Risk Perception in Driving Scenes

Nakul Agarwal
Honda Research Institute, USA
nakul_agarwal@honda-ri.com

Yi-Ting Chen
National Yang Ming Chiao Tung University
ychen@cs.nctu.edu.tw

Abstract

The holy grail of intelligent vehicles is to enable a zero collision mobility experience. This endeavor requires an interdisciplinary effort to understand driver behavior and to assess risks surrounding the vehicle. A driver's perception of risk is a complex cognitive process that is largely manifested by the voluntary response of the driver to external stimuli as well as the apparent attentiveness of participants towards the ego-vehicle. In this work, we examine the problem of risk perception and introduce a new dataset to facilitate research in this domain. Our dataset consists of 4706 short video clips that include annotations of driver intent, road network topology, situation (e.g., crossing pedestrian), driver response, and pedestrian attentiveness using face annotations. We also provide a simple weakly supervised framework to tackle this task which performs favorably against state of the art methods.

1 Introduction

Each year, road traffic accidents are among the leading cause of non-natural death around the world. More than 1.3 million people die in road accidents worldwide every year, approximately 3,700 people per day [20]. Recent research and technological progress in automated and advanced driver assistance systems promise to significantly reduce traffic related collisions in next generation mobility systems. One promising research direction towards the successful deployment of intelligent driving systems is understanding and development of computational models of driver decision processes and risk perception when interacting with surrounding traffic participants.

While risk is generally defined on the basis of collision prediction [6] in the context of intelligent vehicles, this definition does not explicitly capture the notion of risk from a driver's perspective. Such driver-centric modeling of risk has been recently proposed [7], whereby potentially risky objects are defined as those that *influence* driver behavior. The authors propose a new task called risk object identification and develop a weakly supervised computational framework to individually train and evaluate risk using four different reactive scenarios considered in the HDD dataset [14]. While a good starting point, the HDD dataset has limited number of risk situations. Additionally, developing separate models for different risk situations is not practical in the real world. To mitigate these issues, we make the following contributions in this work. First, to address the limitations of existing datasets, we introduce a novel and comprehensive dataset with a diverse set of situations and annotations to enable research for risk perception in driving scenes as shown in Table 1. Second, we provide a weakly supervised framework for risk object identification that performs favorably against state of the art methods and use a single model to benchmark the proposed dataset for future research.

2 Risk Object Identification with Attention (ROI-A) Dataset

The data is collected in the San Francisco Bay Area region using an instrumented vehicle equipped with 3 Point Grey Grasshopper video cameras with a resolution of 1920×1200 pixels, a Velodyne

Machine Learning for Autonomous Driving Workshop at the 36th Conference on Neural Information Processing Systems (NeurIPS 2022), New Orleans, USA.

Table 1: Comparison of our proposed ROI-A with other datasets.

	# Clips	# Risk Situations	Driver intention	Driver response	CAN data	LiDAR	Contextual elements	Road topology	Pedestrian attention	Face annotations	Pedestrian tracklets
JAAD [16]	346	✗	✗	✗	✓	✗	✗	✗	✓	✗	✓
PIE [17]	-	✗	✗	✗	✓	✗	✗	✗	✓	✗	✓
STIP [9]	556	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
LOKI [2]	644	✗	✓	✗	✓	✓	✗	✗	✗	✗	✓
HDD [14]	-	4	✓	✓	✓	✓	✗	✗	✗	✗	✗
ROI-A	4706	10	✓	✓	✓	✓	✓	✓	✓	✓	✓

HDL-64E S2 LiDAR sensor, and high precision GPS. It captures diverse set of traffic scenes including different environments such as urban, suburban, and highway.

2.1 Annotation Methodology

Understanding driver and pedestrian behavior is essential for development of robust intelligent driving systems. Modeling both driver and pedestrian behavior is complex and involves different levels of cognitive processes, particularly in complicated interactive scenes. The data selection and annotation protocol must be carefully designed. The first step involves manual selection of short clips from hours of recording that includes appropriate scenarios for the task. Note that automatic situation localization [11, 22] in untrimmed videos can also be explored using the proposed dataset.

We propose a 4-layer representation, i.e. **Driver Intention**, **Topology**, **Situation**, **Driver Response**, to describe driver behavior for risk assessment. Note that the proposed representation is different from the one proposed in [14] since the labeling structure is designed explicitly for the risk assessment task. For **Driver Intention**, we focus on three classes, i.e., *Left-Turn*, *Right-Turn* and *Go-Straight*. While navigating towards their goal (e.g., reach an intersection), drivers encounter different driving situations (e.g., a bicyclist is crossing the street and a truck is parked near the ego-lane). Drivers are also perceptive of the road topology and situation of scene as part of their planning and decision making. The underlying road topology network is annotated in the **Topology** layer that includes *4-Way*, *3-Way*, and *Straight*. While navigating in a road topology network, drivers react to certain agents or objects on the road (e.g, slowing for a stop sign or yielding to a crossing pedestrian). The road agents that directly impact driver behavior are annotated in the **Situation** layer. Specifically, we select a comprehensive set of situations, i.e., *Stop Sign*, *Traffic Light*, *Crossing Pedestrian*, *Crossing vehicle*, *Car Blocking Ego Lane*, *Congestion*, *JayWalking*, *Car Backing Into Parking Space*, *Car on Shoulder Open Door*, and *Cut In*. Then, the response of driver to these road agents is labeled in the **Driver Response** layer. In this work, two types of decision are annotated, i.e., *Influenced* and *Non-influenced*. Here *Influenced* means whenever the driver alters behavior from its regular course. For e.g. *deviate* from parked vehicle, *yield* to crossing pedestrian or vehicle, or *stop* for stop sign or traffic light.

For pedestrian attentiveness, we focus on the attention of pedestrians when the ego-vehicle is approaching. We explicitly select a subset of scenes, i.e., 854 videos, from the larger dataset used for the risk object identification tasks, where the subset includes scenes in which the driver is influenced by pedestrians. This enables the dataset to be studied for jointly analysing pedestrian attention in the context of risk. The pedestrian attentiveness labels are i.e., *Looking*, *Not Looking*, and *Not Sure*. Specifically, we label bounding boxes and occlusion flags around pedestrian faces as well as pedestrian bodies. The design enables to reason pedestrian attentiveness from both faces and bodies instead of purely using body poses as in [16, 17]. Past research has shown that faces play a significant role in understanding pedestrian attentiveness [5, 15].

3 Methodology

Motivated by [7], we formulate the risk object identification problem as a cause and effect problem [12]. Figure 1 depicts the overview of the proposed framework. Given a sequence of video frames observed in the past, the framework extracts image-level and object-level features for objects of interest. An ego-centric spatio-temporal graph is constructed using these features as the representation of the various nodes in the graph.

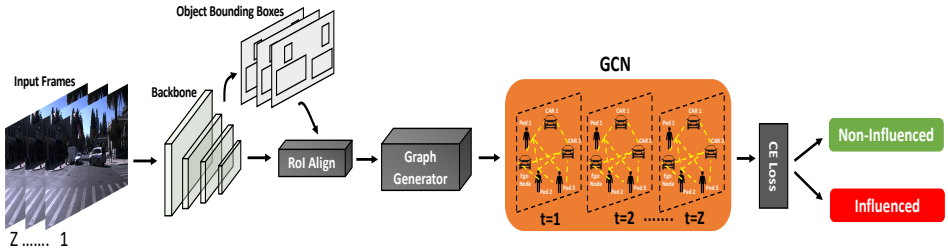

Figure 1: **Proposed Network Architecture.** The proposed algorithm takes as input a sequence of RGB frames, a sequence of binary masks for partial convolution and a set of object tracklets. These inputs are then passed on to the graph convolutional network for getting the scene representation to predict the driver decision of either getting *Influenced* or *Non-influenced* which in turn is used to identify risk objects.

3.1 Graph Based Reasoning

Node Feature Extraction. To obtain node features, Mask R-CNN [3] pretrained on COCO dataset [8] and Deep SORT [19] is applied to detect and track every object. ResNet-50 pretrained for panoptic segmentation on a driving scene dataset [13] along with partial convolution [10] and RoIAlign [3] is used to extract the corresponding object and image level representation. The ego node feature, i.e. representation of the ego-vehicle, is obtained by the same procedure using a frame-size bounding box. This also enables us to capture the scene context. Driving scenes are complicated, and not all objects in the scene influence the driver. Therefore, we limit the objects of interest in our implementation to the following classes: *person*, *bicycle*, *car*, *motorcycle*, *bus*,*truck*, *traffic-light*, and *stop-sign*. Similar to [7], we make use of partial convolution layers [10] to simulate a situation without the presence of an object.

Graph Definition. We utilize the graph structure to explicitly model pair-wise relations between different agents in the driving scene to understand and describe the activities. While previous graph based works [1, 21] consider objects in the graph independently, we make use of tracking to form our graph. Given a set of N agents in the traffic scene with their corresponding tracklets, we construct a spatio-temporal graph $G_t = (V_t, A_t)$, where $V_t = \{v_t^i | \forall i \in \{1, ..., N\}\}$ is the set of vertices of graph G_t and $A_t = \{a_t^{ij} | \forall i, j \in \{1, ..., N\}\}$ is the adjacency matrix $\forall t \in \{1, ..., Z\}$. In our graph, a_t^{ij} models the appearance relation between two agents at time t and is formally defined as:

$$a_t^{ij} = \frac{f_p(v_t^i, v_t^j) exp(f_a(v_t^i, v_t^j))}{\sum_{j=1}^{N} f_p(v_t^i, v_t^j) exp(f_a(v_t^i, v_t^j))}, \tag{1}$$

where $f_a(v_t^i, v_t^j)$ indicates the appearance relation between agents i and j at time t, and $f_p(v_t^i, v_t^j)$ is an indicator function which determines the presence of a tracklet. Softmax function is used to normalize the influence on agent i from other objects. The appearance relation is calculated as below:

$$f_a(v_t^i, v_t^j) = \frac{\theta(v_t^i)^T \phi(v_t^j)}{\sqrt{D}} \tag{2}$$

where $\theta(v_t^i) = \mathbf{w} v_t^i$ and $\phi(v_t^j) = \mathbf{w'} v_t^j$. Both $\mathbf{w} \in \mathbb{R}^{D \times D}$ and $\mathbf{w'} \in \mathbb{R}^{D \times D}$ are learnable parameters which map appearance features to a subspace and enable learning the correlation of two objects, and $\sqrt{D}$ is a normalization factor. While [1, 21] can fill the graph with any random node at time t, we need to take into account missing nodes due to both inconsistencies in tracking and agents entering and leaving the traffic scene at different times. In order to mitigate this issue, we set adjacency matrix values to zero when an object is missing using indicator function f_p as:

$$f_p(v_t^i, v_t^j) = \mathbb{I}(v_t^i = present \textbf{ and } v_t^j = present) \tag{3}$$

Once the nodes and adjacency matrix values are defined, we reason over the Graph Convolutional Network (GCN) [4]. GCN takes a graph as input, performs computations over the structure, and

Table 2: Comparison with baseline methods on the HDD dataset (left) and ROI-A dataset (right).

Method	Crossing Vehicle	Crossing Pedestrian	Parked Vehicle	Cong-estion	mAcc
Random	14.78	6.32	7.21	6.74	8.76
[7]	25.40	**19.88**	21.02	18.58	21.22
[18]	26.52	17.50	22.20	45.05	27.81
Ours	**48.97**	18.21	**35.58**	**58.88**	**40.41**

Method	Crossing Pedestrian	Crossing Vehicle	Car Blocking Ego Lane	Cong-estion	Cut-In	Jay-walking	Traffic Light	Stop Sign	mAcc
Random	5.90	9.56	7.38	7.42	10.26	4.40	2.94	5.67	6.99
[7]	13.96	24.01	15.68	30.46	**33.66**	11.28	4.32	8.00	17.67
[18]	13.27	24.17	14.42	**41.58**	32.30	**12.37**	2.02	6.53	18.33
Ours	**15.64**	**31.05**	**34.48**	27.35	27.69	3.42	2.83	**11.76**	**19.28**

returns a graph as output. For a target node i in the graph, it aggregates features from all neighbor nodes according to values in the adjacency matrix. Formally, one layer of GCN can be written as:

$$Z^{(l+1)} = \sigma A Z^{(l)} W^{(l)} \tag{4}$$

where $A \in \mathbb{R}^{NZ \times NZ}$ is the adjacency matrix for appearance model. $Z^l \in \mathbb{R}^{NZ \times D}$ is the feature representations of nodes in the lth layer. $W^l \in \mathbb{R}^{D \times D}$ is the layer-specific learnable weight matrix. $\sigma(.)$ denotes an activation function, and we adopt ReLU in this work. This layer-wise propagation can be stacked into multi-layers.

Loss Function. We train the network using a standard cross entropy loss given by:

$$\mathcal{L} = -\frac{1}{R}\sum_{i=1}^{R} y_i log(p(y_i)) + (1 - y_i) log(1 - p(y_i)), \tag{5}$$

where y is the label, $p(y)$ is the predicted probability and R is the batch size.

4 Experiments

4.1 Dataset and Metrics

In addition to showing results on our proposed dataset, we also show results on the publicly available HDD dataset [14]. We don't report results on the two classes of *Car Backing Into Parking Space* and *Car on Shoulder Open Door* from ROI-A as they don't have enough samples for training and evaluation when sampled at 3fps. We use the same accuracy metric for evaluation used in [7], i.e., the number of correct predictions over the number of ground truth samples. A prediction is considered accurate if the Intersection over Union (IoU) score is greater than a certain threshold. We report mean accuracy mAcc, which is the average accuracy at 10 IoU thresholds evenly distributed from 0.5 to 0.95 [23].

4.2 Results and Analysis

Baselines. We choose [7, 18] for comparison, in addition to having a random baseline where the risk object is randomly picked. While [7] is the only method which directly deals with our final task and we use their original code, we re-implement [18] to align with our task. Specifically, an object-centric driving model is designed in [18] by learning object level attention weights, which we use as an object selector for identifying risk object.

Results. Table 2 demonstrate that our method outperforms all baselines on two different datasets, highlighting the efficacy of the proposed method. The lower performance of all methods in Table 2 (right) compared to Table 2 (left) also points to the challenging nature of our proposed ROI-A dataset. We also observe that the results of traffic light and stop sign are generally quite low compared to other categories. This is primarily because i) the detection and tracking of these categories is not very consistent and ii) building a reasoning module using these categories is non-trivial as these categories do not physically lie in the path of the ego-vehicle trajectory, and therefore requires a thorough understanding of the scene and layout.

References

[1] F. Baradel, N. Neverova, C. Wolf, J. Mille, and G. Mori. Object Level Visual Reasoning in Videos. In *ECCV*, 2018.

[2] H. Girase, H. Gang, S. Malla, J. Li, A. Kanehara, K. Mangalam, and C. Choi. Loki: Long term and key intentions for trajectory prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9803–9812, 2021.

[3] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask R-CNN. In *ICCV*, 2017.

[4] T. N. Kipf and M. Welling. Semi-supervised Classification with Graph Convolutional Networks. *ICLR*, 2017.

[5] J. Kooij, N. Schneider, F. Flohr, and D. Gavrila. Context-based Pedestrian Path Prediction. In *ECCV*, 2014.

[6] S. Lefèvre, D. Vasquez, and C. Laugier. A Survey on Motion Prediction and Risk Assessment for Intelligent Vehicles. *ROBOMECH Journal*, 1:1, 2014.

[7] C. Li, S. H. Chan, and Y.-T. Chen. Who Make Drivers Stop? Towards Driver-centric Risk Assessment: Risk Object Identification via Causal Inference. In *IROS*, 2020.

[8] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: Common Objects in Context. In *ECCV*, 2014.

[9] B. Liu, E. Adeli, Z. Cao, K.-H. Lee, A. Shenoi, A. Gaidon, and J. C. Niebles. Spatiotemporal relationship reasoning for pedestrian intent prediction. *IEEE Robotics and Automation Letters*, 5(2):3485–3492, 2020.

[10] G. Liu, F. A. Reda, K. J. Shih, T.-C. Wang, A. Tao, and B. Catanzaro. Image Inpainting for Irregular Holes using Partial Convolutions. In *ECCV*.

[11] A. Narayanan, I. Dwivedi, and B. Dariush. Dynamic Traffic Scene Classification with Space-time Coherence. In *ICRA*, 2019.

[12] J. Pearl. *Causality*. Cambridge university press, 2009.

[13] L. Porzi, S. R. Bulo, A. Colovic, and P. Kontschieder. Seamless Scene Segmentation. In *CVPR*, 2019.

[14] V. Ramanishka, Y.-T. Chen, T. Misu, and K. Saenko. Toward Driving Scene Understanding: A Dataset for Learning Driver Behavior and Causal Reasoning. In *CVPR*, 2018.

[15] A. Rasouli and J. K. Tsotsos. Autonomous Vehicles that Interact with Pedestrians: A Survey of Theory and Practice. *IEEE Transactions on Intelligent Transportation Systems*, 21(3):900–918, 2020.

[16] A. Rasouli, I. Kotseruba, and J. K. Tsotsos. Are They Going to Cross? A Benchmark Dataset and Baseline for Pedestrian Crosswalk Behavior. In *ICCV-W*, 2017.

[17] A. Rasouli, I. Kotseruba, T. Kunic, and J. K. Tsotsos. PIE: A Large-scale Dataset and Models for Pedestrian Intention Estimation and Trajectory Prediction. In *ICCV*, 2019.

[18] D. Wang, C. Devin, Q.-Z. Cai, F. Yu, and T. Darrell. Deep Object-centric Policies for Autonomous Driving. In *ICRA*, 2019.

[19] N. Wojke, A. Bewley, and D. Paulus. Simple Online and Realtime Tracking with a Deep Association Metric. In *ICIP*.

[20] World Health Organization. Global Status Report on Road Safety 2018: Summary, 2018.

[21] J. Wu, L. Wang, L. Wang, J. Guo, and G. Wu. Learning actor relation graphs for group activity recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9964–9974, 2019.

[22] M. Xu, M. Gao, Y.-T. Chen, L. Davis, and D. Crandall. Temporal Recurrent Networks for Online Action Detection. In *ICCV*, 2019.

[23] Z. Zhang, C. Yu, and D. Crandall. A self validation network for object-level human attention estimation. In *NeurIPS*, 2019.