Modelling Driver Cognitive Behaviours Under Critical Scenes Based On Deep Learning

Shun Gan, Qingfan Wang, Quan Li, Xizhe Pei, Taisong Cui, Bingbing Nie

I. INTRODUCTION

Drivers' cognitive behaviours (e.g. hazard perception) in pre-crash phases could significantly influence vehicle dynamics (i.e. via braking) and human injury risk (i.e. via posture change) in motor vehicle crashes (MVCs) [1]. In an impending collision situation, drivers read the critical traffic scene information through visual perception, recognize the potential hazard event and exhibit natural avoidance behaviour. Yet, such cognitive behaviours remain underrepresented in current driving safety systems. Modelling of complex cognitive behaviour has undergone a paradigm shift from manual feature engineering towards data-driven, automatic representation learning, owing to the development of the deep learning (DL) approach [2]. Previous work has identified that biologically inspired DL architectures of the primate visual system could locate the spatial allocation of human attention [3]. DL spatial-temporal fusion models have achieved high accuracy on dynamic scene recognition tasks, such as driving lane change intention in complex traffic scenes [4-5]. To quantify the cognitive behaviour of passenger car drivers in near-crash scenarios, we proposed a DL-based modelling approach on the drivers' cognitive behaviours to represent hazard perception, recognition, and avoidance (Fig. 1.).

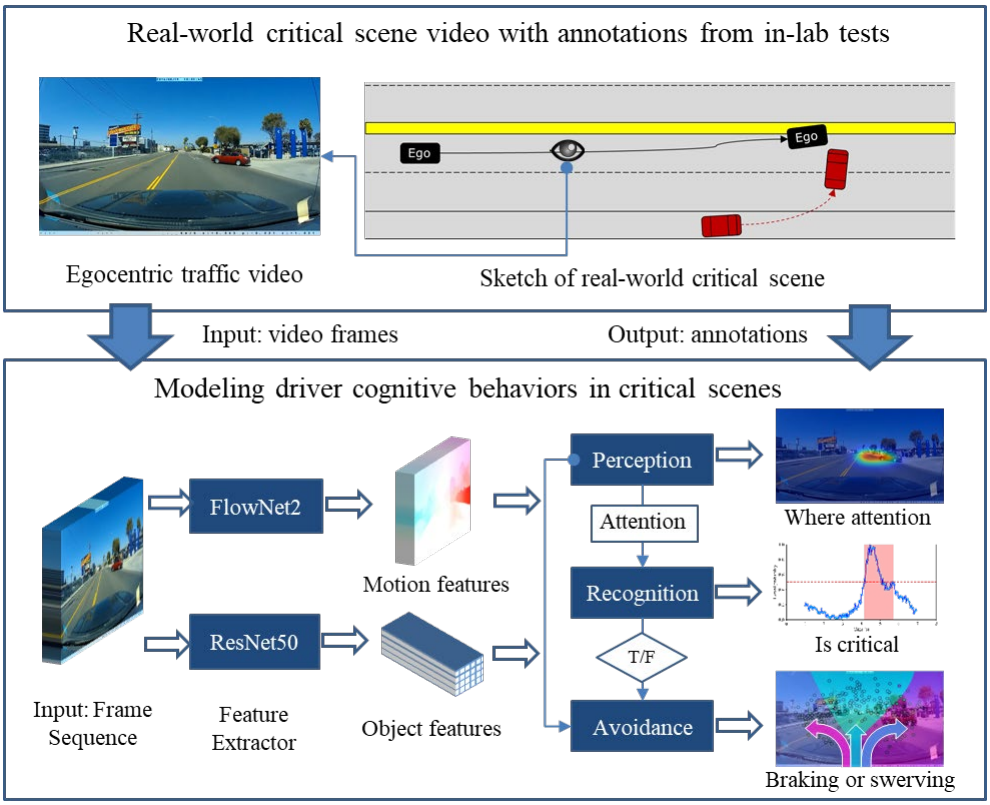

Fig. 1. The framework of modelling cognitive behaviours using large-scale, real-world critical video datasets with human drivers' annotations. (**Hazard perception**: where is the driver's focus of attention in each video frame? **Hazard recognition**: is the video-clip a critical traffic scene? **Hazard avoidance**: would the driver brake or swerve in the critical traffic scene?)

B. Nie (e-mail: nbb@tsinghua.edu.cn; tel: +86-10-6278-8689) is an Associate Professor, S. Gan, Q. Li, Q. Wang and X. Pei are PhD students in Mechanical Engineering, in the School of Vehicle and Mobility at Tsinghua University. T. Cui is a researcher at Chongqing Changan Automobile Company Limited. Affiliations: State Key Laboratory of Automotive Safety and Energy, School of Vehicle and Mobility, Tsinghua University (B. Nie, S. Gan, Q. Li, Q. Wang, X. Pei); State Key Laboratory of Vehicle NVH and Safety Technology, Chongqing, China (B. Nie. S. Gan, T. Cui).

II. METHODS

We established three driver cognitive models (i.e. hazard perception, recognition, and avoidance) via the DL approach, trained by publicly available datasets consisting of large-scale annotated videos. The available traffic video datasets cover diverse critical scenes involving the motorbike, cyclist, truck, and pedestrian. Each critical video involves the events that led to a collision or near-collision. DADA-2000 (driver attention data with driving accident annotation) comprises 2,000 videos with different accident categories and with the driver attention data carefully annotated [7]. DoTA (detection of traffic accident) contains 4,677 videos with temporal (start and end timestamp of the critical event) and spatial (the regions of the hazardous subject in the video frame) annotations [8]. The RHS (road hazard stimuli), published by MIT Agelab, contains 250 normal traffic scene videos and 253 critical scene videos with detailed temporal annotations [9]. In order to validate the model's inferences, the complete physiological signals of driver cognitive behaviours would be needed. As such annotation is absent in the existing datasets, we performed in-lab experiments with a view to building a dataset with complete annotations under critical scenes. In the preliminary in-lab experiments, 50 critical videos were sourced from the RHS dataset with 30 frames per second (fps) and 1280 x 720 pixel resolution. 22 subjects with different driving experiences were recruited and completed all tests. Driver gaze and evasive manoeuvres, captured by the eye-tracker and the driving simulator, were compared with the model inferences in this study.

Object and motion feature extractor: the human visual cortex involves two pathways, i.e. the ventral stream (the perceptual identification of objects) and the dorsal stream (visual guided actions directed at such objects) [10]. Based on these mechanisms, we proposed corresponding two-stream network architecture. Two representative DL models were employed to extract motion and object features, respectively. FlowNet2 could estimate the optical flow (the distribution of apparent velocities of movement of brightness pattern in an image) with high quality [11]. ResNet50 with residual connections and modular structure could facilitate the extraction of object features with different scales [13]. Transfer learning has been widely applied to gain knowledge from previously learned tasks in order to learn new, related tasks [12]. Based on pre-trained models, the object and motion features were extracted as the inputs of the hazard perception, recognition, and avoidance models (Fig. 1.).

Hazard perception model (HPM): HPM is trained to generate the driver gaze distribution maps corresponding to traffic scene video frames. HPM is an ensemble model involving a motion branch and an object branch to process the motion features and object features, respectively (Fig. 2.). The network architecture of both branches comprises a spatiotemporal encoder and a spatial decoder. Each input randomly samples 10 object features and 10 corresponding motion features from the DADA-2000. Driver gaze maps at the same timestamps were treated as the target output of HPM.

Hazard recognition model (HRM): HRM is trained to classify the scene frame sequences and obtain the hazard probability of the video-clip (Fig. 2.). The network architecture of the HRM is also comprised of two branches, consisting of an encoder and a decoder. According to temporal annotations from the DoTA, 10 frames within 2 s after the accident start timestamp were sampled as the critical video-clip and 10 frames within the first 2 s of each video as the normal video-clip. The hazard probability corresponding to the critical video-clip was set as 1, and as 0 for the normal video-clip.

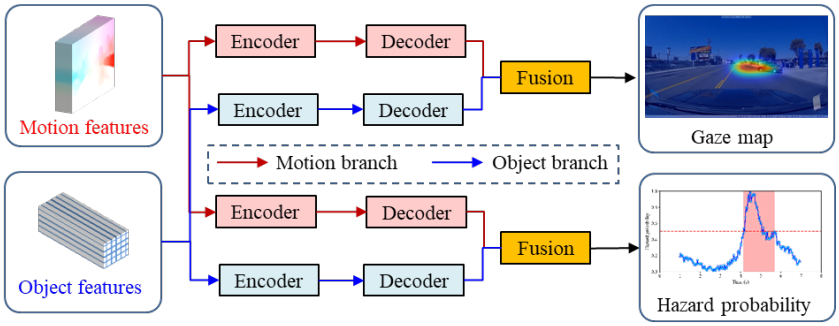

Fig. 2. The two-stream network architecture of HPM and HAM.

***Hazard avoidance model (HAM)*:** HAM is trained to predict the driver's evasive manoeuvres during critical scenes. According to the statistics derived from traffic accident data, in real-world accidents, most drivers (91.9%) swerve towards the same direction as the oncoming obstacle, namely the opposite direction of the obstacle region in the frame [14]. Therefore, we could simplify the HAM as a logic switch based on the output of HPM and HRM. When the hazard probability of HRM inference is greater than the threshold (e.g. 0.5), HAM is triggered and then

determines the swerve direction by the gaze distribution of HPM inference. For instance, the right bias of the gaze distribution indicates that human drivers tend to swerve to the left in this critical scene (Fig. 1.).

III. INITIAL FINDINGS

To quantitatively evaluate the performance of our models, we employed the metrics: KLD (Kullback-Leibler divergence), CC (Pearson's correlation coefficient), and SIM (Similarity) for the HPM and the AUC (Area Under Curve) for HRM. Compared with previous approaches on DADA-2000 and DoTA, we observed a SOTA (new state-of-the-art performance) across the metrics. Specifically, the HPM outperformed the MSAFNet (multi-path semantic-guided attentive fusion network) proposed in DADA-2000 [15] (Table I). Compared with the SOTA approach in DoTA (AUC=78.0), the AUC of HRM reached 87.2.

TABLE I

PERFORMANCE COMPARISON BETWEEN THE PROPOSED HAZARD PERCEPTION AND THE PREVIOUS APPROACHES USING MULTIPLE EVALUATION METRICS

Dataset	KLD↓	CC↑	SIM↑
DADA-2000	2.90	0.34	0.21
Ours	1.63	0.50	0.34

The symbol ↑ estimates a larger value and ↓estimates a smaller value.

A critical video, i.e. No. 013 critical scene from the RHS dataset, was randomly selected to qualitatively compare the human driver's response and the DL model inferences (Fig. 3.). At the beginning of the video, a red sedan on the right side of the road was almost stationary. The first visual indication of a potential hazard was annotated at 3.9 s when the sedan suddenly turned into the driving lane. During the 2 s time window, human drivers relocated their attention from the vanishing point of the road ahead to the turning sedan. Their gaze concentrated on the front end of the sedan, to capture the detailed motion information of the vehicle, until the vehicle disappeared from the frame. Similarly, the gaze maps of HPM inferences significantly shifted to the front end of the sedan during the interval between 3.4 s and 4.4 s. This similarity in attention regions shows that the HPM could locate hazard objects as well as the human driver's visual attention mechanism in the critical scene.

According to our in-lab test results, the mean curves of braking and swerving with standard error after 0-1 normalization demonstrated that the majority of drivers started to take evasive manoeuvres in the time interval between 4 s and 4.5 s, while the red sedan was turning into the driving lane. In the same period, hazard probability increases rapidly and exceeds 0.5 at 4.1 s, reaching the maximum value at 4.5 s. A similar time window interval revealed that the HRM could recognize the hazard pattern in the video-clip as accurately as human drivers.

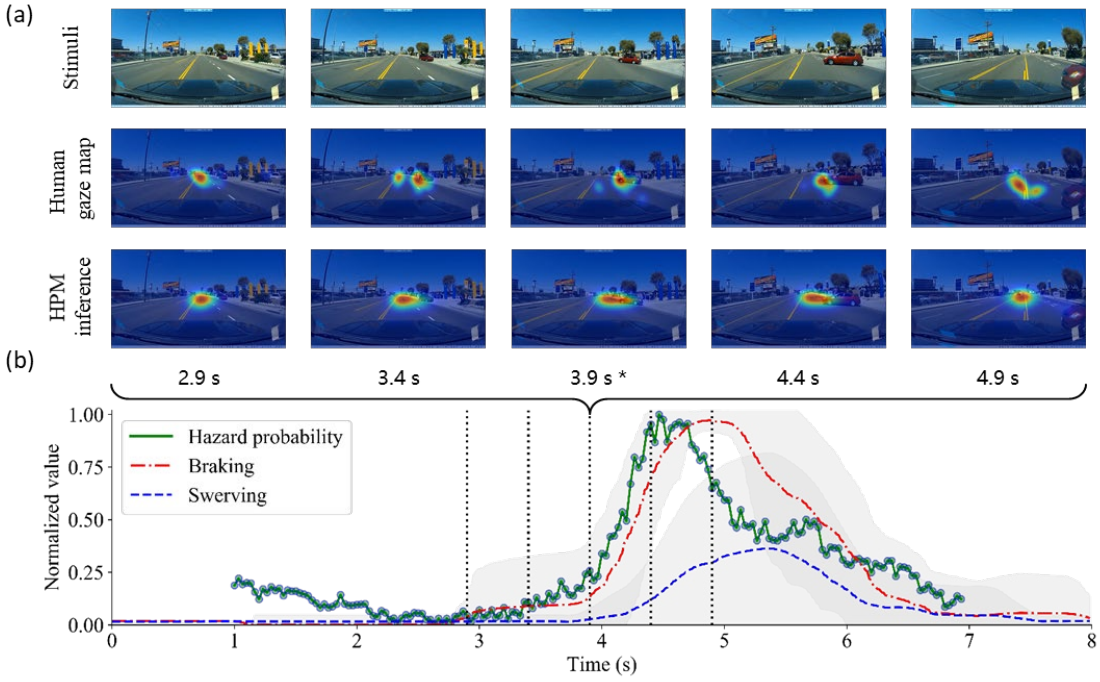

Fig. 3. Human drivers' cognitive behaviours in in-lab test and DL-based inference results: (a) stimuli frames, gaze maps captured from 22 subjects in in-lab experiments, and inference results by HPM at same timestamp; (b) hazard

probability inferred by the HRM, and evasive manoeuvres of human drivers; * denotes the annotated timestamp, first visual indication of a potential hazard; the shaded zone denotes the standard error. Three inference results are available at https://github.com/Shun-Gan/Driver-Cognitive-Behavior.

IV. DISCUSSION

The preliminary results suggest that biologically inspired DL architectures could represent driver cognitive behaviours in diverse critical traffic scenes. A large-scale, real-world critical traffic video dataset with detailed human behaviour annotations could support the DL model training and achieve a higher level of generalization than traditional statistical learning or machine learning models. The cognitive model involving visual attention, traffic scene recognition and evasive manoeuvres could promote understanding of the situation awareness process of human drivers, e.g. the common visual information extraction, the pattern recognition in dynamic scenes, and the decision-making in complicated cognitive behaviours.

As advanced driver assistance systems (ADASs) become widely available on the market, there is an urgent requirement to replicate and predict driver active cognitive behaviours through a generalized representation approach. In safety terms, being able to infer active behaviours in critical traffic scenes contributes to vehicle dynamics prediction controlled by evasive manoeuvres and accurate assessment of injury risk with regard to posture change in pre-crash phases. Based on the inferred state of vehicles and drivers, the humanized decision-making algorithms could be implemented into ADASs to mitigate the injury risk in accident scenes and, eventually, to improve traffic safety.

Several limitations to this study must be noted. First, the available datasets cannot support HAM training through a DL-based approach due to the lack of evasive manoeuvre annotations. A large-scale dataset with complete behaviour annotations should be proposed in subsequent studies. Second, these cognitive behaviour models should integrate with an active human body model with excellent biomechanical characterization to assess injury risk in a digital traffic environment.

V. ACKNOWLEDGEMENTS

This study was in part supported by the National Natural Science Foundation of China (52072216, 51705276) and by the State Key Laboratory of Vehicle NVH and Safety Technology (NVHSKL-202105).

VI. REFERENCES

[1] Nie, B., *et al*., *Traffic Inj Prev*, 2018.
[2] LeCun, Y., *et al., Nature*, 2015.
[3] Itti, L., *et al., IEEE Trans Pattern Anal Mach Intell*, 1988.
[4] Tran, D., *et al., CVPR*, 2018.
[5] Xing, Y., et al., *IEEE trans Intell Transp Syst*, 2019.
[6] Nie, B., *et al., Traffic Inj Prev, 2018.*
[7] Fang, J., *et al., ITSC,* 2019.
[8] Yu, Y., *et al*., https://arxiv.org/abs/2004.03044, 2020.
[9] Wolfe, B., *et al*., *J Exp Psychol Anim Learn Cogn*, 2019.
[10] Goodale, M. A., *et al*., *Trends Neurosci*, 1992.
[11] Ilg, E., *et al*., *CVPR*, 2017.
[12] Pan, S. J., *et al.*, *IEEE Trans Knowl Data Eng*, 2009.
[13] He, K., *et al*., *CVPR*, 2016.
[14] Hu, M., *et al*., *IRCOBI*, 2017.
[15] Fang, J., *et al*., https://arxiv.org/abs/1912.12148, 2019.