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
Motivations

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
Outline

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
  [Rasouli 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.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Go [events]</th>
<th>Stop [events]</th>
<th>Stand</th>
<th>Walk</th>
</tr>
</thead>
<tbody>
<tr>
<td>JAAD</td>
<td>144 [145]</td>
<td>73 [77]</td>
<td>65</td>
<td>416</td>
</tr>
<tr>
<td>PIE</td>
<td>397 [482]</td>
<td>528 [622]</td>
<td>697</td>
<td>483</td>
</tr>
<tr>
<td>TITAN</td>
<td>339 [381]</td>
<td>398 [439]</td>
<td>1077</td>
<td>6233</td>
</tr>
<tr>
<td>TRANS</td>
<td>880 [1008]</td>
<td>999 [1138]</td>
<td>1839</td>
<td>7132</td>
</tr>
</tbody>
</table>
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
Problem Formulation

Taylor Mordan

Pedestrian Stop and Go Forecasting with Hybrid Feature Fusion
Input Modalities

• 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)
### TABLE II

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

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

**Observations:** it helps to use

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

Observations:

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

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