Automatic Interaction Detection Between Vehicles and Vulnerable Road Users During Turning at an Intersection

Hao Cheng[1], Hailong Liu[2], Takatsugu Hirayama[3], Fumito Shinmura[2], Naoki Akai[2] and Hiroshi Murase[2]

***Abstract*— Interaction detection between vehicles and vulnerable road users (e.g. pedestrians and cyclists) is important for e.g. safety control and autonomous driving. However, there are many challenges for automatically detecting interactions, such as the ambiguity of defining when interaction is required in dynamic traffic activities among different road users and the lack of labeled data for training a machine learning detector. To overcome the challenges, we introduce a way to define whether or not interaction is required in various traffic scenes and create a large real–world dataset from a very challenging intersection. A sequence-to-sequence method that uses the object information and motion information of the traffic scenes extracted by a state-of-the-art object detector and from optical flow, respectively, is proposed for automatic interaction detection. The proposed method generates a probability of interaction at each short interval ($< 0.1\,s$) that represents the changing of interaction along a sequence. We obtain a baseline model that differentiates no interaction from interaction on the basis of the location and road user type from the detected object information. Compared with the baseline model, the empirical results of the proposed method demonstrate very accurate predictions for vehicle turning sequences with varying length.**

I. Introduction

Statistics show that accidents between vehicles and vulnerable road users (VRUs, e.g. pedestrians and cyclists) often occur at intersections [1], [2]. Monitoring and understanding how vehicles and VRUs interact with each other at intersections in consideration of collision risks and smoothness are critical for safety control, autonomous driving and traffic management [3]. However, manually analyzing the interactions between them based on observations is not feasible for busy intersections on a daily basis. Nowadays, as the ubiquity of traffic data and the development of computer vision techniques, there is a high demand for automatically recognizing interactions from video data, specifically detecting interactions between vehicles and VRUs from various traffic scenes (see Fig. 1), and then differentiating dangerous behavior (e.g. violation of traffic rules) from normal behavior (e.g. following traffic rules and acting out of courtesy).

[1] Hao Cheng is with Institute of Cartography and Geoinformatics, Leibniz University Hannover, Appelstrasse 9A, Hannover 30167, Germany hao.cheng@ikg.uni-hannover.de

[2] Hailong Liu, Fumito Shinmura, Naoki Akai and Hiroshi Murase are with Graduate School of Informatics, Nagoya University, Furo-cho, Chikusa-ku, Nagoya, Aichi, 464-8601, JAPAN. liuh@murase.m.is.nagoya-u.ac.jp and shinmuraf@murase.is.i.nagoya-u.ac.jp and akai@coi.nagoya-u.ac.jp and murase@i.nagoya-u.ac.jp

[3] Takatsugu Hirayama is with Institutes of Innovation for Future Society, Nagoya University, Furo-cho, Chikusa-ku, Nagoya, Aichi, 464-8601, JAPAN. takatsugu.hirayama@nagoya-u.jp

There are a lot of studies for road user detection using computer vision techniques [4], [5], [6]; pedestrian crosswalking prediction [7], [8], [9]; intention detection at intersections [10], [11], [12] and trajectory prediction [13], [14], [15]. Few studies have been done in the regard to traffic interactions. The most relevant studies are that, for example, [16] uses trajectories extracted from videos to automatically analyze vehicle–pedestrian conflicts and [17] proposes a similar framework to diagnose the safety issues in vehicle–bicycle interactions using computer vision analysis. These studies focus on a more severe interactive level (conflicts). Such conflict interactions only account for a small volume of the total traffic activities and may not cover other less severe interactions. They only consider limited road user types, e.g. vehicle–pedestrian or vehicle–bicycle. In real–world traffic situations at big intersections other heterogeneous road users are often involved. On the other hand, the analyses highly depend on the quality of trajectory data. But the acquisition of trajectory data is often time consuming and costly.

To achieve automatic interaction detection between vehicles and VRUs, the following challenges have to be tackled: (1) how to define a boundary to determine if interactions occur between vehicles and VRUs in traffic scenes of very dynamic activities and various behavior patterns among different types of road users; (2) how to efficiently acquire, process, and label a large amount of video data for training a machine learning model; (3) how to handle the changing of interaction along a sequence of varying duration.

With the consideration of the above challenges, we propose an end-to-end method for detecting interactions among heterogeneous road users at intersection directly from traffic video data. In our work, various activities among all road user types were recorded using a camera at an extremely busy left-turning intersection in Japan. Namely, six types of traffic related objects are considered: pedestrians, cyclists, motorcycles, cars, buses and trucks. As the driving direction in Japan is on the left side, in this intersection, even governed by traffic signals VRUs often need to directly interact with left-turning vehicles. Different from the works mentioned above, there is no need to track the trajectories of the road users in the image processing of the recording. We apply a state-of-the-art object detector called M2Det [6] to extract object information and the dense optical-flow algorithm [18] to extract their motion information in consecutive frames. Our contributions are summarized as follows:

1 Providing a definition for interaction regarding a vehicle left-turning motion with VRUs at an extremely busy intersection in a Japanese metropolis and creating a

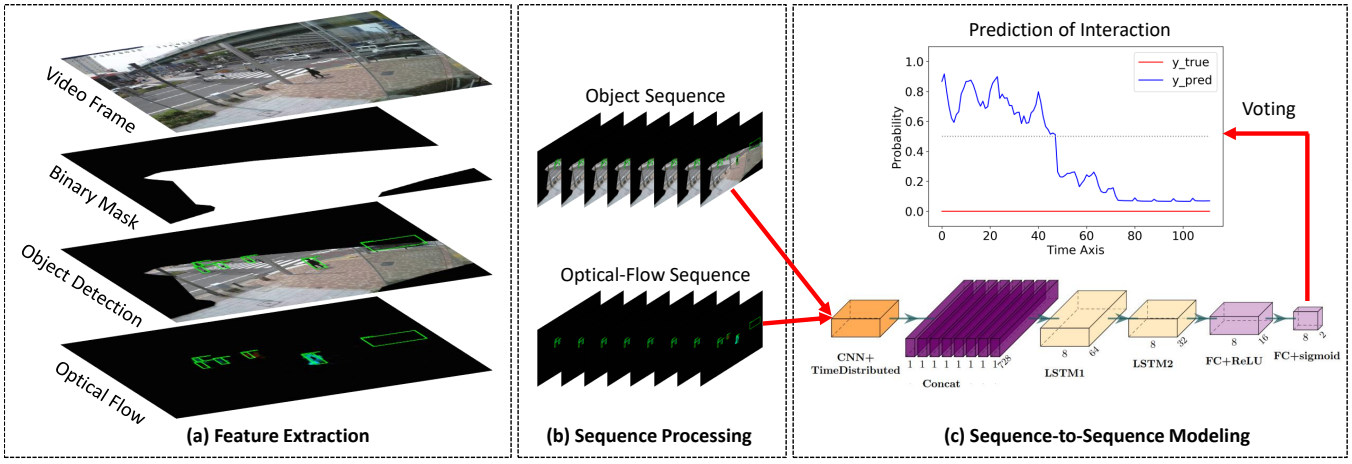


Fig. 1: Automatically detecting interaction between vehicles and vulnerable road users at an intersection using sequence-to-sequence modeling (b) (c) with features extracted by object detection and optical flow from video data (a). At each frame, a probability is generated to predict the interaction and the sequence-level prediction is the voted summation of the frame-level predictions. A binary mask is applied to filter the traffic data out of the left-turning intersection.

large interaction dataset with sequence-level labels;

2 Proposing a sequence-to-sequence method that can automatically detect interaction between vehicles and VRUs at a sequence-level using the object information and motion information extracted from video data;

3 Constructing a hybrid model that uses the road users' location and type from object detection to prevent false positive interaction detection (i.e. no VRUs on site during the vehicle left-turning) in the sequence-to-sequence method.

In this work, several ablative models are carried out to verify the contribution of the object information and the motion information. With the consideration of the dynamics of interactions over time at a short interval ($< 0.1\,s$), three voting schemes are investigated to analyze the impact of the interval interactions on the overall interaction.

II. Methodology

Our framework consists of three major components for detecting interaction between vehicles and VRUs: Feature Extraction, Sequence Processing, and Sequence-to-Sequence Modeling, see Fig.1.

A. Feature Extraction

In the feature extraction component, object information and their motion information are extracted in parallel for each road user at the given intersection from the video data.

The object information contains road users' type and their location. To acquire object information, a state-of-the-art object detector called M2Det that uses multi-level feature pyramid network for detecting objects of different scales [6] is leveraged for detecting all the aforementioned road users at each frame. We use six channels to store the road user type information and each channel is dedicated to one of the types. The location of the detected objects is mapped on each frame using the approximate mass point (the lower middle point of the bounding box).

Without tracking the objects detected from the previous frame to the next frame, the object information alone is not adequate for interaction detection between vehicles and VRUs. Nevertheless, tracking multiple objects from frame to frame is very challenging due to, e.g. abrupt object motion, change appearance and occlusions [19]. Most importantly, the failure of tracking might directly lead to a failure of interaction detection for the objects involved. To circumvent the challenges, we use optical flow to capture the motion of road users. Optical flow describes the distributions of velocities of moving pixels in two consecutive images [18]. In other words, moving objects can be captured by optical flow and static objects or background will be ignored. We use the dense optical-flow algorithm to map the displacement of moving objects and remove the static background information [20]. The orientation of motion is encoded by color and the velocity is encoded by color intensity.

B. Sequence Processing

Following the feature extraction component, the object information and the motion information are fed into a sequence-to-sequence model for interaction detection. Due to the very different duration of each sequence (see Fig. 4), e.g. some vehicles may have to wait for a long time for pedestrians crossing the street, the sequence-to-sequence model should be able to cope with various sequence lengths. To tackle this problem, we resort to a similar approach proposed in [21], [22] that uses a sliding window to divide long sequences into shorter sequences to capture both long and short interactions. We name the sub-sequences divided by the sliding window from a complete sequence as clips.

In order to train a fully sequence-to-sequence model, the sequence-level label is duplicated and associated with each frame in every clip. The modeling problem is defined as: for clip n received $\mathbf{X}^n = \{X_1^n, \cdots, X_w^n\}$ as input of the object information and motion information, and predict the

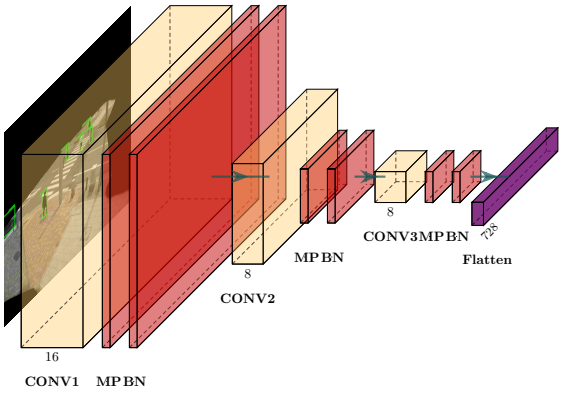


Fig. 2: The CNN in the pipeline for learning latent features from object and optical flow sequence at each frame.

interaction at each frame $\hat{\mathbf{Y}}^n = \{\hat{Y}_1^n, \cdots, \hat{Y}_w^n\}$, where n stands for the index of the clip for the given sequence and $n \in N$, N is the total number of clips, and w stands for the sliding window size. For short,

$$\hat{\mathbf{Y}}^n = f(\mathbf{X}^n). \tag{1}$$

C. Sequence-to-Sequence Modeling

The sequence-to-sequence function $f(\cdot)$ defined by Eq. (1) manipulates both convolutional neural network (CNN) and Long Short-Term Memories (LSTMs) neural networks for learning spatiotemporal features from object information and motion information. The right part of Fig. 1 denotes the overall structure of the pipeline.

At the first step, each frame of the extracted information is passed through the CNN for learning the spatial features. The CNN has three layers with each layer followed by a Maximum Pooling (MP) and a Batch Normalization (BN). The output of the CNN is a flattened feature vector of each frame. Fig. 2 demonstrates the detailed structure of the CNN.

In the following step, all the parsed frames in a given clip are timely distributed and concatenated as the input for the following two LSTMs, which are used to learn the temporal features across the frames.

Finally, the output of the LSTMs are firstly passed through a fully connected layer (FC) with a rectified linear unit (ReLU) activation function and then another FC with a sigmoid activation function. The final output is a probability vector for interaction at each frame, see Fig. 1.

The sequence-to-sequence function $f(\cdot)$ is trained by optimizing the binary cross-entropy loss between the duplicated ground-truth label and the predicted label at each frame.

As the ground-truth label only represents the sequence-level interaction, the interaction at each frame can change and may differ from the ground-truth label. Therefore, the sequence-level prediction is the voted summation of the frame-level predictions, denoted by Eq. (2), where $\delta(t)$ stands for the voting function regarding frame index t and $T = Nw$, which is the total number of frames for the given sequence. We use $\hat{Y}^*$ to denote the voted sequence-level prediction for the whole sequence.

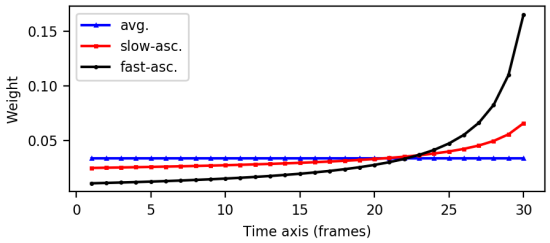


Fig. 3: Voting scheme in relation to time axis, e.g. a vehicle left-turning sequence in 30 frames.

$$\hat{Y}^* = \sum_{n=1}^{N} \sum_{t=nw-w+1}^{nw} \delta(t)\hat{Y}_t. \tag{2}$$

Critical moments often happen when road users come close to each other in time and more attention is required [23]. Hence, we explore different ways to emphasize the impact from the later frames. Besides using an average voting scheme, we investigate two other voting schemes that slowly and quickly increase the weights along the time axis. We call them as *slow-asc.* and *fast-asc.*, see Eq. (3).

$$\delta(t) = \begin{cases} 1/T & \text{avg.} \\ \frac{t/\log(t+e)}{\sum_{t=1}^{T} t/\log(t+e)} & \text{slow-asc.} \\ \frac{1/(T-t+1)}{\sum_{t=1}^{T} 1/(T-t+1)} & \text{fast-asc.} \end{cases} \tag{3}$$

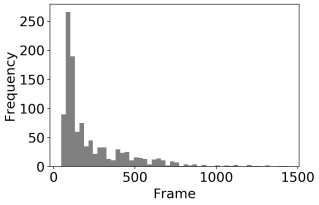
The average voting scheme weights each frame-level prediction equally, and the slow-ascending and fast-ascending voting schemes increase the weights gradually and quickly along the time axis, respectively, see Fig. 3.

III. Dataset

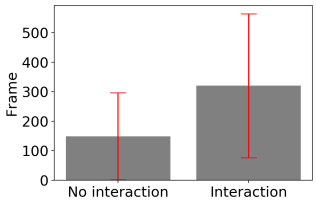
The concept of *interaction* with VRUs is similar to *conflict* [24]. While interaction can range from collision to negligible conflict risk [25]. We define interaction that occurs as if the left-turning vehicle drives in an intersection and any VRUs approach or move on the intersection space, in order to avoid any conflicts that might happen at any time during the vehicle's turning, they adapt their movement (velocity and orientation) accordingly. Otherwise, if the target vehicle drives in an undisturbed manner with VRUs in its neighborhood (if there are any), interaction is not required. It is worth mentioning that in the first step of our study, the interaction for car following and road users' situation awareness are not considered at the moment via observations from static camera. They will be considered in our future work.

To test the interaction detection method, we created a large real-world interaction dataset from an extremely busy intersection in a Japanese metropolis. We recorded approximately 24 hours of traffic footage from an oblique view at one of the major intersections in Nagoya from 11 a.m. 23th to 11 a.m. 24th, April 2019. The video was taken at 1600 × 1200 pixels at 30 fps using a camera[1] installed

[1]Panasonic WV-SF781L



(a) The distribution of sequence length



(b) The sequence lengths of no interaction vs. interaction

Fig. 4: Sequence lengths of the dataset

TABLE I: The distribution of interactions in each set

Class	Training	Validation	Test	Total
Interaction	328	118	98	544
No interaction	328	88	137	553

inside a building adjacent to the intersection and recorded in H.264 format. In this work, we used a 12-hour sub-footage from 11 a.m. to 6 p.m. and from 6 a.m. to 11 p.m. when there was enough traffic and ambient light to perform stable image processing. Two annotators manually detected 1097 scenes where vehicles turned left and extracted the time intervals from when the vehicles entered the intersection until it exited. However, it is difficult to determine whether the interaction actually occurs from external observations. They also subjectively determined whether or not interactions required in each scene, agreed with each other, and labeled each scene "no interaction" or "interaction". We obtained 544 sequences that require interactions and 553 sequences that require no interaction. Without sacrificing the data quality, the video data has been down-sampled from the original frame rate to $15\,\text{fps}$ (the interval is less than $0.1\,s$) with the consideration of a stable performance for optical flow and reducing the computational cost. The processed data will be shared for further research purpose.

The dataset is very challenging for automatic interaction detection between vehicles and VRUs. Due to the varying traffic situations, the length of the left-turning sequences has a very long-tail distribution: ranging from 40 frames to 1500 frames, see Fig. 4a. In addition, the sequence lengths of requiring no interaction (Mean $=$ 148.6, std. $=$ 147.3) are significantly shorter (MannWhitney U-Test, $N =$ 65863, p $<<$ 0.01) than the ones requiring interaction (Mean $=$ 320.1, std. $=$ 243.8). Because of the wide variation of sequence lengths in both classes, shorter sequences do not necessarily indicate no interaction required, see Fig. 4b.

We split the dataset into training, validation and test sets for training and evaluating the performance of the proposed method. The training set has balanced samples for an unbiased training, while the validation and test sets are slightly skewed in a reversed way for a more challenge validation setting. Table I lists the statistics for each set.

IV. Experiment

A. Evaluation Metrics

We apply *Precision*, *Recall*, *F1* score and *Accuracy* to measure the performance of the interaction detection for the two classes on the test set. Given the unbalanced number of sequences of no interaction and interaction in test set, we report the weighted average values for all the measurements.

TABLE II: The detection results for the models using different information and voting schemes

Model	Input	Voting	Prec.	Recall	F1	Accu.
baseline	m2det	—	0.791	0.583	0.536	0.583
s+m2-avg	m2det	avg.	0.767	0.728	0.728	0.728
s+m2-slow	m2det	slow-asc.	0.760	0.736	0.738	0.736
s+m2-fast	m2det	fast-asc.	0.783	0.779	0.780	0.779
s+op-avg	op.	avg.	0.898	0.885	0.886	0.885
s+op-slow	op.	slow-asc.	0.897	0.889	0.890	0.889
s+op-fast	op.	fast-asc.	0.860	0.860	0.860	0.860
s+m2+op-avg	m2det+op.	avg.	0.917	0.911	0.911	0.911
s+m2+op-slow	m2det+op.	slow-asc.	0.926	0.919	0.920	0.919
s+m2+op-fast	m2det+op.	fast- asc.	0.865	0.864	0.862	0.864
hybrid	m2det+op.	slow-asc.	**0.929**	**0.923**	**0.924**	**0.923**

B. Experiment Setting

In all the experiments, after applying a mask (see Fig. 1) to filter other traffic out of the intersection and resizing, the frame size of object information is $(200, 100, 6)$ and the frame size of motion information is $(200, 100, 3)$; The sliding window size for each clip is set to 8; The three-layer CNN has kernel sizes $(4, 4, 4)$ and output dimensions $(16, 8, 8)$; the hidden units of the LSTMs are set to 64 and 32, respectively; The output dimension of the first FC is set to 16.

C. Compared Methods and Hybrid Model

We implement a *baseline* model that detects interaction based on the co-existence of VRUs. If the left-turning vehicle encounters no VRU, there may be no interaction required. Otherwise, interaction is required. It is obvious that the baseline model is $100\,\%$ accurate for detecting no interaction sequences. This leads us to build a hybrid model: calibrating the best sequence-to-sequence model using the baseline model only for sequences with no VRU, if there are any such false positive cases.

Ablative models are carried out to verify the contribution of the object information extracted by object detection using M2Det [6] and motion information extracted by optical flow [18], see Section II-A. The *s+m2* models only take object information as input and the *s+op* models only take motion information as input. The proposed sequence-to-sequence *s+m2+op* models take both object and motion information. Those two types of information are channel-wisely combined. Meanwhile, *avg.*, *slow-asc.* and *fast-asc.* voting schemes (see Section II-C) are compared across all the sequence-to-sequence models.

D. Quantitative Results

Table II lists the detection results for the models using different information and voting schemes. Besides, Table III lists the confusion matrices for the baseline, the sequence-to-sequence models with their respective best voting scheme as well as the hybrid model.

TABLE III: The confusion matrices for the baseline, the sequence-to-sequence models with their respective best voting scheme, as well as the hybrid model. True positive (top left) and true negative (bottom right) values are in boldface.

Class	*baseline*		*s+m2-fast*		*s+op-slow*		*s+m2+op-slow*		*hybrid*	
Int.	**98**	0	**76**	22	**92**	6	**95**	3	**95**	3
No int.	98	**39**	30	**107**	20	**117**	16	**121**	15	**122**

Compared with the baseline model, profound improvement measured by Recall, F1 and Accuracy can be found by the sequence-to-sequence models with all the voting schemes. It can be seen that, according to the confusion matrices, the baseline model correctly detects all the sequences that contain no VRU as no interaction required. On the other hand, the baseline model cannot differentiate the sequences that require no interaction even though some VRUs co-exist with the left-turning vehicle, e.g. pedestrians standing far away on the sidewalk. However, the *s+m2* models that only take the object information as input fall behind the baseline model measured by Precision. Without tracking, no continuous information can be leveraged to deduce the motion of the same object from frame to frame. Moreover, there is a trade-off between high-confidence and failure of the object detection. The upper-bound threshold for detection confidence is set to 0.45 to make sure that the left-turning vehicle is always detected at each frame. This often leads to noisy detection such as wrong type of detected objects, type-swapping and no detection (see Fig. 5). Whereas, the performance for the *s+op* models that take motion information as input is significantly better than the baseline and *s+m2* models by all measurements. The motion information extracted by optical flow captures the dynamics in the interaction between vehicles and VRUs. When using both the object information and motion information, the performance for sequence-to-sequence models is further enhanced.

Across different voting schemes by using the same input information, the ones that put more emphasis on the later frames (*s+m2-fast*, *s+op-slow*, *s+m2+op-slow*) generate more accurate detection and in general, slow-asc. performs better than the fast-asc. voting scheme.

One interesting observation is the comparison between the best sequence-to-sequence model *s+m2+op-slow* and the baseline model. From the confusion matrices we can see that, 82 false positive out of a total of 98 cases detected by the baseline have been correctly classified by *s+m2+op-slow*. It further proves that the co-existence of vehicles and VRUs do not necessarily require immediate interaction, especially when they are far away from each other or remain static. Nevertheless, one of the no-interaction case that involves no VRU is wrongly classified as interaction by *s+m2+op-slow*. Even though it has very high accuracy for interaction detection in situations with no VRU, it is not optimal in this regard. Therefore, we combine the baseline model with the *s+m2+op-slow* into a hybrid model. In the end, the hybrid model prevents the aforementioned false positive interaction detection and increases the accuracy to 0.923.

To sum up, (1) from the ablation study, the object information extracted by M2Det alone is not adequate for interaction detection; (2) Motion dynamics between vehicles and VRUs can be captured by optical flow; (3) The combined information from both M2Det and optical flow using the slow-ascending voting scheme further boosts the performance; (4) False positive interaction detection with no VRU involved can be easily prevented by the combination of the sequence-to-sequence method with the baseline model.

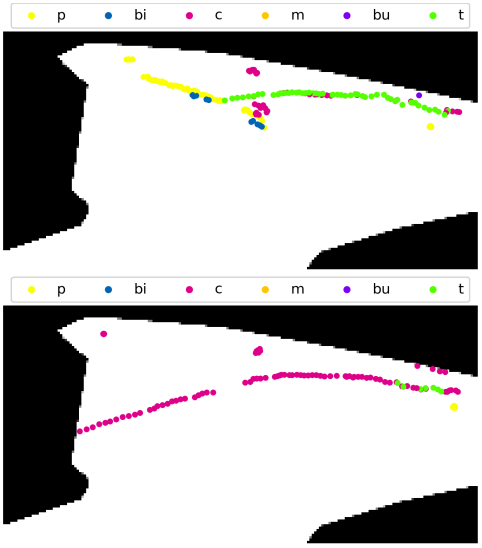


Fig. 5: Examples of M2Det [6] detection over left-turning vehicle sequences denoted by p: pedestrian, bi: bicycle, c: car, m: motorcycle, bu: bus and t: truck. Some steps in the upper figure of bicycle detection are missing and in the lower figure the types of car and truck are swapped.

E. Qualitative Results

Scenarios of different interactions and the corresponding predictions using *s+m2+op-slow* are visualized from Row-1 to Row-4 in Fig. 6. The relevant VRUs and the target vehicle are marked by red and blue bounding boxes, respectively.

In Row-1, several VRUs interact with the left-turning vehicle from frame to frame. The prediction at each frame points to a higher probability (above 0.5) of interaction.

In Row-2, several VRUs approach the left-turning vehicle from a relatively long distance at the beginning. But they close up the gap quickly. The probability of interaction increases as the distance decreases and then stays at a higher level after the vehicle moves close to the VRUs. *s+m2+op-slow* correctly predicts the sequence as interaction by summing up the frame-level predictions.

In Row-3, the vehicle is turning left with no other VRU cross-walking. The prediction at each frame stays at a very low level for interaction.

Row-4 shows the target vehicle behind the leading vehicle approaching the zebra zone which has already been occupied by a cyclist. The probability of interaction is above 0.5 at early frames since both the leading and target vehicles are relatively close to the cyclist. As the cyclist reaches the end of the crossing, the prediction at each frame points to a lower probability of interaction. Throughout the whole sequence,

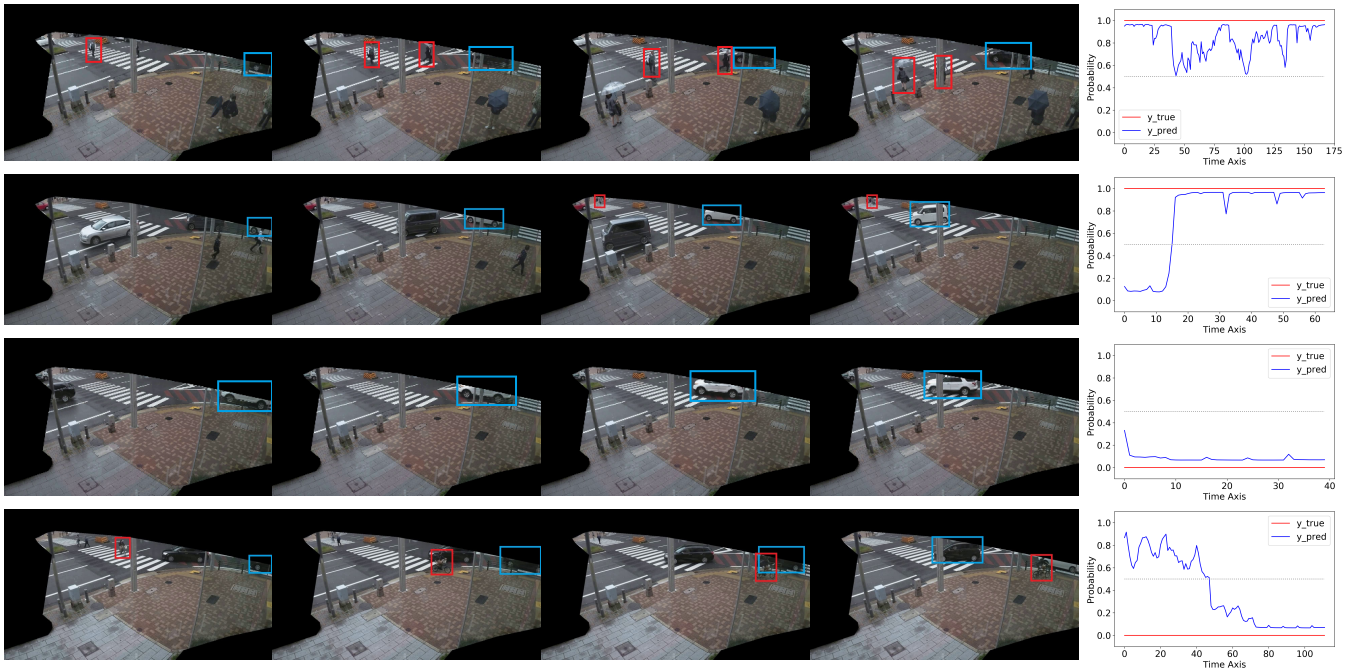


Fig. 6: Scenarios of different interactions and the corresponding predictions denoted by each row: Row-1 and Row-2 denote two vehicle left-turning sequences that require interaction and Row-3 and Row-4 denote two vehicle left-tuning sequences that require no interaction. The relevant VRUs are marked by red bounding boxes and the target vehicle is marked by a blue bounding box. Please note that car-flowing interaction is not considered in this study.

however, there is no direct interaction required between the target vehicle and the cyclist. The summation of the frame-level predictions using the slow-ascending voting scheme correctly predicts this sequence as no interaction.

In summary, from the visualization, it showcases the ability of *s+m2+op-slow* for capturing the interaction dynamics at the short intervals. Even though the sequence-level labels can only represent the overall interaction, the predictions at each frame provide more interactive and accurate information about the changing of the interaction.

V. Discussion

In this section, we will discuss the false detection from the hybrid model. In total, there are 18 cases (7.7 %) out of the test set are wrongly classified, of which 3 are false negative detection (ground truth interaction but prediction no interaction) and 15 are false positive detection (ground truth no interaction but prediction interaction), see Table III.

All the false negative cases are ambiguous for interaction detection even for humans. Namely, in one case the left-turning vehicle reaches the intersection while a pedestrian is approaching the zebra zone (see Fig. 7a), and in the other two similar cases the left-turning vehicle moves in the intersection while a pedestrian is leaving the zebra zone (see Fig. 7b). It is difficult to tell whether the interaction actually occurs from the external observations. In such situations, sensor information might be required to detect the road users' attention (e.g. eye gaze) and situation awareness [26].

Among the false positive cases, most of them are caused by car-following interaction. Direct interaction is required between the leading vehicle and the VRUs involved. The

(a) A pedestrian approaches the zebra zone and the car enters the intersection

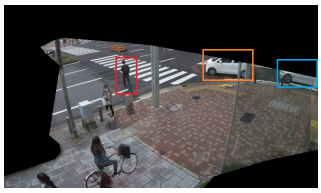
(b) A pedestrian is leaving the zebra zone to be chased by the target car

Fig. 7: False negative cases (ground truth interaction but prediction no interaction) by the hybrid model

target vehicle follows up when the leading vehicle exits the intersection. However, there is often no direct interaction required from the waiting target vehicle with the VRUs (see Fig. 8). Without the consideration of the interaction between vehicles, the hybrid model may have difficulties for interaction detection in the vehicle waiting scenarios. This will be our future work to include interaction between vehicles in vehicle turning sequences.

VI. Conclusion and Future Work

In this paper, we propose a sequence-to-sequence method to automatically detect interaction between vehicles and VRUs at an intersection using video data. A state-of-the-art object detector and optical flow are used to extract object information and motion information, respectively. We create a large dataset of various traffic activities from an extremely busy intersection to evaluate the proposed method. Our method detects interactions in vehicle turning sequences of varying length with an accuracy over 92 % and generates



(a) Interaction required between the leading car (in orange bounding box) and the pedestrian (in red bounding box)

(b) No interaction required as the pedestrian has finished crossing when the target car (in blue bounding box) enters the intersection

Fig. 8: False positive case (ground truth no interaction but prediction interaction) by the hybrid model

a probability of interaction at each frame less than $0.1\,\mathrm{s}$ to represent the dynamics of interaction along a sequence. Via an ablation study we prove that, by avoiding timely and costly work of tracking, the object information and motion information are very beneficial for interaction detection. The comparison of three voting schemes for summing up the frame-level predictions indicates that interaction often changes over time even in the same sequence and later frames weight more in the overall interaction. A hybrid model is used to prevent wrong detection that involves no VRUs to further boost the performance of interaction detection.

In our next step, we will include the interactions between vehicles during the turning and extend the sequence-to-sequence method for classifying interaction patterns at each interval, e.g. differentiating the interaction levels based on collision risks. In comparison to the hard-coded voting scheme investigated in this study, the attention mechanism [27] will be leveraged to automatically learn the location and timing when such interaction patterns occur. More traffic data will be acquired from different interactions for training the proposed model and testing its ability for unseen scenarios.

ACKNOWLEDGMENTS

This work is a collaboration during the first authors research stay in Murase Lab at Nagoya University and supported by Nagoya Toyopet Corporation. The stay is funded by the German Research Foundation (DFG) through the Research Training Group SocialCars (GRK 1931).

REFERENCES

[1] E.-H. Choi, "Crash factors in intersection-related crashes: An on-scene perspective," Tech. Rep., 2010.

[2] A. Habibovic and J. Davidsson, "Requirements of a system to reduce car-to-vulnerable road user crashes in urban intersections," *Accident Analysis & Prevention*, vol. 43, no. 4, pp. 1570–1580, 2011.

[3] M. S. Shirazi and B. T. Morris, "Looking at intersections: A survey of intersection monitoring, behavior and safety analysis of recent studies," *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, no. 1, pp. 4–24, 2016.

[4] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Region-based convolutional networks for accurate object detection and segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 1, pp. 142–158, 2015.

[5] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[6] Q. Zhao, T. Sheng, Y. Wang, Z. Tang, Y. Chen, L. Cai, and H. Ling, "M2det: A single-shot object detector based on multi-level feature pyramid network," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 9259–9266.

[7] J. Hariyono and K.-H. Jo, "Detection of pedestrian crossing road: A study on pedestrian pose recognition," *Neurocomputing*, vol. 234, pp. 144–153, 2017.

[8] A. Rasouli, I. Kotseruba, and J. K. Tsotsos, "Are they going to cross? a benchmark dataset and baseline for pedestrian crosswalk behavior," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 206–213.

[9] F. Camara, O. Giles, R. Madigan, M. Rothmüller, P. H. Rasmussen, S. Vendelbo-Larsen, G. Markkula, Y. M. Lee, L. Garach, N. Merat *et al.*, "Predicting pedestrian road-crossing assertiveness for autonomous vehicle control," in *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2018, pp. 2098–2103.

[10] M. Goldhammer, M. Gerhard, S. Zernetsch, K. Doll, and U. Brunsmann, "Early prediction of a pedestrian's trajectory at intersections," in *16th International IEEE Conference on Intelligent Transportation Systems (ITSC 2013)*. IEEE, 2013, pp. 237–242.

[11] Y. Hashimoto, Y. Gu, L.-T. Hsu, and S. Kamijo, "A probabilistic model for the estimation of pedestrian crossing behavior at signalized intersections," in *2015 IEEE 18th International Conference on Intelligent Transportation Systems*. IEEE, 2015, pp. 1520–1526.

[12] S. Koehler, M. Goldhammer, S. Bauer, S. Zecha, K. Doll, U. Brunsmann, and K. Dietmayer, "Stationary detection of the pedestrian? s intention at intersections," *IEEE Intelligent Transportation Systems Magazine*, vol. 5, no. 4, pp. 87–99, 2013.

[13] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese, "Social lstm: Human trajectory prediction in crowded spaces," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 961–971.

[14] A. Gupta, J. Johnson, L. Fei-Fei, S. Savarese, and A. Alahi, "Social gan: Socially acceptable trajectories with generative adversarial networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2255–2264.

[15] A. Vemula, K. Muelling, and J. Oh, "Social attention: Modeling attention in human crowds," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 1–7.

[16] K. Ismail, T. Sayed, N. Saunier, and C. Lim, "Automated analysis of pedestrian–vehicle conflicts using video data," *Transportation research record*, vol. 2140, no. 1, pp. 44–54, 2009.

[17] T. Sayed, M. H. Zaki, and J. Autey, "Automated safety diagnosis of vehicle–bicycle interactions using computer vision analysis," *Safety science*, vol. 59, pp. 163–172, 2013.

[18] B. K. Horn and B. G. Schunck, "Determining optical flow," *Artificial intelligence*, vol. 17, no. 1-3, pp. 185–203, 1981.

[19] A. Yilmaz, O. Javed, and M. Shah, "Object tracking: A survey," *Acm computing surveys (CSUR)*, vol. 38, no. 4, p. 13, 2006.

[20] G. Farnebäck, "Two-frame motion estimation based on polynomial expansion," in *Scandinavian conference on Image analysis*. Springer, 2003, pp. 363–370.

[21] H. Liu, T. Taniguchi, Y. Tanaka, K. Takenaka, and T. Bando, "Visualization of driving behavior based on hidden feature extraction by using deep learning," *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, no. 9, pp. 2477–2489, 2017.

[22] W. Sultani, C. Chen, and M. Shah, "Real-world anomaly detection in surveillance videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6479–6488.

[23] D. N. Lee, "A theory of visual control of braking based on information about time-to-collision," *Perception*, vol. 5, no. 4, pp. 437–459, 1976.

[24] S. R. Perkins and J. L. Harris, "Traffic conflict characteristics-accident potential at intersections," *Highway Research Record*, no. 225, 1968.

[25] T. Sayed and S. Zein, "Traffic conflict standards for intersections," *Transportation Planning and Technology*, vol. 22, no. 4, pp. 309–323, 1999.

[26] F. Schewe, H. Cheng, A. Hafner, M. Sester, and M. Vollrath, "Occupant monitoring in automated vehicles: Classification of situation awareness based on head movements while cornering," in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 63, no. 1. SAGE Publications Sage CA: Los Angeles, CA, 2019, pp. 2078–2082.

[27] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.