Mixed Traffic Trajectory Prediction Using LSTM–Based Models in Shared Space

Hao Cheng and Monika Sester

Abstract Real–world behaviors of human road users in a non-regulated space (shared space) are complex. Firstly, there is no explicit regulation in such an area. Users self-organize to share the space. They are more likely to use as little energy as possible to reach their destinations in the shortest possible way, and try to avoid any potential collision. Secondly, different types of users (pedestrians, cyclists, and vehicles) behave differently. For example, pedestrians are more flexible to change their speed and trajectory, while cyclists and vehicles are more or less limited by their travel device—abrupt changes might lead to danger. While there are established models to describe the behavior of individual humans (e.g. Social Force model), due to the heterogeneity of transport modes and diversity of environments, hand-crafted models have difficulties in handling complicated interactions in mixed traffic. To this end, this paper proposes using a Long Short–Term Memory (LSTM) recurrent neural networks based deep learning approach to model user behaviors. It encodes user position coordinates, sight of view, and interactions between different types of neighboring users as spatio–temporal features to predict future trajectories with collision avoidance. The real–world data–driven method can be trained with pre-defined neural networks to circumvent complex manual design and calibration. The results show that ViewType-LSTM, which mimics how a human sees and reacts to different transport modes can well predict mixed traffic trajectories in a shared space at least in the next 3 s, and is also robust in complicated situations.

Keywords Shared space · Mixed traffic · Trajectory prediction · Long short–term memory

H. Cheng (✉) · M. Sester
Institute of Cartography and Geoinformatics, Leibniz University, Hannover, Germany
e-mail: hao.cheng@ikg.uni-hannover.de

M. Sester
e-mail: monika.sester@ikg.uni-hannover.de

A. Mansourian et al. (eds.), *Geospatial Technologies for All*,
Lecture Notes in Geoinformation and Cartography,
https://doi.org/10.1007/978-3-319-78208-9_16

1 Introduction

In distinction to classic traffic designs which, in general, separately dedicate road resources to road users by time or space division, an alternative solution—shared space—has been proposed by traffic engineers. This concept was first introduced by the Dutch traffic engineer Hans Monderman in the 1970s (Clarke 2006). It was later formally defined by Reid as "a street or place designed to improve pedestrian movement and comfort by reducing the dominance of motor vehicles and enabling all users to share the space rather than follow the clearly defined rules implied by more conventional designs" (Reid 2009). This design allows mixed types of users (pedestrians, cyclists, and vehicles) to interact with each other and negotiate to take or give their right-of-way.

It is relatively easier and cheaper to construct less or non-regulated spaces than the classic traffic designs and more feasible for urban and crowded places (Karndacharuk et al. 2014). Nevertheless, efficiency and safety in shared spaces need to be fully investigated. At a micro level, understanding how road users behave and how we can foresee their behaviors after a very short observation time (e.g. 3 s) are crucial to traffic planning and autonomous driving in such areas. However, this is not a trivial task. Mixed traffic movement data, especially in shared spaces, contain various spatio–temporal features. The involved geographical space, objects and their associated multidimensional attributes change over time (Andrienko et al. 2011). A simple approach may be sufficient for simple situations, such as the Social Force model for pedestrian dynamics (Helbing and Molnar 1995). Robust approaches are required to handle complex situations when mixed traffic is present.

Human behaviors are affected by lots of factors which are very person dependent (e.g. age, gender, time pressure Kaparias et al. 2012). For this reason, modeling their decision–making process about where and when to go next in the interactions with others is a great challenge. These hidden characteristics of personality, however, will eventually be reflected by the change of their positions, orientations, speeds, acceleration, and deceleration. This phenomenon inspires us to build models which can directly leverage hidden characteristic features and mimic how a human sees and reacts based on his or her explicit motion sequences in the past together with the expected behavior of other traffic participants, and then predict his or her trajectories in the future.

There are models that take movement data as input for trajectory prediction for mixed traffic in shared spaces, but many of them still require domain experts and manual fine–tuning efforts (Schönauer et al. 2012; Rinke et al. 2017; Pascucci et al. 2017). On the other hand, data–driven models, for example, deep learning neural networks, especially recurrent neural networks have achieved massive success for sequence prediction in domains like handwriting and speech recognition (Graves 2013; Graves and Jaitly 2014). In this paper, Long Short–Term Memory (LSTM) recurrent neural network models are proposed for mixed traffic prediction in shared spaces, which circumvent manual model building and calibration procedures.

They are trained by feeding with users' motion sequences in the past along with user type and sight of view using a real–world dataset.

Outline of the paper. In this paper, we first summarize the works that have been published for mixed traffic modeling and prediction in shared spaces and the state–of–the–art data–driven approaches in related domains. Then we introduce our approach motivated by a work for pedestrian modeling in Sect. 3. A real–world dataset and evaluation metrics for our approach are described in Sect. 4. We report our experiments and results in the following sections. In the end, we conclude our paper with some interesting problems that we would like to investigate in future work.

2 Related Work

Mixed traffic in shared spaces The schemes of shared spaces have been a heated topic for an alternative traffic design. However, to the best knowledge of the authors, only a few studies have dealt with shared space modeling: simulating mixed traffic in shared spaces based on game theory (Schönauer et al. 2012) and mixed traffic modeling and prediction using an extended Social Force model with collision avoidance (Pascucci et al. 2015; Schiermeyer et al. 2016; Rinke et al. 2017; Pascucci et al. 2017). Nevertheless, the game proposed by Schönauer et al. for conflict handling is heavily hand-crafted and lacks flexibility—"the type of game, the number of players, the number of games repeated, and whether the game allows cooperation must be specified". On the other hand, in the studies of the extended Social Force model, mixed traffic is analyzed in a categorical fashion regarding involved transport modes, e.g. pedestrian versus pedestrian or pedestrian versus car. Their model does not provide a mechanism that can deal with arbitrary collisions regardless of user types.

Data–driven approaches in trajectory prediction In recent years, with the increased availability of computational power and large-scale datasets, data–driven approaches have been largely used for learning movement data. Long and Nelson summarized possible methods to learn trajectory–related movements (Long and Nelson 2013). Unsupervised learning, for example clustering, and segmentation are applied to recognize similar trajectory patterns (Morris and Trivedi 2009; Pelekis et al. 2011). Due to the divergence of mixed road users and their interdependence in shared spaces, these methods are not reliable when the involved objects and contexts change quickly.

Deep learning approaches in trajectory prediction There are deep learning approaches for behavior modeling, e.g. a conventional neural network based model for pedestrian behavior (Yi et al. 2016) and a recurrent neural network based model for car–following (Wang et al. 2017). But both of these networks are limited to homogeneous user types. Nevertheless, these works shed light on using deep learning approaches for trajectory modeling.

Initially, Long Short–Term Memory recurrent neural networks proved to be powerful for complex sequence generation, e.g. text and handwriting (Graves 2013) and speech recognition (Graves and Jaitly 2014). "They can be trained by processing real data sequences one step at a time and predict what comes next". This process is comparable to trajectory prediction—observing initial steps of movement and trying to forecast the future motion. In comparison to an isolated sequence in a text, a single trajectory cannot be predicted as an independent motion of a road user since there are other road users and factors in the vicinity impacting his or her behavior, the so-called repulsive and attractive effects (Helbing and Molnar 1995). In order to capture these social effects, a centralized bounding grid was introduced in (Alahi et al. 2016) to process the interactions with neighboring users when using LSTM for trajectory prediction (Social-LSTM). Experiments on five open datasets (Lerner et al. 2007; Pellegrini et al. 2009) showed Social-LSTM outperforming the classic model–based approaches, such as Social Force (Yamaguchi et al. 2011) and Iterative Gaussian Process (Trautman et al. 2013), for pedestrian trajectory prediction. In addition, the data–driven approach circumvents complex manual setups needed for fine–tuning these classic models.

However, Social-LSTM was only tested on pedestrians. There are distinctive patterns regarding transport modes. For example, the involved transport modes, environment, and density will impact the intensity of pedestrian reactions to conflicts; cyclists have limited flexibility to deal with collisions due to their bicycles; vehicles may behave prudently to avoid collisions with more vulnerable road users (Rinke et al. 2017). Moreover, equipped with rear mirrors and multiple sensors, vehicles have a larger sight of view compared with pedestrians and cyclists. Lacking a mechanism to handle different transport modes, Social-LSTM could not, up to now, be directly applied to mixed traffic trajectory prediction in shared spaces.

3 Methodology

In order to differentiate transport modes and apply Social-LSTM (Alahi et al. 2016) for mixed traffic trajectory prediction in shared spaces, we introduce a bounding grid which incorporates both user type and sight of view based on Social-LSTM. In the interactions, the regarding user is addressed as an ego-user, which is the same denomination used in (Rinke et al. 2017), and the other users in his or her vicinity are addressed as neighboring users.

Every user is trained as a single LSTM, whereas the interactions with neighboring users are filtered by the bounding grid mentioned above. The basic network structure for our models is derived from Social-LSTM (see Fig. 1). For the input layer, it has a spatial input part to store the user's x and y coordinates and a tensor input part to capture the neighboring users within a predefined bounding grid for each ego-user (see Fig. 3a). Instead of simply pooling a binary indicator to tell the ego-user about the existence of other users in a uniformly sized grid as in the Social-LSTM, the tensor input here also customizes the grid according to the ego-user's sight of

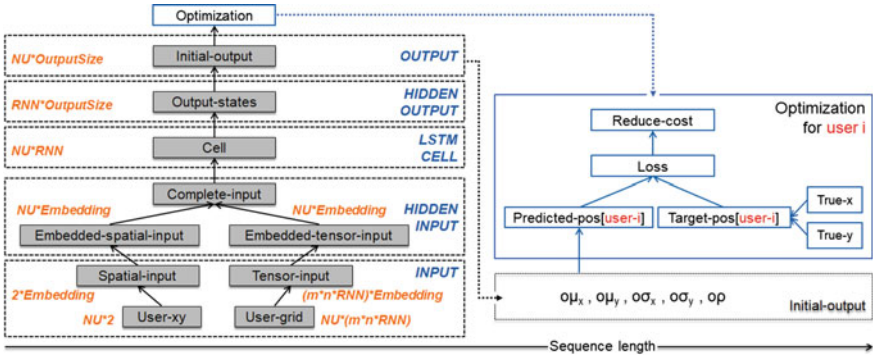

Fig. 1 Basic structure of the long short–term memory network. *NU*: number of users, $2 * Embedding$: embedding dimensions of weight W_S for spatial input, $(m * n * RNN) * Embedding$: embedding dimensions of weight W_T for tensor input, *RNN*: recurrent neural network size, $OutputSize = 5$

view (see Fig. 3c) and distinguishes user types. Equation (1) describes the process. $G_t^i(m, n, :)$ stands for the hidden state at time t for user i with a $m \times n$ cell bounding grid. This grid monitors all neighboring users whose positions are within ego-user i's grid and sight of view, and also stores the user type information for the ego- and neighboring users. Here, user j is from set N_i containing all user i's neighbors within $G_t^i(m, n, :)$. $\text{View}_t^i(\text{pos}_t^j)$ is a binary function that filters the neighboring users based on their positions in ego-user i's sight of view at time t—a value one is assigned if the neighboring user is in the ego-user's sight of view, otherwise zero is assigned. $(\text{type}^i, \text{type}^j)$ stores the pairwise user type information for ego-user i and neighboring user j. In total, there are nine different pairwise user types and they are coded in distinctive numerical values and stored in the $m \times n$ cell.

Since we can easily differentiate these two features—user type and sight of view—in Eq. (1), it empowers us to build controlled experiments to analyze the incorporation of user type and/or sight of view into different models. In order to guarantee valid comparisons, all the models defined in Sect. 5.1 have the same dimensions as described here but only with different pooling values in the bounding grid.

$$G_t^i(m, n, :) = \sum_{j \in N_i} (\text{type}^i, \text{type}^j)[\text{View}_t^i(\text{pos}_t^j)]. \tag{1}$$

From the input layer to the hidden input layer in Fig. 1, the spatial input and tensor input are embedded separately with Rectified Linear Unit (ReLU) as depicted by Eq. 2. W_S and W_T stand for the embedding weights for the spatial input and the tensor input respectively.

$$S_t^i = \text{ReLU}(W_S \cdot (x_t^i, y_t^i)); \quad T_t^i = \text{ReLU}(W_T \cdot G_t^i). \tag{2}$$

The embedded spatial input S_t^i and the embedded tensor input T_t^i are concatenated to form a complete input for the LSTM cell. Equation (3) denotes the forward propagation. h_{t-1}^i is the hidden state at time $t-1$ and W stands for the corresponding weights for the LSTM.

$$h_t^i = \text{LSTM}[h_{t-1}^i, (S_t^i + T_t^i), W]. \tag{3}$$

We apply the same method to train our models as (Alahi et al. 2016), which was initially used in (Graves 2013). Depicted by Fig. 1, the initial output of the neural network is a 5-dimensional vector ($o\mu_x$, $o\mu_y$, $o\sigma_x$, $o\sigma_y$, and $o\rho$) learned at time t, which is used to predict the position of user i for the next time-step $t+1$ using a bivariate Gaussian distribution (see Eq. (4)). μ^i is a 2-dimensional vector for the arithmetic means of the respective distributions in x and y coordinates. σ^i is a 2-dimensional vector for the corresponding standard deviations, and ρ^i is the correlation.

$$(\hat{x}^i, \hat{y}^i)_{t+1} \sim \mathcal{N}(\mu^i, \sigma^i, \rho^i)_t, \tag{4}$$

where

$$\mu^i = (\mu_x^i, \mu_y^i) = (o\mu_x, o\mu_y), \tag{5}$$

$$\sigma^i = (\sigma_x^i, \sigma_y^i) = (\exp(o\sigma_x), \exp(o\sigma_y)), \tag{6}$$

$$\rho^i = \tanh(o\rho). \tag{7}$$

The cost between the predicted position and the target position (true position) is computed by a negative log–likelihood loss function using Eqs. (4) and (8), and the complete loss for the user is the sum of all the costs in predicted time-steps.

$$\text{Loss} = -\sum \log \Pr(x_{t+1}^i, y_{t+1}^i | \mu_t^i, \sigma_t^i, \rho_t^i), \tag{8}$$

where

$$\mathcal{N}(\mu^i, \sigma^i, \rho^i) = \frac{1}{2\pi\sigma_x^i\sigma_y^i\sqrt{1-(\rho^i)^2}} \exp\left[\frac{-Z}{2(1-(\rho^i)^2)}\right], \tag{9}$$

$$Z = \frac{(x_{true}^i - \mu_x^i)^2}{(\sigma_x{}^i)^2} + \frac{(y_{true}^i - \mu_y^i)^2}{(\sigma_y{}^i)^2} - \frac{2\rho(x_{true}^i - \mu_x^i)(y_{true}^i - \mu_y^i)}{\sigma_x^i\sigma_y^i}. \tag{10}$$

To avoid overfitting, least square errors (L2) are used as the regularization to penalize all the learned weights. Hence, the total loss is the sum of the Loss computed by Eq. (8) and L2, and is optimized using Stochastic Gradient Decedent.

4 Dataset and Evaluation Metrics

4.1 Dataset

In this paper, LSTM–based models are evaluated on a real–world dataset provided by Pascucci et al. (2017). The whole area of a shared space is close to a busy train station in the German city of Hamburg and the shared space of a street is 63 m long (see Fig. 2a). There were two cameras positioned at C1 and C2 with an elevation of 7 m towards both directions of the street for incoming vehicles and cyclists. Vehicles are allowed to drive at a maximum speed of 20 km/h with a priority over other types of road users in the shared zone. Meanwhile, pedestrians and cyclists are allowed to cross the street at any point from both sides of the street. However, the captured data shows that rather than strictly followed the regulation, vehicles, cyclists, and pedestrians negotiated the space spontaneously and often gave priority to each other to share the space. More details can be found in Pascucci et al. (2017).

In a 30 min video, there were 1115 pedestrian, 22 cyclist, and 338 vehicle (331 cars and 7 motorcycles) trajectories. Figure 2b shows the corresponding velocity distributions. This video was divided into 1800 time-steps with each time-step lasting 0.5 s. After calibration, all the trajectories were tracked manually and projected onto a 2D plane with the help of video analysis and modeling tool Tracker.[1] After preprocessing, each trajectory contains information of user positions with time-step and user type. The first 10 min (31% of the dataset) are saved as a test set and the last 20 min (69% of the dataset) are used as a training set. 20% of the trajectories in the training set are selected as a validation set for tuning the models. Please note that the number of trajectories were not perfectly evenly distributed and none of the users returned to the shared space in the 30 min video footage.

4.2 Evaluation Metrics

To measure the performance between the predicted and true trajectories of each model, we use four metrics as follows:

1. *Euclidean distance*—The measurement used here is similar to the one used in (Alahi et al. 2016) and (Pellegrini et al. 2009). It is the mean square error (MSE) over all predicted positions and true positions.
2. *Hausdorff distance*—Unlike the Euclidean distance that gives a pointwise average displacement error between each predicted trajectory and true trajectory, the Hausdorff distance measures the largest distance from the set of predicted positions (X_{pred}) to the set of true positions (X_{true}, see Eq. (11)) (Munkres 2000). It can more explicitly show how far a predicted trajectory deviates from the true trajectory and also gives less penalty than the Euclidean distance when errors

[1] http://physlets.org/tracker.

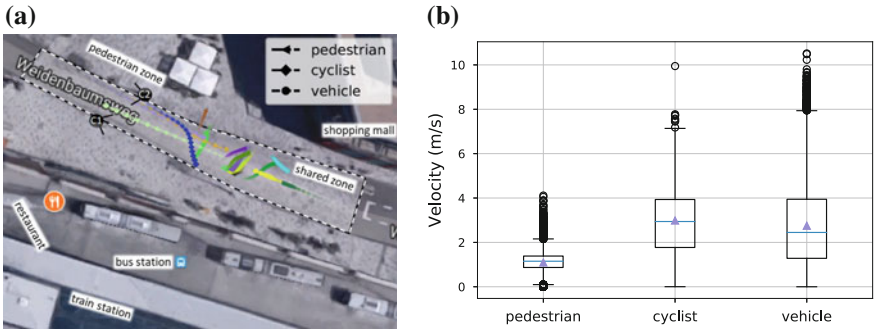

Fig. 2 **a** Layout of the shared space. Trajectories are denoted by color coded dot-lines with respective markers for different types of users. A color with larger size and opacity denotes a later time point. (Background image: Imagery ©2017 Google, Map data ©2017 GeoBasis–DE/BKG (©2009), Google); **b** Velocity distributions

are caused by time offsets. For example, in order to avoid collisions, the predicted trajectory for a user which depicts less accurate deceleration or acceleration compared with the true trajectory should be penalized less if the displacement error is small.

$$d_H(X_{\text{pred}}, X_{\text{true}}) = \max\{\sup_{x_{\text{pred}} \in X_{\text{pred}}} \inf_{x_{\text{true}} \in X_{\text{true}}} d(x_{\text{pred}}, x_{\text{true}})\}. \tag{11}$$

3. *Speed deviation*—Instead of measuring the MSE over all predicted positions and true positions, the speed deviation measures the MSE over all predicted speeds and true speeds in every time-step.
4. *Heading error*—This measurement computes the average degree for the angles between the predicted final heading directions and the true final heading directions over all the trajectories.

Altogether, these metrics allow comprehensive performance analyses for mixed traffic trajectory prediction in terms of positions, speeds, and heading directions.

5 Experiments

To analyze the contributions of incorporating user type and sight of view as described in Sect. 3, five LSTM–based models are tested on the aforementioned real–world dataset with the same configuration.

5.1 LSTM–Based Models

Table 1 lists the features which are used to feed each model respectively. Social-LSTM is the baseline model, which only considers the ego-user's position coordinates and corresponding pre-defined bounding grid. This model does not distinguish user types. In other words, pedestrians, cyclists, and vehicles are treated equally. It also does not consider the effect of the user's sight of view. The grid is the same for all four sides (right, left, front, and back, see Fig. 3a).

The average speeds of cyclists and vehicles in this shared space are about two times faster than the average pedestrian speed. Compared with cyclists and pedestrians, vehicles also occupy larger areas and generate bigger speed deviation (see Fig. 2b). Therefore, the bounding grid should be defined in a way that takes the type of ego-user into account. We call the corresponding model User-LSTM. To be more specific, vehicles and cyclists have twice the distance and 1.5–times the distance to the boundary of their bounding grid than pedestrians, respectively (see Fig. 3b). Given the reality that a user may have different levels of awareness regarding the type of neighbouring users (Rinke et al. 2017), UserType-LSTM not only defines a user-type aware bounding grid for the ego-user, but also accounts for the neighboring users' type. Therefore, the ego-user's interactions with different types of neighboring users are handled differently by this model.

However, the aforementioned models all have ego-user centralized grids. Unlike vehicles which are equipped with rear mirrors and sensors, pedestrians and cyclists normally do not have a good view of their back side. Humans have a maximum horizontal field of view of approximately 190° with two eyes, 120° of which make up the so-called binocular field of view (Henson 1993). As studied in personal space, people tend to preserve an elliptic protective zone around their body. Collision risks in front will be perceived higher than from the side (Gérin-Lajoie et al. 2005). Treating back and front sides of pedestrian or cyclist ego-users equally may lead to noisy information. Hence, on top of User-LSTM and UserType-LSTM, and for computational simplicity, we truncate the bounding grid according to the sight of view with 180° centralized towards the heading direction for pedestrian and cyclist ego-users (see Fig. 3c). The adjusted User-LSTM and UserType-LSTM are then called

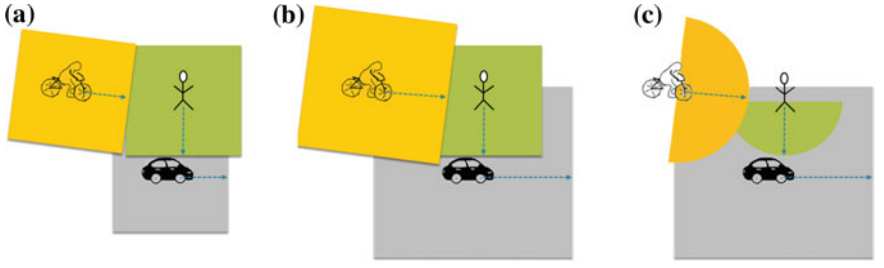

Fig. 3 Bounding grids in different models: **a** *Social-LSTM*, **b** *User-LSTM/UserType-LSTM*, and **c** *View-LSTM/ViewType-LSTM*

Table 1 LSTM–based models

Model name	Input features
Social-LSTM (**baseline**)	Coordinates, bounding grid
User-LSTM	Coordinates, user-type aware bounding grid
UserType-LSTM	Coordinates, user-type aware bounding grid, user-type aware interaction
View-LSTM	Coordinates, user-view aware bounding grid
ViewType-LSTM	Coordinates, user-view aware bounding grid, user-type aware interaction

Table 2 Details of hyper–parameters

Training	Testing
Sequence length: 12 time-steps	Observed sequence: 6 time-steps
Mini–batch size: 16	Predicted sequence: 6 time-steps
Learning rate: 0.003	
Number of epochs: 100	

View-LSTM and ViewType-LSTM, respectively. It is worth mentioning that even the back side is treated passively from the ego-user's perspective for pedestrians and cyclists as this area is still in the sight of approaching users.

5.2 Setup

The experiments were performed on a PC with the Intel(R) Core(TM) i5-6600T CPU and 16 RAM using the framework of TensorFlow.[2] This can be optimized with a more powerful machine with GPU(s) in the future work.

To achieve an optimal configuration for all of the models, 20% of the trajectories in the training set are selected as a validation set to tune hyper–parameters (e.g. learning rate, mini-batch size, the lengths of observed and predicted trajectories), which play an important role in controlling the algorithm's behavior but cannot be directly learned through training (Goodfellow et al. 2016). Table 2 lists the values for the hyper–parameters that are applied in our experiments.

All the models observe six positions in historical trajectories as the input to predict the next six positions. In other words, the models observe 3 s trajectories and try to predict the trajectories of the next 3 s. This can be easily scaled up for longer term prediction by modifying the sequence parameters accordingly. In general, 2.4 s are sufficient for most drivers for a brake reaction (Taoka 1989). Hence, here we only report performances for the next 3 s prediction.

[2] https://www.tensorflow.org.

6 Results

6.1 Evaluation of the Models

Our assumption is that a model that mimics how a human sees and reacts to different transport modes in a shared space (ViewType-LSTM) can well predict human behavior. To validate this assumption, the performance of such a model is compared with other models (Social-LSTM, User-LSTM, UserType-LSTM, and View-LSTM) which do not or do not fully utilize human characteristics in this regard (see Sect. 5.1).

In many situations, road users make decisions based on narrow gaps between the approaching users. For example, a pedestrian may decide to continue crossing the street when the distance of an incoming vehicle is slightly above his or her expected safety distance. Hence, the evaluation metrics should be able to capture small but non-negligible differences of the models. For a close observation of how the models can be used for predicting trajectories of the next 3 s, here we take a look at average values of Euclidean distance, Hausdorff distance, speed deviation, and heading error between the true trajectories and the predicted trajectories (see Sect. 4.2).

From Table 3 we can see the average Euclidean distances from the predicted trajectories to the true trajectories for mixed traffic (all transport modes), pedestrians, cyclists, and vehicles, respectively. UserType-LSTM and User-LSTM generate larger Euclidean distances than the baseline model for all road users, and more profound errors for cyclists and vehicles. On the other hand, ViewType-LSTM gives the best performance, beating the baseline model and View-LSTM. The average Euclidean distance for all transport modes is reduced by 9%, from 0.93 to 0.85 m, for ViewType-LSTM compared with the baseline model. For vehicles, the Euclidean distance is reduced by 11%, from 1.15 to 1.02 m.

The differences of performances are more pronounced when measured by the Hausdorff distance. ViewType-LSTM reduces the error by 13%, from 1.30 to 1.13 m for all transport modes compared with the baseline. Similar improvements can be found for pedestrians, cyclists, and vehicles. However, UserType-LSTM and User-LSTM fall behind the baseline model remarkably.

The average speed deviation of the predicted trajectories to the true trajectories for ViewType-LSTM is 0.25 m/s for all transport modes, which is slightly smaller than the baseline model (0.26 m/s). But more profound improvements can be found for cyclists (ViewType-LSTM 0.34 m/s vs. baseline 0.37 m/s) and vehicles (ViewType-LSTM 0.37 m/s vs. baseline 0.41 m/s). Interestingly, the speed deviations for pedestrians are almost identical for View-LSTM, ViewType-LSTM, and the baseline. This can be explained by pedestrians traveling at a relatively slow and constant speed compared with cyclists and vehicles (see Fig. 2b). In the dataset we use, there are many more pedestrians than cyclists and vehicles (see Sect. 4.1). Therefore, the overall improvement measured by the speed deviation for all transport modes for ViewType-LSTM is not as profound as the ones measured by the Euclidean distance or the Hausdorff distance.

Table 3 Prediction errors for LSTM–based models. Euclidean and Hausdorff distances are measured by meter, speed deviation is measured by meter per second, and heading error is measured by degree. The best values are highlighted in boldface

Metrics	User type	*Social-LSTM*	*User-LSTM*	*UserType-LSTM*	*View-LSTM*	*ViewType-LSTM*
Avg. Euclidean distance (m)	Mixed	0.93	0.98	1.11	0.91	**0.85**
	Pedestrian	0.77	0.87	0.97	0.75	**0.71**
	Cyclist	1.08	1.18	1.23	1.08	**1.01**
	Vehicle	1.15	1.09	1.30	1.11	**1.02**
Avg. Hausdorff distance (m)	Mixed	1.30	1.44	1.65	1.32	**1.13**
	Pedestrian	1.24	1.41	1.66	1.24	**1.08**
	Cyclist	1.39	1.73	1.60	1.33	**1.25**
	Vehicle	1.48	1.56	1.74	1.52	**1.26**
Avg. speed deviation (m/s)	Mixed	0.26	0.27	0.31	0.26	**0.25**
	Pedestrian	**0.17**	0.20	0.22	**0.17**	**0.17**
	Cyclist	0.37	0.37	0.44	0.38	**0.34**
	Vehicle	0.41	0.40	0.47	0.41	**0.37**
Avg. heading error (°)	Mixed	32.72	31.99	38.91	31.68	**27.74**
	Pedestrian	36.79	38.81	45.78	36.44	**31.79**
	Cyclist	6.28	5.49	7.77	5.99	**5.09**
	Vehicle	26.39	**20.95**	28.33	24.90	22.28

The last lines in Table 3 show how far the predicted trajectories rotate from the true trajectories regarding final heading directions. The smallest average errors between the predicted and the true trajectories for all user types, pedestrians, and cyclists are again given by ViewType-LSTM. Interestingly, the best performance for vehicles is given by User-LSTM, which slightly outperforms the second best one—ViewType-LSTM. Overall, the heading errors are much smaller for cyclists than for the other user types across all models. This is caused by the small cyclist set and their similar behaviors (see Sect. 4.1).

To summarize, incorporating user types simply by extending bounding grids, i.e. by increasing the potential influence area for different user types cannot lead to a better performance. To the contrary, it can even degrade the model's performance by including noisy information, especially from the back side of road users. This is further proven by truncating the bounding grids regarding the sight of view for pedestrians and cyclists. Moreover, acknowledgement of neighboring users' transport modes along with sight of view can further boost the accuracy of predictions for the next 3 s trajectories.

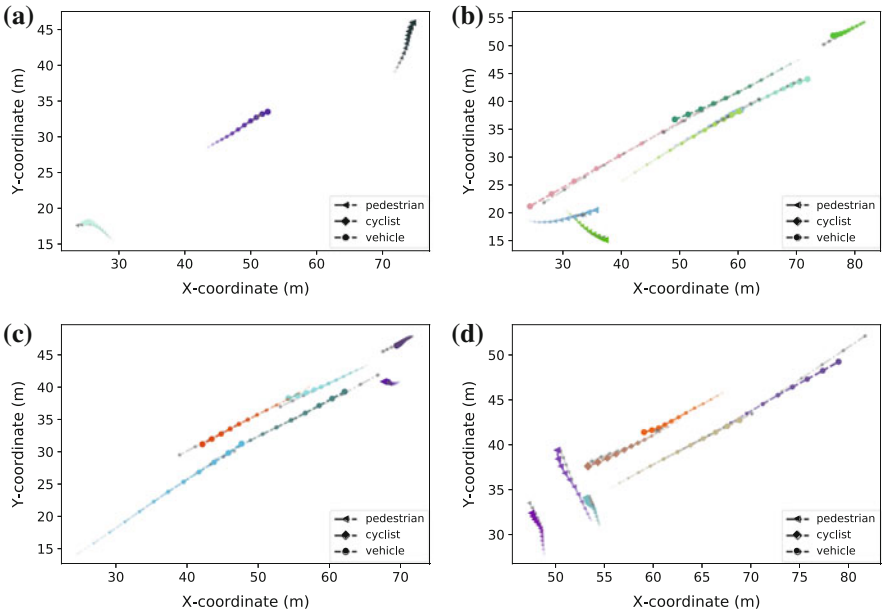

Fig. 4 Trajectory prediction in different situations: **a** Free flows of a vehicle and two pedestrians, **b** complicated situation with multiple users, **c** vehicles avoid an incoming pedestrian, **d** pedestrians avoid incoming vehicles and cyclist. True trajectories are denoted by black dot-lines with respective markers for different types of users. Predicted trajectories are color coded and a color with larger size and opacity denotes a later time point

6.2 Predicted Trajectories

In this section we show that ViewType-LSTM is able to mimic how a human sees and reacts to different transport modes in a shared space in the next 3 s and is also robust in complicated scenarios. Figure 4 shows different scenarios modeled by ViewType-LSTM.

From Fig. 4a we can see that the predicted trajectories for two pedestrians and one vehicle overlay their respective true trajectories. Since they are far from each other, their trajectories are barely impacted by interaction. On the lower left corner, ViewType-LSTM is able to correctly predict a left–turn for the pedestrian using a 3 s observed trajectory.

Figure 4b denotes a more complicated situation with multiple vehicles and pedestrians going in different directions. There is no collision in such busy mixed traffic. With only slight speed deviation and displacement for the upper right corner vehicle, predicted trajectories for the others are very close to the true trajectories.

Figure 4c, d depict different situations of how interactions happen between different road users and how ViewType-LSTM deals with potential collisions. The displacements from the predicted trajectories to the true trajectories in those two situations are barely noticeable, but most of them are caused by collision avoidance. In the upper right corner in Fig. 4c, there is a pedestrian waiting to cross the street. From the prediction, two approaching vehicles decelerate their speed to reduce the risks of hitting this pedestrian. On the lower left corner in Fig. 4d, three pedestrians are crossing the street. As a cyclist and a vehicle are approaching, ViewType-LSTM predicts a detour to the left side for the pedestrian who is very close to the incoming cyclist. It also predicts deceleration and slight left detours for the following two pedestrians to reduce the risks of potential collisions.

To more intuitively show how ViewType-LSTM can predict trajectories that have equal lengths as the observed trajectories, a scenario with mixed road users is depicted second by second in Fig. 5, in which a cyclist overtakes a vehicle from the right side to the left side after a pedestrian crossed the street in front of them. In this case, the prediction is also scaled up from a fixed length (six steps in 3 s) to a range of different lengths (1 s up to 6 s).

After only 1 s or 2 s observations, there are very few historical steps that can be referred to for the prediction. ViewType-LSTM, however, still predicts precise heading directions for each user, with average heading errors being 8.3° and 5.3°, respectively (see Fig. 5a, b).

After an appropriate length of observation (3 s), the performance for predicted trajectories is enhanced further. The predicted trajectories overlap their respective true trajectories (see Fig. 5c).

On the other hand, when the observed and predicted trajectory lengths are further increased to 4 and 5 s, the performances for the predictions of the cyclist and the pedestrian fall down. The reason is that, from the fifth second to the sixth second, both the cyclist and the pedestrian make a small right turn. Without observing the changes (the observed time point is only up to the fifth second), ViewType-LSTM keeps predicting straight trajectories for them (see Fig. 5d, e).

When the observation is extended to 6 s, the changes mentioned above are detected by ViewType-LSTM. In response to the changes, ViewType-LSTM calibrates the predicted trajectories to the right side for the cyclist and the pedestrian. In addition, the deceleration of the vehicle at later time points is also predicted by ViewType-LSTM, with the error for its speed deviation being 0.32 m/s (see Fig. 5f).

In summary, ViewType-LSTM generates reasonable and collision–free predictions in mixed traffic. Even with little information, it can estimate precise heading directions of road users. Moreover, ViewType-LSTM can be easily scaled up for longer term (e.g. up to 6 s) trajectory prediction. However, it is difficult to decide appropriate lengths for observed and predicted trajectories. A long observation for a short prediction might not be feasible in real–world trajectory prediction, but a short

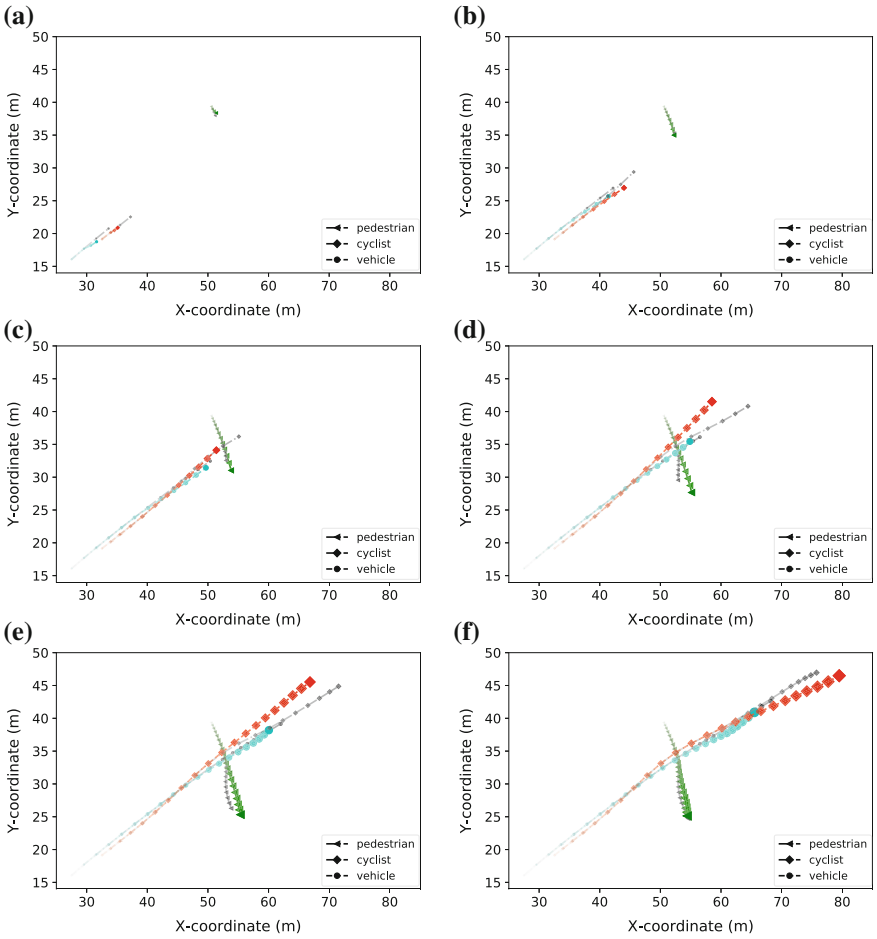

Fig. 5 Predictions of future trajectories that have equal lengths as the observed trajectories from 1 to 6 s: **a–f** Observing 1 s and predicting 1 s trajectories to observing 6 s and predicting 6 s trajectories, respectively. True trajectories are denoted by black dot-lines with respective markers for different types of users. Predicted trajectories are color coded and a color with larger size and opacity denotes a later time point

observation for a long prediction may fail to handle sudden changes made by road users at a later time. Hence, finding optimal observation and prediction lengths needs to be further investigated in future work.

7 Conclusion and Future Work

In this work we showed that LSTM–based models are capable of mixed traffic trajectory prediction in shared spaces. Spatio–temporal features—coordinates, sight of view, and interactions between different types of neighboring users—are encoded to mimic how a human sees and reacts to different transport modes. Instead of manual settings, LSTM–based models can be trained using real–world data for complicated traffic situations and can be easily scaled up for long term trajectory prediction.

In addition to the Spatio–temporal features mentioned above, user behaviors in shared spaces are also impacted by environment and context. An online survey shows that context– and design–specific factors significantly impact the comfort perceived by pedestrians and the willingness of car drivers to share road resources with others in shared spaces (Kaparias et al. 2012). Investigation of context–aware behavior modeling in shared spaces is a promising direction to further increase the accuracy of mixed traffic prediction.

Moreover, in order to extend our models on multiple and more balanced mixed trajectories in shared spaces with divergent space layouts, and make such data available for other studies, object detection and deep learning trajectory tracking techniques (i.e. Fully-Convolutional Siamese Networks, (Bertinetto et al. 2016)) will largely be employed for data acquisition and pre-processing procedures in our future work.

Acknowledgements The authors cordially thank the funding provided by DFG Training Group 1931 for SocialCars and the participants of the research project MODIS (Multi mODal Intersection Simulation) for providing the dataset of road user trajectories used in this work.

References

Alahi A, Goel K, Ramanathan V, Robicquet A, Fei-Fei L, Savarese S (2016) Social LSTM: human trajectory prediction in crowded spaces. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 961–971

Andrienko G, Andrienko N, Bak P, Keim D, Kisilevich S, Wrobel S (2011) A conceptual framework and taxonomy of techniques for analyzing movement. J Vis Lang Comput 22(3):213–232

Bertinetto L, Valmadre J, Henriques JF, Vedaldi A, Torr PH (2016) Fully-convolutional Siamese networks for object tracking. In: European conference on computer vision. Springer, pp 850–865

Clarke E (2006) Shared space: the alternative approach to calming traffic. Traffic Eng. Control 47(8):290–292

Gérin-Lajoie M, Richards CL, McFadyen BJ (2005) The negotiation of stationary and moving obstructions during walking: anticipatory locomotor adaptations and preservation of personal space. Motor control 9(3):242–269

Goodfellow I, Bengio Y, Courville A (2016) Deep learning. MIT Press. http://www.deeplearningbook.org

Graves A (2013) Generating sequences with recurrent neural networks. arXiv:13080850

Graves A, Jaitly N (2014) Towards end-to-end speech recognition with recurrent neural networks. In: Proceedings of the 31st international conference on machine learning (ICML-14), pp 1764–1772

Helbing D, Molnar P (1995) Social force model for pedestrian dynamics. Phys Rev E 51(5):4282

Henson DB (1993) Visual fields. Oxford Medical Publications, Butterworth-Heinemann Ltd (1772)
Kaparias I, Bell MG, Miri A, Chan C, Mount B (2012) Analysing the perceptions of pedestrians and drivers to shared space. Trans Res part F Traffic Psychol Behav 15(3):297–310
Karndacharuk A, Wilson DJ, Dunn R (2014) A review of the evolution of shared (street) space concepts in urban environments. Trans Rev 34(2):190–220
Lerner A, Chrysanthou Y, Lischinski D (2007) Crowds by example. In: Computer graphics forum, vol 26, no 3. Wiley Online Library, pp 655–664
Long JA, Nelson TA (2013) A review of quantitative methods for movement data. Int J Geogr Inf Sci 27(2):292–318
Morris B, Trivedi M (2009) Learning trajectory patterns by clustering: experimental studies and comparative evaluation. In: 2009 IEEE conference on computer vision and pattern recognition CVPR 2009. IEEE, pp 312–319
Munkres JR (2000) Topology. Prentice Hall
Pascucci F, Rinke N, Schiermeyer C, Friedrich B, Berkhahn V (2015) Modeling of shared space with multi-modal traffic using a multi-layer social force approach. Trans Res Procedia 10:316–326
Pascucci F, Rinke N, Schiermeyer C, Berkhahn V, Friedrich B (2017) A discrete choice model for solving conflict situations between pedestrians and vehicles in shared space. arXiv:170909412
Pelekis N, Kopanakis I, Kotsifakos EE, Frentzos E, Theodoridis Y (2011) Clustering uncertain trajectories. Knowl Inf Syst 28(1):117–147
Pellegrini S, Ess A, Schindler K, Van Gool L (2009) You'll never walk alone: modeling social behavior for multi-target tracking. In: 2009 IEEE 12th international conference on computer vision. IEEE, pp 261–268
Reid S (2009) DfT shared space project stage 1: appraisal of shared space. MVA Consultancy
Rinke N, Schiermeyer C, Pascucci F, Berkhahn V, Friedrich B (2017) A multi-layer social force approach to model interactions in shared spaces using collision prediction. Trans Res Procedia 25:1249–1267
Schiermeyer C, Pascucci F, Rinke N, Berkhahn V, Friedrich B (2016) A genetic algorithm approach for the calibration of a social force based model for shared spaces. In: Proceedings of the 8th international conference on pedestrian and evacuation dynamics (PED)
Schönauer R, Stubenschrott M, Huang W, Rudloff C, Fellendorf M (2012) Modeling concepts for mixed traffic: steps toward a microscopic simulation tool for shared space zones. Trans Res Rec: J Trans Res Board 2316:114–121
Taoka GT (1989) Brake reaction times of unalerted drivers. ITE J 59(3):19–21
Trautman P, Ma J, Murray RM, Krause A (2013) Robot navigation in dense human crowds: the case for cooperation. In: 2013 IEEE international conference on robotics and automation (ICRA). IEEE, pp 2153–2160
Wang X, Jiang R, Li L, Lin Y, Zheng X, Wang FY (2017) Capturing car-following behaviors by deep learning. IEEE Trans Intell Trans Syst
Yamaguchi K, Berg AC, Ortiz LE, Berg TL (2011) Who are you with and where are you going? In: 2011 IEEE conference on computer vision and pattern recognition (CVPR). IEEE, pp 1345–1352
Yi S, Li H, Wang X (2016) Pedestrian behavior understanding and prediction with deep neural networks. In: European conference on computer vision. Springer, pp 263–279