Multi-Input Fusion for Practical Pedestrian Intention Prediction

Ankur Singh [*,¶], **Upendra Suddamalla** [*]
[*]Moovita Pte. Ltd., [¶]Indian Institute of Technology Kanpur

2021 ICCV OCTOBER 11-17 VIRTUAL

Objective:

- To determine if a pedestrian is going to cross the road in the future frames.

Figure1: Are the pedestrians going to cross? [1]

Motivation:

- Globally, more than 364,500 pedestrians lose their lives each year, which accounts for 27% of the total deaths in road accidents[1]. Naturally, pedestrian safety becomes important for other road users.
- Pedestrian intention estimation is an essential part of pedestrian safety, especially while crossing the road.

Contributions:

- Intention prediction before the crossing event takes place.
- Evaluating the significance of different combinations of input such as pose, bounding box and surrounding information.
- Extensive experiments on observation length and sampling to study the practical aspects of pedestrian intention prediction.
- Analyzing model predictions during the transition of the pedestrian's intention from not-crossing to crossing and vice-versa.

References:

1. Zhijie Fang et.al. Is the pedestrian going to cross?
2. Kensho Hara et.al. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet?
3. Amir Rasouli, Iuliia Kotseruba, and John K Tsotsos. Are they going to cross? a benchmark dataset and baseline for pedestrian crosswalk behavior.
4. Francesco Piccoli et.al. Fusion of spatio-temporal skeletons for intention prediction network.
5. Bingbin Liu etl.al. Spatiotemporal relationship reasoning for pedestrian intent prediction

Proposed Approach:

- We define the problem of pedestrian intention estimation as a binary classification task - crossing and not-crossing
- Following inputs are used in our approach(adj Figure):
  - Bounding boxes: B
  - Bounding boxes + surrounding information: B^s
  - Pose: P
  - Pose + surrounding information: P^s and
  - Bounding box coordinates: C

- We get the best accuracy of 84.9% (shown in Figure 3) using a pretrained 3D Resnet [2], with a combination of 3 inputs: i) pose with surrounding information, ii) pose and iii) bounding box coordinates.

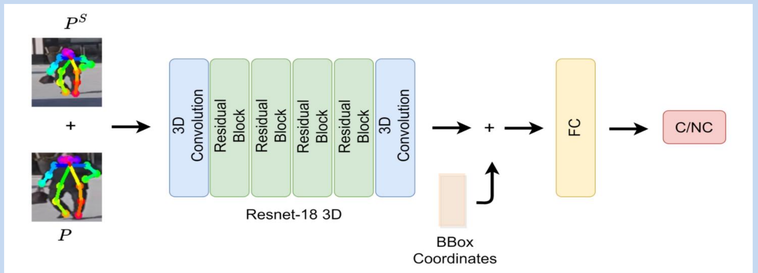

Figure 3: Architecture of the proposed model

Conclusion:

- The proposed method shows that using an implicit pose from the appearance and surrounding information is simple, straightforward, requires less computation and gives a high accuracy of over 84%.
- Computing the human pose explicitly and superimposing on the image boosts the intention detection accuracy to over 84%.
- We also observe that the best results during the transition phase are obtained on using a single frame for observation.

Results:

- We show our results on the JAAD [3] dataset for pedestrian intention prediction.

Input	No. Inputs	Accuracy
B	1	79.8
B^s	1	80.70
P	1	81.14
P^s	1	81.85
B^s, C	2	82.54
P^s, C	2	83.1
P^s, P	2	83.77
P^s, P, C	3	84.89

Table 1: Results on JAAD: Comparison of different input combinations.

Method	Obs. Length	Accuracy
ATGC [3]	1 (0.03s)	63
Fussi-Net [4]	16 (0.533s)	75.6
STIP [5]	30 (1s)	76.98
Ours	16 (0.533s)	84.89

Table 2: Results on JAAD: Comparison with prior works

Input	Obs. Length	TC-Accuracy	T-NC Accuracy
B^s, C	1 (0.03s)	73.71	45.16
B^s, C	8 (0.266s)	71.46	38.54
B^s, C	16 (0.533s)	61.02	36.93

Table 3: Transition State Analysis

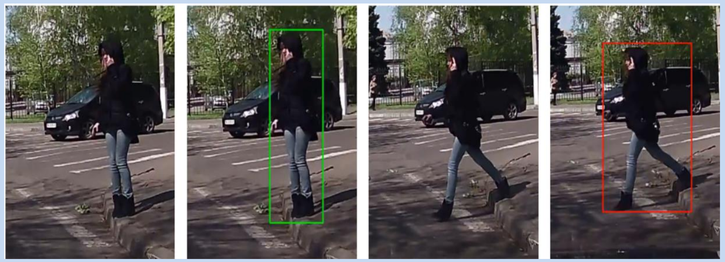

Figure 4: Qualitative Results: Green Bbox – Not Crossing, Red Bbox- Crossing

ankuriit@iitk.ac.in, upendra@moovita.com