Joint Prediction for Kinematic Trajectories in Vehicle-Pedestrian-Mixed Scenes

Huikun Bi[1,2] Zhong Fang[1] Tianlu Mao[1] Zhaoqi Wang[1] Zhigang Deng[2]

[1]Institute of Computing Technology, Chinese Academy of Sciences [2]University of Houston

http://vr.ict.ac.cn/vp-lstm

ICCV 2019 Seoul, Korea

Contributions

We propose a novel multi-task learning architecture called **VP-LSTM** to **jointly predict the kinematic trajectories** of both vehicles and pedestrians simultaneously in vehicle-pedestrian-mixed scenes:

- **Vehicles** are treated as **rigid bodies** (i.e., represented with oriented bounding boxs, OBBs) and optimized by adopting the four-variate Gaussian distribution.
- **Pedestrians** are treated as **particles** and optimized by adopting the bivariate Gaussian distribution.

We introduce a large-scale and high-quality trajectory dataset for both heterogeneous vehicles and pedestrians under different traffic densities (**BJI** and **TJI**).

Motivations

Essential components are not well considered in existing research of trajectory prediction, especially in vehicle-pedestrian-mixed scenes:

- heterogeneous interactions: human-human, human-vehicle, and vehicle-vehicle,
- pedestrians as particles to describe their free movements,
- vehicles as rigid non-particle objects to describe their size,
- combination of positions and orientations to describe the accurate kinematic trajectories of heterogeneous vehicles.

Goal: To predict the **kinematic** trajectories for **all heterogeneous** agents in vehicle-pedestrian-mixed scenes **jointly** and **simultaneously**.

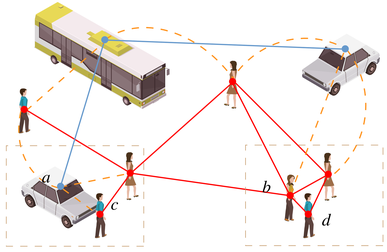
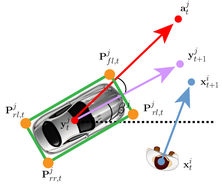


Figure: The vehicle a and pedestrian b in gray dash box have similar interactions with surrounding pedestrians. b walks freely to avoid collisions with d. However, the vehicle a, limited with kinematics, stops to avoid collisions with c.

Formulation

- **Pedestrians.** The input/output trajectory of a pedestrian p^i ($i \in [1,N]$) is a sequence formed of consecutive positions $\mathbf{x}_t^i = (x,y)_t^i$.
- **Vehicles** are represented with OBBs. The input trajectory of vehicle $v^j (j \in [1,M])$ is noted to a temporal sequence of the four vertices on the OBB ($\mathbf{P}_t^j = \{\mathbf{P}_{fl,t}^j, \mathbf{P}_{fr,t}^j, \mathbf{P}_{rr,t}^j, \mathbf{P}_{rl,t}^j\}$). We exploit the positions $\mathbf{y}_t^j = (x,y)_t^j$ and orientations $\mathbf{a}_t^j = (\alpha_x, \alpha_y)_t^j$ as the output trajectory of v^j at step t.



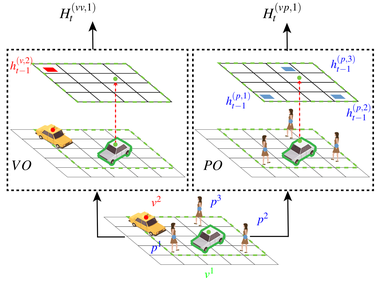
Mixed Social Pooling

$h_t^{(p,i)}$ ($h_t^{(v,j)}$) are the hidden states of p^i (v^j) after LSTMs. We build occupancy map VO and PO for both vehicles and pedestrians to share interactions. The pooling occurs on vehicle v^j as follows:

$$H_t^{(vp,j)}(m,n,:) = \sum_{k \in PO_{t-1}^j} h_{t-1}^{(p,k)}, \quad H_t^{(vv,j)}(m,n,:) = \sum_{l \in VO_{t-1}^j} h_{t-1}^{(v,l)}. \tag{1}$$

$$e_t^{(vp,j)} = \phi(H_t^{(vp,j)}, W_H^{vp}), \quad e_t^{(vv,j)} = \phi(H_t^{(vv,j)}, W_H^{vv}). \tag{2}$$

$e_t^{(vp,j)}$ and $e_t^{(vv,j)}$ are the vehicle-human and vehicle-vehicle interactions. As for pedestrian p^i, the human-human ($e_t^{(pp,i)}$) and human-vehicle ($e_t^{(pv,i)}$) interactions are obtained in a similar way.



Recursion for VP-LSTM. Recursion equations for pedestrian p^i and vehicle v^j are as follows:

$$\begin{aligned} h_t^{(p,i)} &= LSTM(h_{t-1}^{(p,i)}, e_t^{(x,i)}, e_t^{(pp,i)}, e_t^{(pv,i)}, W_{LSTM}^p) \\ h_t^{(v,j)} &= LSTM(h_{t-1}^{(v,j)}, e_t^{(\mathbf{P},j)}, e_t^{(vp,j)}, e_t^{(vv,j)}, W_{LSTM}^v) \end{aligned} \tag{3}$$

Optimization

To train the entire network end-to-end by minimizing respective objectives:

- **Pedestrians** exploit a bivariate Gaussian distribution ($d=2$) to predict the position $\hat{\mathbf{x}}_t^i = (\hat{x},\hat{y})_t^i$. Optimize the mean $\mu_t^{(p,i)} = (\mu_x,\mu_y)_t^{(p,i)}$, the standard deviation $\sigma_t^{(p,i)} = (\sigma_x,\sigma_y)_t^{(p,i)}$, and the correlation coefficient $\rho_t^{(p,i)}$ [2].
- **Vehicles** exploit a four dimensional Gaussian multivariate distribution (d=4) to predict the position $\hat{\mathbf{y}}_t^j = (\hat{x},\hat{y})_t^j$ and orientation $\hat{\mathbf{a}}_t^j = (\hat{\alpha}_x,\hat{\alpha}_y)_t^j$. We use the Cholesky factorization [4] to obtain the distribution by optimizing the values $\theta_{Lt}^{(v,j)}$ in $\mathbf{L}$ (4×4 upper triangular matrix) and mean parameters.

Displacements Prediction. The predicted kinematic trajectories of pedestrians and vehicles at t are respectively given by:

$$(\hat{x},\hat{y})_t^i \sim \mathcal{N}(\mu_t^{(p,i)}, \sigma_t^{(p,i)}, \rho_t^{(p,i)}), \quad (\hat{x},\hat{y},\hat{\alpha}_x,\hat{\alpha}_y)_t^j \sim \mathcal{N}(\mu_t^{(v,j)}, \theta_{Lt}^{(v,j)}). \tag{4}$$

Quantitative Evaluation

- Positions of vehicles/pedestrians: ADE, FDE.
- Orientation of vehicles: $\text{ADE}_\mathbf{O}$, $\text{FDE}_\mathbf{O}$. $\mathbf{P}_{fm}^j$ ($\hat{\mathbf{P}}_{fm}^j$) is the real (predicted) midpoint of the front of OBB oriented.

$$\text{ADE}_\mathbf{O} = \frac{\sum_{j=1}^{M}\sum_{t=T_{\text{obs}}+1}^{T_{\text{obs}}+T_{\text{pred}}} \|\hat{\mathbf{P}}_{fm,t}^j - \mathbf{P}_{fm,t}^j\|}{MT_{\text{pred}}}$$

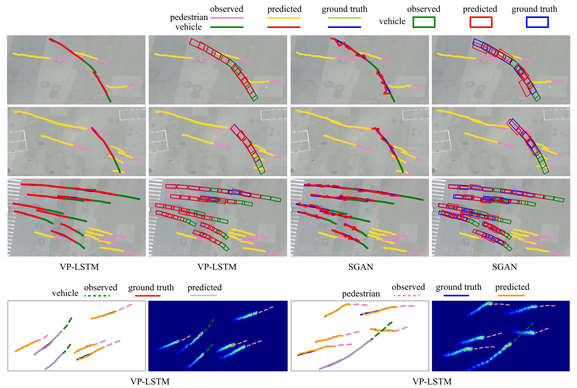
Table: Quantitative results for the vehicles only in NGSIM. Metrics for $T_{\text{pred}} = 8/12$ are reported in feet.

Metric	Dataset	V-LSTM [1]	S-LSTM [1]	SGAN [3]	VP-LSTM
ADE	NGSIM	34.01 / 40.00	11.73 / 15.16	4.56 / 6.52	**2.19 / 2.99**
FDE	NGSIM	43.87 / 52.64	20.03 / 23.64	9.13 / 10.99	**3.70 / 5.20**
$\text{ADE}_\mathbf{O}$	NGSIM	33.89 / 39.87	12.68 / 15.75	5.26 / 7.52	**3.29 / 4.00**
$\text{FDE}_\mathbf{O}$	NGSIM	43.77 / 52.50	22.59 / 24.95	11.07 / 13.10	**4.88 / 6.22**

Table: Quantitative results for objects in BJI and TJI. Metrics for $T_{\text{pred}} = 8/12$ are reported in pixels.

Metric	Dataset	Agent	V-LSTM [1]	S-LSTM [1]	SGAN [3]	VP-LSTM
ADE	BJI	Vehicle	66.69 / 85.48	29.05 / 51.41	20.21 / 24.65	**16.38 / 24.33**
		Pedestrian	34.70 / 48.91	25.26 / 46.89	17.82 / 20.33	4.92 / 6.39
		Average	44.13 / 62.09	26.88 / 48.49	18.64 / 22.53	**8.58 / 12.72**
ADE	TJI	Vehicle	142.17 / 185.93	46.17 / 85.52	26.82 / 39.86	**22.38 / 29.79**
		Pedestrian	115.44 / 135.97	41.19 / 75.55	19.81 / 25.89	**7.42 / 9.12**
		Average	125.27 / 154.31	43.13 / 79.22	24.42 / 34.73	**13.43 / 17.32**
FDE	BJI	Vehicle	114.11 / 152.91	61.49 / 126.03	38.36 / 44.68	**31.27 / 43.60**
		Pedestrian	54.64 / 81.72	56.93 / 111.05	32.57 / 40.52	7.55 / 10.44
		Average	72.17 / 107.38	58.54 / 116.47	35.00 / 42.62	**15.11 / 23.47**
FDE	TJI	Vehicle	215.94 / 303.54	103.67 / 203.90	48.22 / 56.95	**35.38 / 49.31**
		Pedestrian	156.29 / 192.92	92.55 / 177.10	39.98 / 49.31	**10.53 / 13.90**
		Average	178.21 / 233.52	96.91 / 186.97	43.42 / 55.43	**20.51 / 27.95**
$\text{ADE}_\mathbf{O}$	BJI	vehicle	65.51 / 83.78	42.55 / 65.70	27.56 / 33.47	**26.65 / 32.49**
	TJI	vehicle	140.52 / 183.60	50.35 / 88.44	29.69 / 38.65	**26.15 / 33.69**
$\text{FDE}_\mathbf{O}$	BJI	vehicle	112.01 / 149.81	76.18 / 135.49	43.61 / 49.59	**40.61 / 48.02**
	TJI	vehicle	213.04 / 299.39	105.94 / 203.24	50.94 / 64.49	**38.93 / 52.73**

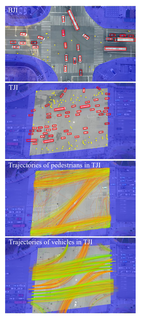
Qualitative Evaluation



Dataset Details

Table: The specifications of our dataset.

Property		Scenario I	Scenario II
Dataset name		BJI	TJI
City		Beijing	Tianjin
Latitude		40.219049N	39.120511N
Longitude		116.220789E	117.173421E
Traffic density		Low	High
Height of drone (meter)		74	121
Resolution (pixel)		3840×2160	3840×2160
Total video duration		39'58''	22'01''
Frame rate (fps)		30	30
Annotated frame number		23498	8000
Annotated frame rate (fps)		10	6
Annotated pedestrian number	Walking	1336	690
	Bike & Motor	1689	2690
	Total	3025	3380
Average pedestrian number per frame		29	46
Max pedestrian number per frame		67	105
Annotated vehicle number	Auto	2581	3523
	Bus & Truck	82	170
	Articulated bus	92	30
	Total	2755	3723
Average vehicle number per frame		19	34
Max vehicle number per frame		33	63



References

[1] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese. Social lstm: Human trajectory prediction in crowded spaces. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2016.

[2] A. Graves.Generating sequences with recurrent neural networks.arXiv preprint arXiv:1308.0850, 2013.

[3] A. Gupta, J. Johnson, L. Fei-Fei, S. Savarese, and A. Alahi.Social gan: Socially acceptable trajectories with generative adversarial networks.In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2018.

[4] I. Hasan, F. Setti, T. Tsesmelis, A. Del Bue, F. Galasso, and M. Cristani. Mx-lstm: Mixing tracklets and vislets to jointly forecast trajectories and head poses. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2018.