build the causal relationship between behavior and states in video, and then predicts the hidden state and behavior [10]. For instance, Li et al. [14] constructs a spatiotemporal and-or graph framework to identify the state of blurred, heavily occluded or deformed vehicle components (such as doors, trunks, etc.) in the video sequence. It especially achieves the state reasoning of components under multi-objective interaction. Wu et al. [15] uses hierarchical And-Or graph to achieve synchronous tracking, learning and parsing of unknown objects in video, and uses spatiotemporal dynamic programming algorithm to infer real-time object position and size.

3 Framework

The framework is shown in Figure 2 and main components are listed as follows.

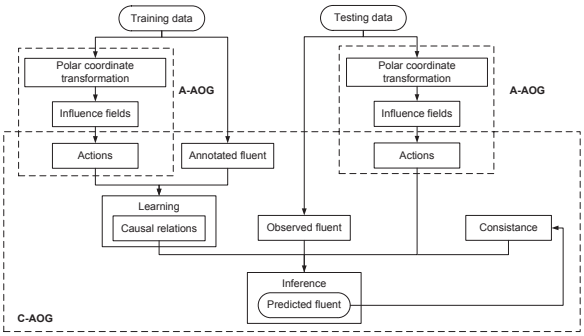

Figure 2 The framework consists of two modules: A-AOG and C-AOG.

Causal learning: In order to get the transition probability of the causal diagram in the C-AOG module, we are inspired by the learning method proposed by Fire et al. [9].

Polar coordinate transformation: We propose a polar coordinate template to transform the synthetic data into the influence fields. The influence fields are the basic units to encode the interaction of multiple objects.

A-AOG: The module decomposes the complex interactions into the pairs of influence fields, which are easy to model objects' interactions.

C-AOG: The module jointly utilizes the weights of causal relations, fluent, actions and consistencies of adjacent frames to make prediction. The consistence makes the choice of the varying states according to the previous fluent.

3.1 Hypotheses

In the paper, the visibility fluent is employed to represent the visible states of objects. We make three hypotheses for the problem formulation.

Hypothesis 1: The fluent changes for the sake of the causal actions.

Hypothesis 2: The fluent maintains unchanged unless the causal actions happen.

Hypothesis 3: The fluent follows the Markov assumption.

The hypotheses indicate that what should be considered is the casual actions but no other potential reasons. Besides, the possible events only depend on the state attained in the previous event.

3.2 Problem Formulation

Four visibility states are used to describe the fluents of observed objects in the road scene: *Visible* (V), *Partly-occluded* (P), *Fully-occluded* (F), and *Lost* (L, it means driving out from the camera's field of view).We assign an individual fluent F_t^i to each object at the frame f_t as follows:

$$F_t^i \in \{V, P, F, L\}, \quad (1)$$

where t and i represent the index of frame and objects at the frame f_t.

There are two types of frames in the sequence V: key frames f_{key} and other frames f_{other}. According to the Hypothesis 1&2, we assume that the change of the fluents and the actions only happen in the key frames. We suppose that no interactive actions happen in the other frames. For this, visibility fluents remain unchanged, which is called 'inertial'. The sequence V can be described as follows:

$$V = \{f_1, f_2, \cdots, f_t\}, \quad (2)$$

$$f_t = \begin{cases} \{F_t^1, F_t^2, \cdots, F_t^i, A_t^1, A_t^2, \cdots, A_t^j\}, & \text{for } f_t \in f_{key} \\ \{F_t^1, F_t^2, \cdots, F_t^i\}, & \text{for } f_t \in f_{other} \end{cases}, \quad (3)$$

where F_t^i and A_t^i denote the fluent and the action in the frame f_t.

Overall, the state in the other frame depends on the previous key frame. In our work, the unknown fluents and actions can be estimated by jointly utilizing the detected fluents and actions, the previous states of objects, the causal relations and the prior knowledge.

3.3 And-or Graph

In this paper, we employ a grammar model based on a hierarchically bottom-up and-or graph to represent the causal relations between the fluents and the actions. There are three types of nodes in the AOG: **And-node**, **Or-node**, **Leaf-node**. The And-nodes represent the decompositions of the top level entities. The Or-nodes represent variations or choices [11]. The Leaf-nodes are the basic units that can be directly acquired in the synthetic data sequence. There exist edges linking the nodes, which represents the causal relation.

The hierarchical and-or graph is a *bottom-up* structure as shown in Figure 3. There are 6 layers in the module of C-

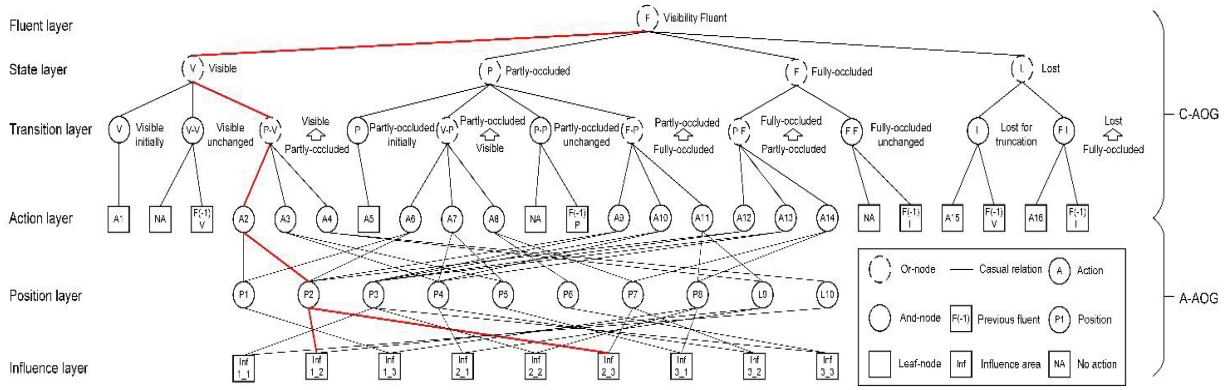

Figure 3 The hierarchical and-or graph of the grammar model.

AOG and A-AOG. The C-AOG including the top 4 layers indicates the causal relations between actions and fluents. The A-AOG contains the bottom 3 layers, which decompose the actions into interactions of objects that can be directly detected in the sequence.

In the AOG, a parse graph is composed of linked nodes. An active parse graph is indicated by the red linking path as illustrated in Figure 3. At the frame f_t, the parse graph pg_t^i involves several function layers as follows:

Fluent layer: At the frame f_t, we assign a visibility fluent to each object.

State layer: At the parent node, fluent would only be one of the four states: *Visible* (V), *Partly-Occluded* (P), *Fully-occluded* (F) and *Lost* (L).

Transition layer: If the state changes, there are only 11 specific pairs, such as V-V, V-P, and P-F. For example, the state cannot jump from *Visible* to *Fully-occluded*, because it should change to *Partly-occluded* before being *Fully-occluded*.

Action layer: The layer links C-AOG and A-AOG. The actions are the evidences to reason the varying fluent. However, the actions are hard to be directly described. Hence, we decompose the actions into the two following layers.

Position layer: The layer denotes the relative position of objects. There are 9 types of relative position as illustrated in Figure 5.

Influence layer: We propose the influence field to score the influence of interaction between multiple objects, which is presented in Section 3.6.

3.4 Probabilistic Representation

We raise a Bayesian probability distribution as

$$P(pg_t^i|f_{1:t}) \propto P(V_t|pg_t^i;\Theta)P(pg_t^i|f_{1:t-1};\Theta)$$
$$= \frac{1}{Z}\exp\{-\varepsilon(V_t|pg_t^i;\Theta) - \varepsilon(pg_t^i|f_{1:t-1};\Theta)\}, \quad (4)$$

where pg_t^i represents the best parse graph for object i at the frame f_t. $f_{1:t}$ represents the sequence from frame 1 to t. Θ denotes the parameters in our method. $\varepsilon(V_t|pg_t^i;\Theta)$ is the energy function. It is combined by the following three components as:

$$\varepsilon(V_t|pg_t^i;\Theta) = \varepsilon(F_t^i) + \sum_{d_m^i \in D_{and}(pg_t^i)} \varepsilon(d_m^i) + \sum_{d_m^i \in D_{or}(pg_t^i)} \varepsilon(d_m^i) + \sum_{r_m^i \in R(pg_t^i)} \varepsilon(r_m^i), \quad (5)$$

where $\varepsilon(F_t^i)$ represents the energy of F_t^i. If the fluent F_t^i is unable to be detected at the frame f_t, it is set as the default value. $\varepsilon(d_m^i)$ represents the energy of the and-nodes and or-nodes. $D_{and}(pg_t^i)$ and $D_{or}(pg_t^i)$ represent the active and-nodes and or-nodes in the parse graph pg_t^i. $\varepsilon(r_m^i)$ denotes the energy of the causal relation r_m^i. The causal relations in the A-AOG have no energy representation.

Furthermore, the energy of and-nodes and or-nodes is calculated according to their child nodes. $\varepsilon(d_m^i) = \sum_{d_n^i \in child(d_m^i)} \varepsilon(d_n^i)$, $\varepsilon(d_m^i) \in \{\varepsilon(child(d_m^i))\}$, $child(d_m^i)$ 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:

$$\sum_{d_m^i \in D_{leaf}(pg_t^i)} \varepsilon(d_m^i|A_t^{1:j}) = \sum_{d_m^i \in D_{and}(pg_t^i)} \varepsilon(d_m^i) + \sum_{d_m^i \in D_{or}(pg_t^i)} \varepsilon(d_m^i) \quad (6)$$

where $A_t^{1:j}$ denotes all the j actions at the frame f_t. $D_{leaf}(pg_t^i)$ represents the chosen leaf-nodes in the parse graph pg_t^i.

In addition, $\varepsilon(pg_t^i|f_{1:t-1};\Theta)$ stands for the prior energy. It can be formulated as:

$$\varepsilon(pg_t^i|f_{1:t-1};\Theta) = \varepsilon(pg_0^i|\Theta) + \sum_{t=1}^{N} \varepsilon(pg_t^i|pg_{t-1}^i), \quad (7)$$

where $\varepsilon(pg_t^i|pg_{t-1}^i)$ is the transition energy which depends only on the previous parse graph pg_{t-1}^i

according to the hypothesis 3.$\varepsilon(pg_0^i|\Theta)$ is the initial energy based on the learning knowledge.

3.5 Learning of Causal Relations

The probability of causal relations is iteratively updated during training. The training process is denoted as:

$$p_0(V) \to p_1(V) \to \cdots \to p_n(V) \to p_{n+1}(V) \to \cdots \to p_k(V) \approx \hat{p}(V), \quad (8)$$

where $p_0(V)$ to $p_n(V)$ are the probability in the first step where the possible causal relations is added. $p_{n+1}(V)$ to $p_k(V)$ represent the probability in the second step where the redundant causal relations is removed. Finally, $\hat{p}(V)$ represent the learned probability of causal relations.

KL-divergence is employed to pursuit the causal relations. The best causal relation, denoted as cr_+, is selected through a greedy pursuit which leads to the maximum reduction of the KL-divergence:

$$cr_+ = \text{argmax}_{cr}(KL(f||p) - KL(f||p_+)), \quad (9)$$

where cr_+ denotes the best causal relation. p and p_+ represent the current probability and the updated probability.

3.6 Influence Field

In order to score the nodes of the bottom influence layer in the A-AOG in Figure 3, we adopt the influence field to quantify the interaction as shown in Figure 4. Taking account of the complexity, we only focus on the vehicle without other types of objects.

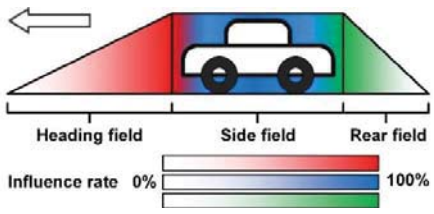

Figure 4 Three influence fields in the vehicle

With the definition of three influence fields, the relative positions of observed vehicles are illustrated in Figure 5. There are 9 pairs of influence fields, corresponding to the 9 influence nodes in Figure 3. The left column denotes three influence fields of ne object. The top row denotes three influence fields of another object.

With the definition of influence fields, we use a polar coordinate transformation to describe the interactions in the bird-eye view. The field of view of ego-car is divided into cells according to the angles and radius. As shown in Figure 6, there are several cars with different positions and orientations. The cells are differently marked according to the influence fields.

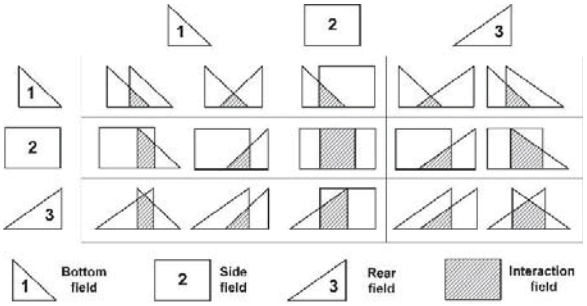

Figure 5 Pairs of influence fields.

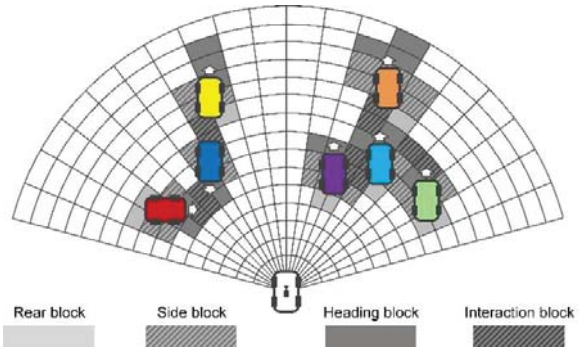

Figure 6 The polar coordinate transformation.

3.7 Inference

In the inference stage, the maximum a posteriori probability (MAP) is employed to compute the parse graph. The energy function of Eq. 4. can be re-written as:

$$P(pg_t^i|f_{1:t}) \propto P(V_t|pg_t^i;\Theta)P(pg_t^i|f_{1:t-1};\Theta)$$
$$= \frac{1}{Z}\exp(-\varepsilon(F_t^i) - \varepsilon(pg_0^i|\Theta) - \sum_{t=1}^{N}\varepsilon(pg_t^i|pg_{t-1}^i)$$
$$- \sum_{d_m^i \in D_{leaf}(pg_t^i)}\varepsilon(d_m^i|A_t^{1:j}) - \sum_{r_m^i \in R(pg_t^i)}\varepsilon(r_m^i)), \quad (10)$$

where $\varepsilon(F_t^i)$ and $\varepsilon(d_m^i|A_t^{1:j})$ denote the detection score of fluent and causal actions. $\varepsilon(r_m^i)$ represents the score of causal relations. $\varepsilon(pg_0^i|\Theta)$ denotes the initial parse graph generated by the priori probability. $\varepsilon(pg_t^i|pg_{t-1}^i)$ denotes the punishment energy checking the consistency between the frames.

Then, we use the Viterbi algorithm to make the dynamic programming for the inference. A best parse graph chain is a sequence of the previous parse graphs. The best parse graph chain decides the states of objects in the key frames. The fluents maintain unchanged at the other frames.

4 Experiment and analysis

We use the synthetic data of multi-object interaction to evaluate our method. Three cases are defined as shown in Figure 7. In each case, the simulation area is

$48\text{m} \times 12\text{m}$ involving three 3.5m-width lanes. The sizes of cars range from $3.2\text{m} \times 1.6\text{m}$ to $3.4\text{m} \times 1.8\text{m}$. The Gaussian noise is added to the simulated data considering the uncertainty of observation. In each case, there are two observed cars *A* & *B* and the ego-car *E*. In the case I, *A* overtakes B within its lane beside the ego-car *E*. In the case II, *A* changes the lane to overtake *B*. In the case III, *A* overtakes B and drives into the lane of *B*.

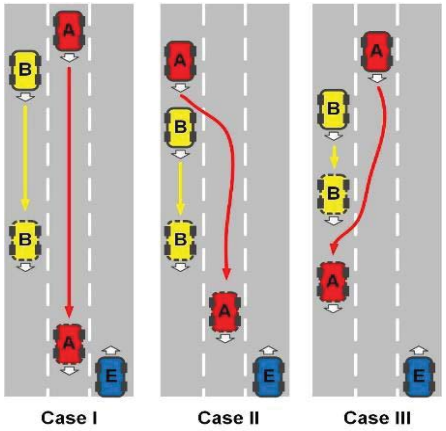

Figure 7 Three cases of synthetic data on road scenes

For each case, 200 examples are involved where the cars have different positions, speeds and accelerations. The period of each example is around 200 frames. The simulator annotates the ground truth. For each frame, our method focuses on the prediction of the state of the car *B*, since it is occluded by the car A. We evaluate our method by using the bias δ_t defined as:

$$\delta_t = t_{gt} - t_p , \qquad (11)$$

where t_{gt} represents the time when the state transition happens. t_p represents the transition time predicted by the method.

In addition, the biases of the entire sample set are ascendingly sorted and then presented by three levels: 10 percentile, 50 percentile, 90 percentile. The N-percentile means N% of the entire biases below the certain frames. It can evaluate the performance of our method to predict the probable state transitions.

The performance of our method is shown in the Table I, II, III and Figures 8, 9, 10. Here, we focus on the car *B* whose visibility fluent changes due to the motion of car A.

In the case I, the two cars drive straight and regularly. It takes the median 7-8 frames for our method to predict the probable transitions. Most of the predictions are correctly made within 2 to 12 frames. The predictions of F-P and F-L are more effective. Particularly, the transition of F-L means the visual state of the car *B* can only be inferred by the car *A*'s actions. However, it should be noticed that over 20% predictions of V-P and P-F takes over 10 frames, which is not good enough.

Table I The biases in case I of 200 samples

Transitions	Percentile		
	10	50	90
V-P	3-4	8-9	14-15
P-F	7-8	9-10	11-12
F-P	1-2	2-3	3-4
F-L	2-3	3-4	5-6
All transitions	2-3	7-8	12-13

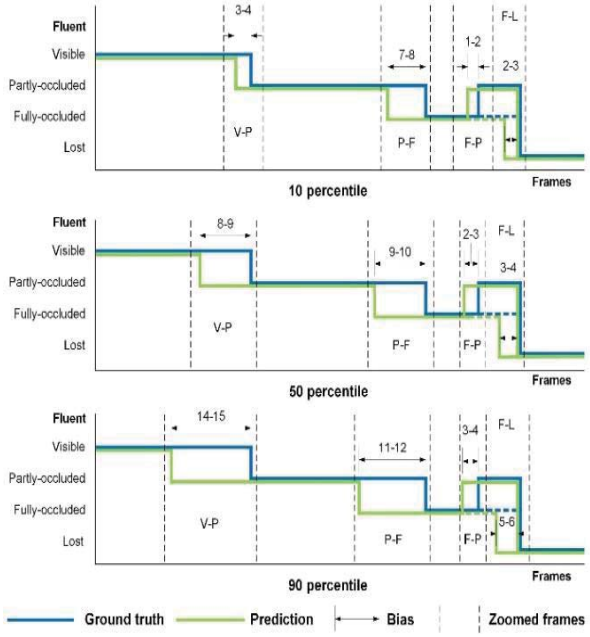

Figure 8 The fluents and biases car *B* in case I

Compared with the case I, there are the same four transitions in the case II. Similarly, they have optimal predictions in F-P and F-L. However, the performance in V-P and P-F improves a lot. The reason is that the action of changing the lane for overtaking is identical enough for tracking.

Table II The biases in case II of 200 samples

Transitions	Percentile		
	10	50	90
V-P	4-5	6-7	7-8
P-F	6-7	7-8	8-9
F-P	1-2	2-3	2-3
F-L	1-2	2-3	4-5
All transitions	2-3	5-6	7-8

In the case III, the predictions of V-P, P-F and F-P have the similar performance with those in case II. The new transition P-V denotes that the observed appears after being occluded by the car *B*. The performance varies a lot for that the car *A* have different orientations and speeds to cut in the lane.

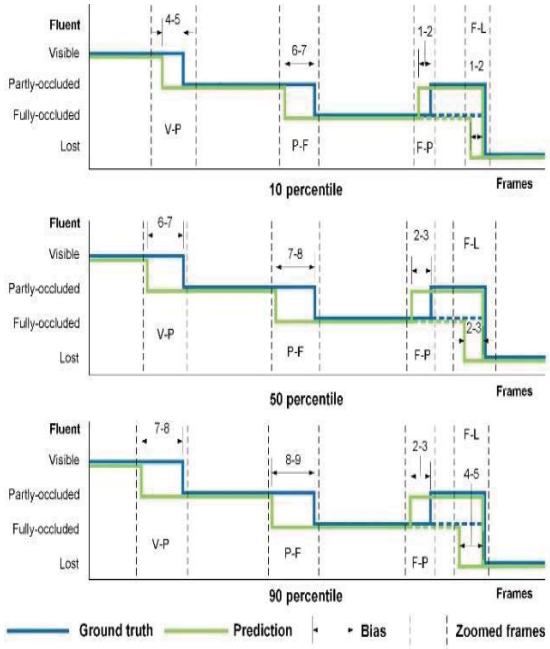

Figure 9 The fluents and biases car *B* in case II.

Table III The biases in case III of 200 samples

Transitions	Percentile		
	10	50	90
V-P	4-5	5-6	9-10
P-F	3-4	4-5	5-6
F-P	1-2	1-2	2-3
P-V	0-1	1-2	7-8
All transitions	1-2	4-5	7-8

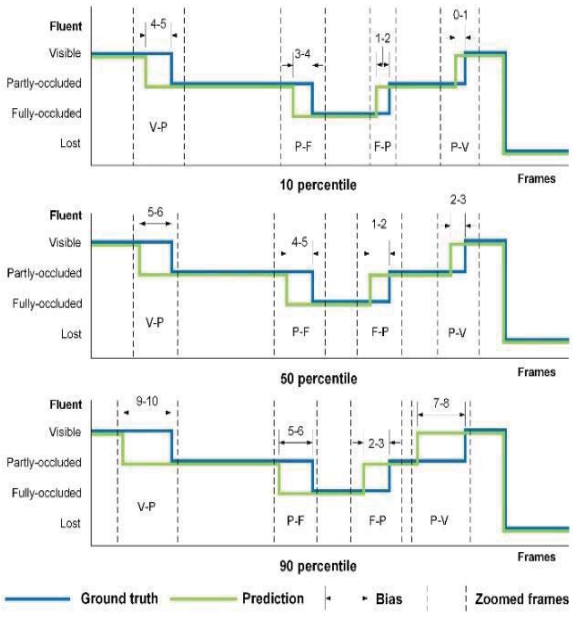

Figure 10 The fluents and biases of car *B* in case III.

Overall, under common circumstances, our method can give the prediction of transitions within 10 bias frames ahead of time. Especially, the states of the car *B* are estimated completely depending on its fluent and the car *A*'s actions. However, the performance in the actions with weak tendentiousness still remains to be improved.

5 Conclusions

In this paper, we propose a hierarchical and-or graph model to jointly predict the object visibility. We use the causal and-or graph (C-AOG) to represent the causal relations between the fluents and the actions. The influence field and the action and-or graph (A-AOG) are proposed to decompose the comprehensive actions into the overlay of influence fields. In addition, we adopt a polar coordinate transformation to describe the interaction between observed objects. It has proven a reasonable performance on the simulated data. In the further step, it remains to be improved and applied to the real data.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (No. 61906038) the Natural Science Foundation of Jiangsu Province (No. BK20160700) and the Fundamental Research Funds for the Central Universities (No. 2242019K40039).

References

[1] Chabot Florian, Chaouch Mohamed, Rabarisoa Jaonary et al. Deep MANTA: A Coarse-to-Fine Many-Task Network for Joint 2D and 3D Vehicle Analysis from Monocular Image. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 2040-2049.

[2] Zhou Yi, Liu Li, Shao Ling et al. DAVE: A Unified Framework for Fast Vehicle Detection and Annotation. European Conference on Computer Vision. 2016: 278-293.

[3] Deng R, Di B, Song L. Cooperative Collision Avoidance for Overtaking Maneuvers in Cellular V2X-Based Autonomous Driving. IEEE Transactions on Vehicular Technology, 2019, 68(5): 4434-4446.

[4] Fan Yang, Houjin Chen, Jupeng Li et al. Single Shot Multibox Detector With Kalman Filter for Online Pedestrian Detection in Video. IEEE Access, 2019, 7: 15478-15488.

[5] Du Yu, ShangGuan Wei, Chai LinGuo. Particle Filter Based Object Tracking of 3D Sparse Point Clouds for Autopilot. 2018 Chinese Automation Congress, 2018:1102-1107

[6] Thoemmes F, Mohan K. Graphical representation of missing data problems. Structural Equation Modeling: A Multidisciplinary Journal, 2015, 22(4): 631-642.

[7] Mohan K, Pearl J. Graphical models for processing missing data. arXiv preprint arXiv:1801.03583, 2018.

[8] Bareinboim E, Pearl J. Causal inference and the data-fusion problem. Proceedings of the National Academy of Sciences, 2016, 113(27): 7345-7352.

[9] Fire A, Zhu S C. Learning perceptual causality from video. ACM Transactions on Intelligent Systems and Technology, 2016, 7(2): 23.

[10] Fire A, Zhu S C. Inferring Hidden Statuses and Actions

in Video by Causal Reasoning. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2017: 48-56.
[11] Qin L, Xu Y, Liu X, et al. A Causal And-Or Graph Model for Visibility Fluent Reasoning in Human-Object Interactions. arXiv preprint arXiv:1709.05437, 2017.1
[12] Ayazoglu M, Yilmaz B, Sznaier M, et al. Finding causal interactions in video sequences. Proceedings of the IEEE International Conference on Computer Vision. 2013: 3575-3582.
[13] Xie Z, Wu T, Yang X, et al. Jointly social grouping and identification in visual dynamics with causality-induced hierarchical Bayesian model. Journal of Visual Communication and Image Representation, 2019, 59: 62-75.
[14] Li B, Wu T, Xiong C, et al. Recognizing car fluents from video. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 3803-3812.
[15] Wu T, Lu Y, Zhu S C. Online Object Tracking, Learning and Parsing with And-Or Graphs. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(12): 2465-2480.