International Journal of Automation and Computing 15(3), June 2018, 249-266
DOI: 10.1007/s11633-018-1126-y

A Survey of Scene Understanding by Event Reasoning in Autonomous Driving

Jian-Ru Xue[1] Jian-Wu Fang[1,2] Pu Zhang[1]

[1]Institute of Artificial Intelligence and Robotics, Xi′an Jiaotong University, Xi′an 710049, China

[2]Chang′an University, Xi′an 710064, China

Abstract: Realizing autonomy is a hot research topic for automatic vehicles in recent years. For a long time, most of the efforts to this goal concentrate on understanding the scenes surrounding the ego-vehicle (autonomous vehicle itself). By completing low-level vision tasks, such as detection, tracking and segmentation of the surrounding traffic participants, e.g., pedestrian, cyclists and vehicles, the scenes can be interpreted. However, for an autonomous vehicle, low-level vision tasks are largely insufficient to give help to comprehensive scene understanding. What are and how about the past, the on-going and the future of the scene participants? This deep question actually steers the vehicles towards truly full automation, just like human beings. Based on this thoughtfulness, this paper attempts to investigate the interpretation of traffic scene in autonomous driving from an event reasoning view. To reach this goal, we study the most relevant literatures and the state-of-the-arts on scene representation, event detection and intention prediction in autonomous driving. In addition, we also discuss the open challenges and problems in this field and endeavor to provide possible solutions.

Keywords: Autonomous vehicle, scene understanding, event reasoning, intention prediction, scene representation.

1 Introduction

- *Automation is one of the hottest topics in transportation research and could yield completely driverless cars in less than a decade. — Nature in 2015*[1].

Can the driverless cars be completely yielded in less than a decade? Manifestly, it is still decades away based on the observations of current progresses and remaining challenges in autonomous vehicles. So far, no one is close to develop a fully autonomous vehicle. The fleet testing by Uber and Google operates under tightly controlled conditions[2].

The reasons are from four aspects: 1) The existing methods of environment perception, e.g., detection[3], tracking[4, 5] and segmentation[6] of participants in traffic scenes, still produce inevitable errors in real environment; 2) The driving environment is rather complex, unpredictable, dynamic, and uncertain; 3) Deep traffic scene understanding, such as understanding the geometry/topology structure of scene, and spatio-temporal evolution of participants (pedestrian, vehicle, etc.), is studied far from sufficient, whose ultimate goal is to semantically reasoning the scene evolvement so as to provide clues for behavior decision and autonomous vehicle control. Actually, it is difficult to study because these elements are implicitly contained in the driving environment and cannot be directly observed; 4) The deployment of autonomous vehicle faces social dilemma and involves moral issue[7]. Complementary to our survey, Janai et al.[8] exhaustively reviewed the traffic participant recognition, detection and tracking, scene reconstruction, motion estimation, semantic segmentation, and many other vision-based tasks. Xue et al.[9] made an overview on autonomous vehicle systems from the perspectives of self-localization and multi-sensor fusion for obstacle detection and tracking, and emphasized vision-centered fusion of multiple sensors. Zhu et al.[10] studied the latest progresses on lane detection, traffic sign/light recognition in the perception of intelligent vehicles. These surveys, to a great extent, give a comprehensive and detailed investigation concerning with the first reason mentioned above.

In this paper, we focus on the third aspect: survey on the deep understanding of traffic scene for autonomous vehicles. This paper aims to explore the evolution of traffic scene from an event reasoning view. That is because event can reflect the dynamic evolution process of scene with tractable reasoning strategy[11]. In order to provide a clear and logical investigation, this paper reasons the event from its representation, detection, as well as prediction stages. In the representation stage, the main goal is to obtain high-level clues for the following stages. In this stage, we expound the saliency, the contextual layout, and the topology rules for autonomous driving. As for the detection stage, we review the event detection with respect to different participants, such as pedestrian and vehicles. For the prediction

Review
Manuscript received October 16, 2017; accepted March 9, 2018; published online April 18, 2018
This work was supported by National Key R&D Program Project of China (No. 2016YFB1001004), National Natural Science Foundation of China (Nos. 61751308, 61603057, 61773311), China Postdoctoral Science Foundation (No. 2017M613152) and Collaborative Research with MSRA.
Recommended by Associate Editor Matjaz Gams

stage, this paper elaborates the intention of autonomous vehicles with regard to the expected time span for future prediction. We classify the prediction of intention as long-term intention prediction and short-term prediction. Fig. 1 demonstrates the surveying flowchart in this paper. At distinct stages, we discuss open problems and challenges, and endeavour to provide possible solutions.

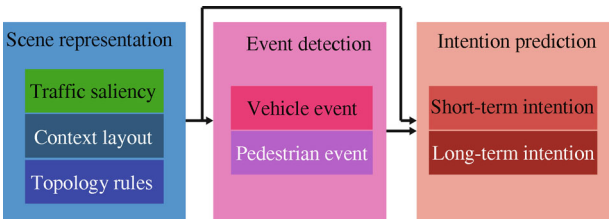


Fig. 1 The scene understanding flowchart by event reasoning framework for autonomous driving

Actually, beyond those stages, some end-to-end approaches emerge recently for scene understanding facing autonomous driving[12−14]. They rely on a large-scale data-driven mechanism, and formulate the scene to decide with deep layers or recursive perception, such as fast recurrent fully convolutional networks (FCN) for direct perception in autonomous driving[12] and FCN-LSTM[13] for a future motion action feasibility distribution. We specially take a section to present this category. We hope that our survey can sweep away some entry barriers of deep scene understanding for autonomous driving, and draw forth meaningful insights and solutions for this field.

1.1 Autonomy pursuit in driving

Developing autonomous systems aim to assist humans in handing everyday tasks. Autonomous driving system, a system for closely related to humans′ everyday trips, has become people′s one of the most typical pursuits. It can free hands from the steering wheel, and spare time for tackling many other things. Meanwhile, the equipped sensors of autonomous vehicle can also recognize the surrounding condition immediately and ensure safe driving, thus decreasing traffic accidents. Encouraged by those merits, researchers are diligently pursuing autonomous driving all the time. There are two kinds of driving force in the development of autonomous driving. One is the projects launched and challenges posed by different governments, research institutes and vehicle manufacturers. The other we want to emphasize is the publicly available benchmarks.

Projects and launched challenges. Since 1986, Europe started an intelligent transportation system project, named as PROMETHEUS, involving more than 13 vehicle manufacturers and research institutions around 19 countries. Thorpe et al.[15] in Carnegie Mellon University launched the first autonomous driving project in the United States. This project made breakthrough in 1995 that autonomously drove a car from Pittsburgh, Pennsylvania to San Diego, California. Supported by many related studies, the US government established the National Automated Highway System Consortium (NAHSC) in 1995. Motivated by these projects, highway scenarios has been intensively studied for a long time, while urban scene remained as an uncultivated area. Actually, urban scene is closely related to human′s daily lives. At that time, a famous "DARPA Grand Challenge (DUC)" launched by Defense Advanced Research Projects Agency (DARPA) largely accelerated the progress of autonomous vehicle. Among them, "Urban Challenge"[16], the third challenge launched by DARPA (others had been held in 2004 and 2005 respectively, aiming to test the self-driving performance in the Mojave Desert of the United States[17, 18], took place on November 3, 2007 at the now-closed George Air Force Base in Victorville, California. Rules included obeying all traffic regulations while negotiating with other vehicles and obstacles and merging into traffic. There were 4 teams completed the route within 6 hours. In 2009, National Natural Science Foundation of China launched the China Intelligent Vehicle Future Challenge (iVFC). Up to now, the ninth contest was held in November 2017. Google started their self-driving car project in 2009, and completed over 5 million miles driving test until March 2018[1]. In 2016, the project was evolved into an independent self-driving technology company Waymo. Tesla Autopilot[2], by equipping cameras, twelve ultrasonic sensors and a forward-facing radar, all the vehicles can have the self-driving ability since October 2016. As a matter of fact, more and more vehicle manufacturers, such as Audi, BMW, Benz, have begin their projects to develop their self-driving vehicles.

Benchmarks. In 2012, Geiger et al.[19] introduced the KITTI vision benchmark, which contained six different urban scenes, and had 156 video sequences with time span from 2 minutes to 8 minutes. Within this benchmark, they launched several typical vision tasks, such as pedestrian/vehicle detection, optical flow, stereo flow, road detection, lane detection, etc. The benchmark was collected by an ego-vehicle equipped with color and gray cameras, and Velodyne 3D laser scanner and high-precision GPS/IMU inertial navigation systems. At the same time, Cambridge University released CamVid dataset[20], which provided a semantic segmentation evaluation benchmark containing only four video sequences on urban scene. Another popular benchmark is the Cityscapes dataset[21] released in 2016. Urban scene was collected in 50 cities, and have 5 000 fine-annotated images and 20 000 coarse-annotated images. Cityscapes has become the most challenging dataset for semantic segmentation task. Actually, annotation is time and labor consuming. Based on that, Gaidon et al.[22] constructed a large-scale KITTI-like virtual dataset[3] by computer graphic technology. The benefit of virtual dataset

[1] https://www.theverge.com/2018/2/28/17058030/waymo-self-driving-car-360-degree-video
[2] https://www.tesla.com/autopilot
[3] http://www.europe.naverlabs.com/Research/Computer-Vision/Proxy-Virtual-Worlds

is that it can generate every wanted task, for those that are encountered rarely. However, these benchmarks have a short time span, which is difficult to the diversity and complexity of scenes. To solve this problem, Maddern et al.[23] constructed RobotCar dataset[4] by travesing 1 000 km in central Oxford in the UK for one year. Images, Lidar and GPS data were collected. This dataset presents larger variations in scene appearance, illumination, and weather. The downside of RobotCar is that it does not provide sufficient annotation. Besides various kinds of sensor equipment systems, some researchers focus on a full view calibration by equipping multiple cameras to cover different perception view around the ego-vehicle, such as LISA-Trajectory[24] and PKU-POSS dataset[25]. Fig. 2 gives a glance for the KITTI and Cityscapes dataset.

Fig. 2 A glance at the KITTI and Cityscapes dataset. The left part is the KITTI data demonstration and relating sensors (adapted from [19]), and the right part is the fine-annotated city scene of a frame (adapted from [21]).

Most datasets focus on the development of algorithm comparison for autonomous driving with respect to different vision based tasks, and the ranking of each task changes more and more frequently. Actually, it is the same for the general computer vision domain, such as PASCAL VOC[26], ImageNet challenge[27], Middlebury for stereo and optical flow[5] , MOT for tracking[28], ActivityNet[29], etc. As several decades passed, there is a consensus that a more diverse and challenging dataset would make the designed method be better and better. However, is it true? Is the best method in the comparison list also the best for practical application? Actually, autonomous driving should not only focus on individual task. When driving, the surrounding scene may produce various tasks simultaneously, such as detection, tracking, segmentation, behavior judgement, event detection, intention prediction, etc. A successful autonomous driving system must realize long-term driving under various environments. Even so, autonomous driving still have a long and zigzag way to go because of the ever changing traffic scenes. There is least one thing for sure that the above two of driving forces for autonomous driving will coexist in the future.

The rest of this paper is organized as follows. Section 2 presents the representation of scene for following event reasoning. Section 3 provides the review for the pedestrian and vehicle event detection, followed by the intention prediction overview in Section 4. In Section 5, we also present the end-to-end frameworks for the direct reasoning by deep learning architectures. Then, we elaborate the evaluation metrics and relating datasets for the event reasoning in Section 6. This paper is finally concluded in Section 7.

2 Scene representation for event reasoning in autonomous driving

Defining event is a difficult problem in cognition science. What kind of scene variation should be taken as an event? Why does the event occur? We attempt to answer these questions from scene representation. Specifically, this paper focuses on the aspects of traffic saliency, content layout and topology rules for self-driving. Reasons are that: 1) Traffic saliency formulates where the scene should or may be looked when driving in different traffic situations[30]. An event always influences and changes the attention of human drivers. In other words, traffic saliency can provide locally instantaneous clue for event reasoning. 2) Context layout specifies the relationship of traffic elements of scenes, such as geometrical layout of road scene, providing prior knowledge for the event definition. That is to say that context layout supplies globally spatial. 3) Topology rules[31] intuitively denote the operational logic of traffic flow and the reasonable running rules with a relatively long time accumulation. Bluntly speaking, topology rules generate the spatial-temporally logical clue for event reasoning.

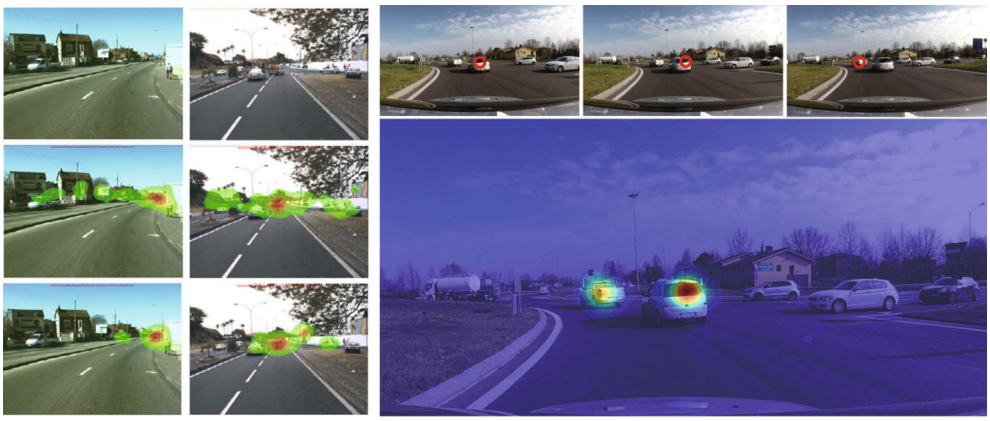
2.1 Traffic saliency for driving

Saliency mechanism, as a critical region extraction and information simplification technology, has been widely used for attractive region selection in images. Over the past few decades, saliency has been generally formulated as bottom-up and top-down modes. Bottom-up modes[32−34] are fast, data-driven, pre-attentive and task-independent. Top-down approaches[35−38] often entail supervised learning with pre-collected task labels by a large set of training examples, and are task-oriented and vary in different environments.

[4] http://robotcar-dataset.robots.ox.ac.uk/
[5] http://vision.middlebury.edu/stereo/

Driving has clear destination and path, and is manifestly the task-driven case. This derives a question: Where should we look when driving in different environments? For seeking the answer, most of the works focus on detecting the obvious traffic sign or light[39−41]. For example, Wang et al.[41] proposed a fast traffic sign detection method based on a cascade method with saliency test and neighboring scale awareness. The saliency was utilized to prune the obtained target window by previous cascade procedure. Kim et al.[42] utilized top-down importance information, such as pedestrian and traffic light detection results, to arouse drivers′ attentions. John et al.[43] generated a saliency map by a convolutional neural network in offline mode, which was used to extract the region of interest for traffic light detection in images. Kuang et al.[44] established a Bayes-based saliency proposal applied to nighttime scenes. They exploited the edge prior, luminance, local contrast and vehicle taillight map to infer the probability of belongingness of a vehicle to a bounding box.

The aforementioned utilizations of saliency mainly focus on one kind of task in driving, i.e., to detect the important traffic sign/light or participants. However, for a practical driving, the tasks always switch frequently, and the saliency assumption may be compromised in many scene conditions[45] stated by Luc Van Gool, the head of the computer vision lab in ETH Zürich. Actually, we need to seek the saliency mechanism in different driving environments. From the view of human vision, a recommendable work contributed by Deng et al.[30] exploited the top-down saliency detection in a general driving environment. They collected the drivers′ attentions data by an eye tracker (Eyelink2000, SR Research Ltd.). They restricted the participants′ head movements by a forehead and chin rest. Then the pupil of the left eye was tracked at a sample rate of 1 000 Hz and a spatial resolution of approximately 0.1°. By their efforts, they found that a driver′s attention mostly concentrates on the end of the road in front of the vehicle, and they treated the road′s vanishing point[6] as the salient regions of interest. When we look back to the data of [30], we find that the collected scenes are with quite few traffic participants and under simple traffic scenes. It is largely insufficient to imitate the saliency mechanism of humans driving. Alletto et al.[46] established a DR(eye)VE video dataset[7], devoting to attention formulation for autonomous and assisted driving. This dataset is composed of more than 500 000 frames, containing drivers′ gaze fixations and their temporal integration which provides task-specific saliency maps. An exemplar of traffic saliency stated in [30] and [46] are shown in Fig. 3, respectively. The DR(eye)VE dataset was firstly used in [47] by Palazzi et al., and they concluded that the gazing frequency and gazing time were different, which correspond to distinct semantic categories. Fig. 4 offers a demonstration.



(a) (b)

Fig. 3 An exemplar of demonstrating traffic saliency. (a) is adapted from [30], which represents an image-based traffic saliency. (b) is adapted from [46], which specifies a video-based traffic saliency.

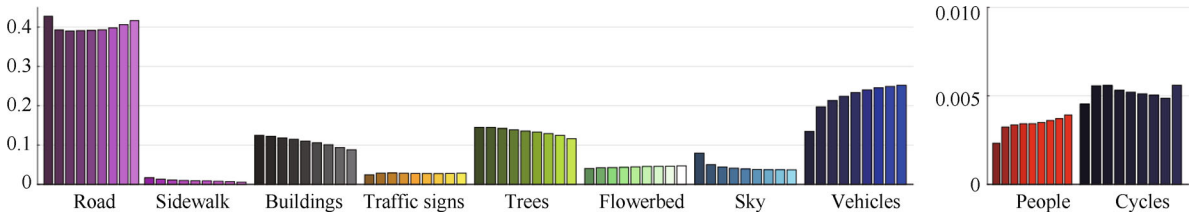


Fig. 4 Proportion of semantic categories hit by gazing maps with an increased value for thresholding. The descending trend means a circumstantial gaze, and increasing trend indicates a focus of gaze. This figure is adapted from [47].

[6] In graphical perspective, a vanishing point is an abstract point on the image plane where 2D projections (or drawings) of a set of parallel lines in 3D space appear to converge. In road plane, vanishing point commonly represents the converging point at the end of the road.

[7] http://imagelab.ing.unimore.it/dreyeve

In fact, traffic saliency can formulate intuitional representation for different traffic scenes, where different situations have distinct landmarks and task demands. Meanwhile, task-driven saliency sometimes may conflict with bottom-up saliency when making a driving decision, such as the road place in scenic areas. Therefore, to pursue a promising traffic saliency representation, we need to explore a mechanism so as to collaborate the task-driven demand and the bottom-up stimulation.

2.2 Context layout for driving

Context layout aims to represent spatially geometrical relationships[48] among different traffic elements with certain semantic label. It is different from the semantic segmentation frameworks[49, 50]. Context layout not only contains the static components of traffic scene (Typical technique for this aspect is simultaneous localization and mapping (SLAM)[51, 52]), such as road, the type of traffic lanes, traffic direction, and participant orientation, but also consists of many kinds of dynamic elements, e.g., motion correlation of participants. The study[8, 53] has given a detailed review on semantic segmentation. Here we only review the literatures which take the traffic geometry inferring into consideration.

In traffic geometry inferring, vanishing point detection is a typical task for autonomous driving. Vanishing point represents the end of the road which provides a guidance for automatic driving. A milestone work[54], proposed by Kong et al., firstly extracted the road region and the road boundaries. Then they took the junction point of the road boundaries as the vanishing point. Later, they found that it was difficult to obtain the road region under unstructured road scene. They creatively analyzed the road texture direction[55] by a Gaussian filter. The intersection location of the extending line of the texture direction was denoted as the road vanishing point. Inspired by that, Shi et al.[56] proposed a fast and robust vanishing point detection method for unstructured road scene. They boosted the robustness of the vanishing points by a temporal tracking.

Only the vanishing point obviously cannot sufficiently represent the context layout of driving scene. More and more studies want to give an overall representation of context layout information. Alvarez et al.[57] attempted to extract the 3D contextual information of roads. They realized it by horizonal lines, vanishing points, 3D scene layout and 3D road geometry, and combined these clues by a Bayesian framework. Casapietra et al.[58] proposed a grid-based road representation, which worked on a road terrain representation and assigned a lane and a driving direction to each patch of road. Seff and Xiao[59] firstly collected a large-scale dataset by gathering one million Google Street View and label the road layout by OpenStreetMap[8]. Then they trained a road layout classification model by deep convolutional networks. Liu et al.[60] represented the street scene with 4-layer interpretation that compactly contains the ground, the participants, the building and the sky. A representative institute for context layout inferring is the Department of Measurement and Control in Karlsruhe Institute of Technology (KIT). They are also the founder of the KITTI dataset. The tracklet of participants, vanishing points, scene layout and traffic lane are combined to construct the directional graphic model and to generate the context layout information by Bayesian inference[61].

[8] www.openstreetmap.org

2.3 Topology rules for driving

Studying the topology rules is a high-level task in scene understanding. Topology studying specifically is the history of a region as indicated by its topography, referring to the definition in the Dictionary of Merriam-Webster. It is mathematically concerned with the properties of space that are preserved under continuous deformations, such as stretching, crumpling and bending, but not tearing or gluing. Accordingly, topology rules for driving are the historically collected road type by traffic state modeling. It serves as an apparent guidance and reminder for safe driving. On one hand, it can be obtained by pre-collected GPS information or digital earth maps. On the other hand, it can be learned by the visual observation of traffic states over a period of time. In this paper, we focus on the latter category.

The difficulty of learning the topology rules is to rule out the noisy observations and overcome the on-board camera motion when driving. In this point, the spatial 3D scene layout and tracklets are commonly adopted. For example, Ess et al.[62] aimed to obtain the road type (straight, left/right curve, crossing, etc.), as well as the simultaneously encountered participants. They firstly represented the road scene with a meta feature representation, specifying 12 components, such as pedestrian, vehicle, zebra line, etc. Then they fed these features into an Adaboost classifier to classify 13 kinds of road types. This method was a single image based approach, and the temporal correlation of scene was not considered efficiently. Later, Geiger et al.[31] proposed a principled generative model to estimate the varying road topology of intersections as well as the 3D content of the scene. They fused the dynamic 3D scene flow[63] with static occupancy grids[64] to construct the interring feature. From the perspective of birds′ eyes, they modeled a directional graphic model and inferred it by a Bayesian estimation. Inspired by this work, Zhang et al.[65] aimed to design a finite traffic pattern with strong ability of topology representation and focused on a high-order dependence modeling for participants. By that, they can detect the illegal traffic situation. Specifically, they adopted the tracklets instead of 3D scene flow, so as to provide a convenience for dependence modeling of participants. This work was further extended in [61]. Fig. 5 shows the utilized topology rules in the related works. Motivated by this kind of topology rules, Chen et al.[66] proposed a direct perception approach, which was modeled by directly perceiving the angle of the car against the road, the distance to the lane markings and the distance

to cars in the current and adjacent lanes. These clues were learned with a ConvNet to construct an affordance map.

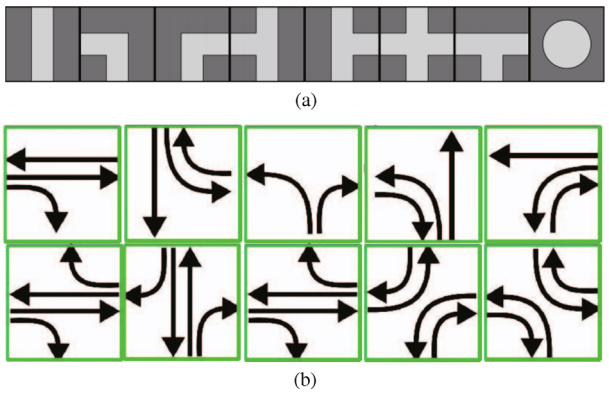


Fig. 5 The utilized topology rules in (a)[61] and (b)[62]

2.4 Discussion

Traffic saliency, context layout and topology rules provide a scene representation with different time spans, space spans and interpretational levels. They, to some extent, set a constraint for autonomous driving and assist in judging the occurrence of driving event. In fact, the scene representation has large impact on driving. We all know that the experience of drivers plays an important role in completing current driving task in different environments. Based on the study[67], drivers with richer driving experience can handle sudden and accidental events. The reasons are two-fold: 1) Experienced drivers have better driving consciousness; 2) They are familiar with the scene encountered. In other words, the experienced drivers stored the scene representation and transferred it for current decisions. Therefore, for autonomous driving, how to model the scene representation better and how to transfer the modeled representations to current scene understanding are two problems to be solved. Geiger et al.[31, 61] have made some attempts. However, it is largely insufficient for real driving situations. In addition, transfer learning[68] recently attracted the attention in many computer vision tasks. For the scene representation, structure information should be focused when using transfer learning because of the representation type of context layout and topology rules. Therefore, tree structure based transfer learning[69] may be a good choice.

In the following, we will closely present the event detection frameworks in autonomous driving.

3 Scene event detection in autonomous driving

As said before, event definition is a difficult problem in cognition science because of the dynamic and unpredictable behaviors of different participants around the ego-vehicle. After doing a literature review, the existing event detection methods in driving consist of two categories: ego-vehicle event and scene event. Ego-vehicle event detection[70, 71] mainly collects records for its maneuvers, such as braking, accelerating, turning, etc. Scene event aims to perceive the actions or behaviors of other participants around the ego-vehicle. In this paper, we concentrate on the second category, i.e., scene event. We will elaborate it from vehicle event and pedestrian event detection. That is because pedestrian and vehicle manifestly have different event type in the road and demonstrate entirely different behaviors.

3.1 Vehicle event detection

Lane change[72, 73], overtaking[74, 75] and rear-ending are three mostly focused events in previous studies and the highway scenarios are the concentration.

Among them, a German study reported a turn signal usage of 55% for lane changes on urban roads and 75% on highways[76]. Similar results were obtained in a second observational study including almost 400 000 vehicles with a turn signal usage of 71% on German highways[77]. For lane change detection, Kasper et al.[78] defined 27 kinds of maneuvers with a reasoning by an object-oriented Bayesian networks. They exploited a lane-related coordinate system together with individual occupancy schedule grids of all vehicles and assigned all the vehicles coordinates within this system. The coordinate system can efficiently model the vehicle-lane and vehicle-vehicle relations. Yao et al.[79] proposed an automatic method for lane-change trajectory segmentation and constructed a trajectory dataset with more than 1 000 samples from historical driving. They then modeled the lane-change event by inferring the cross-correlating spatial-temporal features of ego-vehicle w.r.t. other vehicles on the trajectory′s end location and speed.

Additionally, Gindele et al.[80] decomposed the overtaking event into five steps: following, acceleration phase, overtake, sheer out and free ride and constructed a dynamic Bayesian network to combine the longitudinal distance (the distance of a vehicle to the next vehicle ahead), lateral distance (measurement of the displacement between the position of a car and the centerline) and relative velocity (difference between their velocities of vehicles). Sivaraman et al.[81] learned the trajectories of other vehicles around the ego one. They firstly tracked the surrounding vehicles by Kalman filter and then learned the primary trajectories by a Gaussian mixture model zone selection and hidden Markov process after a tracklet clustering. Satzoda and Trivedi[82] proposed an appearance-based method for detecting both overtaking and receding vehicles with respect to ego-vehicle. They firstly detected the vehicles with Adaboost cascade classifier and then tracked them to find the overtaking and receding behavior. Specifically, this work selected the left and right boundary regions to conduct a detection.

The above works are all based on the view in front of the ego-vehicle. Recently, some works aim to achieve a more comprehensive understanding of surrounding vehicles′ behaviors by taking the panorama vision of ego-vehicle into account. For example, Kritoffersen et al.[83] equipped with six on-board cameras and conducted detection and tracking

on surrounding vehicles in a panorama vision. Then, the tracked trajectories were projected into a ground plane for a more intuitional inferring of relationships between them. This work defined 14 kinds of events, such as overtaking and cut-ins. They extended this work in [84] by making a concise definition of event classes, i.e., five classes.

In addition to the highway scenarios, urban scene event recently has gradually caught attentions. Different from the highway scenarios, urban event mostly concentrates on the intersection scene. That is because intersection has the most complex traffic situation, involving a mixture of participants, such as pedestrians, vehicles, cyclists, motorbikes, etc. The studies in this domain adopt the public datasets, such as KITTI, cityscapes, as evaluation benchmarks. For instance, Khosroshahi et al.[85] built a 3-layer long short-term memory (LSTM) to learn the temporal representation of the motion of surrounding vehicles and utilized the LSTM architecture to classify 12 kinds of events based on different driving directions and road directions. Ernst et al.[86] exploited the velocity and acceleration of surrounding vehicles by a Ladar sensor. Then they incorporated Markov model and fuzzy logic to infer the state of examined vehicles. They defined six kinds of scene events, i.e., accelerating, decelerating, straight driving, scram, lateral move, start.

3.2 Pedestrian event detection

Compared with the vehicle event detection, pedestrian event detection is more complicated. Pedestrian has higher mobility, more uncertain movement, different gender and ages and commonly cooperates with each other. Therefore, pedestrian event owns diverse forms under different conditions. Actually, the detection of pedestrian event mainly focuses on unsafe behaviors or hazards, such as sudden crossing, dart action and overtaking. These behaviors are big threats for ego-vehicle. By the latest reports of traffic fatalities[9] released by National Highway Traffic Safety Administration (NHTSA), about 90.4% pedestrian fatalities occurred when the pedestrian steps in front of the ego-vehicle and with a crossing behavior.

For reasoning the pedestrian event, the trajectory and moving state (velocity, direction, orientation, etc.) are commonly used clues. For example, the work of [87] took the pedestrian trajectory to map the traffic patterns and constructed a scene graph by clustering the trajectories and involved a temporal move of the graph to judge the pedestrian behavior. Hariyono and Jo[88] recognized the pedestrian crossing event by checking the pedestrian pose, lateral speed, motion direction and spatial layout of the environment. Mueid et al.[89] explored optical flow and histogram of gradient of pedestrian to classify eight kinds of events, including walking, running, turn/return, tumbling, etc. Quintero et al.[90] divided the trajectory of pedestrians into walking, stopping, starting and standing behaviors and classified them by a balanced Gaussian process dynamical models. Ogawa et al.[91] aimed to detect the sudden appearance change of pedestrians before consecutive moving. They computed the appearance change by Kullback-Leibler divergence (KLD) between temporal frames and then detected the non-periodic sequence by a smoothing method which was conducted by summing up the previous KLD values of previous frames. Chan et al.[92] directly modeled the interaction of various participants with a recurrent neural network (RNN) and detected collision. The work of [93] addressed the occlusion problem of participants by using RGB-D data. They segmented the participants into different layers and determined the state of overtaking by the variation of bounding boxes of targets. In addition to common dangerous behavior of pedestrian, the work[94] recognized the undertaking evasive actions based on permutation entropy which can identify the deviations from the normal free walking.

3.3 Discussion

Based on the above literature review, we find that the studies mainly focused on the events which are unsafe and can be well-defined in autonomous driving. Of course, these events manifestly cannot include all of the scene events that may be encountered. In the meantime, existing works aim to tackle one kind of participants in scene event detection. In fact, the road scene involves different participants at the same time. For example, in driving, road-crossing has direct influence on the behavior of vehicles to give the way to pedestrians. Therefore, an accident always involves different participants. In addition, existing works handle the same events under different names and definitions in their works, while have the same name for different kinds of events. It is urgent to construct a unifying framework for event type definition and naming. Besides, the datasets in each work are not publicly available, which cannot fairly reflect the performance of the methods.

4 Intention prediction in autonomous driving

In autonomous driving, the purpose of the aforementioned scene representation and scene event detection is to provide an accurate reasoning clue for future driving, i.e., to make a precise movement prediction of surrounding participants in the future, so that the autonomous vehicle can smoothly pass the observed scene and reach the destination as fast as possible. Actually, for future movement prediction, it has a more exact description: intention prediction. One aspect of it is to predict the future movement of participants as long as possible, so as to spare enough time for comfortable maneuver. Another aspect is to discover the moving

[9] https://crashstats.nhtsa.dot.gov/Api/ Public/ViewPublication/811888

patterns of participants based on historical observation. According to different time spans, we classify the intention prediction as short-term intention prediction and long-term prediction.

4.1 Short-term intention prediction

Short-term intention prediction researches mainly take the demonstrated transient movement of participants to estimate the behavior in next milliseconds or seconds. With effective detection, the optical flow, contour variation, gait of pedestrians are commonly optional to represent the appearance of transient movement, followed by efficient classification and prolongation of movement to predict short-term future. For short-term intention prediction, the intention categories of pedestrian, cyclists[95] and vehicles are different. Crossing, starting, stopping, running are the pedestrian′s common manifestation patterns, while lane-change, overtaking, starting, stopping, turning and approaching are the intentions of vehicles. Fig. 6 demonstrates some typical vehicle intentions and pedestrian crossing intention in urban scenes.

To predict the intention of pedestrians, common practice is to collect the appearance variation, velocity variation, orientation variation to represent the transient movement and feeds the features into a reasoning framework to estimate future movement. For example, Fugger et al.[98] novelly took the average velocity, average acceleration of four steps of human into account to estimate the intention. Schneider and Gavrila[99] defined four kinds of moving patterns, including crossing, stopping, turning and re-starting and estimated the short-term walking, stopping and turning intention by a recursive Bayesian filters. Goldhammer et al.[100, 101] predicted the displacement between the starting and heeling off the ground. Incorporating with the acceleration, timestamp range of accelerating and average velocity, they achieved a trajectory prediction in next three seconds. Keller and Gavrila[102] adopted the optical flow feature to represent the transient state of pedestrian and linked the states with a state trajectory which was prolonged to make an estimation. Kohler et al.[103] exploited the contour variation to express the movement by a motion contour image based on HOG-like descriptor. Then the stopping and walking intention was predicted by a support vector machine. Kooij et al.[104] explored the movement rules of head by a dynamic Bayesian networks. Quintero et al.[90] predicted the walking intention in next one second by estimating the movement of the parts of body with the balanced Gaussian process dynamical models. In order to predict the intention of pedestrians in nighttime, the work[105] captured the scene with an infrared camera and employed the dynamic

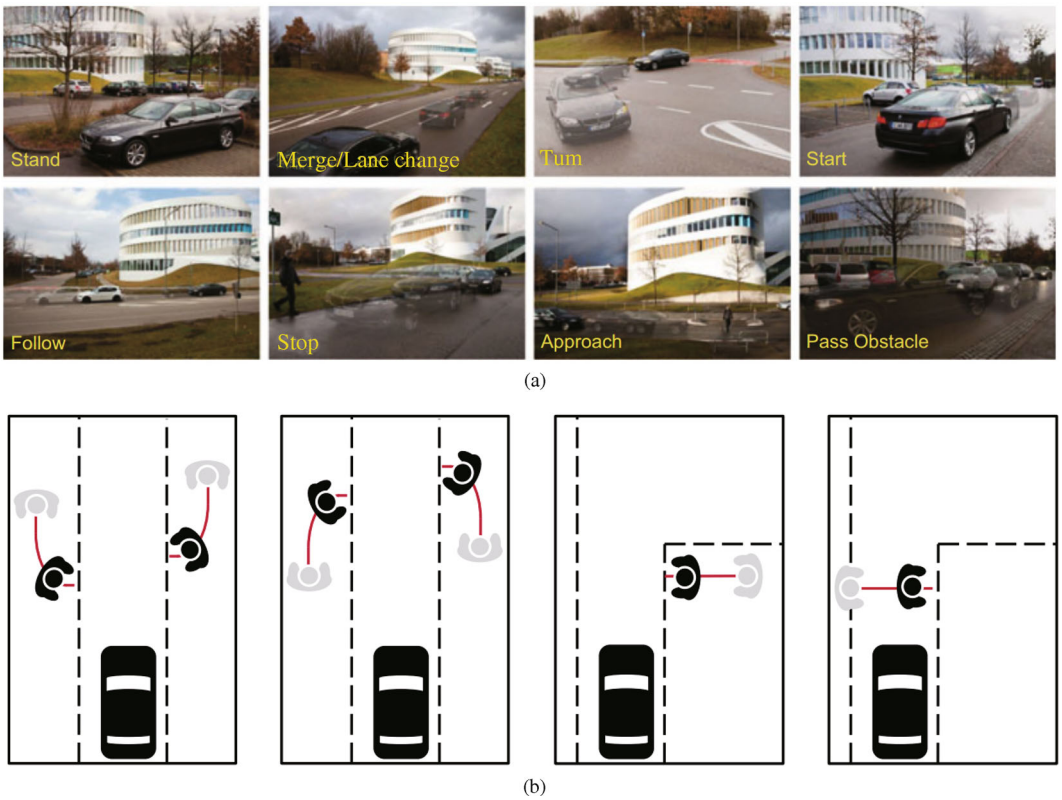


Fig. 6 Some typical (a) vehicle intentions and (b) pedestrian crossing intentions in urban scene. These figures are adapted from [96] and [97], respectively.

fuzzy automata (DFA) to overcome prediction uncertainty. In their work, four kinds of intentions were focused, including, standing on sidewalk, walking along the sidewalk, walking-crossing and running-crossing.

For vehicle intention prediction, the instantaneous velocity, acceleration, turning move are commonly used clues[106, 107]. For example, the vehicle posture state and velocity of front vehicle were exploited by the work of [108], with a hidden Markov model for reasoning the state transition. In this work, the starting, stopping and lane change intentions were focused. Hou et al.[109] predicted the afflux intention of vehicles by incorporating the Bayesian classifier with a decision trees. In this approach, the velocity difference between the ego-vehicle and other surrounding vehicles, and the distance between them were employed to represent the transient information. A graphical modeling and unsupervised learning techniques are used to construct a model for inferring the intention of drivers in surrounding vehicles[110]. This work focused on the urban scene and synthesized a hidden Markov model with the aid of unscented Kalman filtering. The expectation maximization technique is employed to predict the lane change intention of vehicles.

4.2 Long-term intention prediction

Different from the short-term intention prediction, long-term intention prediction can be exactly understood as path planning. In other words, long-term intention prediction takes the historical observed trajectory with incorporated context information so as to give a possible driving planning.

In this category, Bayesian filter is a customary reasoning framework. Typical ones are Kalman filter (KF), extended Kalman filter (EKF), multiple model filter (MMF) and particle filter (PF). For instance, Gu et al.[111] took the historical trajectory of pedestrian as a reference and used PF to estimate the future state. The estimated moving state was further adopted to compute the passable possibility for ego-vehicle. Additionally, the first-order EKF was taken into the trajectory estimation of surrounding vehicles, with the help of time to collision (TTC) and minimum distance between vehicles[106]. The work[112] employed KF to predict the trajectory of other vehicles and plan the path for ego-vehicle with linear-quadratic Gaussian.

In addition to the Bayesian filters, the context layout and topology information aforementioned are effectively used in this domain. Specifically, the transformation of coordinate systems beyond the global coordinates is popular. For example, Cartesian coordinates[113, 114] treated the ego-vehicle as the origin and can more intuitionally compute the relative distance, velocity and trajectory of other vehicles. Curvilinear coordinates[115] can transform the driving path into an orthogonal space and generate a more convenient calculation. Under the coordinate systems, some works projected the observed scene into topology space. For example, Gu et al.[113] posed the trajectory of other vehicles into different topology space, such as overtaking and car-following. Then, they projected the observed trajectory into different topology space and make a prediction of future moving. Pool et al.[116] exploited the local road topology to obtain better predictive distributions for cyclists, where the tracklets of the cyclists were extracted and spatially aligned to the road curves and crossings. Then the KF was used to make a prediction. Evestedt et al.[117] used the intelligent driver model (IDM) to estimate the trajectory of vehicles in T-type intersection. Curvilinear coordinates was employed by Jo et al.[115] for a path planning in a curved road. Fig. 7 gives a geometry demonstration on different coordinate systems.

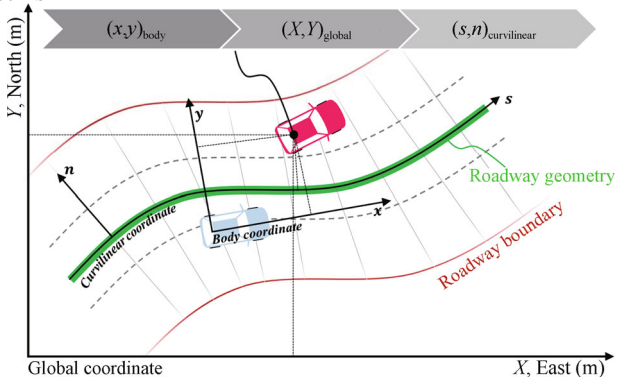


Fig. 7 A geometry demonstration on different coordinate systems (This figure is adapted from [115])

4.3 Discussion

Intention prediction aims to provide an acceptable path and information for safe and comfortable future autonomous driving. Existing works on intention prediction either focus on the locally short-term prediction or globally long-term prediction. The acceptable time span for driving in local mode is limited, while global mode loses the local intention information. It is promising to combine local mode with global mode. In the meanwhile, the local mode can be provided by the scene context layout with adequate geometrical information and topology rules can provide a clue for global planing. However, the information was not applied to the previous works. In addition, as similar to the scene event detection, there is no unified metric and dataset to evaluate performance of intention prediction.

5 End-to-end reasoning

The aforementioned formulations for event reasoning undergo multiple stages and their ultimate goal is to realize fully-automatic driving. With the aid of a large amount of driving data acquisition, some studies focus on the direct reasoning by deep learning architectures[13, 14, 118]. These works mainly focus on constructing data-driven driving model for automatic driving planning and steering wheel control.

Actually, these formulations bypass the complicated modeling for the aforementioned reasoning frameworks. For

example, Xu et al.[13] proposed novel FCN-LSTM architecture to predict a distribution over future vehicle ego motion from instantaneous monocular camera observations, and to predict previous vehicle state from large scale crowd-sourced video data. They learn the countermeasure from 10 000 hours of driving dash-cam video streams at different places around the world, which consists of much human driving experience. Eraqi et al.[118] proposed a convolutional long short-term memory recurrent neural network (C-LSTM) that earned both visual and dynamic temporal dependencies of driving, which was further introduced to solving the steering angle regression problem for steering decision. Caltagirone et al.[119] carried out the perception and path generation from real-world driving sequences simultaneously by developing a method to generate driving paths that integrates Lidar point clouds, GPS-IMU information and Google driving together through fully convolutional neural network directions.

By a large amount of data, end-to-end architectures give a direct driving strategies. They comprehend the traffic scene automatically by learning the patterns contained in the large-scale data. These formulations rely on the quality of training data and labeling labors. Certainly, we can see the progress on the scene understanding by these end-to-end architectures. However, the process of understanding is difficult to interpret. For example, the true reason of the action and the underlying reason for the progress is still unverifiable, whereas it has been verified at the aforementioned stages. In addition, the current end-to-end architectures are not good enough for the deep traffic scene understanding. The situations in a certain traffic scene vary in many ways because of the dynamic participants.

6 Evaluation datasets and metrics

To validate the performance of every method for each task aforementioned, it is indispensable to compare the performance of different approaches. Dataset and evaluation metrics are the two most important elements to be firstly examined. In the following part, we will present the dataset and corresponding evaluation metrics of each task in autonomous driving.

6.1 Evaluation on scene representation

As mentioned above, we introduced traffic saliency, context layout and topology rules for scene representation. The corresponding evaluation setups are presented below.

6.1.1 Evaluating traffic saliency

Dataset. Traffic saliency attracts the attention of researchers in recent years. Actually, it is difficult to acquire the ground-truth of traffic saliency in general driving situations. Based on the investigation, we found that DR(eye)VE video dataset[10] may be useful for traffic saliency evaluation. That is because the car-mounted view, drivers′ point of view, gaze map overlay and geo-reference were captured when constructing the dataset. Besides, the videos were recorded in different places, including downtown, countryside and highway scenarios and covered a broad range of traffic conditions from free traffic to crowded situations. Fig. 8 demonstrates an exemplar frame of DR(eye)VE video dataset. This dataset was released recently and adopted in [47].

Fig. 8 An exemplar frame from DR(eye)VE video dataset. From left to right, from up to bottom: car-mounted view, driver′s point of view, gaze map overlay and geo-referenced course (This figure is adapted from [46])

[10] http://imagelab.ing.unimore.it/dreyeve

Metrics. For the performance evaluation of traffic saliency, the metrics is the same as the saliency detection domain for general images. Commonly, there are eight metrics usually adopted[120] (6 similarity metrics and 2 dissimilarity metrics). They are area under receiver operating characteristic (ROC) curve (AUC), shuffled AUC (sAUC), normalized scanpath daliency (NSS), Pearson′s correlation coefficient (CC), earth mover′s distance (EMD), similarity or histogram intersection (SIM), Kullback-Leibler divergence (KL) and information gain (IG).

AUC is widely used for saliency map evaluating. The saliency map is treated as a binary classifier of fixations at different thresholds, where the ROC curve is computed by measuring the true positive rates and false positive rates under the binary classifier.

SAUC aims to overcome the center bias in AUC computation and it compensates for the central fixation bias[121].

Normalized scanpath saliency (NSS) was introduced as a simple correspondence measure between saliency maps and ground truth, and computed as the average normalized saliency at fixated locations[122]. The calculating procedure is

$$NSS(A,B^b) = \frac{1}{N}\sum_i \bar{A}_i \times B_i^b \tag{1}$$

where A is the saliency map, B^b denotes a binary map of fixation locations, $N = \sum_i B_i^b$, $\bar{A} = \frac{A-\mu(A)}{\sigma(A)}$, i indexes the i-th pixel and N represents the number of fixated pixels.

IG was recently introduced by Kümmerer et al.[123] which measured the information gain of a saliency map. Given a saliency map A, the binary map of fixation locations B^b and baseline C, the information gain is calculated by

$$IG(A,B^b) = \frac{1}{N}\sum_i B_i^b[\log_2(\varepsilon + A_i) - \log_2(\varepsilon + C_i)] \tag{2}$$

where ε is a very small constant for avoiding $-\infty$.

Similarity or histogram intersection (SIM) measures the similarity between two distributions, viewed as histograms[124]. Given a saliency map A and a continuous fixation map B^c, the SIM is computed as

$$SIM(A,B^c) = \sum_i \min(A_i, B_i^c) \tag{3}$$

where $\sum_i A_i = \sum_i B_i^c = 1$.

CC treats the saliency and fixation maps, A and B^c, as random variables and measures the linear relationship between them[125]. CC is computed as

$$CC(A,B^c) = \frac{\sigma(A,B^c)}{\sigma(A) \times \sigma(B^c)} \tag{4}$$

where $\sigma(A,B^c)$ denotes the covariance of A and B^c.

Kullback-Leibler divergence (KL) evaluates the loss of information when distribution saliency map A approximates the distribution of B^c and is calculated by

$$KL(A,B^c) = \sum_i B_i^c \log\left(\varepsilon + \frac{B_i^c}{\varepsilon + A_i}\right). \tag{5}$$

EMD measures the spatial distance between the saliency map A and B^c over a region by computing the minimum cost of morphing from one distribution into the other. The EMD has many variants. Detailed content can be seen from [126, 127].

All of these metrics aim to evaluate the similarity between human fixation and obtained saliency maps. Choosing appropriate evaluation metrics remains as an open research question because this choice depends on how saliency and fixation data are defined and represented. The inherent ambiguity leads to different choices for evaluation.

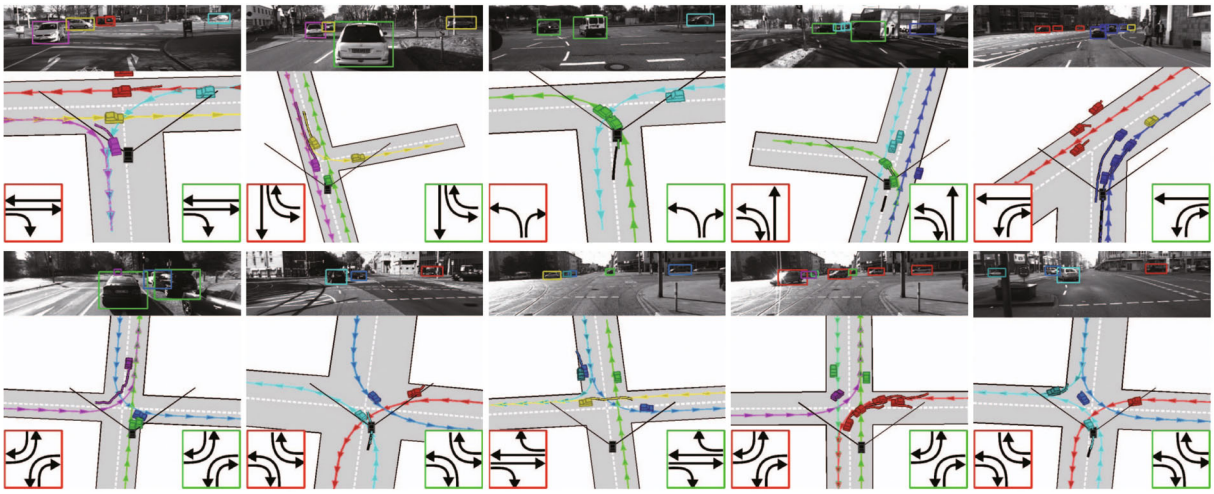


Fig. 9 The inferring results of topology rules by the tracklets. The left-bottom topology rules bounded by red box are the inferred results and the green box in the right-bottom is the ground-truth (This figure is adapted from [65]).

6.1.2 Evaluating context layout

Dataset. To evaluate the performance of context layout learning for autonomous driving, the frequently selected dataset is KITTI which provides color, gray-scale images, 3D laser scanning and geo-inference. The context layout information is represented by geometry clues, such as the intersection center, road region and orientation of street or road. In addition to KITTI, most related works collected the datasets by themselves and did not open them. Therefore, for context layout evaluation, KITTI is the only publicly available one.

Metrics. As for performance comparison, the absolute error between the truly road orientation or intersection location and the predicted ones are commonly adopted to compare the line or point based clues[61]. In terms of the road region, overlapping rate is usually used, which is computed as[61]:

$$Overlap = \frac{B_T \cap B_P}{B_T \cup B_P} \times 100\% \tag{6}$$

where B_T and B_P are the binary road mask of ground-truth and predicted one, respectively. $\bigcap$ and $\bigcup$ are the intersection and union operator, respectively. Actually, the metrics here are inspired by the result representation structure, such as point, line, or plane forms for representing layout geometry.

6.1.3 Evaluating topology rules

Dataset. In terms of measuring topology rules, the publicly available dataset is also KITTI released by Zhang et al.[65]. They labeled the topology rules as many traffic patterns, including turning left, straight driving, etc. Fig. 9 demonstrates the traffic patterns labeled in [65].

Metrics. For evaluating the performance of topology learning, confusion matrix is a specific table layout that allows visualization of the performance of an algorithm evaluation, which is commonly utilized in supervised learning. Each row of the confusion matrix represents the instances in a predicted class, while each column represents the instances in an actual class and vice versa. Corresponding to the topology rule, the "class" is the pattern. Within confusion matrix, researchers can derive many metrics for evaluation, such as precision, recall, true positive rate (TPR), false positive rate (FPR), etc. As for the confusion matrix, readers can refer to [128] and the wikipage website[11].

6.2 Evaluation on event detection

Dataset. To evaluate the event detection in autonomous driving, there is only one publicly available dataset for pedestrian crossing event detection called Daimler dataset[99], which was collected by Schneider and Gavrila. This dataset contains 68 pedestrian sequences collected from a stationary and moving vehicles. Four different pedestrian motion types are considered, including crossing, stopping, starting to walk and bending-in. There is only one pedestrian without occlusion in each sequence. Fig. 10 demonstrates two frames in Daimler dataset.

Fig. 10 Two examples in Daimler dataset for pedestrian crossing detection (This figure is adapted from Daimler dataset[65]).

Metrics. As for the evaluating metrics, owning to that different works detect the event at distinct levels, the metrics are different in all of the works. For example, the work in [80, 84] models the event with many trajectories and computes the detection error by calculating the average root mean square error (RMSE), which is computed by

$$RMSE = \frac{1}{n}\sqrt{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \tag{7}$$

where $\hat{y}$ and y are the attribute value of the estimated one and ground-truth, respectively. The attribute value may include velocity, position and gear angle of vehicles. Additionally, some works defined such kind of event, like overtaking, crossing, etc. Therefore, similar to the evaluating methods for topology categories mentioned above, confusion matrix is also utilized in this domain.

6.3 Evaluation on intention prediction

In terms of the evaluation of intention prediction, there is no publicly available dataset till now. Therefore, in this subsection, we only review the evaluation metrics of this task. For intention prediction, in addition to the same evaluating metrics as the event detection for the prediction of certain intention class, such as crossing, overtaking, etc., TTC is another important index for safety evaluation, which is used to examine whether an interaction with a certain observing vehicle could be predicted by the ego one. In the work of [79], they compute the TTC as

$$TTC_0^{OV,EGO} = \frac{\left(s_0^{OV} - s_0^{EGO}\right)}{\left(\dot{s}_0^{OV} - \dot{s}_0^{EGO}\right)} \tag{8}$$

[11] https://en.wikipedia.org/wiki/Confusion-matrix

where s_0 and $\dot{s}_0$ denote the location and speed of the vehicles in initial time. It is assumed that we have obtained the location s_0^{OV} of the observed vehicles (OV) and we know the location s_0^{EGO} of ego vehicle (EGO). Meanwhile, the speed of OV is estimated as $\dot{s}_0^{OV}$. We can compute TTC by (8).

6.4 Discussion

The publicly available benchmarks for autonomous driving, such as KITTI and Cityscapes, concentrate on the low-level vision tasks. There is no large-scale publicly available benchmark for deeper scene understanding. Actually, it is just like there is no strong power to propel deeper understanding. To construct this kind of benchmark with an acceptance and deeper understanding of scene, we should clarify two problems: 1) What should be annotated? 2) How to annotate?

What should be annotated? Low-level vision tasks which clear annotation goal, such as annotating pedestrian, vehicles, cyclists, road, etc. Differently, in deeper understanding of scene, the dynamics, logicality, causality and other high-level semantic relationship of participants in the traffic scene may be the focuses. Of course, these high-level relationships should be built on certain participants. Therefore, the construction of benchmark should have a multi-level annotation, including low-level participants annotation, mid-level trajectory annotation and high-level relationship annotation. As for low-level participant annotation, what kind of information should be contained: color, distance to ego vehicle, size, or semantic class? For deeper understanding of scene, it is necessary to contain attributes of participants as much as possible.

How to annotate? Based on the exhaustive efforts on the low-level vision tasks and mid-level trajectory annotation, low-level and mid-level annotation can be achieved with the aid of the state-of-the art automatic methods. For inevitably generated errors, it can be corrected by humans. In other words, hybrid human-machine annotation may be the mode for low-level participant annotation. For multiple attribute annotation, the collaboration of multiple sensors might be realized by the road, such as KITTI[19]. However, calibration of different sensors should be taken into account, but it is different from KITTI which has only six video sequences. Much larger data should be collected and labeled in various driving environments. The high-level relationship annotation is the most important component for deeper scene understanding. Thus, we think event and intention are two stages and manual annotation may be the only way to obtain an accurate labeling. These insights are all need further study.

7 Concluding remarks

In this paper, we comprehensively reviewed the works on the scene understanding under an event reasoning framework. Specifically, we presented the contents from three stages: scene representation, scene event detection and intention prediction. At each stage, we firstly gave a taxonomy to classify each approach and then conclusively reviewed relevant literatures as well as the state-of-the-arts methods. Besides, we discussed open problems at each stage of scene event reasoning and tried our best to provide a possible solution. We hope that this survey can encourage new research and insight for this field and provide basic knowledge for beginners.

References

[1] M. M. Waldrop. Autonomous vehicles: No drivers required. *Nature*, vol. 518, no. 7537, pp. 20–23, 2015. DOI: 10.1038/518020a.

[2] J. Mervis. Are We Going Too Fast on Driverless Cars? http://www.sciencemag.org/news/2017/12/are-we-going-too-fast-driverless-cars, December 14, 2017.

[3] Y. Y. Zheng, J. Yao. Multi-angle face detection based on DP-adaboost. *International Journal of Automation and Computing*, vol. 12, no. 4, pp. 421–431, 2015. DOI: 10.1007/s11633-014-0872-8.

[4] H. G. Ren, W. M. Liu, T. Shi, F. J. Li. Compressive tracking based on online Hough forest. *International Journal of Automation and Computing*, vol. 14, no. 4, pp. 396–406, 2017. DOI: 10.1007/s11633-017-1083-x.

[5] J. W. Fang, H. K. Xu, Q. Wang, T. J. Wu. Online hash tracking with spatio-temporal saliency auxiliary. *Computer Vision and Image Understanding*, vol. 160, pp. 57–72, 2017. DOI: 10.1016/j.cviu.2017.03.006.

[6] S. Arumugadevi, V. Seenivasagam. Color image segmentation using feedforward neural networks with FCM. *International Journal of Automation and Computing*, vol. 13, no. 5, pp. 491–500, 2016. DOI: 10.1007/s11633-016-0975-5.

[7] J. F. Bonnefon, A. Shariff, I. Rahwan. The social dilemma of autonomous vehicles. *Science*, vol. 352, no. 6293, pp. 1573–1576, 2016. DOI: 10.1126/science.aaf2654.

[8] J. Janai, F. Güney, A. Behl, A. Geiger. Computer vision for autonomous vehicles: Problems, datasets and state-of-the-art. arXiv:1704.05519, 2017.

[9] J. R. Xue, D. Wang, S. Y. Du, D. X. Cui, Y. Huang, N. N. Zheng. A vision-centered multi-sensor fusing approach to self-localization and obstacle perception for robotic cars. *Frontiers of Information Technology & Electronic Engineering*, vol. 18, no. 1, pp. 122–138, 2017. DOI: 10.1631/FITEE.1601873.

[10] H. Zhu, K. V. Yuen, L. Mihaylova, H. Leung. Overview of environment perception for intelligent vehicles. *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, no. 10, pp. 2584–2601, 2017. DOI: 10.1109/TITS.2017.2658662.

[11] D. L. Waltz. Understanding scene descriptions as event simulations. In *Proceedings of the 18th Annual Meeting on Association for Computational Linguistics*, Association for Computational Linguistics, Philadelphia, USA, pp. 7–11, 1980. DOI: 10.3115/981436.981439.

[12] Y. Q. Hou, S. Hornauer, K. Zipser. Fast recurrent fully convolutional networks for direct perception in autonomous driving. arXiv:1711.06459, 2017.

[13] H. Z. Xu, Y. Gao, F. Yu, T. Darrell. End-to-end learning of driving models from large-scale video datasets. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, Honolulu, USA, pp. 3530–3538, 2017. DOI: 10.1109/CVPR.2017.376.

[14] T. Fernando, S. Denman, S. Sridharan, C. Fookes. Going deeper: Autonomous steering with neural memory networks. In *Proceedings of IEEE International Conference on Computer Vision Workshop*, IEEE, Venice, Italy, pp. 214–221, 2017. DOI: 10.1109/ICCVW.2017.34.

[15] C. Thorpe, M. H. Hebert, T. Kanade, S. A. Shafer. Vision and navigation for the Carnegie-Mellon Navlab. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 10, no. 3, pp. 362–372, 1988. DOI: 10.1109/34.3900.

[16] M. Buehler, K. Iagnemma, S. Singh. *The DARPA Urban Challenge: Autonomous Vehicles in City Traffic*, Berlin, Heidelberg, Germany: Springer, 2009. DOI: 10.1007/978-3-642-03991-1.

[17] J. Hooper. From DARPA Grand Challenge 2004DARPA′s Debacle in The Desert. https://www.popsci.com/scitech/article/2004-06/darpa-grand-challenge-2004darpas-debacle-desert, June 4, 2004.

[18] S. Thrun, M. Montemerlo, H. Dahlkamp, D. Stavens, A. Aron, J. Diebel, P. Fong, J. Gale, M. Halpenny, G. Hoffmann, K. Lau, C. Oakley, M. Palatucci, V. Pratt, P. Stang, S. Strohband, C. Dupont, L. E. Jendrossek, C. Koelen, C. Markey, C. Rummel, J. Van Niekerk, E. Jensen, P. Alessandrini, G. Bradski, B. Davies, S. Ettinger, A. Kaehle, A. Nefian, P. Mahoney. Stanley: The robot that won the DARPA grand challenge. *The 2005 DARPA Grand Challenge*, M. Buehler, K. Iagnemma, S. Singh, Eds., Berlin, Heidelberg, Germany: Springer, 2007. DOI: 10.1007/978-3-540-73429-1_1.

[19] A. Geiger, P. Lenz, R. Urtasun. Are we ready for autonomous driving? The KITTI vision benchmark suite. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, Providence, USA, pp. 3354–3361, 2012. DOI: 10.1109/CVPR.2012.6248074.

[20] G. J. Brostow, J. Fauqueur, R. Cipolla. Semantic object classes in video: A high-definition ground truth database. *Pattern Recognition Letters*, vol. 30, no. 2, pp. 88–97, 2009. DOI: 10.1016/j.patrec.2008.04.005.

[21] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, B. Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, Las Vegas, USA, pp. 3213–3223, 2016. DOI: 10.1109/CVPR.2016.350.

[22] A. Gaidon, Q. Wang, Y. Cabon, E. Vig. Virtual worlds as proxy for multi-object tracking analysis. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4340–4349, 2016. DOI: 10.1109/CVPR.2016.470.

[23] W. Maddern, G. Pascoe, C. Linegar, P. Newman. 1 year, 1000 km: The Oxford RobotCar dataset. *International Journal of Robotics Research*, vol. 36, no. 1, pp. 3–15, 2017. DOI: 10.1177/0278364916679498.

[24] J. V. Dueholm, M. S. Kristoffersen, R. K. Satzoda, E. Ohn-Bar, T. B. Moeslund, M. M. Trivedi. Multi-perspective vehicle detection and tracking: Challenges, dataset, and metrics. In *Proceedings of the 19th International Conference on Intelligent Transportation Systems*, IEEE, Rio de Janeiro, Brazil, pp. 959–964, 2016. DOI: 10.1109/ITSC.2016.7795671.

[25] C. Wang, Y. K. Fang, H. J. Zhao, C. Z. Guo, S. Mita, H. B. Zha. Probabilistic inference for occluded and Multiview on-road vehicle detection. *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 1, pp. 215–229, 2015. DOI: 10.1109/TITS.2015.2466109.

[26] D. Hoiem, S. K. Divvala, J. H. Hays. Pascal VOC 2008 challenge. *World Literature Today*, 2009.

[27] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. H. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, F. F. Li. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015. DOI: 10.1007/s11263-015-0816-y.

[28] A. Milan, L. Leal-Taixé, I. Reid, S. Roth, K. Schindler. MOT16: A benchmark for multi-object tracking. arXiv:1603.00831, 2016.

[29] F. C. Heilbron, V. Escorcia, B. Ghanem, J. C. Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, Boston, USA, pp. 961–970, 2015. DOI: 10.1109/CVPR.2015.7298698.

[30] T. Deng, K. F. Yang, Y. J. Li, H. M. Yan. Where does the driver look? Top-down-based saliency detection in a traffic driving environment. *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 7, pp. 2051–2062, 2016. DOI: 10.1109/TITS.2016.2535402.

[31] A. Geiger, M. Lauer, R. Urtasun. A generative model for 3D urban scene understanding from movable platforms. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, Colorado Springs, USA, pp. 1945–1952, 2011. DOI: 10.1109/CVPR.2011.5995641.

[32] J. M. Zhang, S. Sclaroff. Exploiting surroundedness for saliency detection: A Boolean map approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 5, pp. 889–902, 2016. DOI: 10.1109/TPAMI.2015.2473844.

[33] L. Zhou, Y. F. Ju, J. W. Fang, J. R. Xue. Saliency detection via background invariance in scale space. *Journal of Electronic Imaging*, vol. 26, no. 4, Article number 043021, 2017. DOI: 10.1117/1.JEI.26.4.043021.

[34] Q. Wang, Y. Yuan, P. K. Yan, X. L. Li. Saliency detection by multiple-instance learning. *IEEE Transactions on Cybernetics*, vol. 43, no. 2, pp. 660–672, 2013. DOI: 10.1109/TSMCB.2012.2214210.

[35] S. F. He, R. W. H. Lau. Exemplar-driven top-down saliency detection via deep association. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, Las Vegas, USA, pp. 5723–5732, 2016. DOI: 10.1109/CVPR.2016.617.

[36] J. M. Yang, M. H. Yang. Top-down visual saliency via joint CRF and dictionary learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 3, pp. 576–588, 2017. DOI: 10.1109/TPAMI.2016.2547384.

[37] J. T. Pan, E. Sayrol, X. Giro-I-Nieto, K. McGuinness, N. E. O′Connor. Shallow and deep convolutional networks for saliency prediction. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, Las Vegas, NV, USA, pp. 598–606, 2016. DOI: 10.1109/CVPR.2016.71.

[38] Y. Xia, D. Q. Zhang, A. Pozdnoukhov, K. Nakayama, K. Zipser, D. Whitney. Training a network to attend like human drivers saves it from common but misleading loss functions. arXiv:1711.06406, 2017.

[39] Y. Xie, L. F. Liu, C. H. Li, Y. Y. Qu. Unifying visual saliency with hog feature learning for traffic sign detection. In *Proceedings of IEEE Intelligent Vehicles Symposium*, IEEE, Xi′an, China, pp. 24–29, 2009. DOI: 10.1109/IVS.2009.5164247.

[40] W. J. Won, M. Lee, J. W. Son. Implementation of road traffic signs detection based on saliency map model. In *Proceedings of IEEE Intelligent Vehicles Symposium*, IEEE, Eindhoven, Netherlands, pp. 542–547, 2008. DOI: 10.1109/IVS.2008.4621144.

[41] D. D. Wang, X. W. Hou, J. W. Xu, S. G. Yue, C. L. Liu. Traffic sign detection using a cascade method with fast feature extraction and saliency test. *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, no. 12, pp. 3290–3302, 2017. DOI: 10.1109/TITS.2017.2682181.

[42] J. Kim, S. Kim, R. Mallipeddi, G. Jang, M. Lee. Adaptive driver assistance system based on traffic information saliency map. In *Proceedings of International Joint Conference on Neural Networks*, IEEE, Vancouver, Canada, pp. 1918–1923, 2016. DOI: 10.1109/IJCNN.2016.7727434.

[43] V. John, K. Yoneda, Z. Liu, S. Mita. Saliency map generation by the convolutional neural network for real-time traffic light detection using template matching. *IEEE Transactions on Computational Imaging*, vol. 1, no. 3, pp. 159–173, 2015. DOI: 10.1109/TCI.2015.2480006.

[44] H. L. Kuang, K. F. Yang, L. Chen, Y. J. Li, L. L. H. Chan, H. Yan. Bayes saliency-based object proposal generator for nighttime traffic images. *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. 3, pp. 814–825, 2017. DOI: 10.1109/TITS.2017.2702665.

[45] R. Timofte, K. Zimmermann, L. V. Gool. Multi-view traffic sign detection, recognition, and 3D localisation. *Machine Vision and Applications*, vol. 25, no. 3, pp. 633–647, 2014. DOI: 10.1007/s00138-011-0391-3.

[46] S. Alletto, A. Palazzi, F. Solera, S. Calderara, R. Cucchiara. DR(eye)VE: A dataset for attention-based tasks with applications to autonomous and assisted driving. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshops*, IEEE, Las Vegas, USA, 2016. DOI: 10.1109/CVPRW.2016.14.

[47] A. Palazzi, F. Solera, S. Calderara, S. Alletto, R. Cucchiara. Where should you attend while driving? arXiv:1611.08215, 2016.

[48] C. Landsiedel, D. Wollherr. Road geometry estimation for urban semantic maps using open data. *Advanced Robotics*, vol. 31, no. 5, pp. 282–290, 2017. DOI: 10.1080/01691864.2016.1250675.

[49] E. Levinkov, M. Fritz. Sequential Bayesian model update under structured scene prior for semantic road scenes labeling. In *Proceedings of IEEE International Conference on Computer Vision*, IEEE, Sydney, Australia, pp. 1321–1328, 2013. DOI: 10.1109/ICCV.2013.167.

[50] Z. Y. Zhang, S. Fidler, R. Urtasun. Instance-level segmentation for autonomous driving with deep densely connected MRFs. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, Las Vegas, USA, pp. 669–677, 2016. DOI: 10.1109/CVPR.2016.79.

[51] T. Cavallari, Semantic Slam: A New Paradigm for Object Recognition and Scene Reconstruction, Ph. D. dissertation, University of Bologna, Italy, 2017.

[52] S. C. Zhou, R. Yan, J. X. Li, Y. K. Chen, H. J. Tang. A brain-inspired SLAM system based on ORB features. *International Journal of Automation and Computing*, vol. 14, no. 5, pp. 564–575, 2017. DOI: 10.1007/s11633-017-1090-y.

[53] B. Zhao, J. S. Feng, X. Wu, S. C. Yan. A survey on deep learning-based fine-grained object classification and semantic segmentation. *International Journal of Automation and Computing*, vol. 14, no. 2, pp. 119–135, 2017. DOI: 10.1007/s11633-017-1053-3.

[54] H. Kong, J. Y. Audibert, J. Ponce. Vanishing point detection for road detection. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, Miami, USA, pp. 96–103, 2009. DOI: 10.1109/CVPR.2009.5206787.

[55] H. Kong, S. E. Sarma, F. Tang. Generalizing Laplacian of Gaussian filters for vanishing-point detection. *IEEE Transactions on Intelligent Transportation Systems*, vol. 14, no. 1, pp. 408–418, 2013. DOI: 10.1109/TITS.2012.2216878.

[56] J. J. Shi, J. X. Wang, F. F. Fu. Fast and robust vanishing point detection for unstructured road following. *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 4, pp. 970–979, 2016. DOI: 10.1109/TITS.2015.2490556.

[57] J. M. Alvarez, T. Gevers, A. M. Lopez. 3D scene priors for road detection. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, San Francisco, USA, pp. 57–64, 2010. DOI: 10.1109/CVPR.2010.5540228.

[58] E. Casapietra, T. H. Weisswange, C. Goerick, F. Kummert. Enriching a spatial road representation with lanes and driving directions. In *Proceedings of the 19th International Conference on Intelligent Transportation Systems*, IEEE, Rio de Janeiro, Brazil, pp. 1579–1585, 2016. DOI: 10.1109/ITSC.2016.7795768.

[59] A. Seff, J. X. Xiao. Learning from maps: Visual common sense for autonomous driving. arXiv:1611.08583, 2016.

[60] M. Y. Liu, S. X. Lin, S. Ramalingam, O. Tuzel. Layered interpretation of street view images. arXiv:1506.04723, 2015.

[61] A. Geiger, M. Lauer, C. Wojek, C. Stiller, R. Urtasun. 3D traffic scene understanding from movable platforms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 5, pp. 1012–1025, 2014. DOI: 10.1109/TPAMI.2013.185.

[62] A. Ess, T. Mueller, H. Grabner, L. Van Gool. Segmentation-based urban traffic scene understanding. In *Proceedings of British Machine Vision Conference*, London, UK, 2009. DOI: 10.5244/C.23.84.

[63] B. Kitt, A. Geiger, H. Lategahn. Visual odometry based on stereo image sequences with RANSAC-based outlier rejection scheme. In *Proceedings of IEEE Intelligent Vehicles Symposium*, IEEE, San Diego, USA, pp. 486–492, 2010. DOI: 10.1109/IVS.2010.5548123.

[64] S. Thrun, W. Burgard, D. Fox. *Probabilistic Robotics (Intelligent Robotics and Autonomous Agents)*, Cambridge, Mass, UK: MIT, 2005.

[65] H. Y. Zhang, A. Geiger, R. Urtasun. Understanding high-level semantics by modeling traffic patterns. In *Proceedings of IEEE Conference on Computer Vision*, IEEE, Sydney, Australia, pp. 3056–3063, 2013. DOI: 10.1109/ICCV.2013.379.

[66] C. Y. Chen, A. Seff, A. Kornhauser, J. X. Xiao. DeepDriving: Learning affordance for direct perception in autonomous driving. In *Proceedings of IEEE International Conference on Computer Vision*, IEEE, Santiago, Chile, pp. 2722–2730, 2015. DOI: 10.1109/ICCV.2015.312.

[67] P. Stahl, B. Donmez, G. A. Jamieson. Anticipation in driving: The role of experience in the efficacy of pre-event conflict cues. *IEEE Transactions on Human-Machine Systems*, vol. 44, no. 5, pp. 603–613, 2014. DOI: 10.1109/THMS.2014.2325558.

[68] S. J. Pan, Q. Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, 2010. DOI: 10.1109/TKDE.2009.191.

[69] N. Segev, M. Harel, S. Mannor, K. Crammer, R. El-Yaniv. Learn on source, refine on target: A model transfer learning framework with random forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 9, pp. 1811–1824, 2017. DOI: 10.1109/TPAMI.2016.2618118.

[70] D. Mitrovic. Reliable method for driving events recognition. *IEEE Transactions on Intelligent Transportation Systems*, vol. 6, no. 2, pp. 198–205, 2005. DOI: 10.1109/TITS.2005.848367.

[71] B. F. Wu, Y. H. Chen, C. H. Yeh, Y. F. Li. Reasoning-based framework for driving safety monitoring using driving event recognition. *IEEE Transactions on Intelligent Transportation Systems*, vol. 14, no. 3, pp. 1231–1241, 2013. DOI: 10.1109/TITS.2013.2257759.

[72] A. Ramirez, E. Ohn-Bar, M. Trivedi. Integrating motion and appearance for overtaking vehicle detection. In *Proceedings of IEEE Intelligent Vehicles Symposium Proceedings*, IEEE, Dearborn, USA, pp. 96–101, 2014. DOI: 10.1109/IVS.2014.6856598.

[73] J. D. Alonso, E. R. Vidal, A. Rotter, M. Muhlenberg. Lane-change decision aid system based on motion-driven vehicle tracking. *IEEE Transactions on Vehicular Technology*, vol. 57, no. 5, pp. 2736–2746, 2008. DOI: 10.1109/TVT.2008.917220.

[74] Y. Zhu, D. Comaniciu, M. Pellkofer, T. Koehler. Reliable detection of overtaking vehicles using robust information fusion. *IEEE Transactions on Intelligent Transportation Systems*, vol. 7, no. 4, pp. 401–414, 2006. DOI: 10.1109/TITS.2006.883936.

[75] F. Garcia, P. Cerri, A. Broggi, A. De La Escalera, J. M. Armingol. Data fusion for overtaking vehicle detection based on radar and optical flow. In *Proceedings of IEEE Intelligent Vehicles Symposium*, IEEE, Alcala de Henares, Spain, pp. 494–499, 2012. DOI: 10.1109/IVS.2012.6232199.

[76] Deutscher Verkehrssicherheitsrat. DVR-Report: Fachmagazin für Verkehrssicherheit. https://www.dvr.de/presse/dvr-report/2017-04.

[77] Auto Club Europa (ACE). Reviere der blinkmuffel. http://www.ace-online.de/fileadmin/user_uploads/Der_Club/Dokumente/10.07.2008_Grafik_Blinkmuffel_1.pdf.

[78] D. Kasper, G. Weidl, T. Dang, G. Breuel, A. Tamke, A. Wedel, W. Rosenstiel. Object-oriented Bayesian networks for detection of lane change maneuvers. *IEEE Intelligent Transportation Systems Magazine*, vol. 4, no. 3, pp. 19–31, 2012. DOI: 10.1109/MITS.2012.2203229.

[79] W. Yao, Q. Q. Zeng, Y. P. Lin, D. H. Xu, H. J. Zhao, F. Guillemard, S. Geronimi, F. Aioun. On-road vehicle trajectory collection and scene-based lane change analysis: Part II. *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, no. 1, pp. 206–220, 2017. DOI: 10.1109/TITS.2016.2571724.

[80] T. Gindele, S. Brechtel, R. Dillmann. A probabilistic model for estimating driver behaviors and vehicle trajectories in traffic environments. In *Proceedings of the 13th International Conference on Intelligent Transportation Systems*, IEEE, Funchal, Portugal, pp. 1625–1631, 2010. DOI: 10.1109/ITSC.2010.5625262.

[81] S. Sivaraman, B. Morris, M. Trivedi. Learning multi-lane trajectories using vehicle-based vision. In *Proceedings of IEEE Conference on Computer Vision Workshops*, IEEE, Barcelona, Spain, pp. 2070–2076, 2011. DOI: 10.1109/ICCVW.2011.6130503.

[82] R. K. Satzoda, M. M. Trivedi. Overtaking & receding vehicle detection for driver assistance and naturalistic driving studies. In *Proceedings of the 17th International Conference on Intelligent Transportation Systems*, IEEE, Qingdao, China, pp. 697–702, 2014. DOI: 10.1109/ITSC.2014.6957771.

[83] M. S. Kristoffersen, J. V. Dueholm, R. K. Satzoda, M. M. Trivedi, A. Mogelmose, T. B. Moeslund. Towards semantic understanding of surrounding vehicular maneuvers: A panoramic vision-based framework for real-world highway studies. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshops*, IEEE, Las Vegas, USA, pp. 1584–1591, 2016. DOI: 10.1109/CVPRW.2016.197.

[84] J. V. Dueholm, M. S. Kristoffersen, R. K. Satzoda, T. B. Moeslund, M. M. Trivedi. Trajectories and maneuvers of surrounding vehicles with panoramic camera arrays. *IEEE Transactions on Intelligent Vehicles*, vol. 1, no. 2, pp. 203–214, 2016. DOI: 10.1109/TIV.2016.2622921.

[85] A. Khosroshahi, E. Ohn-Bar, M. M. Trivedi. Surround vehicles trajectory analysis with recurrent neural networks. In *Proceedings of the 19th International Conference on Intelligent Transportation Systems*, IEEE, Rio de Janeiro, Brazil, pp. 2267–2272, 2016. DOI: 10.1109/ITSC.2016.7795922.

[86] S. Ernst, J. Rieken, M. Maurer. Behaviour recognition of traffic participants by using manoeuvre primitives for automated vehicles in urban traffic. In *Proceedings of the 19th International Conference on Intelligent Transportation Systems*, IEEE, Rio de Janeiro, Brazil, 2016. DOI: 10.1109/ITSC.2016.7795674.

[87] S. Busch, T. Schindler, T. Klinger, C. Brenner. Analysis of Spatio-temporal traffic patterns based on pedestrian trajectories. In *Proceedings of International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, Prague, Czech Republic, vol. XLI-B2, pp. 497–503, 2016. DOI: 10.5194/isprsarchives-XLI-B2-497-2016.

[88] J. Hariyono, K. H. Jo. Detection of pedestrian crossing road: A study on pedestrian pose recognition. *Neurocomputing*, vol. 234, pp. 144–153, 2017. DOI: 10.1016/j.neucom.2016.12.050.

[89] R. M. Mueid, C. Ahmed, M. A. R. Ahad. Pedestrian activity classification using patterns of motion and histogram of oriented gradient. *Journal on Multimodal User Interfaces*, vol. 10, no. 4, pp. 299–305, 2016. DOI: 10.1007/s12193-015-0178-3.

[90] R. Quintero, I. Parra, D. F. Llorca, M. A. Sotelo. Pedestrian intention and pose prediction through dynamical models and behaviour classification. In *Proceedings of the 18th International Conference on Intelligent Transportation Systems*, IEEE, Las Palmas, Spain, pp. 83–88, 2015. DOI: 10.1109/ITSC.2015.22.

[91] M. Ogawa, H. Fukamachi, R. Funayama, T. Kindo. CYKLS: Detect pedestrian′s dart focusing on an appearance change. In *Proceedings of the 12th International Conference on Computer Vision*, Springer-Verlag, Florence, Italy, pp. 556–565, 2012. DOI: 10.1007/978-3-642-33868-7_55.

[92] F. H. Chan, Y. T. Chen, Y. Xiang, M. Sun. Anticipating accidents in dashcam videos. In *Proceedings of the 13th Asian Conference on Computer Vision*, Springer, Taipei, China, pp. 136–153, 2016. DOI: 10.1007/978-3-319-54190-7_9.

[93] Y. J. Xia, W. W. Xu, L. M. Zhang, X. M. Shi, K. Mao. Integrating 3D structure into traffic scene understanding with RGB-D data. *Neurocomputing*, vol. 151, pp. 700–709, 2015. DOI: 10.1016/j.neucom.2014.05.091.

[94] A. Tageldin, M. H. Zaki, T. Sayed. Examining pedestrian evasive actions as a potential indicator for traffic conflicts. *IET Intelligent Transport Systems*, vol. 11, no. 5, pp. 282–289, 2017. DOI: 10.1049/iet-its.2016.0066.

[95] F. Westerhuis, D. De Waard. Reading cyclist intentions: Can a lead cyclists behaviour be predicted? *Accident Analysis & Prevention*, vol. 105, pp. 146–155, 2017. DOI: 10.1016/j.aap.2016.06.026.

[96] D. Manstetten. Behaviour prediction and intention detection in UR:BAN VIE – overview and introduction. *UR:BAN Human Factors in Traffic*, K. Bengler, J. Drüke, S. Hoffmann, D. Manstetten, A. Neukum, Eds., Wiesbaden, Germany: Springer, 2018. DOI: 10.1007/978-3-658-15418-9_8.

[97] F. Schneemann, P. Heinemann. Context-based detection of pedestrian crossing intention for autonomous driving in urban environments. In *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems*, IEEE, Daejeon, South Korea, pp. 2243–2248, 2016. DOI: 10.1109/IROS.2016.7759351.

[98] T. Fugger, B. Randles, A. Stein, W. Whiting, B. Gallagher. Analysis of pedestrian gait and Perception-reaction at signal-controlled crosswalk intersections. *Transportation Research Record: Journal of the Transportation Research Board*, vol. 1705, no. 1, pp. 20–25, 2000. DOI: 10.3141/1705-04.

[99] N. Schneider, D. M. Gavrila. Pedestrian path prediction with recursive Bayesian filters: A comparative study. In *Proceedings of the 35th German Conference on Pattern Recognition*, Springer, Saarbrücken, Germany, pp. 174–183, 2013. DOI: 10.1007/978-3-642-40602-7_18.

[100] M. Goldhammer, M. Gerhard, S. Zernetsch, K. Doll, U. Brunsmann. Early prediction of a pedestrian′s trajectory at intersections. In *Proceedings of the 16th International IEEE Conference on Intelligent Transportation Systems*, IEEE, The Hague, The Netherlands, pp. 237–242, 2013. DOI: 10.1109/ITSC.2013.6728239.

[101] M. Goldhammer, K. Doll, U. Brunsmann, A. Gensler, B. Sick. Pedestrians trajectory forecast in public traffic with artificial neural networks. In *Proceedings of the 22nd International Conference on Pattern Recognition*, IEEE, Stockholm, Sweden, pp. 4110–4115, 2014. DOI: 10.1109/ICPR.2014.704.

[102] C. G. Keller, D. M. Gavrila. Will the pedestrian cross? A study on pedestrian path prediction. *IEEE Transactions on Intelligent Transportation Systems*, vol. 15, no. 2, pp. 494–506, 2014. DOI: 10.1109/TITS.2013.2280766.

[103] S. Koehler, M. Goldhammer, S. Bauer, S. Zecha, K. Doll, U. Brunsmann, K. Dietmayer. Stationary detection of the Pedestrians intention at intersections. *IEEE Intelligent Transportation Systems Magazine*, vol. 5, no. 4, pp. 87–99, 2013. DOI: 10.1109/MITS.2013.2276939.

[104] J. F. P. Kooij, N. Schneider, F. Flohr, D. M. Gavrila. Context-based pedestrian path prediction. In *Proceedings of the 13th European Conference on Computer Vision*, Springer, Zurich, Switzerland, pp. 618–633, 2014. DOI: 10.1007/978-3-319-10599-4_40.

[105] J. Y. Kwak, B. C. Ko, J. Y. Nam. Pedestrian intention prediction based on dynamic fuzzy automata for vehicle driving at nighttime. *Infrared Physics & Technology*, vol. 81, pp. 41–51, 2017. DOI: 10.1016/j.infrared.2016.12.014.

[106] G. Q. Xu, L. Liu, Y. S. Ou, Z. J. Song. Dynamic modeling of driver control strategy of lane-change behavior and trajectory planning for collision prediction. *IEEE Transactions on Intelligent Transportation Systems*, vol. 13, no. 3, pp. 1138–1155, 2012. DOI: 10.1109/TITS.2012.2187447.

[107] R. N. Dang, J. Q. Wang, S. E. Li, K. Q. Li. Coordinated adaptive cruise control system with lane-change assistance. *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 5, pp. 2373–2383, 2015. DOI: 10.1109/TITS.2015.2389527.

[108] W. Liu, S. W. Kim, K. Marczuk, M. H. Ang. Vehicle motion intention reasoning using cooperative perception on urban road. In *Proceedings of the 17th International Conference on Intelligent Transportation Systems*, IEEE, Qingdao, China, pp. 424–430, 2014. DOI: 10.1109/ITSC.2014.6957727.

[109] Y. Hou, P. Edara, C. Sun. Modeling mandatory lane changing using Bayes classifier and decision trees. *IEEE Transactions on Intelligent Transportation Systems*, vol. 15, no. 2, pp. 647–655, 2014. DOI: 10.1109/TITS.2013.2285337.

[110] D. Lee, A. Hansen, J. K. Hedrick. Probabilistic inference of traffic participants lane change intention for enhancing adaptive cruise control. In *Proceedings of IEEE Intelligent Vehicles Symposium*, IEEE, Los Angeles, USA, pp. 855–860, 2017. DOI: 10.1109/IVS.2017.7995823.

[111] Y. L. Gu, Y. Hashimoto, L. T. Hsu, S. Kamijo. Motion planning based on learning models of pedestrian and driver behaviors. In *Proceedings of the 19th International Conference on Intelligent Transportation Systems*, IEEE, Rio de Janeiro, Brazil, pp. 808–813, 2016. DOI: 10.1109/ITSC.2016.7795648.

[112] W. D. Xu, J. Pan, J. Q. Wei, J. M. Dolan. Motion planning under uncertainty for on-road autonomous driving. In *Proceedings of IEEE International Conference on Robotics and Automation*, IEEE, Hong Kong, China, pp. 2507–2512, 2014. DOI: 10.1109/ICRA.2014.6907209.

[113] T. Y. Gu, J. M. Dolan, J. W. Lee. Automated tactical maneuver discovery, reasoning and trajectory planning for autonomous driving. In *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems*, IEEE, Daejeon, South Korea, pp. 5474–5480, 2016. DOI: 10.1109/IROS.2016.7759805.

[114] N. Nagasaka, M. Harada. Towards safe, smooth, and stable path planning for on-road autonomous driving under uncertainty. In *Proceedings of the 19th International Conference on Intelligent Transportation Systems*, IEEE, Rio de Janeiro, Brazil, pp. 795–801, 2016. DOI: 10.1109/ITSC.2016.7795646.

[115] K. Jo, M. Lee, J. Kim, M. Sunwoo. Tracking and behavior reasoning of moving vehicles based on roadway geometry constraints. *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, no. 2, pp. 460–476, 2017. DOI: 10.1109/TITS.2016.2605163.

[116] E. A. I. Pool, J. F. P. Kooij, D. M. Gavrila. Using road topology to improve cyclist path prediction. In *Proceedings of IEEE Intelligent Vehicles Symposium*, IEEE, Los Angeles, USA, pp. 289–296, 2017. DOI: 10.1109/IVS.2017.7995734.

[117] N. Evestedt, E. Ward, J. Folkesson, D. Axehill. Interaction aware trajectory planning for merge scenarios in congested traffic situations. In *Proceedings of the 19th International Conference on Intelligent Transportation Systems*, IEEE, Rio de Janeiro, Brazil, pp. 465–472, 2016. DOI: 10.1109/ITSC.2016.7795596.

[118] H. M. Eraqi, M. N. Moustafa, J. Honer. End-to-end deep learning for steering autonomous vehicles considering temporal dependencies. arXiv:1710.03804, 2017.

[119] L. Caltagirone, M. Bellone, L. Svensson, M. Wahde. Simultaneous perception and path generation using fully convolutional neural networks. arXiv:1703.08987, 2017.

[120] Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, F. Durand. What do different evaluation metrics tell us about saliency models? arXiv:1604.03605, 2016.

[121] B. W. Tatler. The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. *Journal of Vision*, vol. 7, no. 14-17, pp. 4.1-17, 2007. DOI: 10.1167/7.14.4.

[122] R. J. Peters, A. Iyer, L. Itti, C. Koch. Components of bottom-up gaze allocation in natural images. *Vision Research*, vol. 45, no. 18, pp. 2397–2416, 2005. DOI: 10.1016/j.visres.2005.03.019.

[123] M. Kümmerer, T. S. A. Wallis, M. Bethge. Information-theoretic model comparison unifies saliency metrics. In *Proceedings of the National Academy of Sciences of the United States of America*, vol. 112, no. 52, pp. 16054–16059, 2015. DOI: 10.1073/pnas.1510393112.

[124] M. J. Swain, D. H. Ballard. Color indexing. *International Journal of Computer Vision*, vol. 7, no. 1, pp. 11–32, 1991. DOI: 10.1007/BF00130487.

[125] O. Le Meur, P. Le Callet, D. Barba. Predicting visual fixations on video based on low-level visual features. *Vision Research*, vol. 47, no. 19, pp. 2483–2498, 2007. DOI: 10.1016/j.visres.2007.06.015.

[126] O. Pele, M. Werman. A linear time histogram metric for improved sift matching. In *Proceedings of the 10th European Conference on Computer Vision: Part III*, Marseille, France, pp. 495–508, 2008. DOI: 10.1007/978-3-540-88690-7_37.

[127] Y. Rubner, C. Tomasi, L. J. Guibas. The earth movers distance as a metric for image retrieval. *International Journal of Computer Vision*, vol. 40, no. 2, pp. 99–121, 2000. DOI: 10.1023/A:1026543900054.

[128] C. Sammut, G. I. Webb. *Encyclopedia of Machine Learning*, Boston, MA: Springer, 2010.

Jian-Ru Xue received the M. Sc. and Ph. D. degrees from Xi′an Jiaotong University (XJTU), China in 1999 and 2003, respectively. He was with FujiXerox, Japan from 2002 to 2003, and visited the University of California at Los Angeles, USA from 2008 to 2009. He is currently a professor with the Institute of Artificial Intelligence and Robotics at XJTU. He served as a co-organization chair of the *Asian Conference on Computer Vision and Virtual System and Multimedia Conference*. He also served as a PC member of the *Pattern Recognition Conference* in 2012, and *Asian Conference on Computer Vision* in 2010 and 2012.

His research interests include computer vision, visual navigation, and scene understanding for autonomous system.

E-mail: jrxue@mail.xjtu.edu.cn

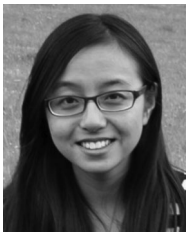
ORCID iD: 0000-0002-4994-9343

Jian-Wu Fang received the Ph. D. degree in signal and information processing from Univerisity of Chinese Academy of Sciences, China in 2015. He is currently an assistant professor in School of Electronic and Control Engineering, Chang′an University, China, and is also a postdoctor in Institute of Artificial Intelligence and Robotics, Xi′an Jiaotong University, China.

His research interests include computer vision, pattern recognition and scene understanding.

E-mail: j.w.fangit@gmail.com (Corresponding author)

ORCID iD: 0000-0002-0300-6892



Pu Zhang received the B. Sc. degree in automation from Southeast University, China in 2016. She is currently a Ph. D. degree candidate at Institute of Artificial Intelligence and Robotics, Xi′an Jiaotong University, China.

Her research interests include computer vision and on-road scene understanding.

E-mail: zhangpu2016@stu.xjtu.edu.cn