Pedestrian Stop and Go Forecasting with Hybrid Feature Fusion

Dongxu Guo, **Taylor Mordan**, Alexandre Alahi

Sunday 12th September, 2021

EPFL VITA

Visual Intelligence for Transportation

Improve safety for autonomous vehicles in urban areas: better predict pedestrian trajectories

- Pedestrian safety: one major challenge for deploying autonomous vehicles in urban environments
- Learning human motion patterns in traffic: crucial for avoiding collisions

Stop and Go:

- Transitions between *standing still* and *walking*
- Important aspect of human movement patterns, highly non-linear
- Help making trajectory prediction more robust: current methods react poorly to abrupt changes

Task: predicting the pedestrians' stop-and-go behaviors around vehicles

- Introduce TRANS, a new dataset for pedestrian transitions
- Propose a new model using pedestrian and scene attributes
- Evaluate multiple baselines to setup a benchmark

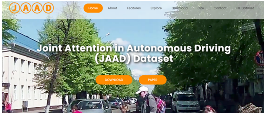
1. TRANS Dataset
2. Hybrid Feature Fusion
3. Experiments and Results
4. Conclusions

TRANS Dataset

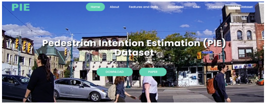
Goal: explicitly study the stop-and-go behaviors of pedestrians in traffic

Benchmark selection:

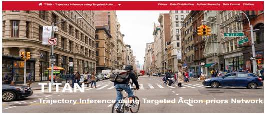
- large scale driving dataset, diversity
- ego-centric view (on-board front camera)
- localization and motion information



JAAD
crossing and attributes
[Rasouli et al., ICCV'17]



PIE
crossing intention
[Rasoulie et al., ICCV'19]



TITAN
action recognition
[Malla et al., CVPR'20]

1. Detect stop and go transitions based on the changes in pedestrian motion states (walking/standing)
2. Remove 'hesitations' (very short transitions)
3. Index examples, all unique pedestrians can be categorized into:
   - *walk*, *stand* (no transitions in video)
   - *stop*, *go* (show transitions)

TABLE I
STATISTICS OF OUR TRANS DATASET. *Go*, *Stop*, *Stand*, *Walk* INDICATE THE NUMBER OF UNIQUE PEDESTRIANS IN CORRESPONDING CATEGORIES. IN BRACKETS, WE ALSO COUNT THE NUMBER OF EVENTS, I.E., STOP AND GO TRANSITIONS.

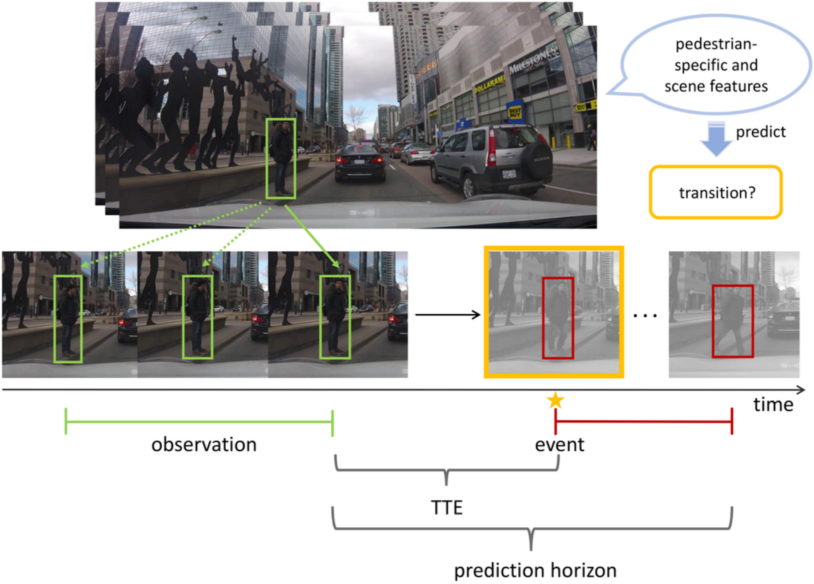
Dataset	Go [events]	Stop [events]	Stand	Walk
JAAD	144 [145]	73 [77]	65	416
PIE	397 [482]	528 [622]	697	483
TITAN	339 [381]	398 [439]	1077	6233
TRANS	880 [1008]	999 [1138]	1839	7132

Binary classification problem (*transition* vs. *no-transition*):

- Given:
    - T time steps of past observation of a walking/standing pedestrian
    - fine-grained attributes of the scene
- Objective: predict whether the pedestrian will stop or go within 2 seconds

Notes:

- We assume the motion state is known (walking/standing)
- Stop and go predictions use separate models


pedestrian-specific and scene features
predict
transition?
time
observation
event
TTE
prediction horizon

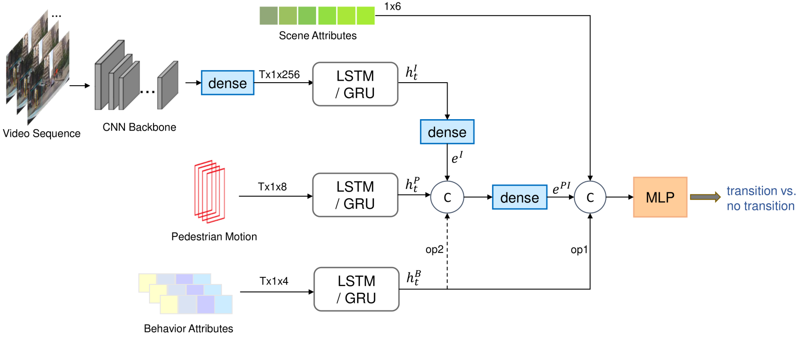
Hybrid Feature Fusion

- Visual encoding: RGB image frames
- Motion encoding: pedestrian dynamics from bounding boxes
- Behavior encoding: fine-grained attributes of 4 atomic behaviors: walking, looking, nodding, hand-gesture
- Scene encoding: 6 fine-grained attributes of the traffic scene, number of lanes, intersection, designated, signalized, traffic direction, motion direction

Behavior and Scene attributes not available in TITAN

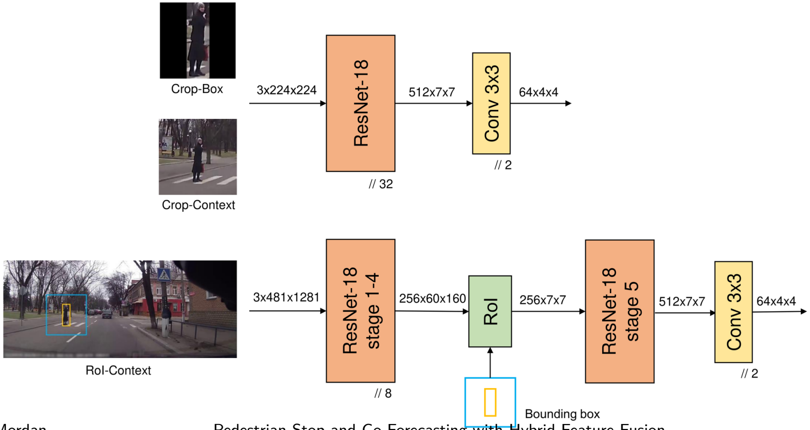
Idea:

- Progressively fuse all features and attributes
- Use LSTMs for temporal processing



Different sizes of context for visual encoding:

- No context (just bounding box)
- Local context (enlarged bounding box)
- Global context (full image)



Experiments and Results

TABLE II
EVALUATION RESULTS IN AVERAGE PRECISION (AP) FOR BASELINES AND OUR MODEL ON TRANS DATASET. BLANK LINES SEPARATE DIFFERENT TYPES OF ARCHITECTURES: STATIC, VIDEO AND HYBRID.

	Go			Stop		
Model	JAAD	PIE	TITAN	JAAD	PIE	TITAN
Crop-Box	54.3	52.0	56.2	52.5	53.1	56.4
Crop-Context	70.4	59.1	61.4	57.3	61.1	60.3
RoI-Context	73.3	61.2	60.9	58.7	62.5	59.1
CB-LSTM	60.6	56.4	58.6	57.2	59.4	58.7
CC-LSTM	73.6	61.8	63.2	61.4	63.3	61.5
RC-LSTM	76.4	64.7	62.9	62.9	64.2	61.7
PVI-LSTM	80.6	66.5	**65.1**	64.7	64.9	**63.6**
PVIBS-LSTM	**85.9**	**70.2**	-	**67.8**	**65.4**	-

Observations: it helps to use

- More visual context
- Temporal processing with sequential models (LSTMs)
- Fine-grained semantic attributes

TABLE III
ABLATION STUDY ON THE CHOICE OF FEATURES

Features	Go		Stop	
	JAAD	PIE	JAAD	PIE
PV	61.5	59.8	59.4	60.6
S	74.2	55.1	53.3	54.2
PVB	68.4	63.7	61.6	62.1
PVS	82.6	64.9	62.1	61.7
PVBS	84.7	67.3	62.5	64.7
PVI (Crop-Context)	78.4	65.1	63.4	63.5
PVI (RoI-Context)	80.6	66.5	64.7	64.9
PVIBS (Crop-Context)	85.2	69.5	67.2	**65.7**
PVIBS (RoI-Context)	**85.9**	**70.2**	**67.8**	65.4

Observations:

- Adding modalities improve results
- High-level attribute contain rich information

Conclusions

Contributions of the paper:

- Introduce the task of pedestrian stop-and-go forecasting from ego-centric view of the vehicle
- Build a novel dataset specially for this problem, based on three exiting datasets
- Propose a hybrid model utilizing multi-modal input features for transition forecasting
- Implement several baselines to create a task benchmark
- Analyze the impacts of various design choices and contributions of different features

Future work:

- Incorporate more input feature modalities: keypoints, semantic maps, spatial distances...
- Predict fine-grained TTE

Thank you for your attention

Questions?

Pedestrian Stop and Go Forecasting with Hybrid Feature Fusion

Dongxu Guo, **Taylor Mordan**, Alexandre Alahi

EPFL VITA