TSINGHUA SCIENCE AND TECHNOLOGY
ISSN 1007-0214 12/16 pp599–609
DOI: 10.26599/TST.2020.9010061
Volume 27, Number 3, June 2022

Spatial-Temporal ConvLSTM for Vehicle Driving Intention Prediction

He Huang, Zheni Zeng, Danya Yao, Xin Pei*, and Yi Zhang

Abstract: Driving intention prediction from a bird's-eye view has always been an active research area. However, existing research, on one hand, has only focused on predicting lane change intention in highway scenarios and, on the other hand, has not modeled the influence and spatiotemporal relationship of surrounding vehicles. This study extends the application scenarios to urban road scenarios. A spatial-temporal convolutional long short-term memory (ConvLSTM) model is proposed to predict the vehicle's lateral and longitudinal driving intentions simultaneously. This network includes two modules: the first module mines the information of the target vehicle using the long short-term memory (LSTM) network and the second module uses ConvLSTM to capture the spatial interactions and temporal evolution of surrounding vehicles simultaneously when modeling the influence of surrounding vehicles. The model is trained and verified on a real road dataset, and the results show that the spatial-temporal ConvLSTM model is superior to the traditional LSTM in terms of accuracy, precision, and recall, which helps improve the prediction accuracy at different time horizons.

Key words: driving intention prediction; lane change intention; ConvLSTM

1 Introduction

Driving intention prediction plays an important role in driving safety. In the real world, the vehicle's driving state, particularly its speed and position, changes rapidly. Predicting the driving intention of a vehicle can effectively predict its future state, which helps make correct decisions.

In terms of predicted scenarios, most of the research on vehicle driving intentions focused on highway scenarios[1–3], high-speed entrance and exit ramp scenarios[4], and urban intersection scenarios[5] using the Next Generation Simulation (NGSIM) dataset[6]. This study extends to urban road scenarios to predict vehicle driving intentions. BDD100K[7] is used as the original dataset to extract vehicle driving data from urban road scenarios. In terms of prediction results, most of the research focused on the vehicle's lane change intention, i.e., whether to change lanes or further refine it into one of three categories, namely, maintaining, changing lanes to the left, and changing lanes to the right. Driving intentions in urban road scenarios, such as acceleration and deceleration of vehicles, are frequent and important. We divide a vehicle's typical intention into lateral and longitudinal intentions. Lateral intentions include lane changing to the left and right. Longitudinal intentions include holding, sharp acceleration, sharp deceleration, and stopping. We use information from a vehicle's historical motion and surrounding environment to predict the lateral and longitudinal intentions simultaneously.

With the development of machine learning, early researchers used support vector machine (SVM) for predictive modeling[8, 9]. Vehicle motion state is used as the input to predict whether a vehicle will change lanes. Because the SVM model is generally used to classify

• He Huang, Danya Yao, Xin Pei, and Yi Zhang are with Department of Automation, Tsinghua University, and also with National Laboratory for Information Science and Technology (TNList), Tsinghua University, Beijing 100084, China. E-mail: huangh14@mails.tsinghua.edu.cn; yaody@tsinghua.edu.cn; peixin@tsinghua.edu.cn; zhyi@mail.tsinghua.edu.cn.
• Zheni Zeng is with Department of Computer Science, Tsinghua University, and also with National Laboratory for Information Science and Technology (TNList), Tsinghua University, Beijing 100084, China. E-mail: zzn20@mails.tsinghua.edu.cn.
* To whom correspondence should be addressed.
Manuscript received: 2020-11-19; revised: 2020-12-07; accepted: 2020-12-14

high-dimensional vectors to two results, the results depend on the construction of features. Thus, effective features need to be manually constructed, which is a difficult task. Subsequently, neural networks were gradually applied to classification problems, which can extract effective features internally. Dou et al. used multilayer perceptron (MLP) to predict lane change intentions at highway lane drops[10]. Because MLP networks use vectors as input, capturing the sequential relations in time series data is insufficient. To predict the intention of the target vehicle, historical data of the vehicle need to be used, thus requiring the model to be able to process time series data. Liu et al. proposed the use of the hidden Markov model, which is based on probability graph theory, to predict driving intention[11]. Xing et al. proposed the use of a bidirectional long short-term memory (LSTM) network to predict the vehicle's lane change intention[12]. Bidirectional LSTM could take into account the dependence on long-term relationships. Girma et al. used a deep bidirectional LSTM with an attention mechanism model based on a hybrid-state system framework to predict driving intention at intersections[13]. Mozaffari et al. reviewed the latest developments in vehicle driving intention prediction research from three perspectives, namely, input representations, output results, and prediction algorithms[14]. With the further use of deep neural networks, Dai et al. used a modified LSTM network to deal with multidimensional time series data for trajectory prediction[15]. Su et al. used LSTM modeling to predict lane change intentions in highway scenarios[1]. They considered the influence of surrounding vehicles but concatenated the information of the target vehicle with the information of surrounding vehicles into high-dimensional vectors as input[1]. However, this approach did not help model the influence of surrounding vehicles in a clear and specific manner. The vehicles surrounding the target vehicle are also in a state of rapid change during driving. We analyzed the influence of surrounding vehicles from a macroscopic perspective, wherein the surrounding vehicles are regarded as a whole. Vehicles in different positions within the whole exhibit spatial interactions. At the same time, the whole exhibits time dependence. We propose the use of the convolutional long short-term memory (ConvLSTM) network to model the spatial interactions of surrounding vehicles, capture the time dependence of surrounding vehicle motion states, and obtain a better representation of the target vehicle and environmental information, which can be used to predict future intentions. The contributions of this study are as follows. (1) We extend the application scenarios to urban road scenarios to predict both lateral and longitudinal intentions simultaneously. (2) The proposed model not only considers the motion information of the target vehicle itself but also models the influence and spatiotemporal relationship of surrounding vehicles from a macroscopic perspective, considering the spatiotemporal interactions of surrounding vehicles. (3) This study elaborates the relationship between the accuracy of prediction results and different time horizons.

In Section 2, we introduce related research, including the differences between detection and prediction and between LSTM and ConvLSTM for time series prediction. In Section 3, we introduce the extraction of data samples, the proposed model framework, and the key parts of the model in detail. In Sections 4 and 5, we introduce the preparation and division of the experimental dataset, baseline model, evaluation metrics, model implementation, training details, comparative experiments, and results analysis. In Section 6, we summarize the results of this study and propose future work.

2 Related Work

2.1 Recognition and prediction

The study of vehicle driving intention has two modes. The first mode is detection[8, 16], i.e., the vehicle has obvious typical characteristics, e.g., the vehicle has crossed the lane line that the model needs to identify the intention. The second mode is prediction, i.e., the vehicle has not yet exhibited key typical characteristics, e.g., the vehicle has not crossed the lane line. At this time, using only the motion information of the target vehicle itself is insufficient to predict the future intention of the vehicle. Thus, more information needs to be considered. In addition to the vehicle's information, the prediction results are affected by road information, such as the speed and direction of travel defined by the lane. At the same time, the vehicle's intention is affected by the movement of surrounding vehicles, such as the sudden braking of the front vehicle. Road information requires the support of high-precision maps, which is not yet perfect. Thus, we must focus on the relevant motion information of surrounding vehicles to predict the driving intention of the target vehicle in advance.

2.2 LSTM and ConvLSTM for time series data

In early research, the recurrent neural network (RNN) was used to process time series data[17]. However, the RNN encountered the vanishing and exploding gradient problems. Subsequently, the LSTM was proposed to model sequence data processing[18]. The LSTM is based on the gating mechanism, which solves the vanishing and exploding gradient problems. The LSTM has been an effective tool for modeling sequential and time series data in recent years[19]. Sundermeyer et al. used LSTM for language modeling[20]. Since then, more studies have used LSTM to build models with encoder and decoder architectures to solve the trajectory prediction problem[21–23]. When the LSTM is applied to driving intention prediction, the input of the model includes the historical information of the vehicle and environmental information of surrounding vehicles. The information of surrounding vehicles is directly concatenated into the target vehicle as new features of the target vehicle information[1]. This study analyzes the influence of surrounding vehicles from a macroscopic perspective. The ConvLSTM network was originally used for precipitation nowcasting[24]. Compared with LSTM, ConvLSTM can model spatiotemporal information simultaneously[25]. In recent years, ConvLSTM has been widely used in spatiotemporal data modeling, such as travel demand forecasting[26] and traffic accident prediction[27]. Inspired by these ideas, this study considers the dynamic change process of surrounding vehicles from a macroscopic perspective, which involves both highly interactive spatial interactions and dynamic change in time. This study proposes an ensemble model, including LSTM and ConvLSTM subnetworks, which process the information of the target and surrounding vehicles, respectively.

3 Data Extraction and Process

This section introduces the dataset used in this study. First, a typical driving scenario on an urban road is taken as an example to explain the problem of predicting driving intention. Then, we extracted data from urban driving scenarios and constructed data samples for predicting driving intention.

3.1 Problem illustration

The University of California, Berkeley released a large-scale urban scene driving dataset, i.e., BDD100K, in 2018[7], providing a wealth of driving data in various urban road scenarios. A typical driving scenario of a vehicle on an urban road is shown in Fig. 1.

As shown in Fig. 1, the target vehicle can move along the longitudinal direction of the road, such as accelerating, decelerating, and holding. The target vehicle can also move in the lateral direction, such as changing lanes to the left or right. We want to predict its longitudinal and lateral driving intentions simultaneously.

To construct a dataset for supervised driving intention prediction, we extracted the required data from the BDD100K dataset with a sampling frequency of 5 Hz. The characteristics of each driving intention are shown in Table 1.

3.2 Data extraction

In the scenario shown in Fig. 1, when predicting the future driving intention of the target vehicle, we use its historical motion information to make inferences, although the motion data may not show the typical characteristics. More information, such as the existence of adjacent lanes and relative motion information of surrounding vehicles, has a significant influence on predicting the driving intention of the target vehicle. Therefore, the extracted data contain three parts, namely, historical motion data of the target vehicle (Table 2), constraint information of the lane (Table 3), and historical motion information of surrounding vehicles

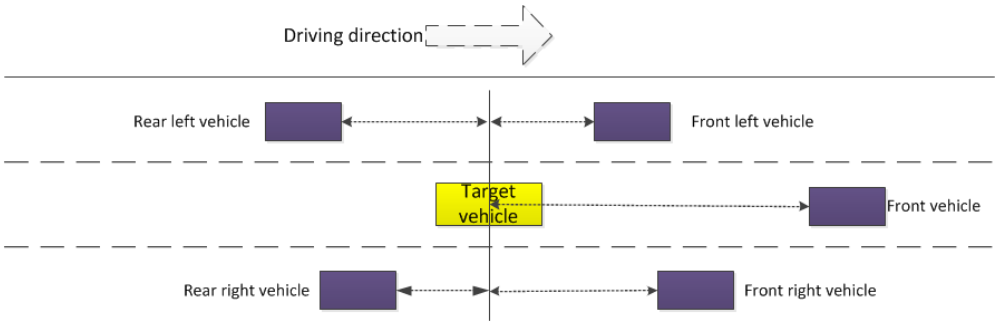


Fig. 1 A typical urban road driving scenario—we are interested in the target vehicle's driving intention.

Table 1 Classification results of driving intention used in this study.

Driving intention		Typical characteristic
Lateral intention	Changing to the left	Approaching, crossing, and adjusting
	Changing to the right	Approaching, crossing, and adjusting
Longitudinal intention	Rapid acceleration	Moving forward and Acc$> 0.2g$
	Rapid deceleration	Moving forward and Acc$< -0.2g$
	Keeping	Moving forward and $-0.2g <$ Acc $< 0.2g$
	Stopping	Moving forward and going to stop

Table 2 Extracted features of the target vehicle.

Feature	Description
Speed	Target vehicle speed, m/s
Acceleration	Target vehicle acceleration, m/s^2
Relative distance to the left lane	Distance between the left border of the vehicle and the left side of the lane
Relative distance to the right lane	Distance between the right border of the vehicle and the right side of the lane
Driving angle with respect to the road	Angle between the speed direction and the direction of the lane

Table 3 Extracted features for adjacent lane information.

Feature	Description
Existence of the left lane	1 if it exists, 0 if not
Existence of the right lane	1 if it exists, 0 if not

(Table 4).

The surrounding vehicles have three typical locations, namely, front, left front, and right front. If a vehicle does not exist at a certain location, then the corresponding virtual vehicle is filled in this study, as has been previously reported[1]. In this situation, the relative speed is set to 0, and the relative distance is set to have a large value.

3.3 Sample construction for prediction

To predict driving intention, the sample data need to be a historical window with the length of His_t, and this window must be taken before the driving intention starts. The corresponding prediction horizon is Fut_t. The sample construction process involves taking the lane changing process as an example, as shown in Fig. 2.

Table 4 Extracted features of surrounding vehicles.

Feature		Description
Left front/ front / right front	Relative distance	Longitudinal distance between the target vehicle and surrounding vehicles
	Relative speed	Longitudinal speed between the target vehicle and surrounding vehicles
	Stop light	1 if it is on, 0 if it is off
	Turn light	1 means the left turn light is on, 2 means the right turn light is on, 0 means off

The sample construction process shown in Fig. 2 has two important points that need to be clarified and explained:

(1) This study aims to analyze the performance of driving intention prediction under different prediction horizons. Therefore, only one sample is extracted from the file corresponding to each driving intention, rather than constructing multiple samples by sliding windows. Thus, the number and proportion of various types of samples in the complete dataset used for training and testing are the same as those shown in Table 5.

(2) Two key parameters are used in this study and should be described. The first key parameter is

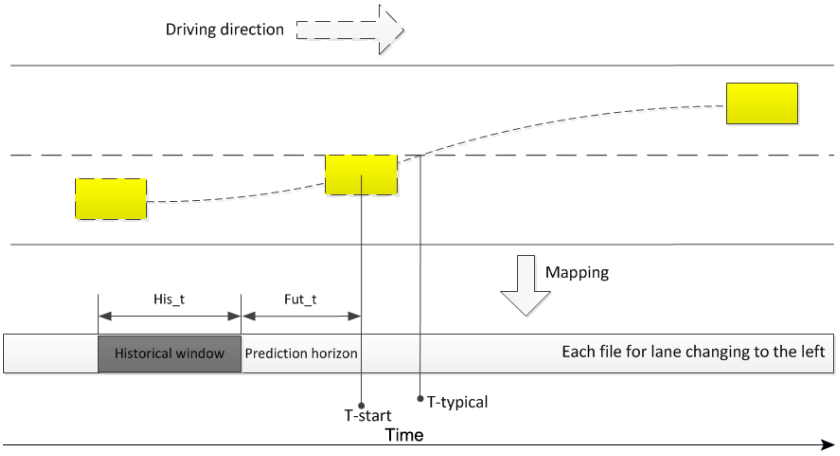


Fig. 2 Sample construction process using the moment when the target vehicle first touches the lane divider line as the intention start time denoted as T-start.

Table 5 Quantity distribution of extracted driving intention files.

Driving intention	Number	Percent (%)
Lane changing to the left	750	7.65
Lane changing to the right	685	6.98
Keeping straight	2874	29.31
Rapid accelerating	1554	15.85
Rapid decelerating	2399	24.46
Stopping	1544	15.75
Total	9806	100.00

the prediction horizon, which is how early a driving intention can be predicted. The second key parameter is the historical window size, which is how much historical information should be used to make a prediction.

To model the spatiotemporal relationship of surrounding vehicles, the historical motion data of each surrounding vehicle need to be extracted according to the process shown in Fig. 2 and organized in a grid manner, as illustrated in Fig. 3. First, we divide the left, middle, and right lanes and the front positions into 3×3 grids, where each column corresponds to a single lane and the rows are separated by a distance of 5 m, which is approximately equal to one car length. Each surrounding vehicle has multiple features (defined as channels here) at each time step t_i. Then, the motion data of surrounding vehicles at certain time steps are stacked along the channel axis to form a three-dimensional tensor. Finally, the three-dimensional tensors generated at different time steps are stacked into four-dimensional tensors, which are used as the input of the ConvLSTM network.

4 Methodology

4.1 Input, output, and model architecture

4.1.1 Input

The input for driving intention prediction has two parts, namely, the target vehicle's motion and road constraint information and the relative motion information of surrounding vehicles, as described in Section 3.3. We denote the target vehicle information as Tar_info and the surrounding vehicle information as Sur_info. Therefore,

$$\begin{aligned} \text{Tar\_info} &= [x_{t_1}, x_{t_2}, \ldots, x_{t_h}]^{\mathrm{T}}, \\ x_{t_i} &= [x^{f_1}, x^{f_2}, \ldots, x^{f_m}] \end{aligned} \tag{1}$$

where t_h denotes the window size and f_m denotes all of the features of the target vehicle shown in Tables 2 and 3.

4.1.2 Output

The output is the prediction result of driving intention, including the four longitudinal driving intentions and two lateral driving intentions shown in Table 1. Given that the results are formulated as category data, they need to be numerically processed and one-hot encoded as Label Y.

4.1.3 Model architecture

The model needs to process both the historical motion information of the target vehicle and the historical information of surrounding vehicles. To mine the historical information of the target vehicle, the LSTM layer is used for feature extraction. To model the influence of surrounding vehicles, we consider the surrounding vehicles as environmental context. This study proposes a network model based on ConvLSTM, which can learn to capture spatial interactions of surrounding vehicles and their time dependence. The ConvLSTM-based model framework is shown in Fig. 4.

The model shown in Fig. 4 is mainly composed of two parts. The first part is the input layer, which represents the motion information of the target vehicle as a two-dimensional matrix. The historical motion information of the target vehicle is processed through an LSTM encoder with a peephole. According to the

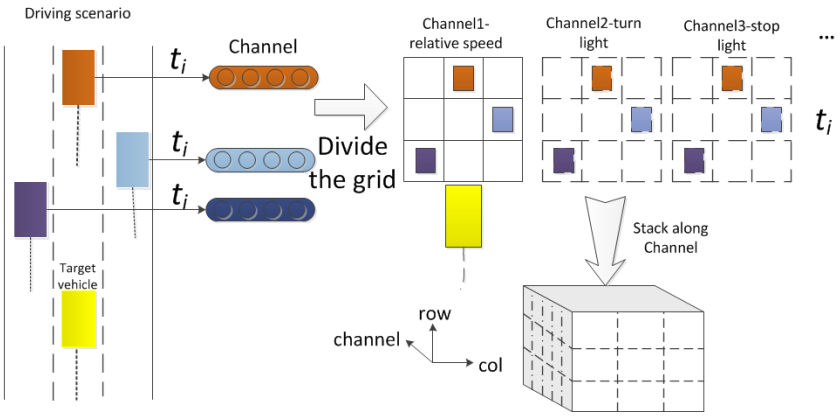


Fig. 3 Reorganizing the data of surrounding vehicles.

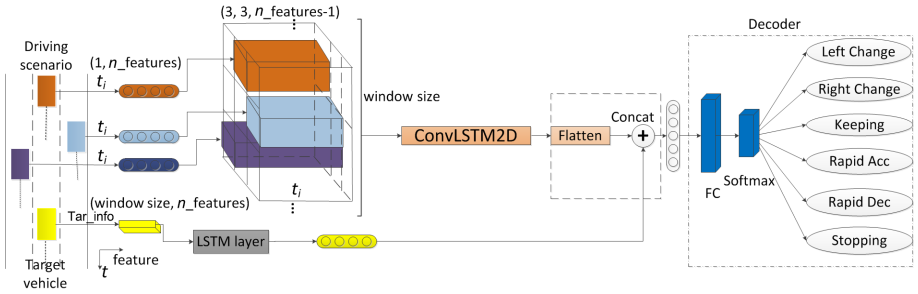


Fig. 4 ConvLSTM-based model framework—the encoder part includes the LSTM and ConvLSTM networks, decoder part uses fully connected layers, and activation function is softmax.

grid position occupied by the surrounding vehicles, the data of surrounding vehicles at different time form a three-dimensional tensor, including row positions, column positions, and channels. Three-dimensional tensors at different time steps are stacked together in a time sequence to form a four-dimensional tensor. The ConvLSTM layer is used to capture the spatial interactions of surrounding vehicles. Then, the vector obtained using the ConvLSTM layer and compression vector of the target vehicle are concatenated together as a final representation of the original input data.

The second part is the output layer. The last fully connected layer uses the softmax activation function to output the probability of each driving intention. Six output neurons correspond to the longitudinal and lateral driving intentions. The neuron with the largest probability value is used as the final prediction result. The calculation and derivation of the key parts of the model are described in Sections 4.2 and 4.3.

4.2 LSTM encoder with a peephole

Taking into account the differences in vehicles, four different LSTM encoders are used in this model. The basic LSTM unit is shown in Fig. 5.

The LSTM unit processes inputs sequentially to obtain the cumulative representation for the input X_t. The LSTM unit uses the forget gate, input gate, and output gate to update its cell and hidden states.

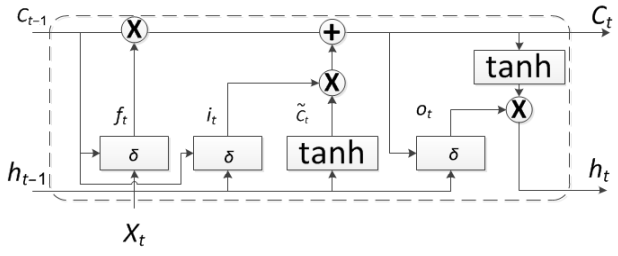


Fig. 5 Basic LSTM unit with a peephole—peephole denotes that the cell state is also used as an input for each gate.

The forget gate is used to control how much of the cell state information at the previous time step can be retained. The forget gate contains three inputs, namely, x_t, h_{t-1}, and C_{t-1}. The output of the forget gate is calculated as follows:

$$f_t = \sigma\left(W_f \times [C_{t-1}, h_{t-1}, x_t] + b_f\right) \tag{2}$$

where W_f is the parameter matrix, C_{t-1} is the cell state at the last time step, h_{t-1} is the hidden state of the previous time step, x_t is the input vector at the current time step, b_f is the bias of every neuron, and σ is the sigmoid activation function.

The input gate is used to control the addition of new information. The equation for adding information at the current time step is expressed as follows:

$$i_t = \sigma\left(W_i \times [C_{t-1}, h_{t-1}, x_t] + b_i\right) \tag{3}$$

The equation for calculating the candidate cell states is expressed as follows:

$$\tilde{C}_t = \tanh\left(W_C \times [h_{t-1}, x_t] + b_C\right) \tag{4}$$

where tanh is the activation function.

After calculating the forget and input gates, the cell state, C_t, is updated using the following equation:

$$C_t = f_t \times C_{t-1} + i_t \times \tilde{C}_t \tag{5}$$

The output gate regulates how much of the cell's state can be used to update the output of the hidden state. The output of the output gate is calculated as follows:

$$o_t = \sigma\left(W_o\left[C_t, h_{t-1}, x_t\right] + b_o\right) \tag{6}$$

After completing the calculations, the unit is ready to update the hidden state, h_t, using the following equation:

$$h_t = o_t \times \tanh\left(C_t\right) \tag{7}$$

4.3 ConvLSTM layer for modeling surrounding vehicles

The ConvLSTM unit also uses the gating mechanism but differs from the LSTM unit in that it uses convolution operations. The mechanism of the ConvLSTM unit is shown in Fig. 6.

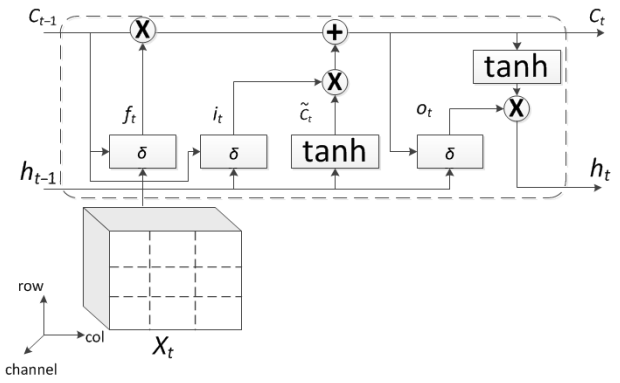


Fig. 6 Mechanism of the ConvLSTM unit.

Similar to the calculation of the LSTM unit described in Section 4.2, the calculation process of the ConvLSTM unit is as follows:

$$f_t = \sigma\left(W_{xf} \odot X_t + W_{hf} \odot h_{t-1} + W_{cf} \circ C_{t-1} + b_f\right) \tag{8}$$

$$i_t = \sigma\left(W_{xi} \odot X_t + W_{hi} \odot h_{t-1} + W_{ci} \circ C_{t-1} + b_i\right) \tag{9}$$

$$\tilde{C}_t = \tanh\left(W_{xc} \odot X_t + W_{hc} \odot h_{t-1} + b_c\right) \tag{10}$$

$$C_t = f_t \times C_{t-1} + i_t \times \tilde{C}_t \tag{11}$$

$$o_t = \sigma\left(W_{xo} \odot X_t + W_{ho} \odot h_{t-1} + W_{co} \circ C_t + b_o\right) \tag{12}$$

$$h_t = o_t \circ \tanh\left(C_t\right) \tag{13}$$

where $\odot$ denotes the convolution operations and $\circ$ denotes Hadamard product.

5 Experiment and Analysis

5.1 Division and normalization of the dataset

Based on the sample construction process described in Section 3.3, we divided the obtained dataset into the training and testing sets. The proportions of the sample sizes that the training and testing sets contain are 80% and 20%, respectively. The min-max normalization method was used to normalize the training and testing sets separately, so that each feature is within a fixed range. Finally, the model was trained on the training set and verified on the testing set.

5.2 Baseline model

We compared the performance of our method with that of the previously published vanilla LSTM method[1] on lane change intention prediction. The vanilla LSTM model uses the dynamic information of surrounding vehicles as additional input features of the target vehicle. We extended the output of the model to the prediction of longitudinal and lateral driving intentions.

5.3 Model implementation details and evaluation metrics

The model proposed in this study is described in Section 4.1. The network includes the LSTM encoder, the ConvLSTM layer, a fully connected layer, and a softmax output layer. The parameters of each layer are listed in Table 6.

The proposed and baseline models are implemented using TensorFlow[28]. The loss function is the cross-entropy loss with six classes. The optimizer is Adam, and its initial learning rate is 0.0001. The model is trained in batch mode, and the batch size is 64.

We determined the accuracy, precision, recall, and confusion matrix to evaluate the performance of the model from multiple perspectives.

5.4 Results and analysis

5.4.1 Comparison among different historical window sizes in the Vanilla LSTM model

To determine how much historical information should be used to predict driving intentions, we obtained and compared the predicted accuracy rates for different historical time. In our dataset, the sampling frequency of time series data is fixed at 5 Hz, i.e., the time step interval is 0.2 s. The total historical time is equal to the product of the interval and the number of time steps (historical window size). We can control the length of time by controlling the historical window size. Specifically, we set the historical window sizes of the LSTM structure to be 8, 10, 12, and 14, which are equivalent to the lengths of the historical time set to be 1.6, 2.0, 2.4, and 2.8 s. The results are presented in Fig. 7.

We compared the prediction accuracy at different historical time. The prediction accuracy first increased and then decreased as the historical time increased. The baseline model has the highest accuracy with the historical time set to 2.4 s.

Table 6 Parameters of each layer.

LSTM encoder		ConvLSTM layer				FC layer		Softmax layer	
Number of time steps	Number of units	Number of time steps	Number of rows	Number of cols	Number of channels	Input size	Output size	Input size	Output size
12	20	12	3	3	5	50	10	10	6

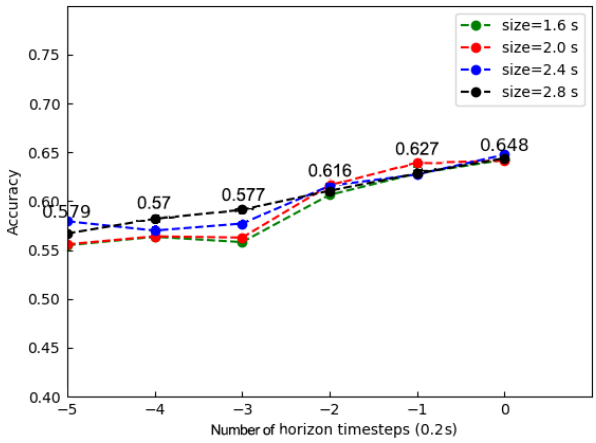


Fig. 7 Accuracy under different sizes and different time advances.

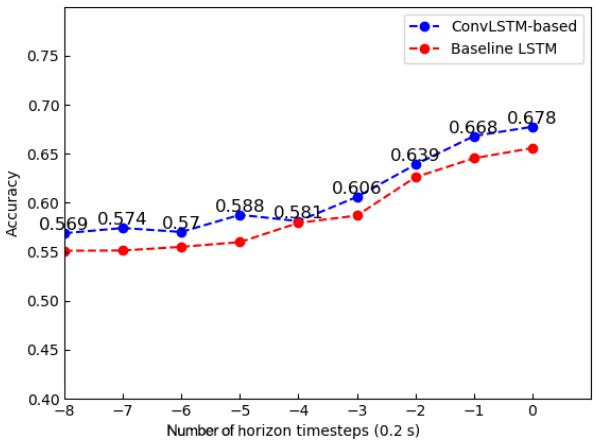


Fig. 8 Accuracy compared under different horizons.

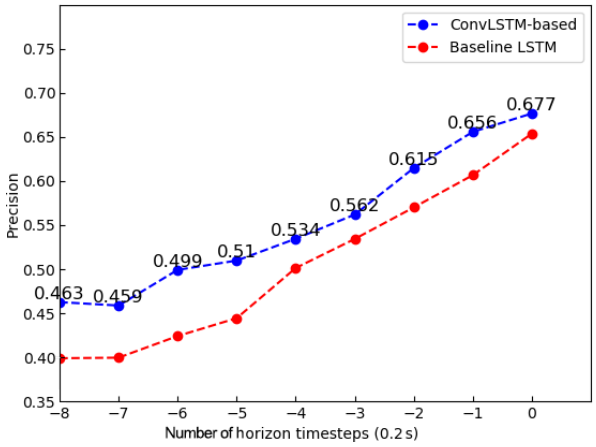


Fig. 9 Precision compared under different horizons.

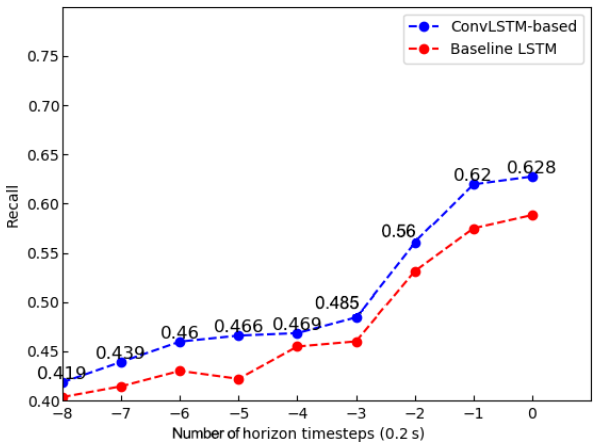


Fig. 10 Recall compared under different horizons.

To obtain a general understanding of the change trends of curves, we compared the results of the testing set. As the historical window size increases, the prediction accuracy first increases and then decreases. This finding indicates the trade-off between window size and prediction accuracy. As expected, the longer the historical window size is, the more information will be obtained, and the more accurate the prediction will be in the final results. However, when the historical window size exceeds a certain threshold, factors that are not relative to the current driving intention will be introduced into the input, leading to a decrease in prediction accuracy.

5.4.2 Comparison between proposed and baseline models

After determining the best historical window size to be 2.4 s, we obtained and compared the results of our model with those of the vanilla LSTM model to show the advantage of using ConvLSTM to model the influence of surrounding vehicles. We compared the accuracy, precision, and recall at different time horizons. The results are shown in Figs. 8–10 and Table 7.

Figures 8–10 show the comparisons of the accuracy, precision, and recall of two models, respectively. All three metrics show that the use of ConvLSTM to model the influence of surrounding vehicles can improve the performance of longitudinal and lateral driving intention predictions.

Qualitative analysis showed that the accuracy, precision, and recall of our model is superior to that of the vanilla LSTM model in predicting the longitudinal and lateral driving intentions at different time horizons.

Quantitative analysis was performed, and the results are shown in Table 7. Compared with LSTM, in terms of accuracy, our model exhibits an average increase of 3.1%, with a maximum increase of 4.9%. In terms of precision, our model exhibits an average increase of 10.5%, with a maximum increase of 17.7%. In terms of

Table 7 Prediction results of the baseline LSTM and ConvLSTM-based models.

(%)

Number of horizon timesteps	Accuracy		Precision		Recall	
	LSTM	Conv LSTM	LSTM	Conv LSTM	LSTM	Conv LSTM
−8	55.10	56.89	39.89	46.26	40.36	41.86
−7	55.13	57.42	39.96	45.87	41.45	43.94
−6	55.50	57.02	42.41	49.92	43.02	45.99
−5	55.98	58.77	44.43	50.97	42.21	46.61
−4	57.95	58.13	50.12	53.45	45.49	46.86
−3	58.68	60.60	53.42	56.19	46.01	48.45
−2	62.60	63.89	57.00	61.46	53.14	56.01
−1	64.53	66.83	60.66	65.57	57.49	61.96
0	65.58	67.75	65.36	67.66	56.84	62.76

recall, our model exhibits an average increase of 6.1%, with a maximum increase of 10.4%.

5.4.3 Prediction performance of the proposed model at different time horizons

To analyze the performance of our model in predicting the longitudinal and lateral driving intentions comprehensively, we further compared the confusion matrix at 0.5 and 1 s advances, respectively, as shown in Figs. 11 and 12, where "Keep" denotes maintaining the current speed and direction, "Acc" denotes acceleration, "Dec" denotes deceleration, "Stop" denotes slowing down and stopping, "Lch" denotes changing lanes to the left, "Rch" denotes changing lanes to the right. The confusion matrix in the testing set shows that, as the time advance decreases, the prediction results become more accurate.

6 Conclusion

For driving intention prediction from a bird's-eye view, because the target vehicle does not exhibit typical

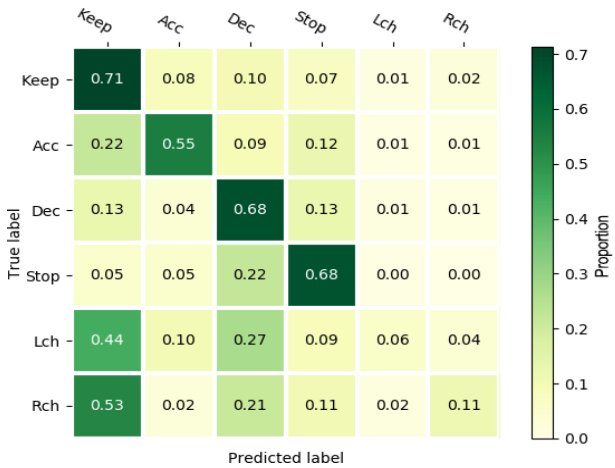


Fig. 11 Confusion matrix at 0.5 s advance.

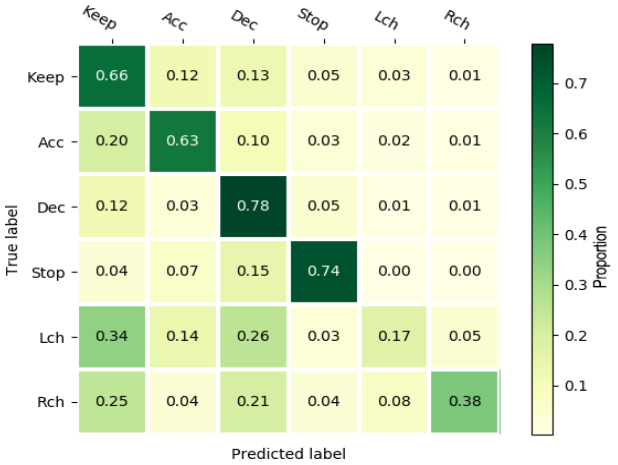


Fig. 12 Confusion matrix at 1 s advance.

characteristics, more effective information is needed to predict its driving intention. In this study, a novel ConvLSTM-based model was proposed to model the overall time dependence of surrounding vehicles and the spatial interactions of surrounding vehicles at the same time. This model can also predict the target vehicle's longitudinal and lateral driving intentions simultaneously. First, we determined that the optimal time segment required is 2.4 s and too long or too short time segments are inconducive to prediction. Second, in the experiment to explore the correlation between different time advances and forecast accuracies, we determined that, as the time advance decreases, the forecast accuracy of our model gradually increases. When the forecast is 1 s ahead, the forecast accuracy is 58.8%. Finally, we verified the effectiveness of the model proposed in this study by comparing it with the baseline LSTM model in a real dataset. Our analysis showed that our model improves accuracy by 3.1% on average, precision by 10.5% on average, and recall by 6.1% on average, which helps make effective predictions earlier. In future studies, we will investigate how to adapt our predictive model to different road scenarios, such as roundabouts and various intersections.

Acknowledgment

The work was supported by the National Key Research and Development Program of China (No. 2017YFB0102601), the National Natural Science Foundation of China (No. 71671100), and the Key Laboratory of Road and Traffic Engineering of the Ministry of Education, Tongji University.

References

[1] S. Su, K. Muelling, J. M. Dolan, P. Palanisamy, and P.

Mudalige, Learning vehicle cooperative lane-changing behavior from observed trajectories in the NGSIM dataset, in *Proc. 2018 IEEE Intelligent Vehicles Symp.*, Suzhou, China, 2018, pp. 1412–1417.

[2] N. Deo and M. M. Trivedi, Convolutional social pooling for vehicle trajectory prediction, in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition Workshops*, Salt Lake City, UT, USA, 2018.

[3] G. Weidl, A. L. Madsen, V. Tereshchenko, W. Zhang, S. Wang, and D. Kasper, Situation awareness and early recognition of traffic maneuvers, in *Proc. 9th EUROSIM & 57th SIMS*, Oulu, Finland, 2016, pp. 8–18.

[4] C. Y. Dong, J. M. Dolan, and B. Litkouhi, Intention estimation for ramp merging control in autonomous driving, in *Proc. 2017 IEEE Intelligent Vehicles Symp.* (*IV*), Los Angeles, CA, USA, 2017, pp. 1584–1589.

[5] D. J. Phillips, T. A. Wheeler, and M. J. Kochenderfer, Generalizable intention prediction of human drivers at intersections, in *Proc. 2017 IEEE Intelligent Vehicles Symp. (IV)*, Los Angeles, CA, USA, 2017, pp. 1665–1670.

[6] Federal Highway Administration, Next generation simulation fact sheet, https://www.fhwa.dot.gov/publications/research/operations/its/06135/index.cfm, 2011.

[7] F. Yu, H. F. Chen, X. Wang, W. Q. Xian, Y. Y. Chen, F. C. Liu, V. Madhavan, and T. Darrell, BDD100K: A diverse driving dataset for heterogeneous multitask learning, arXiv preprint arXiv:1805.04687, 2018.

[8] H. M. Mandalia and M. D. D. Salvucci, Using support vector machines for lane-change detection, *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 49, no. 22, pp. 1965–1969, 2005.

[9] S. B. Amsalu, A. Homaifar, F. Afghah, S. Ramyar, and A. Kurt, Driver behavior modeling near intersections using support vector machines based on statistical feature extraction, in *Proc. 2015 IEEE Intelligent Vehicles Symp.* (*IV*), Seoul, Republic of Korea, 2015, pp. 1270–1275.

[10] Y. L. Dou, F. J. Yan, and D. W. Feng, Lane changing prediction at highway lane drops using support vector machine and artificial neural network classifiers, in *Proc. 2016 IEEE Int. Conf. Advanced Intelligent Mechatronics (AIM)*, Banff, Canada, 2016, pp. 901–906.

[11] S. W. Liu, K. Zheng, L. Zhao, and P. Z. Fan, A driving intention prediction method based on hidden Markov model for autonomous driving, *Computer Communications*, vol. 157, pp. 143–149, 2020.

[12] Y. Xing, C. Lv, H. J. Wang, D. P. Cao, and E. Velenis, An ensemble deep learning approach for driver lane change intention inference, *Transportation Research Part C*: *Emerging Technologies*, vol. 115, p. 102615, 2020.

[13] A. Girma, S. Amsalu, A. Workineh, M. Khan, and A. Homaifar, Deep learning with attention mechanism for predicting driver intention at intersection, arXiv preprint arXiv: 2006.05918, 2020.

[14] S. Mozaffari, O. Y. Al-Jarrah, M. Dianati, P. Jennings, and A. Mouzakitis, Deep learning-based vehicle behavior prediction for autonomous driving applications: A review, *IEEE Transactions on Intelligent Transportation Systems*, doi: 10.1109/TITS.2020.3012034.

[15] S. Z. Dai, L. Li, and Z. H. Li, Modeling vehicle interactions via modified LSTM models for trajectory prediction, *IEEE Access*, vol. 7, pp. 38287–38296, 2019.

[16] J. Schlechtriemen, A. Wedel, J. Hillenbrand, G. Breuel, and K. D. Kuhnert, A lane change detection approach using feature ranking with maximized predictive power, in *2014 IEEE Intelligent Vehicles Symp. Proc.*, Dearborn, MI, USA, 2014, pp. 108–114.

[17] J. T. Connor, R. D. Martin, and L. E. Atlas, Recurrent neural networks and robust time series prediction, *IEEE Transactions on Neural Networks*, vol. 5, no. 2, pp. 240–254, 1994.

[18] S. Hochreiter and J. Schmidhuber, Long short-term memory, *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[19] F. Karim, S. Majumdar, H. Darabi, and S. Harford, Multivariate LSTM-FCNs for time series classification, *Neural Networks*, vol. 116, pp. 237–245, 2019.

[20] M. Sundermeyer, R. Schlüter, and H. Ney, LSTM neural networks for language modeling, in *Proc. 13th Annual Conf. Int. Speech Commun. Association*, Portland, OR, USA, 2012.

[21] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese, Social LSTM: Human trajectory prediction in crowded spaces, in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, 2016, pp. 961–971.

[22] F. Altché and A. de La Fortelle, An LSTM network for highway trajectory prediction, in *Proc. 2017 IEEE 20th Int. Conf. Intelligent Transportation Systems* (*ITSC*), Yokohama, Japan, 2017, pp. 353–359.

[23] S. H. Park, B. Kim, C. M. Kang, C. C. Chung, and J. W. Choi, Sequence-to-sequence prediction of vehicle trajectory via LSTM encoder-decoder architecture, in *Proc. 2018 IEEE Intelligent Vehicles Symp.* (*IV*), Changshu, China, 2018, pp. 1672–1678.

[24] X. J. Shi, Z. R. Chen, H. Wang, D. Y. Yeung, W. K. Wong, and W. C. Woo, Convolutional LSTM network: A machine learning approach for precipitation nowcasting, in *Proc. 28th Int. Conf. Neural Information Processing Systems*, Cambridge, MA, USA, 2015, pp. 802–810.

[25] L. Zhang, G. M. Zhu, P. Y. Shen, J. Song, S. A. Shah, and M. Bennamoun, Learning spatiotemporal features using 3DCNN and convolutional LSTM for gesture recognition, in *Proc. 2017 IEEE Int. Conf. Computer Vision Workshops*, Venice, Italy, 2017, pp. 3120–3128.

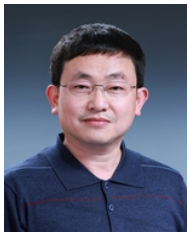
[26] D. J. Wang, Y. Yang, and S. M. Ning, DeepSTCL: A deep spatio-temporal convLSTM for travel demand prediction, in *Proc. 2018 Int. Joint Conf. Neural Networks* (*IJCNN*), Rio de Janeiro, Brazil, 2018, pp. 1–8.

[27] Z. N. Yuan, X. Zhou, and T. B. Yang, Hetero-convLSTM: A deep learning approach to traffic accident prediction on heterogeneous spatio-temporal data, in *Proc. 24th ACM SIGKDD Int. Conf. Knowledge Discovery & Data Mining*, London, UK, 2018, pp. 984–992.

[28] M. Abadi, P. Barham, J. M. Chen, Z. F. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, et al., TensorFlow: A system for large-scale machine learning, in *Proc. 12th USENIX Conf. Operating Systems Design and Implementation*, Berkeley, CA, USA, 2016, pp. 265–283.

He Huang received the bachelor degree from China Agricultural University, Beijing, China, in 2014. He is currently a PhD student at Tsinghua University. His research interests include deep learning and multi-sensor fusion.

Zheni Zeng received the BEng degree from Tsinghua University, Beijing, China, in 2020. She is currently a PhD student at Tsinghua University. Her research interests include natural language processing and cross-modal learning.



Danya Yao is currently a professor at Tsinghua University. He received the BEng, MEng, and PhD degrees from Tsinghua Unversity in 1988, 1990, and 1994, respectively. His active research areas include vehicle infrastructure cooperation systems, advanced detection technology, and systems engineering. He has published more than 100 papers. He is PI/co-PI of more than 5 national research programs including 863 program, 973 program, and NSFC project. He was ever the chief expert of the National High-Tech Research and Development Program (863 Program) Research on Key Technologies of Intelligent Vehicle-Infrastructure Co-operation System.

Xin Pei received the BS and MS degrees from Tsinghua University, China, in 2005 and 2007, respectively, and the PhD degree from the University of Hong Kong in 2011. She is currently a research associate professor with the Department of Automation, Tsinghua University. Her current research interests include road safety evaluation and driving behavior analysis. She is the Principal Investigator (PI)/co-PI of 3 road safety related NSFC projects. She has published more than 50 SCI/SSCI/EI indexed papers, especially, there are 7 papers published on the *Journal of Accident Analysis and Prevention*, which is a top journal for road safety analysis.

Yi Zhang received the BS and MS degrees from Tsinghua University, China, in 1986 and 1988, respectively, and the PhD degree from the University of Strathclyde, UK, in 1995. He is currently a professor in control science and engineering at Tsinghua University. His current research interests include intelligent transportation systems, intelligent vehicle-infrastructure cooperative systems, analysis of urban transportation systems, urban road network management, traffic data fusion and dissemination, urban traffic control and management, advanced control theory and applications, advanced detection and measurement, and systems engineering.