Research on Behavior Recognition of Disyielding Pedestrians Based on LSTM

Rui Zhang [1,*], Haosen Qin [1], Ling Wang [1], Ziyang Wu [2]

[1] School of traffic and Transportation Engineering, Dalian Jiaotong University, Dalian, Liaoning, 116028

[2.] School of electrical and Automation Engineering, Dalian Jiaotong University, Dalian, Liaoning, 116028

* Corresponding Author Email: ns3o4uy1yz65d@outlook.com

Abstract. Serious traffic accidents caused by vehicles' disobedience to pedestrians occur frequently, so it is very important to supervise and regulate the behavior of disobedient pedestrians and explore the identification methods of disobedient behaviors. Compared with the existing research, this paper proposes a recognition algorithm based on Long Short Memory (LSTM) network for vehicles' indecisive pedestrian behavior. This algorithm is more practical and adaptable. First, it predicts the future continuous kinematics of the vehicle through the past transverse and longitudinal characteristics of the vehicle, and obtains the driver's manipulation intention. Then, the manipulation intention is combined with the past pedestrian trajectory and pedestrian head direction information, and the joint modeling is carried out to obtain the probability of the occurrence of vehicles' behavior that does not yield to pedestrians. Using real driving data to test, the test results show good recognition accuracy, and the goodness of fit of the model to the sample reaches 0.93.

Keywords: LSTM; Behavior recognition; Courtesy to pedestrians; Trajectory prediction

1. Introduction

In the context of today's fast-paced society, many drivers drive their vehicles at will, often giving way to pedestrians. Discourtesy to pedestrians may result in points deduction and fines, or serious traffic accidents. According to the statistics of the Ministry of Public Security, 14000 traffic accidents have occurred on zebra crossings in the past three years. Therefore, it is particularly important to explore the judgment and identification methods of non comity behaviors in order to avoid the occurrence of non comity behaviors of pedestrians due to drivers' artificial reasons and to supervise and regulate them.

At present, it has been studied at home and abroad. Based on video analysis, Zhou et al. designed and implemented a detection system for illegal behaviors before the zebra crossing. Papageorgiou et al. developed a target detection system, which detects people and vehicles through SVM. Large et al. obtained pedestrian motion patterns through clustering and estimated the probability of a pedestrian vehicle collision. Havasi et al. also recognizes pedestrians through the leg feature information of pedestrians. Track the tracks of pedestrians and classifies them.

However, these studies are based on pedestrian behavior characteristics or vehicle behavior characteristics to identify, classify, judge, and do not combine the two. This paper divides the behavior recognition method of whether the vehicle is courteous to pedestrians into the following steps: First, build the LSTM short-term memory artificial neural network (hereinafter referred to as LSTM) in the first part, encode the coordinate data obtained from radar sensor measurement as input, and then generate the future position information of the vehicle after network processing and decoding, so as to predict the driver's intention and the complex dynamics of vehicle movement. Then, the second part LSTM is constructed. Using image data including pedestrians collected by the front end of the vehicle, the interaction information between pedestrians and vehicles is injected into the predicted vehicle trajectory information to obtain the future trajectory of pedestrians. Finally, a layer of BP neural network is established to analyze whether there is any behavior of offensive pedestrians. The method is developed based on nuScenes dataset of natural driving data, and shows good prediction accuracy. Identification of offensive pedestrians' driving behavior depends on the information observed from their own vehicles. Due to the limitations of sensors and the occlusion of objects. These data may not be completely observed. Therefore, this paper assumes that the data are available in the actual scene.

2. Model framework

2.1. System architecture

In order to overcome the data imbalance between information constraints and other driving actions and the challenge of model generalization, this paper first proposes a LSTM framework to predict driving intentions. Then, based on the influence of pedestrian motion on vehicle dynamics, that is, how they affect their decision-making when crossing the street, a LSTM framework for predicting pedestrian trajectory is proposed. Finally, a BP neural network framework is established based on the future trajectories of vehicles and pedestrians, as well as the driving status of vehicles. The neural network proposed in this paper is shown in Figure 1. It consists of two LSTM networks and a BP network. Each LSTM network is composed of an encoder and a decoder. The encoder of the first layer LSTM network is used to learn the dynamic characteristics of vehicle motion and reduce its dimensions. The decoder recovers its dimensions and predicts the probability distribution of different horizontal and vertical driving intentions, such as horizontal intentions: left and right lane changes and lane keeping, Longitudinal intention: get the probability distribution of future vehicle motion when driving and braking normally. The second layer LSTM network encoder encodes the pedestrian position, head direction and driving intention, connects the output and uses it as the input of LSTM decoder, which outputs the future trajectory distribution of pedestrians. The last layer of the network is an ordinary BP neural network, which classifies the different combinations of the future movement tracks and driving states of vehicles and pedestrians, establishes the criteria for determining the unacceptable behavior, and finally obtains the probability distribution of vehicles' unacceptable behavior to pedestrians after learning from the network (Figure 1).

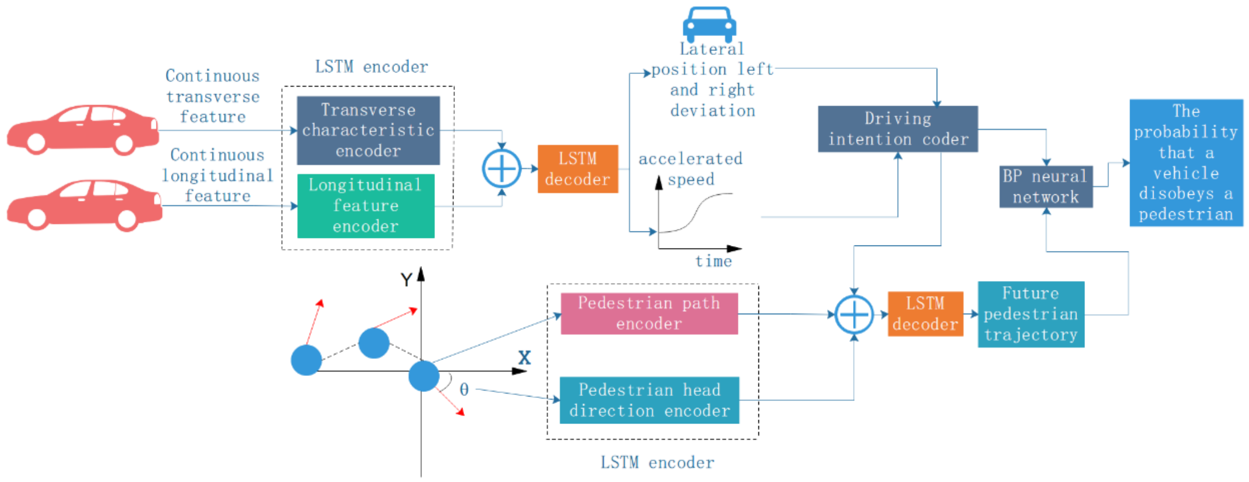


Figure 1. System architecture

2.2. Introduction to LSTM Network

The recurrent neural network (RNN) is different from the traditional feedforward network. It can analyze the sequence information and learning time characteristics through its internal state cycle. LSTM is a special implementation of RNN, which is developed to avoid the gradient of loss function disappearing with time.

LSTM has a storage state called "element", which is used to store the interpretation of past input data [1]. Update cells based on current input and previous cell status. There are different dates between the input and the unit. The unique control mechanism LSTM can learn when to forget the past state and update the state when new input is obtained. Let c_t be the state of the storage unit with a time step of t, h_t be the state of the hidden layer, and c_t and h_t are updated by the following equation.

$$i_t = \sigma\left(w_{x^i} x_t + w_{h^i} h_{t-1} + b_i\right) \tag{1}$$

$$f_t = \sigma\left(w_x f x_t + w_{h^f} h_{t-1} + b_f\right) \tag{2}$$

$$o_t = \sigma(w_{x^o} x_t + w_{h^o} h_{t-1} + b_o) \tag{3}$$

$$g_t = \tanh(w_x c x_t + w_{h^c} h_{t-1} + b_c) \tag{4}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot g_t \tag{5}$$

$$h_t = o_t \odot \tanh(c_t) \tag{6}$$

Where $\sigma(x) = \frac{1}{1+e^{-x}}$ is the activation function; i_t, f_t, o_t and g_t are input gate vector, forgetting gate vector, output gate vector and state update vector respectively; $w_x i$, $w_x f$, $w_x o$, $w_x c$, $w_h i$, $w_h f$, $w_h o$, $w_h c$ is the weight of linear combination. b_i, b_f, b_o, b_c is the relative offset constant, and $\odot$ represents the product of the corresponding position elements of two matrices.

Figure 2 shows the internal structure of the LSTM network. Input gate i_t and output gate o_t The data flow from input to output can be controlled separately. Forgotten Gate f_t can decide whether to forget the information stored in the memory unit. The gate control mechanism is used to learn data, which can overcome the problem caused by exponential attenuation or increase of input over time in general RNN. As time goes on, LSTM is a deep neural network, which can extract higher level elements and make the network deeper by adding additional LSTM layers. However, in this paper, similar results were found using LSTMs of different layers, so only one layer of LSTMs was used to save storage and computing resources. In order to generate different forms of output related to a given task, namely, driving intention recognition and behavior recognition of offensive pedestrians, an additional full connection dense output layer with hidden state is added.

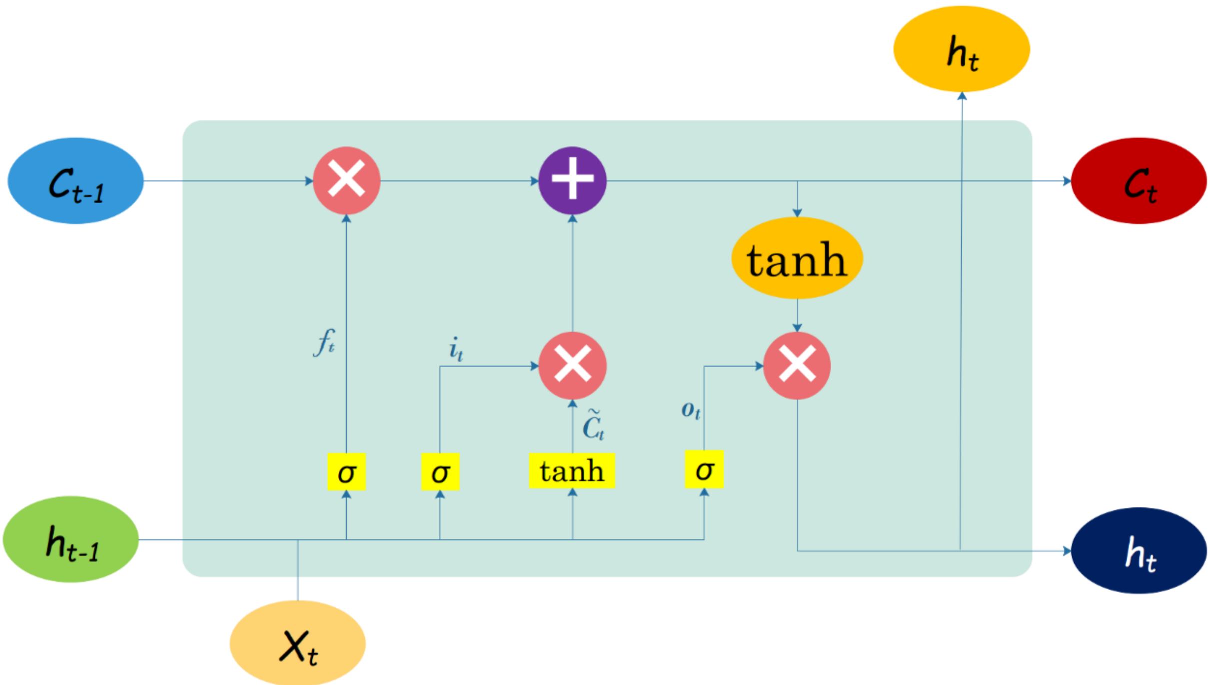


Figure 2. LSTM Element Structure

3. Input characteristics and the predicted output of the model

3.1. Part I LSTM

It is actually a regression problem [2] to use the limited historical observation data from the vehicle sensors to achieve a more accurate prediction of the future vehicle trajectory. More accurately, the goal of predicting the vehicle trajectory is to train a prediction model, which input the observable feature $\mathrm{X} = \{x_k^t\}_{k \in T_{pre}, x^t \in \mathrm{F}}$, where， $T_{\text{pre}} = \{-t_{\text{prs}}, \dots, 0\}$represents the time interval of historical input data, and F represents the observed and calculated feature set. Output vehicle track information in sequence $Y = \{y_k\}_{k \in T_{\text{pos}}, y \in \mathrm{O}}$, where $T_{pos} = \{1, \dots, t_{pos},\}$ represents the future time interval, and O represents the set to be output. When $x^t \in F$ and $k \in T_{pre}$, When x_k^t represents the observation data of k time intervals earlier than the current time. When $y \in \mathrm{O}$ and $k \in T_{\text{pos}}$, y_k represents the predicted output value after k time intervals in the future.

For the first part of the network, the longitudinal and transverse characteristics of the vehicle that needs to be input can be measured through radar, lidar and other on-board sensors [3]. When the vehicle moves, all transverse and longitudinal input features are measured and represented in a local coordinate system parallel to the global coordinate system. Define the lateral position of the vehicle as:

$$x_{re} = x - x_{pe} \tag{7}$$

Where, x represents local transverse position, x_{pe} indicates the local transverse position of the lane marking line closest to the vehicle. Analyze x_{re} in time series The first and second derivative x of re ̇_ Re "and" x ̈_ Re helps to identify lateral driving intentions. Transverse feature x_ re、x ̇_ Re "and" x ̈_ Re and longitudinal characteristic v_ Y and a_ Y as the input of the first LSTM encoder, these features are selected not only because the sensor can measure these data, but also because they reflect the driver's decision intention.

Where, x represents the local transverse position, and represents the local transverse position of the lane marking line closest to the vehicle. Analyzing the first and second derivatives $\dot{x}_{re}$ and $\ddot{x}_{re}$ of x_{re} in the time series is helpful to identify the horizontal driving intention. The horizontal characteristics x_{re}, $\dot{x}_{re}$,$\ddot{x}_{re}$, as well as the longitudinal characteristics v_y and a_y are the inputs of the first LSTM encoder. These characteristics are selected not only because the sensor can measure these data, but also because they reflect the driver's decision-making intention.

For the predicted output of the first part of the network, the vertical span is about 100 meters. The value of longitudinal position may be quite large. In order to predict the future trajectory and driving intention of the vehicle, output the future longitudinal acceleration $\{\hat{a}_y(k)\}_{k \in T\text{pos}}$ of the vehicle on the predicted line. Accordingly, the future longitudinal trajectory is calculated as:

$$\hat{v}_y(k) = \hat{v}_y(k-1) + \hat{a}_y(k) * \delta t \tag{8}$$

$$\hat{y}(k) = \hat{y}(k-1) + \hat{v}_y(k) * \delta t \tag{9}$$

Where, δt represents the time step. Since the change of lateral position is usually limited, the lateral deviation $\{\hat{x}_{re}(k)\}_{k \in T_{\text{pos}}}$ on the prediction line is used as the output, and the lateral position of the vehicle can be calculated as:

$$\hat{x}(k) = \hat{x}_{re}(k) + \hat{x}_{pe} \tag{10}$$

Where, $\hat{x}_{pe}$ represents the local transverse position of the lane marking line nearest to the vehicle predicted by the network.

3.2. Part II LSTM

The LSTM network in the second part uses an encoder to encode the pedestrian track, the vehicle track predicted in the first part and the pedestrian's head direction within a 1-second time interval. The final state of the encoder is connected and transmitted to the LSTM decoder. The decoder obtains the prediction result of the pedestrian trajectory 2 seconds later. Then it analyzes whether the predicted trajectory of the vehicle 2 seconds later intersects the predicted trajectory of the pedestrian. If the predicted trajectory intersects, calculate the transverse and longitudinal characteristics of the vehicle 2 seconds later to obtain the probability distribution of offensive behavior. Specifically, the input X of the LSTM network is the information in the past one second of the k-time interval Δk is taken as 0.2 seconds:

$$X = [x^{k-4\Delta k}, \dots, x^{k-\Delta k}, x^k] \tag{11}$$

The position coordinate $x^k = \left[x_{\mathrm{ps}}^k, y_{\mathrm{ps}}^k\right]$ of the pedestrian in the reference system at time k and the pedestrian head direction $\Theta_{pd} = \left[\theta_{pd}^{k-4\Delta k}, \dots, \theta_{pd}^{k-\Delta k}, \theta_{pd}^k\right]$ are input into the X at time k, and the prediction information $\hat{x}(k)$ and $\hat{y}(k)$ of the vehicle at time k are also input. The output Y of the network is the predicted future position of pedestrians in 2 seconds:

$$Y = [x^{k+\Delta k}, x^{k+2\Delta k}, \dots, x^{k+10\Delta k}] \tag{12}$$

4. Judgment criteria for irreverence

The input of the second LSTM network depends on the output of the first LSTM network. At the same time, the decision criteria are established based on the output results of the two networks at a certain time and are learned through a layer of ordinary BP reverse network. If the position coordinate $\left[x_p^k, y_p^k\right]$ at time t belongs to both the output of the first LSTM network and the output of the second LSTM network, it indicates that there is an intersection between the predicted trajectory of vehicles and the predicted trajectory of pedestrians, and vehicles may behave irreverently. At this time, if the driver makes the vehicle decelerate and brake, it is considered that there is no indecent behavior. This article defines that when the vehicle speed is less than 0.8 times of that 2 seconds ago, it means that the braking operation is in progress. Finally, the probability distribution of vehicles giving way to pedestrians is obtained as follows:

$$P\left[\left(V_y^k > 0.8V_y^{k-4\Delta k}\right) \mid \left(x_p^k = \left[x_{ps}^k, y_{ps}^k\right], x_c^k = [x_{re}^k, y_{re}^k]\right)\right] \tag{13}$$

5. Result analysis

5.1. Data selection

The data used in this article is from the nuScenes dataset. The nuScenes data set is a public large-scale automatic driving data set developed by the Motive team. The team used two Renault electric vehicles in Boston and Singapore to collect data on 1000 driving scenes. The nuScenes data set is the first large-scale data set that provides data from the entire sensor kit (6 cameras, 1 lidar, 5 radars, GPS, IMU) of the vehicle. Compared with the KITTI dataset, nuScenes contains more than 7 times more object annotations. For the nuScenes dataset, the part of interest in this paper is the images of pedestrians collected by the front camera of the vehicle and the vehicle speed and position coordinates measured by the radar. In this paper, each object containing the key frame of pedestrians is labeled at a rate of 5Hz (0.2s), and a total of 23 types of objects are labeled with 3D bbox. As showed in Figure 3, the yellow box represents vehicles and the blue box represents pedestrians. On this basis, pedestrian head position, head direction and other information are also marked as data labels. The extraction rate of vehicle track information is also 5Hz, including instantaneous speed, acceleration, transverse and longitudinal position information. This paper analyzes the interaction between vehicle dynamics and pedestrian trajectory in these annotations. In order to train and verify the model, this paper divides the data into three types of input networks: verification, testing and training. In order to improve the accuracy, it also uses the data that pedestrians are not ready to cross the road (Figure 3).

Figure 3. Image annotation

5.2. Network training details

The network in this paper is programmed in Python language and trained through 25 time step keyframes, that is, the interval between each input is 0.2 seconds, and 25 time steps represent the key information in the past 5 seconds. At the same time, in order to improve the training efficiency, the input information is updated at a rate of 5Hz. For the LSTM network in the first part, it is actually a multi classification problem of driving intentions. It includes six categories of horizontal driving intentions: lane keeping, left lane change, right lane change and longitudinal driving intentions: normal driving and braking. During training, the input features are coded into 64 dimensions, and the previously hidden state H of the output layer_ T uses the softmax function to decode and recover data to 128 dimensions, in the form of:

$$P(y = j \mid H_t) = \frac{e^{H_t^T W_j}}{\sum_{i=1}^{6} e^{H_t^T W_j}} \text{ for } j = 1, \dots, 6 \tag{14}$$

Where y is the predicted driving intention; w_j is the weight of each driving intention category.

For the second part of LSTM network, the predicted vehicle trajectory and pedestrian information are used as input and coded, and the quadratic form of the difference between the actual pedestrian trajectory and the predicted pedestrian trajectory is set as the loss function. The weights are updated through time backpropagation, and finally the predicted trajectory of pedestrians is decoded. Finally, a layer of ordinary BP backpropagation network is added. Based on the predicted trajectory of vehicles, the predicted trajectory of pedestrians and the current speed of vehicles, the weight of backpropagation is updated to learn the probability distribution of offensive pedestrians.

5.3. Analysis of training results

For the vehicle trajectory prediction of the first LSTM network, this paper uses real vehicle trajectory data to evaluate the accuracy of network prediction. LSTM network is used on the actual vehicle track of 3.5 seconds, and each time step is predicted 3.5 seconds in advance to obtain the future vehicle track. The second LSTM network uses the same evaluation method for pedestrian trajectory prediction. Figure 4 shows the prediction and evaluation results of the two trajectories. It can be seen from the figure that the prediction accuracy decreases nonlinearly with the extension of the prediction time. Figure 4(b) shows the fitting between the predicted value and the real value of the last layer of ordinary BP network, which can be used to evaluate the accuracy of the model in judging the unacceptable behavior. It can be seen from Figure 5 that the goodness of fit of the model for training set and verification set is above 0.89, and the overall goodness of fit reaches 0.939, showing good performance; The goodness of fit of the test sample is 0.61, which also shows good recognition accuracy.

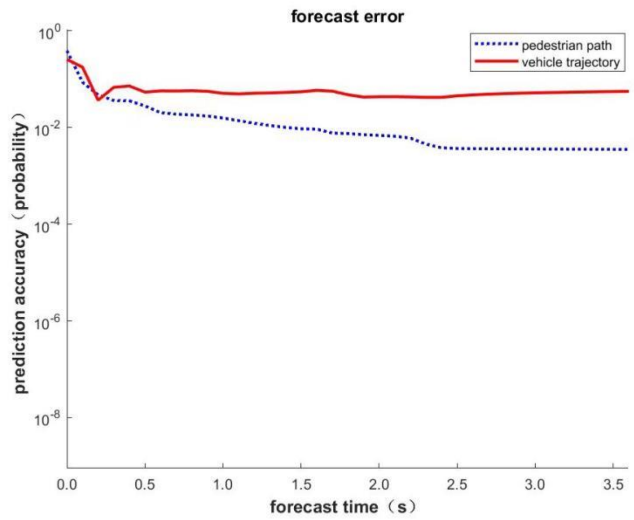


Figure 4. Track error

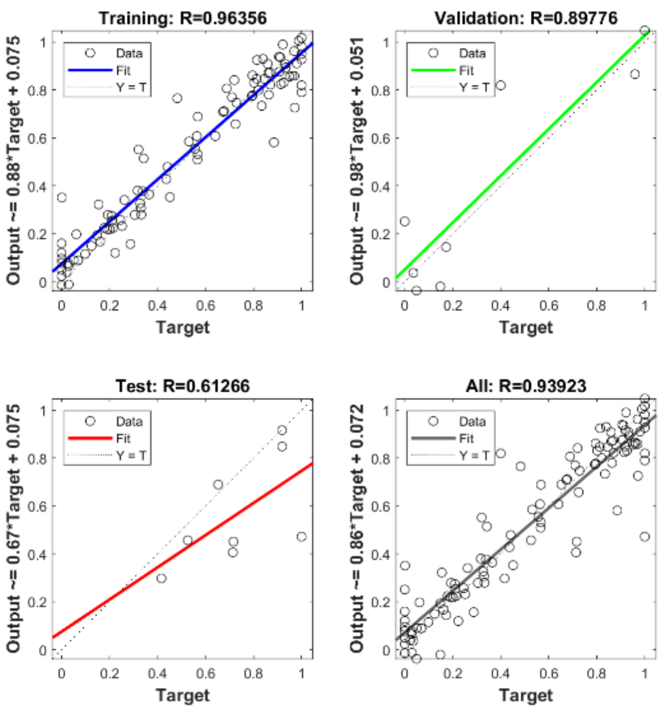
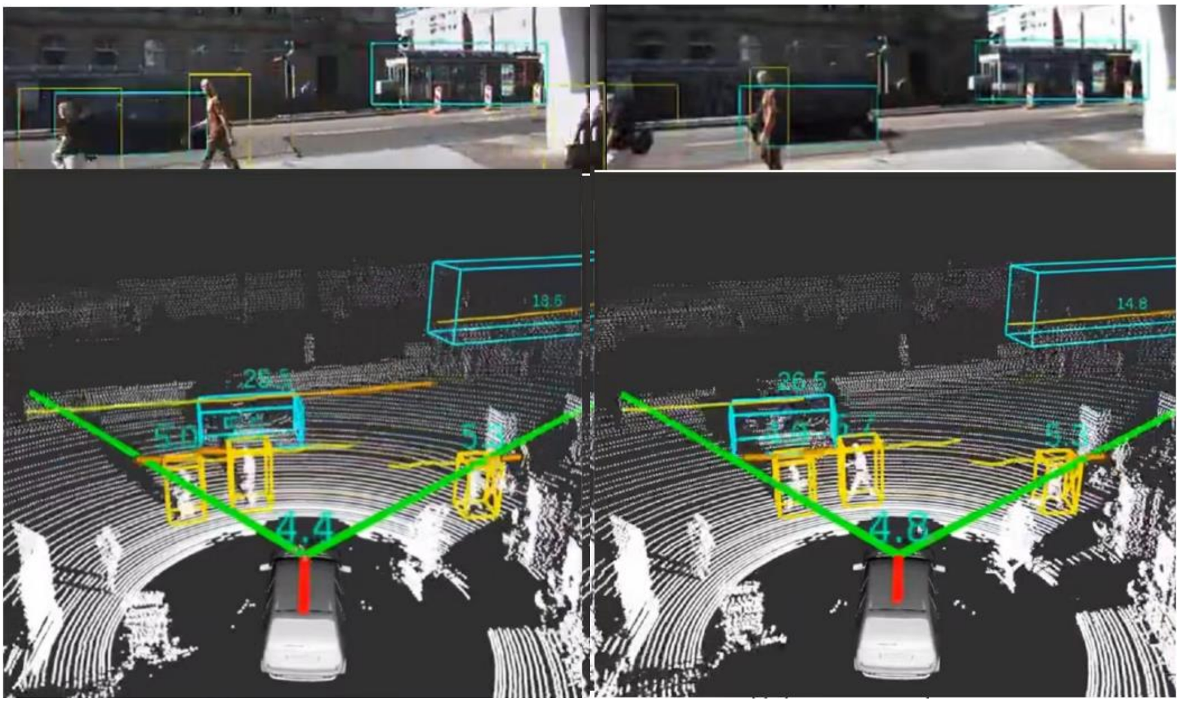


Figure 5. Fitting error

5.4. Empirical analyses

Figure 6 shows the prediction results obtained by using the model in the actual scenario. The blue number on the top of the vehicle is the current speed, the red line is the historical track of the vehicle, the blue box represents other vehicles, the yellow box represents pedestrians, the number above the pedestrians is the current walking speed of pedestrians, the yellow line is the historical track of pedestrians, and the arrow is the predicted pedestrian track. As showed in Figure 6, there are pedestrians passing in front of the vehicle, and the pedestrians are about to cross the street. The model calculates that the probability of the intersection between the predicted trajectory of the vehicle and the predicted trajectory of the pedestrians is 0.83, so the vehicles may behave irreverently. At this time, the current speed is 4.4 km/h, while the speed 2 seconds ago was 4.8 km/h. Since the current speed is greater than 0.8 times the speed two seconds ago, the model judges that the vehicle may behave irreverently to pedestrians.



(a) After 2 seconds (b) 2 seconds ago

Figure 6. Forecast Example

6. Conclusion and prospect

Based on LSTM, this paper studies the recognition of the interaction between vehicles and pedestrians, and the recognition of the behavior of vehicles that do not yield to pedestrians. It realizes the prediction of the future trajectory of vehicles and pedestrians, and judges whether to yield to pedestrians based on the predicted and actual vehicle trajectory. The data comes from the nuScenes dataset, 70% for training and 30% for verification. Use LSTM's Python framework, with Adam as the optimizer. When testing the network, input the transverse and longitudinal characteristic data of the vehicle, as well as the pedestrian position and head direction data, the network can use the previous learning of training samples to analyze the driving intention, the future trajectory of the vehicle, and the future trajectory of the pedestrian in a probabilistic way, and empirically analyze whether there is a phenomenon that the vehicle does not yield to pedestrians.

The difference between this paper and previous studies is that the vehicle behavior characteristics and pedestrian behavior characteristics are combined to establish a joint model, based on which the occurrence probability of vehicle indecency behavior is obtained. The traffic control department can embed sensors and cameras in the vehicle to identify the vehicle's non comity based on the model in this paper. Thus realizing the supervision and regulation of drivers. The work in this paper is based on the image features of the front end of the vehicle. However, if there is more image information, the model can be trained through the top view, which may improve the accuracy of the model.

References

[1] Xia Jiawei, Zhu Xufang, Luo Yasong, Wu Zhaodong. Research on UAV trajectory tracking algorithm based on deep reinforcement learning [J]. Journal of Huazhong University of Science and Technology (Natural Science Edition): 1-8 [202-09-03].

[2] Li Wenli, Zhang Yinan, Wang Mengxin, Li An. Pedestrian trajectory prediction based on field of view mechanism [J]. Computer application research: 1-7 [2022-09-03].

[3] Liu Jinming, Zhang Yuyan, Zhang Biling. UAV track tracking algorithm based on LSTM-KF [J]. Journal of Beijing University of Posts and Telecommunications: 1-7 [2022-09-03].

[4] Min Chao, Zhan Wei, Zhang Yuqi, Lv Jianhua, Hong Shengbing, Dong Tianyu, She Jinhui, Huang Huazi. Track tracking and behavior research of stored grain pests based on Hungarian algorithm and LSTM network [J]. Chinese Journal of Cereals and Oils: 1-13 [2022-09-03].

[5] Yufan Ji, Lunwen Wang, Weilu Wu,Hao Shao,Yanqing Feng. A Method for LSTM-Based Trajectory Modeling and Abnormal Trajectory Detection[J]. IEEE Access,2020,8.

[6] Zhai Jingyu, Chen Jinli. Prediction method of pedestrian trajectory based on LSTM Attention millimeter wave radar [J]. Journal of Chinese Academy of Electronic Sciences, 2022,17 (06): 534-541

[7] Oberkirsch Laurin, Grobbel Johannes,Maldonado Quinto Daniel,Schwarzbözl Peter,Hoffschmidt Bernhard. Controlling a solar receiver with multiple thermochemical reactors for hydrogen production by an LSTM neural network-based cascade controller[J]. Solar Energy,2022,243.

[8] Yang Fan, Sang Yongsheng, Lv Jiancheng,Cao Jun. Prediction of gasoline yield in fluid catalytic cracking based on multiple level LSTM[J]. Chemical Engineering Research and Design,2022,185.

[9] Baloch Zartasha, Shaikh Faisal Karim, Unar Mukhtiar Ali. CNN-LSTM-Based Late Sensor Fusion for Human Activity Recognition in Big Data Networks[J]. Wireless Communications and Mobile Computing,2022,2022.

[10] Liu Yuan, Yu Wangyang, Gao Cong,Chen Minsi. An Auto-Extraction Framework for CEP Rules Based on the Two-Layer LSTM Attention Mechanism: A Case Study on City Air Pollution Forecasting[J]. Energies,2022,15(16).