Pedestrian trajectory prediction method based on the Social-LSTM model for vehicle collision

OXFORD

Transportation Safety and Environment, 2024, 6: tdad044

DOI: 10.1093/tse/tdad044
Advance access publication date: 19 December 2023
Research Article

Pedestrian trajectory prediction method based on the Social-LSTM model for vehicle collision

Yong Han[1,2,*], Xujie Lin[1], Di Pan[1,2], Yanting Li[1], Liang Su[1,3], Robert Thomson[4] and Koji Mizuno[5]

[1]School of Mechanical and Automotive Engineering, Xiamen University of Technology, Xiamen 361024, China;
[2]Fujian Province Key Laboratory of Advanced Design and Manufacturing of Buses, Xiamen 361024, China;
[3]Engineering Research Institute of Xiamen Kinglong United Automobile Industry Co., Ltd., Xiamen 361000, China;
[4]Chalmers University of Technology, Gothenburg 41296, Sweden;
[5]Department of Mechanical Systems Engineering, Nagoya University, Nagoya 464-8603, Japan.
*Corresponding author. E-mail: Yonghan@xmut.edu.cn

Abstract

Techniques for predicting the trajectory of vulnerable road users are important to the development of perception systems for autonomous vehicles to avoid accidents. The most effective trajectory prediction methods, such as Social-LSTM, are often used to predict pedestrian trajectories in normal passage scenarios. However, they can produce unsatisfactory prediction results and data redundancy, as well as difficulties in predicting trajectories using pixel-based coordinate systems in collision avoidance systems. There is also a lack of validations using real vehicle-to-pedestrian collisions. To address these issues, some insightful approaches to improve the trajectory prediction scheme of Social-LSTM were proposed, such methods included transforming pedestrian trajectory coordinates and converting image coordinates to world coordinates. The YOLOv5 detection model was introduced to reduce target loss and improve prediction accuracy. The DeepSORT algorithm was employed to reduce the number of target transformations in the tracking model. Image Perspective Transformation (IPT) and Direct Linear Transformation (DLT) theories were combined to transform the coordinates to world coordinates, identifying the collision location where the accident could occur. The performance of the proposed method was validated by training tests using MS COCO (Microsoft Common Objects in Context) and ETH/UCY datasets. The results showed that the target detection accuracy was more than 90% and the prediction loss tends to decrease with increasing training steps, with the final loss value less than 1%. The reliability and effectiveness of the improved method were demonstrated by benchmarking system performance to two video recordings of real pedestrian accidents with different lighting conditions.

Keywords: vehicle-to-pedestrian collisions; pedestrian trajectory prediction; YOLOv5; DeepSORT; Social-LSTM

Highlights

- Deep-learning algorithms based on YOLOv5, DeepSORT and Social-LSTM can achieve real-time detection, tracking and trajectory prediction of pedestrians in accident videos, and the trajectory prediction results have small errors and are consistent with pedestrian trajectories in real accident videos.
- Based on the theory of perspective transformation and direct linear transformation, the impact of video distortion on pedestrian prediction trajectory can be reduced, and the conversion between pixel coordinates and world coordinates of pedestrian prediction trajectory can be realized.
- By combining the world of pedestrian prediction trajectory and vehicle motion trajectory, the location of the pedestrian collision point can be accurately predicted for vehicle collision, providing a reference basis for intelligent vehicle collision avoidance sensing and decision fusion.

1. Introduction

According to a report by the World Health Organization [1], more than half of road traffic deaths are among vulnerable road users, including pedestrians and cyclists as well as motorcyclists and persons with disabilities or reduced mobility and orientation. A staggering estimated 1.35 million people die each year globally due to road crashes, of which 23% are pedestrians. Improvements in automotive active safety technologies have played an important role in reducing pedestrian accidents. Refs. [2, 3] examined the benefits of active safety systems in preventing accidents and reducing injury severity. These systems consist of four main components: environment sensing, crash risk assessment, decision-making and evasive measures [4]. Therefore, achieving accurate detection, tracking and trajectory prediction of vulnerable road users (VRUs) is very important to avoid accidents in the research of developing sensing, decision-making and control technologies for advanced assisted-driving vehicles and self-driving vehicles.

Currently, pedestrian detection relies on two main approaches: traditional machine-learning methods and deep-learning detection methods. Since the accuracy of the traditional target detection methods is not very high, the recognition effect is not very good, and the overall transportation speed will be slow when the computational volume is large. Nowadays, machine-learning methods are maturing and, compared to traditional manual feature extraction methods, they are able to convert pixel information in the input image into deeper, more abstract features. Machine-learning-based target detection methods are also much

Received: August 2, 2023. **Revised:** October 18, 2023. **Accepted:** November 27, 2023

better than traditional methods in terms of speed, accuracy and robustness. Among the machine-learning methods, the most representative ones are AdaBoost [5]; Log AdaBoost [6] and Support Vector Machines (SVM) [7], which express the target through artificially designed features, such as classical Haar-like features [8], histogram of oriented gradients (HOG) features [9, 10] and creativity support systems (CSS) features [11]. Felzenszwalb et al. [12, 13] proposed a HOG-based deformable part model (DPM) algorithm, and then improved the DPM algorithm, which was the most effective method in the field of traditional target detection. However, these traditional methods are unable to meet the demand for pedestrian detection and cannot contribute to the development of autonomous driving technologies. Deep-learning target detection methods can obtain high-level abstract features, which can be divided into one-stage and two-stage target detection algorithms according to the diffident processing framework. The one-stage detection algorithms include YOLO [14], Retina-Net [15] Single Shot Detector (SSD) [16], among others. In particular, the algorithms of the YOLO family have advantages regarding speed and accuracy. Among them, YOLOv5 is widely used because it is faster, more accurate and lighter in weight among the YOLO series [17].

In pedestrian tracking, there are traditional and deep-learning approaches. The earliest algorithms to appear in the field of target tracking were traditional target-tracking methods, and although the method is difficult to apply to today's changing environment, these methods have laid a certain foundation for subsequent research. Traditional multi-target tracking methods include nearest neighbour criteria filtering (NNSF) [18], joint probability density data association (JPDA) [19, 20], network flow data association (NFDA) [21] and multiple hypothesis tracking (MHT) [22, 23]). Deep-learning-based multi-target tracking methods include the SORT (Simple Online Realtime Tracking) algorithm proposed by Bewley et al. [24]. Wojke et al. [25] designed the DeepSORT algorithm based on SORT. The algorithm used a deep association metric containing the target appearance features learned by a generalized residual network instead of the original association metric obtained by Kalman filtering frame data. The algorithm added cascade matching to the matching module and introduced a state update strategy to further improve the performance of the tracking algorithm [26].

In the realm of trajectory prediction, traditional methods often utilize artificial features to model pedestrian behaviour. Among these methods, the Social Force (SF) model [27], proposes that human motion is shaped by social forces. Bera et al. [28] understood global and local motion patterns from two-dimensional trajectories for predicting pedestrian motion in crowds. There are also Kalman filter-based methods [29] and dynamic Bayesian networks [30] methods for pedestrian motion prediction. Due to the complexity and variability of the pedestrian's motion, it is difficult to fully express it with manual rules, which can be well addressed by deep-learning-based methods. Therefore, Alahi et al. [31] proposed a Social Long and Short-Term Memory Neural Network (Social-LSTM), which considers that the trajectories generated by pedestrians are influenced by two constraints, namely obstacles to be avoided and other pedestrians. Based on these two conditions, the trajectory prediction problem was considered as a sequence generation problem and a Social-LSTM model was proposed. This data-driven approach achieved better results. Yagi et al. [32] first proposed perspective-based trajectory prediction, which combined three conditions of self-motion, target human scale and target pose to improve the accuracy of predicted pedestrian trajectories. Zhou et al. [33] improved the LaneGCN algorithm in several ways to obtain the trajectory prediction of vehicles. Ref. [34] employed an enhanced perspective change technique for vehicle detection, 3D bounding box estimation, tracking and subsequent velocity estimation. Meng et al. [35] used LSTM to accomplish the prediction of transverse longitudinal trajectories and thus lane-change trajectories. Palsodkar et al. [36] used perspective change to calculate human-to-human distance. Wang et al. [37] used direct linear transformation to conduct pedestrian velocity analysis. Wang et al. [38] researched trajectory data to derive crash prediction and Li et al. [39] obtained prediction models by conflicting trajectory data. Taken together, these studies of predictive modelling have shown that predicting pedestrian trajectories can help vehicles make decisions in traffic accidents, which shows the relevance of trajectory prediction data for traffic accident research and the prospective nature of our work in proposing trajectory prediction and collision prediction.

The above methods are important for pedestrian motion trajectory prediction. Through detection and tracking, the historical trajectories of pedestrians are recorded, and based on their historical trajectories, a prediction model predicts future trajectories. However, the following problems still exist. According to the limitations of pedestrian detection pointed out by Ref. [40], in real traffic scenarios, the accuracy and speed of detection are not only dependent on hardware devices but also highly disturbed by external environmental factors. 1) In pedestrian detection, the current popular pedestrian detection algorithms lack the speed and accuracy for the real-time processing required in real traffic scenarios and cannot be achieved with the extremely complex algorithms many researchers use to improve detection performance. 2) In pedestrian tracking, the existing trajectory tracking algorithms ignore the surface features of pedestrians and the targets are easily lost during tracking, resulting in data redundancy, which greatly reduces the effectiveness of tracking. 3) In trajectory prediction, pedestrian trajectory prediction models are mostly developed in simple situations and cannot be used to analyse transient behaviour in complex traffic accidents, which are still difficult to incorporate in VRU collision avoidance control strategies.

In the current study, a trajectory prediction model based on Social-LSTM was developed and assessed with actual traffic accidents. Fig. 1 shows the analysis flow of the pedestrian trajectory prediction. The main objectives were:

1) To combine the deep-learning algorithms of YOLOv5 and DeepSORT to achieve real-time detection and trac king of pedestrians, improve the accuracy of the prediction model and reduce data redundancy.
2) To establish pedestrian prediction trajectories from the first viewpoint of the vehicle using perspective transformation and direct linear transformation theory to identify possible collision points between the vehicle and pedestrians, to provide a reference basis for the development of an advanced in-vehicle pedestrian sensing system, and to avoid accidents.

2. Methods

2.1. Pedestrian detection model

The VRU detection algorithm model of YOLOv5 comprises four main parts in its structural diagram: input, backbone, neck layer and output layer (refer to Fig. 2). For the input layer, three techniques are employed: the Mosaic data enhancement method, adaptive anchor frame calculation and adaptive image scaling.

The backbone layer consists of the Focus structure and CSP (cross-stage local network) structure algorithms. The Focus

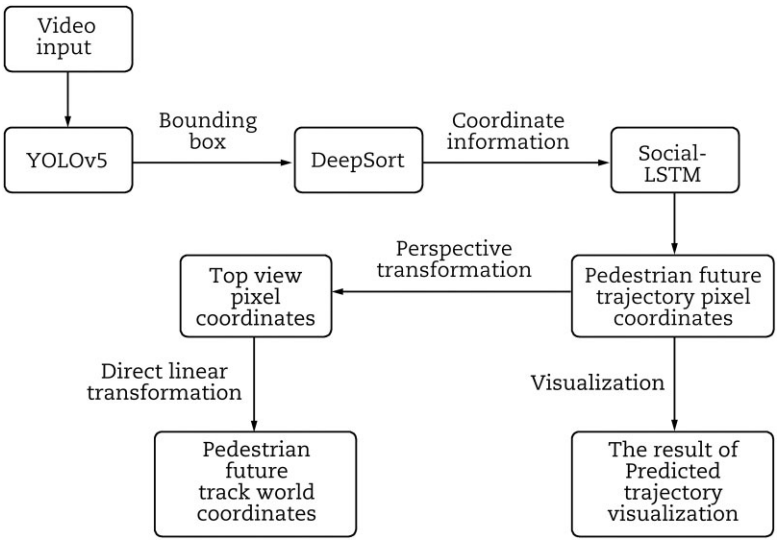

Fig. 1. Pedestrian trajectory prediction technology analysis flow.

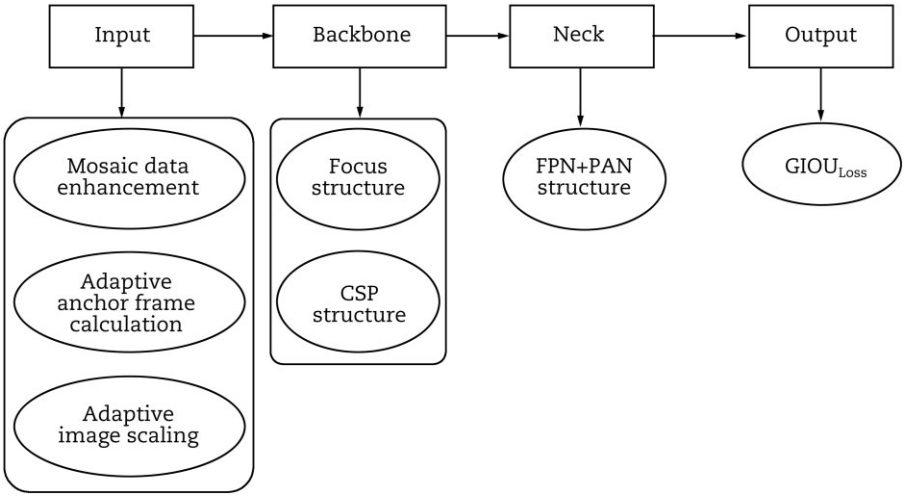

Fig. 2. YOLOv5 algorithm model structure.

structure is a slicing operation on an image structure with an original resolution of 608×608×3, which was first converted into a feature map with a resolution of 304×304×12. The 304×304×32 feature maps were obtained by convolution operation with 32 kernels.

Feature Pyramid Network (FPN) [41] and Path Aggregation Network (PAN) [42] used different structures to extract information from images. FPN used a top-down up-sampling approach so that the bottom feature map carried important semantic information. On the other hand, PAN used a bottom-up down-sampling approach so that the top features contained robust positional information. By fusing these two features, the resulting feature map contained robust semantic and positional information, which enabled accurate prediction of images of different sizes. In pedestrian detection, the extraction of language information and location information of target pedestrians is especially important, while FPN can extract semantic information of pedestrians and PAN can extract pedestrian location information to improve detection accuracy. The neck layer adopted the FPN+PAN structure to achieve the transmission of semantic and positional information.

The output layer calculated the loss of the detected frame by GIOU Loss. The loss consists of three components: bounding-frame regression loss, target confidence prediction loss and category prediction loss [34]. The minimum outer frame is introduced based on the IoU (Intersection over Union) feature to solve the problem of loss equal to 0 when there is no overlap between the detection frame and the real frame.

The binary cross-entropy losses in category prediction were defined as:

$$y_i = \text{Sigmoid}(x_i) = \frac{1}{1+e^{-xi}} \tag{1}$$

$$L_{\text{class}} = -\frac{1}{N}\sum_{n=1}^{N} y_i^* \log(y_i) + (1 - y_i^*)\log(1 - y_i) \tag{2}$$

where, N was the total number of categories, x_i was the predicted value of the current category and y_i was the probability of the current category obtained after the resultant activation function. The term y_i^* indicated the true value of the current category (0 or 1) and L_{class} indicated the category predicted loss.

The goal of image tracking was to define a bounding box around the object of interest. Defining the predicted bounding box as B

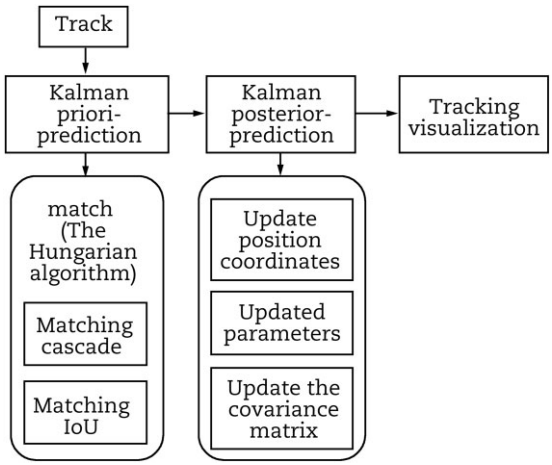

Fig. 3. DeepSORT algorithm model structure.

and B_{gt} as the real bounding box, a loss can be determined by the Intersection over Union defined in equation (3).

$$IoU\left(B, B_{gt}\right) = \frac{\left|B \cap B_{gt}\right|}{\left|B \cup B_{gt}\right|} \tag{3}$$

2.2. Trajectory-tracking model

Fig. 3 shows the model structure of the DeepSORT algorithm, which was centred on prediction, observation and updating information describing the target being tracked. The Kalman filter was utilized for position prediction, and the Hungarian algorithm was employed to match the predicted trajectory with the real trajectory. The matching methods included Cascade and IoU matching [43].

2.3. Trajectory-prediction model

Tracking objects and predicting their future position requires knowledge of their previous state and determining the characteristics of the motion patterns. Recurrent Neural Network (RNN) has been found to have insufficient long-term dependence due to recurrence disappearance [44]. Long Short-Term Memory (LSTM) could not obtain the interaction information between different pedestrians in the same scene [45]. However, Social-LSTM has been successfully used to predict the future trajectory of pedestrians in real time [46]. Fig. 4 shows the structure of LSTM [47]. The algorithm created information sharing between different pedestrians in a video scene by adding a pooling layer for each LSTM on the adjacent space. LSTM had three types of gate composition: forgetting gate f_t, input gate i_t and output gate o_t, where the role of the forgetting gate f_t was to determine the size of the previous moment cell state into the current moment, and was defined as:

$$f_t = \text{sigmoid}\left(W_f \times [h_{t-1}, x_t] + b_f\right) \tag{4}$$

where W_f denoted the weight of the forgetting gate, h_{t-1} was the input value at the previous moment, x_t was the input value at the current moment and b_f denoted the deviation term of the forgetting gate.

The role of the input gate i_t was to determine the size of the network output into the cell state at the current moment and was defined by:

$$i_t = \text{sigmoid}\left(w_i \times [h_{t-1}, x_t] + b_i\right) \tag{5}$$

where w_i was the weight value of the input gate and b_i was the deviation term of the input gate.

The role of the output gate o_t was to determine the size of the unit state into the current output value, whose expression was

$$o_t = \text{sigmoid}\left(w_o \times [h_{t-1}, x_t] + b_o\right) \tag{6}$$

where w_o was the weight value of the output gate and b_o was the deviation term of the output gate.

According to the forgetting gate f_t, the input gate i_t, and the output gate o_t, the output of the LSTM of a certain layer could be obtained

$$c_t = f_t \times c_{t-1} + i_t \times \tilde{c} \tag{7}$$

$$h_t = o_t \times \tanh\left(c_t\right) \tag{8}$$

where c_{t-1} denoted the state value at the previous moment, and $\tilde{c}$ denoted the current input cell and was defined by

$$\tilde{c} = \tanh\left(w_c \times [h_{t-1}, x_t] + b_c\right) \tag{9}$$

where w_c was the weight of the state value and b_c was the deviation term of the state value.

The main task of the Social-Pooling layer was to collect the hidden state information of neighbouring targets through spatial information. As shown in Fig. 5, the black dots and their surrounding parts represented the domain of the tracking target. The hidden state of the tracking target was gathered around a certain spatial distance, and the hidden state of the tracking target was calculated according to equation (10).

$$H_t^i(m, n, :) = \sum_{j \epsilon N_i} 1_{mn}\left[x_t^j - x_t^i,\ y_t^j - y_t^i\right] h_{t-1}^j \tag{10}$$

where h_{t-1}^j denoted the hidden state of the jth target in the LSTM at moment t-1, $1_{mn}[x - y]$ denoted whether (x, y) was within the grid (m, n) and N_i denoted the neighbour of the ith tracking target.

As shown in Fig. 6, where LSTM denoted LSTM neural network, the LSTM network used unsupervised learning, and the motion trajectory of the target in the video may be affected by the motion of the neighbouring targets. For this reason, the S-Pooling layer connected the LSTM network of the target to be detected with the LSTM networks of its neighbouring targets to form a new network called Social-LSTM, which was then used to predict the target's motion trajectory. Based on Social-LSTM to predict the future trajectory of pedestrians, we analyse the future motion trajectory of target pedestrians in traffic accidents, and then derive the location of the pedestrian–vehicle collision point, which improves the reference basis for the research of vehicle collision avoidance for pedestrians.

2.4. Coordinate mapping model

To achieve real-time matching of the predicted future trajectory of the pedestrian with the perceived position information in the vehicle collision avoidance system, a new coordinate mapping model was developed based on Direct Linear Transform (DLT) and Perspective Transformation (PT). The model is shown in Fig. 7. The mapping process consists of two parts: perspective transformation and direct linear transformation. The perspective transformation used the condition that the three points of the perspective centre, image point and target point were co-linear to the perspective from one plane to another plane, which could still keep the shadow-bearing surface unchanged (Kocur et al. [34]). The direct linear transformation was a processing method of images that defined a relationship between the image coordinates and their corresponding object space coordinates [37]. Video images for pedestrian trajectory prediction were mostly collected by road surveillance cameras, and the height and viewing angle caused a perspective distortion of the captured scene. To solve the problem of

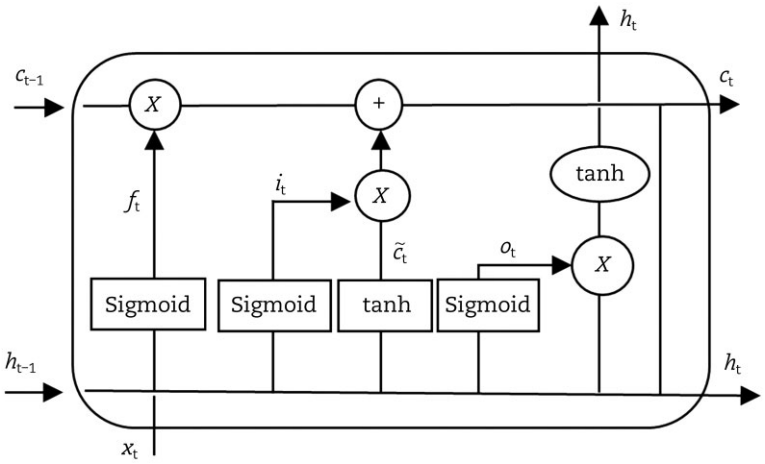

Fig. 4. LSTM structure.

Fig. 5. Social-Pooling structure.

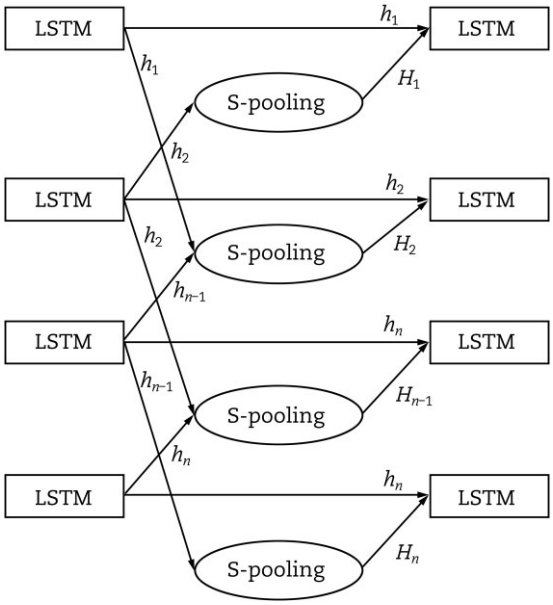

Fig. 6. Social-LSTM structure.

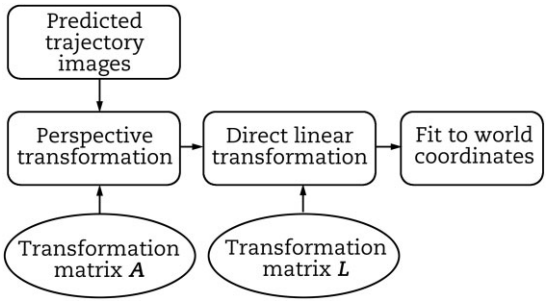

Fig. 7. Coordinate mapping model.

image distortion based on the surveillance camera viewpoint, it was transformed into a top view.

The perspective transformation model consists of the following perspective transformation matrix **A**.

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \tag{11}$$

where $\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$ denoted linear transformations, such as scaling, staggering and flipping, $[\, a_{31} \;\; a_{32} \,]$ denoted translation and $\begin{bmatrix} a_{13} \\ a_{23} \end{bmatrix}$ denoted the resulting perspective transformation. The perspective transformation model in the following equation was formed from the transformation matrix.

$$\begin{bmatrix} x' & y' & w' \end{bmatrix} = \begin{bmatrix} u & v & w \end{bmatrix} \mathbf{A} \tag{12}$$

where u and v were the pixel coordinates of the original image, x' and y' were the pixel coordinates after perspective transformation, w was the coordinate in three dimensions and, since the image was two-dimensional, w is 1.

$$x = \frac{x'}{w'} = \frac{a_{11}u + a_{21}v + a_{31}}{a_{13}u + a_{23}v + a_{33}} \tag{13}$$

$$y = \frac{y'}{w'} = \frac{a_{12}u + a_{22}v + a_{32}}{a_{13}u + a_{23}v + a_{33}} \tag{14}$$

The direct linear transformation theory was used to transform pixel coordinates of the predicted future motion trajectory of pedestrians to world coordinates. First, the relationship between the pixel coordinate system and the world coordinate system was analysed and the transformation matrix was obtained based on the linear transformation expression (11) and the selected control point coordinates. The transformation from pixel coordinates to the world coordinate system was completed by the transformation matrix.

U and V were pixel coordinates corresponding to world coordinates (X, Y), $\boldsymbol{L}$ was the transformation matrix and equation (11) was the direct linear transformation model between pixel coordinates and world coordinates.

$$\begin{bmatrix} U \\ V \end{bmatrix} = \boldsymbol{L} \begin{bmatrix} X \\ Y \end{bmatrix} \tag{15}$$

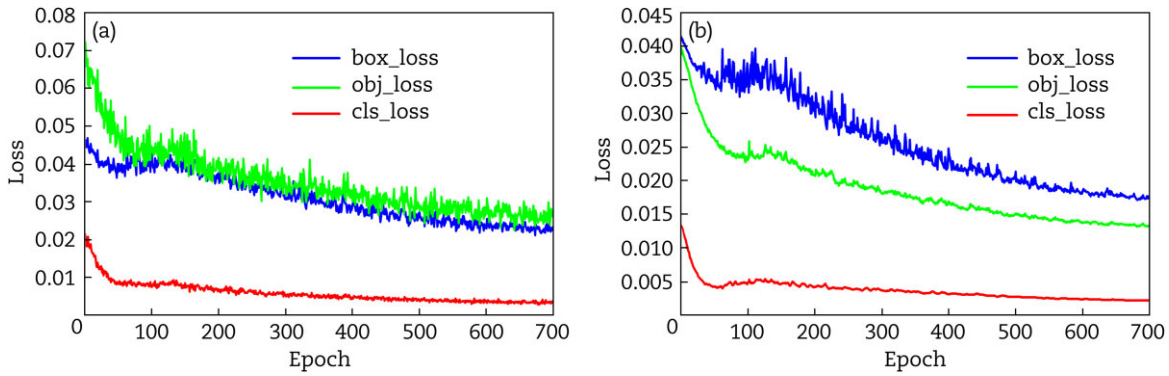

Fig. 8. YOLOv5 loss parameter variation: (a) training loss parameter variation diagram and (b) validation loss parameter variation diagram.

Table 1. YOLOv5 training and validation results.

COCO128	box_loss	obj_loss	cls_loss
Train	0.023	0.026	0.003
Test	0.017	0.013	0.002

where the transformation matrix $\boldsymbol{L}$ was:

$$\boldsymbol{L} = \begin{bmatrix} l_1 & l_2 & l_3 \\ l_4 & l_5 & l_6 \\ l_7 & l_8 & 1 \end{bmatrix} \tag{16}$$

Combining equations (11) and (12) yields the following equation

$$\begin{cases} U - \frac{l_1 X + l_2 Y + l_3}{l_7 X + l_8 Y + 1} = 0 \\ V - \frac{l_4 X + l_5 Y + l_6}{l_7 X + l_8 Y + 1} = 0 \end{cases} \tag{17}$$

2.5. Training and evaluation of the algorithm

The algorithm was trained using the COCO128 dataset [48]. This database has a small tutorial dataset consisting of the first 128 images of the entire COCO dataset. The COCO dataset could be used for image detection, semantic segmentation and image captioning. It contained 1.5 million targets, 80 target classes, 91 material classes and 250,000 pedestrians with key point annotation. The parameters of the training process were set: the learning rate was 0.01 for the initial learning rate and 0.2 for the cycle learning rate. The image size was 640×640×3. The training iteration period was 700 times. After the training, the algorithm was validated with another database, the ETH/UCY pedestrian trajectory dataset [49]. This database contained five subsets, namely eth, ucy, hotel, zara1 and zara2, which represent five scenes and 2206 human motion trajectories. Each subset includes an aerial view and the two-dimensional location of each individual. These subsets contain a variety of challenging scenarios, including human collision avoidance, humans crossing each other and group behaviour. The effectiveness of the procedure in dealing with real accidents under different lighting conditions was evaluated through cases in VRU-TRAVi ([50–52]; [40]). The videos were previously processed to derive the 'true' trajectories of the pedestrians and vehicles. The videos were then processed with the previously validated tracking and prediction algorithm to derive the pedestrian motion prediction data that would be available in a future pedestrian detection system. The two accident cases are described below.

Case 1 (Daytime): The pedestrian was walking on a pedestrian crossing. The pedestrian crossed without looking at the black SUV approaching from the left, and the driver of the black SUV did not see the pedestrian. The driver did not slow down or brake until after hitting the pedestrian.

Case 2 (Nighttime): The pedestrian was running on the pedestrian crossing. The driver of the vehicle had overtaken another vehicle when approaching the intersection, and failed to slow down as it entered the crosswalk, and struck the pedestrian.

3. Results

3.1. Detection and tracking algorithm validation

Fig. 8 shows the results of the training process of the YOLOv5 and DeepSORT systems. Figs. 8(a) and (b) show the trends of training loss and validation loss with increasing training steps (times), respectively. After 700 training steps, the loss generally tended to decrease and, according to Table 1, the final loss of the bounding box regression (box_loss) was 0.023, the loss of the target confidence prediction (obj_loss) was 0.026 and the loss of the category prediction (cls_loss) was 0.003. The loss functions of the training and validation sets remain low.

Fig. 9 shows the trend of each accuracy metric with the increase in step length, where the horizontal coordinate indicated the training step length and the vertical coordinate indicated the accuracy value.

Precision is calculated by the formula:

$$precision = \frac{TP}{TP + FP} \tag{18}$$

Recall is calculated by the formula:

$$recall = \frac{TP}{TP + FN} \tag{19}$$

where *TP*, *FP*, *FN* and *TN* are the actual categories in the binary confusion matrix, *TP* (True Positive) means actual positive samples and predicted positive samples, *FP* (False Positive) means actual negative samples and predicted positive samples, *FN* (False Negative) means actual positive samples and predicted negative samples and *TN* (True Negative) means actual negative samples and predicted negative samples. The mAP is the area enclosed after plotting with Precision and Recall as the two axes, m denotes the average and 0.5 denotes that the IoU threshold for determining positive and negative samples is taken as 0.5.

The figure shows that for the first 50 epochs, the indicators increased sharply, and with increasing training, the indicators gradually tended to be stable, in which the accuracy, recall and average

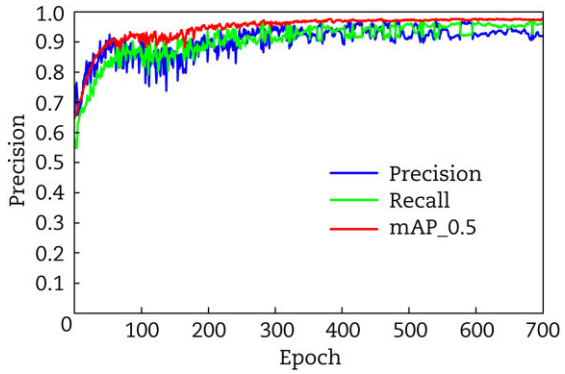

Fig. 9. Precision index change.

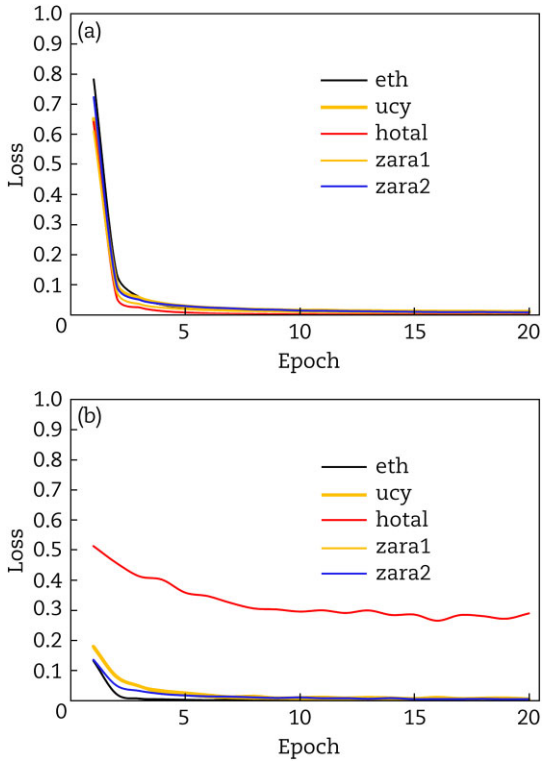

Fig. 10. Dataset training and validation results: (a) training results for each subset and (b) testing results for each subset.

accuracy tended to 91.9%, 96.0% and 97.5% of the average value, respectively, and the overall accuracy was higher than 90%, which indicated that the model had a good detection effect.

3.2. Trajectory prediction algorithm validation

The validation results of the Social-LSTM trajectory prediction algorithm are shown in Fig. 10(a). After 20 training steps, the overall loss showed a decreasing trend, where the loss of each of the subsets decreased to values less than 0.011 (ucy). The losses in the training and validation sets level off after five steps.

Fig. 10(b) shows that the training algorithm was tested using different subsets, four of which achieved good results. However, the subset hotels had a loss of 0.29 at the time of testing and required more iterations, but this tended to converge. This is due to the improved prediction model being more lightweight and the fact that this dataset was larger, with some subsets having larger errors in subsequent tests.

The prediction model is evaluated by evaluating the metrics average displacement error (ADE) and final displacement error (FDE), and it can be obtained that the average displacement error is 0.087 and the final displacement error is 0.092 (see Table 2), which are lower than the original mode prediction model, in which the average displacement error is reduced by 18.3% and the final displacement error is reduced by 51.9%. Therefore, using YOLOv5 and DeepSORT to detect and track pedestrians, combined with the historical trajectory of pedestrians, can reduce the error of the prediction model.

3.3. Non-collision scenarios track alignment

Two video examples of people traversing a pedestrian crossing in a road section in Xiamen, China were randomly captured using different monitoring perspectives, and the detection, tracking, prediction and perspective transformations of pedestrians in the video were performed. As can be seen in Fig. 11, the historical and predicted trajectories of pedestrians can be rendered, and the picture distortion due to the problem of the shooting angle is reduced after the perspective transformation.

3.4. Real accident video track alignment

The YOLOv5-DeepSORT model was used to detect and track the pedestrians (day and night) crossing the street in the two typical accidents from VRU-TRAVi. The trajectories of the pedestrians in the accident videos were then predicted by Social-LSTM and the results of detection tracking and predicted trajectories are visualized.

Table 2. Comparative analysis of indicators for the assessment of predictive models.

Metric	Datasets	Social-LSTM	Our-Social-LSTM	Error reduction rate/%
ADE	eth	0.50	0.0742	42.48
	hotel	0.11	0.0969	1.31
	zara1	0.22	0.0826	13.74
	zara2	0.25	0.0994	15.06
	ucy	0.27	0.0820	18.80
	Average	0.27	**0.0870**	**18.30**
FDE	eth	1.07	0.0886	98.14
	hotel	0.23	0.0986	13.14
	zara1	0.48	0.0877	39.23
	zara2	0.50	0.0973	40.27
	ucy	0.77	0.0889	68.11
	Average	0.61	**0.0920**	**51.90**

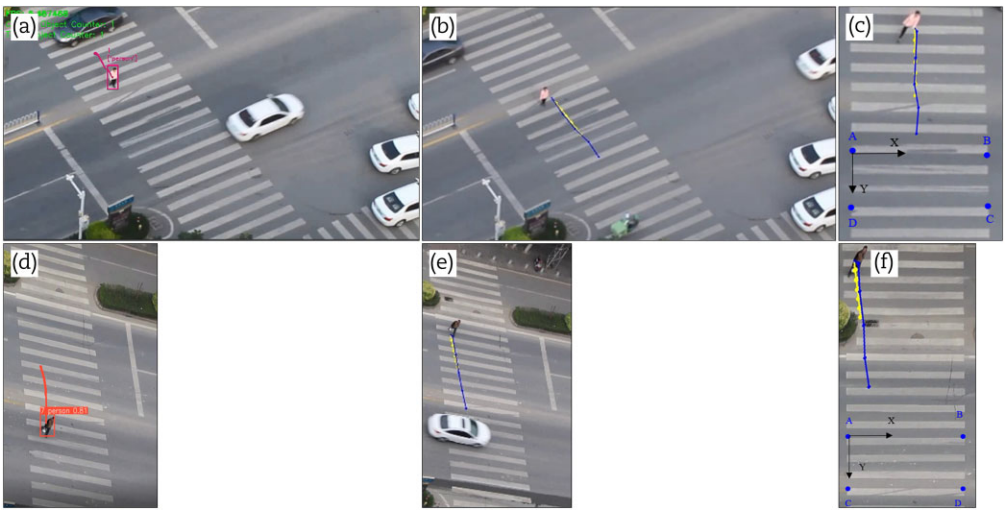

Fig. 11. Detection and tracking displayed on non-collision scenarios (a)~(f).

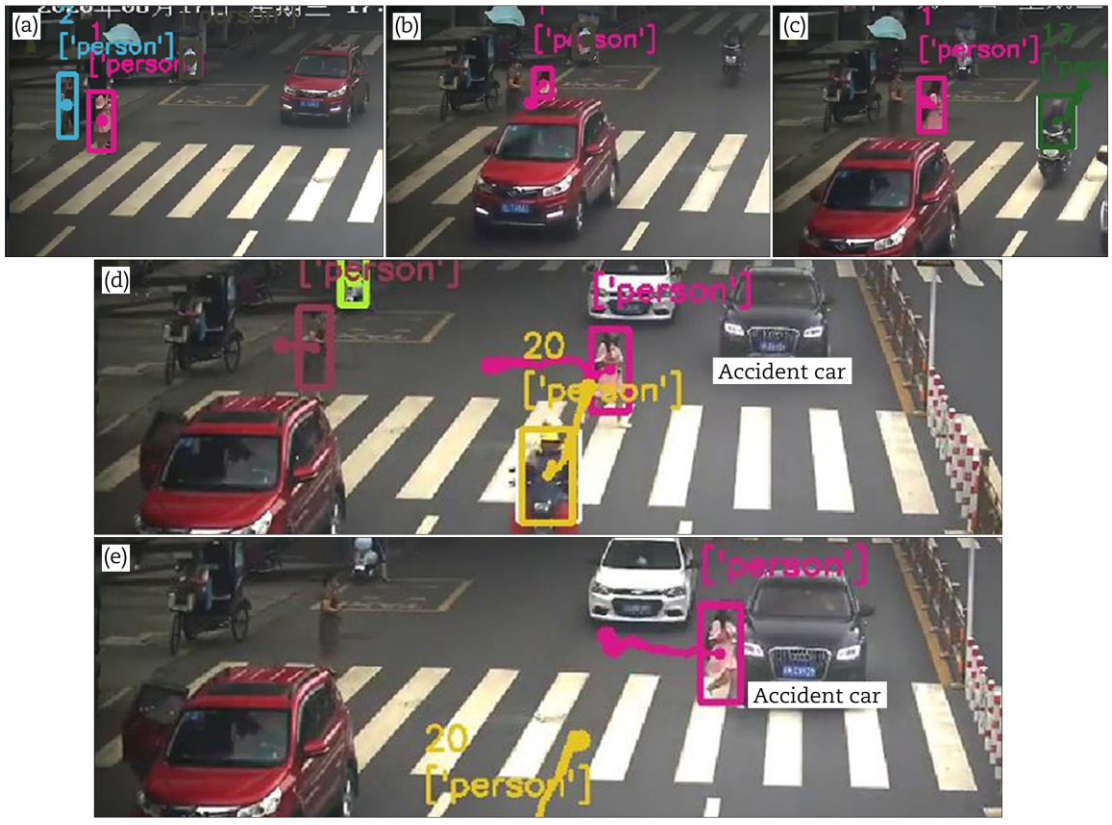

Fig. 12. Case 1. Detection and tracking displayed on accident video (a)~(e).

Fig. 12 shows the tracking effect of the pre-crash pedestrian trajectory in Accident Case 1. As can be seen from the first three images, after the red SUV occlusion the struck pedestrian was still represented by the pink detection frame, and there was no ID reassignment due to the occlusion. The algorithm's identification number of the struck pedestrian remained unchanged throughout the pre-crash phase (the colour of the detection frame remained unchanged). This shows that the YOLOv5-DeepSORT model was effective in detecting and tracking pedestrians in accident videos, even in the presence of occlusions.

The pedestrian trajectory prediction in the Case 1 video based on Social-LSTM is shown in Fig. 13. The yellow line was the real trajectory and the blue line was the predicted trajectory of the pedestrian. The blue line was consistent with the yellow line, indicating that the prediction model successfully predicted the future trajectory of the pedestrian.

Figs. 14(a) and (b) show the real-time tracking results for Case 2. The tracked trajectory was consistent with the real trajectory of the pedestrians. Figs. 14(c) and (d) show the results of pedestrian trajectory prediction, and the blue line was consistent with the

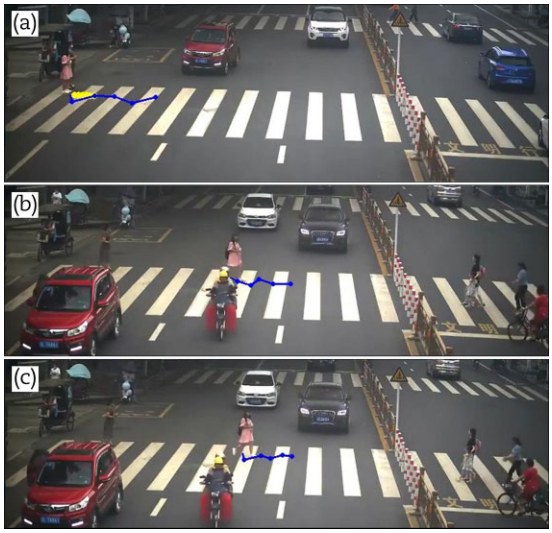

Fig. 13. Case1. Predicted trajectory before crashing (a)~(c).

yellow line, indicating that the prediction model predicted the future trajectory of pedestrians better.

3.5. Coordinate mapping of predicted pedestrian trajectories

Figs. 15(a) and (b) show the comparisons of pedestrian trajectory prediction before and after perspective transformation, respectively. From Figs. 14(c) and (d), it can be seen that the overall length and width distortion of the pedestrian crossing was significantly reduced after the perspective transformation of the video image, which indicated that the perspective transformation was more effective, and the pixel coordinates of the pedestrian trajectories predicted using the perspective transformation could be directly converted to world coordinates. The perspective-transformed image was used to establish a two-dimensional world coordinate system based on the pedestrian crossing, as shown in Fig. 15(c). The world coordinates of the future trajectories of the pedestrians could be known by a direct linear transformation.

Fig. 16 shows the results of fitting the pedestrian's predicted trajectory to the vehicle's motion trajectory in the world coordinate system. It can be observed that the transformation between the pixel coordinates of the pedestrian predicted trajectories and the world coordinates could be achieved by a direct linear transformation. The predicted pedestrian and vehicle motion trajectories and collision points were consistent with the real accident collision points, indicating that the pedestrian predicted trajectory coordinate mapping model was effective.

The same operation was performed for Case 2. After the perspective transformation of the original image, the overall distortion of the pedestrian crossing was reduced and the trajectory of the pedestrians from the top view was more suitable for the coordinate transformation (see Fig. 17(b)). Four control points (E, F, G and H) were distributed on the pedestrian crossing, and the pixel coordinates of the pedestrians in the figure could be transformed into world coordinates using the DLT (see Fig. 17(c)).

Fig. 18 shows the results of the pedestrian tracking, prediction and transformation process. The pedestrian's trajectory was predicted from the start of detection until the pedestrian was struck, and when combined with the vehicle's trajectory, the resulting collision point matched the location of the collision point in the original accident.

4. Discussion

Deep-learning based trajectory prediction drives the development of autonomous driving safety teleology. Machine vision prediction methods can minimize the probability of an accident by extracting more pedestrian features as well as other factors such as pedestrians' posture, surroundings and human–vehicle distance. This paper introduced the concepts of detection, tracking and prediction algorithms as a basis for solving the problem of pedestrian trajectory prediction in traffic accidents. A pedestrian trajectory prediction model for vehicle collision accident scenarios was proposed. Among them, considering the distortion caused by the shooting angle of the accident video and the limitation of the pixel coordinates, a perspective transformation was added to the trajectory prediction to correct the distortion of the pedestrian crossing in the accident as the benchmark of the direct linear transformation, and then the pixel coordinates of the pedestrians were converted to the world coordinates and the validity of the trajectory prediction was verified, and the pedestrian trajectory was fitted to the vehicle trajectory under the world coordinate system to predict the location of the pedestrian–vehicle collision point.

The target detection (YOLOv5) and multi-target tracking algorithm (DeepSORT) were utilized to obtain better historical

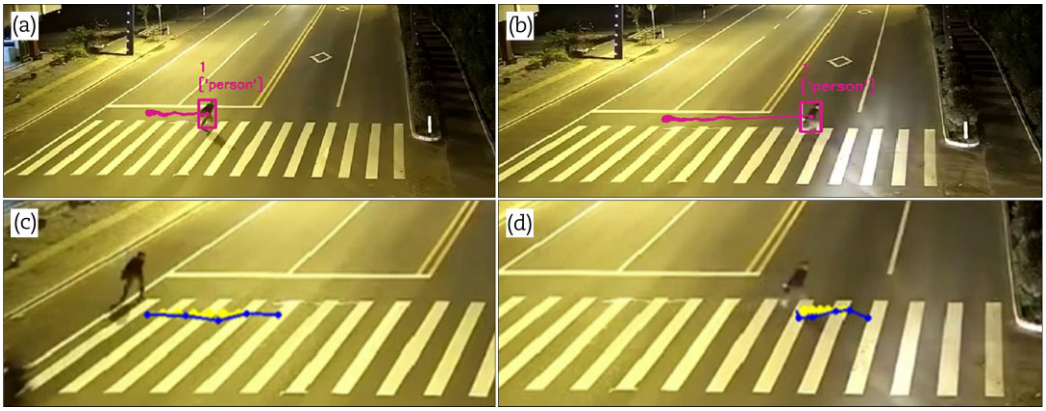

Fig. 14. Case 2. Detection tracking and predicted trajectory: (a) and (b) detection tracking of the pedestrian, (c) and (d) the predicted trajectory of the pedestrian.

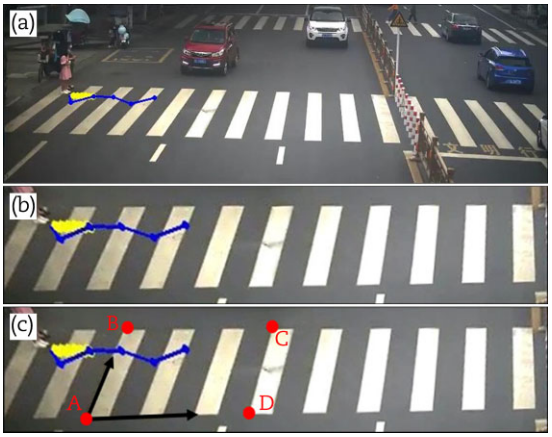

Fig. 15. Case1. Predicted trajectory before crashing: (a) no perspective transformation, (b) perspective transformation and (c) coordinate system references.

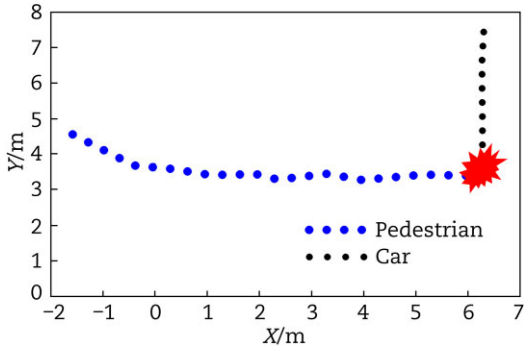

Fig. 16. Case 1. Pedestrian and vehicle trajectory mapping.

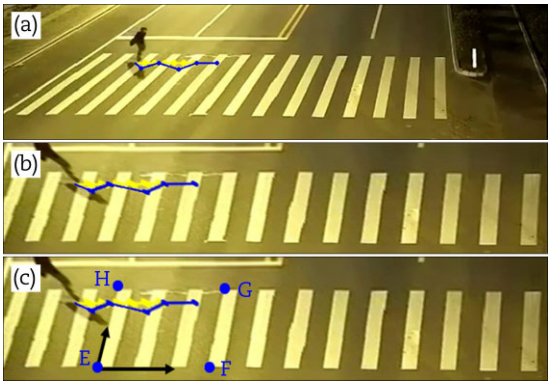

Fig. 17. Case 2. Predicted trajectory before crashing: (a) no perspective transformation before touching, (b) perspective transformation before touching and (c) coordinate system reference.

pedestrian trajectories. The results show that the accuracy of pedestrian detection was 93.9%. The error was reduced by 1.83% compared to the original Social-LSTM prediction model proposed by Alahi et al. [31]. In the trajectory prediction problem, Social-LSTM was utilized to predict the trajectories of pedestrians before the collision in the video, and the results show that most of its losses could be reduced to less than 1% after training. The predicted trajectories were transformed by two mathematical methods (Perspective Transformation and Direct Linear Transformation), and the predictions are more effective in improving the effective data of the vehicle collision avoidance systems by comparing with the trajectory prediction model of Yagi et al. [32].

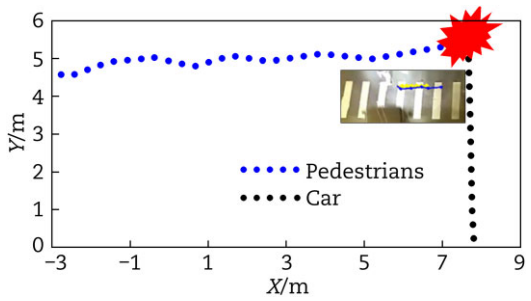

Fig. 18. Case 2. Pedestrian and vehicle trajectory mapping.

Finally, based on a real accident scene, the practical effects of YOLOv5, DeepSORT and Social-LSTM were verified. After tracking and predicting the pedestrian trajectories in two accident videos with different light scenes, the results show that the trajectories of the pedestrians before hitting in both accidents could be predicted, and the predicted trajectories were the same as the real trajectories. Combining pedestrian prediction and vehicle motion trajectory, the location of pedestrian–vehicle collision could be accurately predicted to provide a basis for the perception and decision-making of vehicle intelligent collision avoidance.

The current research method was still immature, and some shortcomings need to be further improved, such as that in Fig. 11, apart from the trajectory of the pedestrian hit (which was tracked continuously), there were individual trajectories that were not tracked effectively. The algorithm encountered difficulties with the lack of colour contrast between the pedestrian and the road surface, resulting in intermittent tracking of their trajectories. Therefore, the tracking model, or camera resolution, needs to be improved in further study. In addition, the posture of the pedestrian and the surrounding environmental factors could affect the movement trend of the pedestrian, and this paper has not yet addressed these factors. The algorithm models should improve in further study, considering more influential factors. This paper is currently using a single model; in subsequent research multiple models will be used, which will then be compared and analysed to enhance the credibility of this paper.

5. Conclusions

This paper presents a pedestrian trajectory prediction model for vehicle collision accident scenarios by applying a combination of detection, tracking and prediction algorithms to track and predict the trajectories of pedestrians in real traffic accidents. The following conclusions were obtained:

1) The historical pedestrian trajectories could be obtained using target detection (YOLOv5) and multi-objective tracking algorithms (DeepSORT), and the accuracy of pedestrian detection was higher than 90%.
2) The pre-crash trajectory of pedestrians in the video was predicted using a Social-LSTM, which could be trained to reduce most of the loss to less than 1%.
3) The effectiveness of the pedestrian trajectory prediction model was verified by using videos of vehicle-pedestrian col-

lision accidents with different illumination levels. The predicted pre-crash and future trajectories of pedestrians, and collision locations in both accidents, were generally consistent with the real accidents. The pedestrian trajectory prediction model could provide a reference for the development of sensing and decision-making technologies for intelligent vehicle collision avoidance.

4) Future work should further improve the accuracy of predictions and consider the impact of multiple factors on pedestrian trajectories, such as pedestrian posture and surrounding environmental conditions. Also, more different accident scenarios can be analysed in conjunction with each other to make the trajectory prediction model widely applicable in avoiding pedestrian accidents.

Acknowledgements

The authors would like to acknowledge the support of the Natural Science Foundation of China (Grant No. 51775466) and the Xiamen City Natural Science Foundation (No. 3502Z20227223).

Conflict of interest statement

None declared.

References

1. World Health Organization. *Global Status Report on Road Safety*. Geneva: World Health Organization, 2018.
2. Dong X, Yan S, Duan C. A lightweight vehicles detection network model based on YOLOv5. *Eng Appl Artif IntellIntelligence* 2022; **113**:104914.
3. Rosen E. Autonomous emergency braking for vulnerable road users. In: *IRCOBI Conference*, 2013, 618–27.
4. Pan D, Han Y, Jin Q et al. Probabilistic prediction of collisions between cyclists and vehicles based on uncertainty of cyclists' movements. *Transp Res Rec* 2023; **2677**:1151–64.
5. Jiang J, Xiong H. Fast Pedestrian Detection Based on HOG-PCA and Gentle AdaBoost. In: *2012 International Conference on Computer Science and Service System*, Nanjing, China, 2012, 1819–22.
6. Lin M, Luo H. Log AdaBoost: optimizing polylog loss function to improve the generalization performance of AdaBoost. In: *37th Youth Academic Annual Conference of Chinese Association of Automation (YAC)*, Beijing, China, 2022, 958–61.
7. Meus B, Kryjak T, Gorgon M. Embedded vision system for pedestrian detection based on HOG+ SVM and use of motion information implemented in Zynq heterogeneous device. In: *Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA)*. IEEE, 2017; 406–11.
8. Xing W, Zhao Y, Cheng R et al. Fast pedestrian detection based on Haar pre-detection. *International Journal of Computer and Communication Engineering* 2022; **1**:207.
9. Nagajyothi D, Charan PS, Zeeshan M et al. Image enhancement for pedestrian detection at night time. In: *2nd International Conference for Innovation in Technology (INOCON)*, IEEE, Bangalore, India, 2023, 1–7.
10. Pei WJ, Zhang YL, Zhang Y et al. Pedestrian detection based on HOG and LBP. In: *International Conference on Intelligent Computing*, Springer, Cham, 2014, 715–20.
11. Cosmo DL, Salles EOT, Ciarelli PM. Pedestrian detection system based on HOG and a modified version of CSS. In: *Seventh International Conference on Machine Vision (ICMV 2014)*, 2014, 97–101.
12. Felzenszwalb P, McAllester D, Ramanan D. A discriminatively trained, multiscale, deformable part model. In: *IEEE Conference on Computer Vision and Pattern Recognition*, 2008, 1–8.
13. Felzenszwalb P F, Girshick R B, McAllester D et al. Object detection with discriminatively trained part-based models. *IEEE Trans Pattern Anal Mach Intellintelligence* 2010; **32**:1627–45.
14. Redmon J, Farhadi A. YOLO9000: better, faster, stronger. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, 7263–71.
15. Lin T Y, Goyal P, Girshick R et al. Focal loss for dense object detection. In: *Proceedings of the IEEE International Conference on Computer Vision*, 2017, 2980–8.
16. Liu W, Anguelov D, Erhan D et al. SSD: single shot multibox detector. In: Springer, *European Conference on Computer Vision*, Cham, 2016, 21–37.
17. Yao J, Qi J, Zhang J et al. A real-time detection algorithm for Kiwifruit defects based on YOLOv5. *Electronics* 2021; **10**:1711.
18. Zhang J, Huang X, Shen Y et al. Nearest neighbor method to estimate internal target for real-time tumor tracking. *Technol Cancer Res Treat* 2018; **17**:1533033818786597.
19. Ainsleigh PL, Luginbuhl TE, Willett PK. A sequential target existence statistic for joint probabilistic data association. *IEEE Trans Aerosp Electron Syst* 2020; **57**: 371–81.
20. Fortmann T, Bar-Shalom Y, Scheffe M. Sonar tracking of multiple targets using joint probabilistic data association. *IEEE J Oceanic Eng* 1983; **8**:173–84.
21. Schulter S, Vernaza P, Choi W et al. Deep network flow for multi-object tracking. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, 6951–60.
22. Cox IJ, Hingorani SL. An efficient implementation of Reid's multiple hypothesis tracking algorithm and its evaluation for the purpose of visual tracking. In: *Transactions on pattern analysis and machine intelligence*, 1996, 138–50.
23. Xing W, Xie D, Wang J. GNN-guided track branch formation for multiple hypothesis tracking. In: *4th International Conference on Communications, Information System and Computer Engineering (CISCE)*, IEEE, 2022; 57–60.
24. Bewley A, Ge Z, Ott L et al. Simple online and real-time tracking. In: *IEEE International Conference on Image Processing (ICIP)*, Phoenix, AZ, USA, 2016, 3464–8.
25. Wojke N, Bewley A, Paulus D. Simple online and real-time tracking with a deep association metric. In: *IEEE International Conference on Image Processing (ICIP)*, IEEE, 2017, 3645–9.
26. Zagoruyko S, Komodakis N. Wide residual networks. 2016. arXiv preprint arXiv:1605.07146.
27. Helbing D, Molnar P. Social force model for pedestrian dynamics. *Phys Rev E* 1995; **51**:4282.
28. Bera A, Kim S, Randhavane T et al. GLMP-realtime pedestrian path prediction using global and local movement patterns. In: *2016 IEEE International Conference on Robotics and Automation (ICRA)*, 2016, 5528–35.
29. Koehler S, Goldhammer M, Bauer S et al. Stationary detection of the pedestrian? S intention at intersections. *IEEE Intell Transp Syst Mag* 2013; **5**:87–99.
30. Kooij JFP, Schneider N, Flohr F et al. Context-based pedestrian path prediction. In: *European Conference on Computer Vision*, Springer, Cham, 2014, 618–33.
31. Alahi A, Goel K, Ramanathan V et al., Social LSTM: human trajectory prediction in crowded spaces. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, 961–71.

32. Yagi T, Mangalam K, Yonetani R et al., Future person localization in first-person videos. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, 7593–602.
33. Zhou B, Zou JJ, Wu XJ et al. Research on the improvement of the LaneGCN trajectory prediction algorithm. *Transp Saf Environ* 2022; **4**:tdac034.
34. Kocur V, Ftáčnik M. Detection of 3D bounding boxes of vehicles using perspective transformation for accurate speed measurement. *Mach Vis Appl* 2020; **31**:1–15.
35. Meng XW, Tang JJ, Yang F et al. Lane-changing trajectory prediction based on multi-task learning. *Transp Saf Environ* 2023; **5**:tdac073.
36. Palsodkar P, Palsodkar P, Dubey Y et al. Pandemic surveillance through perspective transformation using YOLO and mobile net. *Intelligent Systems for Social Good*, Singapore: Springer, 2022, 193–205.
37. Wang G, Li J, Zhang P et al. Pedestrian speed estimation based on direct linear transformation calibration. In: *International Conference on Audio, Language and Image Processing*, IEEE, Shanghai, China, 2014, 195–9.
38. Wang J, Luo T, Fu T et al. Crash prediction based on traffic platoon characteristics using floating car trajectory data and the machine learning approach. *Accid Anal Prev* 2019; **133**:105320.
39. Li Y, Dalhatu S, Youn C et al. Analyzing freeway diverging risks using high-resolution trajectory data based on conflict prediction models. *Transp Saf Environ* 2024; **6**:tdad002.
40. Pan D, Han Y, Jin Q et al. Study of typical electric two-wheelers pre-crash scenarios using K-medoids clustering methodology based on video recordings in China. *Accid Anal Prev* 2021; **160**:106320.
41. Lin TY, Dollár P, Girshick R et al. Feature pyramid networks for object detection. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, 2117–25.
42. Liu S, Qi L, Qin H et al. Path aggregation network for instance segmentation[C]. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, 8759–68.
43. Ling L, Tao J, Wu G. Pedestrian detection and feedback application based on YOLOv5s and DeepSORT. In: *34th Chinese Control and Decision Conference (CCDC)*, IEEE, 2023, 5716–21.
44. Fragkiadaki K, Levine S, Felsen P et al. Recurrent network models for human dynamics. In: *Proceedings of the IEEE International Conference on Computer Vision*, 2015, 4346–54.
45. Charan S, Saravanan MS, Surendran R. Prediction of sufficient accuracy for human activity recognition using novel long short term memory in compared with decision tree. In: *2023 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS)*, IEEE, 2023, 609–14.
46. Si Z, Zhao H, Tang H et al. Pedestrian trajectory prediction by modeling the interactions using social LSTM extensions. *2022 China Automation Congress (CAC)*, IEEE, 2022, 4159–64.
47. Qiao Y, Xu K, Zhou K. Research on time series based on improved LSTM. In: *2023 IEEE 3rd International Conference on Power, Electronics and Computer Applications (ICPECA)*, IEEE, 2023, 951–8.
48. Lin TY, Maire M, Belongie S et al. Microsoft COCO: Common Objects in Context. In: *European Conference on Computer Vision*, Springer, Cham, 2014, 740–55.
49. Pellegrini S, Ess A, Schindler K et al. You'll never walk alone: modeling social behavior for multi-target tracking. In: *2009 IEEE 12th International Conference on Computer Vision*, IEEE, Kyoto, Japan, 2009, 261–8.
50. Han Y, Li Q, Wu H et al. Analysis of vulnerable road user kinematics before/during/after vehicle collisions based on video recordings. In: *Proceedings of the IRCOBI Conference*, 2017.
51. Han Y, Li Q, Wang F et al. Analysis of pedestrian kinematics and ground impact in traffic accidents using video recordings. *Int J Crashworthiness* 2018; **24**:211–20.
52. Li Q, Han Y, Mizuno K. Ground Landing Mechanisms in Vehicle-To-Pedestrian Impacts Based on Accident Video Records. In: *SAE Technical Paper*, 2018.
53 Eidehall A, Petersson L. Statistical threat assessment for general road scenes using Monte Carlo sampling. *IEEE Trans Intell Transp Syst* 2008; **9**:137–47.

Received: August 2, 2023. **Revised:** October 18, 2023. **Accepted:** November 27, 2023