Pedestrian Intention Prediction with Multi-Input Concatenation

Ankur Singh[1,2], Upendra Suddamalla[1], Anthony Wong[1], Dilip Kumar Limbu[1]
[1]Moovita Pte. Ltd. [2]Indian Institute of Technology Kanpur
ankuriit@iitk.ac.in, {upendra, anthonywong, diliplimbu}@moovita.com

Abstract

Responding safely to the pedestrians on the road is one of the critical challenges for autonomous vehicles. For smooth navigation of autonomous vehicles in urban environments, it is crucial to predict the pedestrians' road crossing intention accurately and respond safely. Though motion analysis is a key feature for estimating future trajectories, it may be inconsistent for the small variable actions of humans. For reliable early prediction of the future trajectory of a pedestrian, visual pose and surrounding information are helpful. In this work, we propose a novel approach to determine the intention of a pedestrian by using pose, surrounding context, and bounding box information over a small duration of half a second (last 16 frames). We study the significance of different combinations of these features. We adopt 3D convolution networks, that have shown remarkable performance in activity recognition tasks. In our experiments using the popular pedestrian intention dataset, JAAD, the proposed method achieved over 84% accuracy in estimating the intention of a pedestrian to cross.

1. Introduction

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. An essential aspect in the context of pedestrian safety is pedestrian intention estimation, especially while crossing the road. Pedestrian intention estimation refers to determining whether the pedestrian is going to cross the road in the next few seconds. Timely and accurate prediction of pedestrian's intention is vital in safer maneuvering of autonomous vehicles, thus avoiding potential accidents.

In the past few years, pedestrian intention estimation has attracted significant attention in the computer vision community. This has been made possible largely because of the availability of richly annotated pedestrian intention datasets such as the Daimler dataset[17], Joint Attention for Autonomous Driving (JAAD)[13, 14], Pedestrian Intention Estimation (PIE)[12].

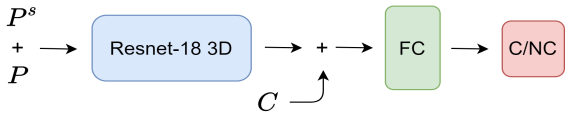


Figure 1. The architecture of our best performing model: Here P^s(pose having surrounding information) is concatenated with P(pose having local context only). The concatenated input is fed to a pretrained Resnet-18 3D. The features extracted from the last convolutional layer of Resnet 3D are then concatenated with the Bounding box coordinates C. This is finally fed to the fully-connected layer to make the crossing prediction.

Previous works on pedestrian intention prediction have relied on bounding boxes[14, 16, 18], pose[2, 10], semantic segmentation maps[8, 11] as their input. However, experimenting with different input combinations has largely been unexplored on the JAAD dataset. We set the key objectives of this work as 1. Intention prediction before the crossing event takes place. 2. Evaluating the significance of different combinations of pose, bounding box and surrounding information.

In our work, we propose an approach that uses pose, surrounding context and bounding box information for pedestrian intention prediction. We experiment with different inputs to determine the best possible input combination for this task. Through experiments, we show that our best performing model, shown in Figure 1, outperforms other methods on pedestrian crossing prediction task on the JAAD dataset.

2. Proposed Approach

We define the problem of pedestrian intention estimation as a binary classification task, the two classes being crossing and not crossing. The objective is to determine whether the pedestrian will start crossing the road at time t when provided with the observations for some frames n before time t. Formally, given the sequence of observations

[1]WHO Global status report on road safety 2018

(a) (b) (c) (d)

Figure 2. Inputs: (a) Cropped Bounding Box. (b) Bounding Box with surrounding information (c) Pose (d) Pose with surrounding information

.

$X = \{x_1, x_2, ..., x_n\}$ before time t, we want to learn parameters θ to predict the probability $p(y|X, \theta)$ of the pedestrian crossing the road at time t.

We leverage the spatio-temporal information of the frames to make the predictions. We experiment with different sources of information in our approach. These include bounding boxes $B = \{b_1, b_2, ..., b_n\}$, bounding boxes with surrounding information $B^s = \{b_1^s, b_2^s, ..., b_n^s\}$, pose $P = \{p_1, p_2, ..., p_n\}$, pose with surrounding information $P^s = \{p_1^s, p_2^s, ..., p_n^s\}$ and the bounding box coordinates $C = \{c_1, c_2, ..., c_n\}$.

2.1. Input Information

We now give a detailed explanation of the sources of information that we experiment with in our approach:

Bounding Boxes: Given the ground truth bounding coordinates, we crop the bounding box around the pedestrian in a frame as shown in Figure 2(a). Bounding boxes are cheaper to compute and can help in determining the pedestrian's gait(walking/standing).

Bounding Boxes with Surrounding Information: These are obtained by scaling the 2D bounding boxes to 1.5 times their original size. This is shown in Figure 2(b). Apart from providing knowledge about the pedestrian's gait, they also give an idea about the pedestrian's surroundings such as curb, road, etc.

Pose: Given the cropped bounding box we use OpenPose[1] to generate pose. The generated pose is then superimposed on the pedestrian, Figure 2(c). Pose has been widely used in the past for action recognition and action anticipation tasks. Pose information simplifies learning for action recognition by providing head and body orientations.

Pose with Surrounding Information: The cropped bounding boxes are scaled to 1.5 times their original size before the pose is superimposed on the pedestrian as shown in Figure 2(d).

Bounding Box coordinates: Like [15], we believe the bounding box coordinates give a sense of the relative displacement of the pedestrian and can also be seen as the pedestrian's velocity.

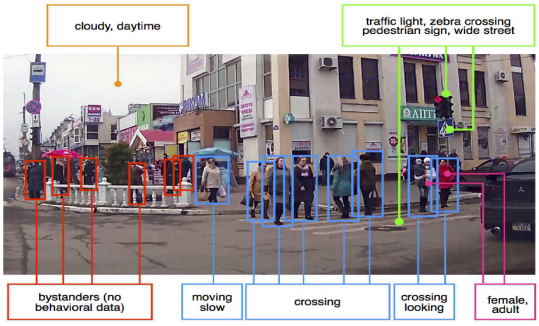


Figure 3. Examples of the annotations provided in the dataset. Image from [14]

2.2. Classification

Owing to the success of 3D-CNNs[4] in video classification tasks in the recent past, we use a 3D Resnet-18[3] pre-trained on Kinetics-400[5] as the classification network in our experiments.

We concatenate the inputs before passing them to our classification network. Except for the bounding box coordinates, all the other inputs are passed into the first layer of the network. In experiments where bounding box coordinates are used, they are concatenated with the feature output of the last convolution layer and then passed to the fully-connected layer.

3. Experiments

In this section, we describe the dataset we use for our experiments and report our results.

3.1. Dataset

We use the Joint Attention in Autonomous Driving (JAAD) Dataset[13, 14] for all our experiments. JAAD dataset is a dataset for studying pedestrian and driver behavior at the point of crossing the road. It has a collection of 346 videos each 5-10 seconds long. The videos are recorded at 30 FPS with a resolution of 1920 x 1080 pixels. Each video comes with rich ground truth annotations which include bounding box annotations, behavioral tags and scene annotations, shown in Figure 3.

3.2. Training Details

We use a pretrained Resnet3D-18 as the classification network. A batch size of 16 is used during training and optimization is done using Adam[6] with a learning rate of 0.0001. We use the NVIDIA GeForce GTX1080 GPU to train our networks. All the experiments are performed using the Pytorch[9] deep learning framework.

3.3. Evaluation Technique

We train on the first 250 videos and evaluate on the remaining 96 videos from JAAD. We observe sequences of

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 dataset: Comparison of different input combinations. Different inputs used are: B Bounding Box, B^s Bounding box having surrounding context, P Pose, P^s Pose with surrounding context, C Bounding box coordinates.

Method	Obs. Length(s)	Acc.	Pred. Horizon
ATGC[14]	0.03	63	next frame
Fussi-Net[10]	0.533	75.6	next frame
STIP [7]	-	79.28	1-30 frames
Ours	0.533	84.9	next frame

Table 2. Results on JAAD dataset: Comparison with prior works

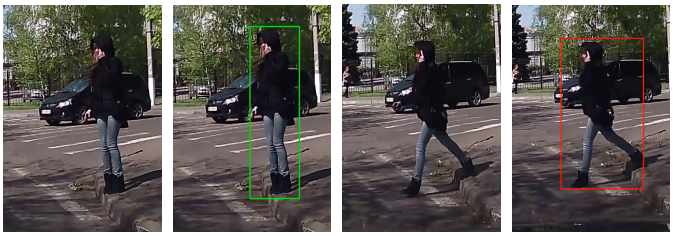
0.53 seconds(16 frames) before making a prediction. The prediction horizon in our experiments is the next frame. The train set consists of 93545 such observations of which 55006 belong to crossing and 38539 belong to not crossing. Similarly for the test set we have 39155 observations of which 20041 are crossing and 19114 are not crossing.

3.4. Results

We experiment with various input information in our approach. The results of experiments involving different inputs are summarised in Table 1. We observe that increasing the number of modalities of information improves the results. Using multiple input sources allows the network to learn discriminative features better than with one single source.

Using bounding boxes as the only input to the classification network proves to be a good baseline for the rest of our experiments. Next, we see experiments that improve upon this baseline. Looking at the results we observe that using bounding boxes with surrounding information improves the accuracy by 0.9%. In the single input case, pose with surrounding information gives the best results with an accuracy of 81.85%.

We also observe that incorporating bounding box coordinates along with other inputs seems to boost results drastically. For instance, we see an improvement of 1.84% in the case where bounding box coordinates are used along with bounding boxes having surrounding information. We get the best accuracy of 84.9% , shown in Figure 1, with a combination of 3 inputs: i) pose with surrounding information, ii) pose and iii) bounding box coordinates.



(a) (b) (c) (d)

Figure 4. Results of intention prediction (a) Pedestrian is standing on the curb (b) The green bounding box around the pedestrian generated by our network shows that the pedestrian is not crossing the road (c) The pedestrian is intending to cross the road (d) The red bounding box around the pedestrian signifies the crossing intention of the pedestrian.

.

Table 2 shows the comparison of our approach against the state of the art methods on the JAAD dataset. For a fair comparison, we only compare against methods where the observation endpoint is before the event, and the observation length is less than 1 second. ATGC[14] uses a single frame of pedestrian information for intention prediction and achieves an accuracy of 63%. Fussi-Net[10] uses 16 frames of pose sequence as input and then feeds it to a Spatio-temporal Densenet[16] for classification. STIP[7] uses a graph-based network to interact with different objects in the surrounding and achieves a prediction accuracy of 79.28 averaged over the next 1-30 frames. From the results, we can see that our approach is able to outperform other methods on the dataset. This is mainly because of the multiple input modalities used in our approach. The output generated by our network is shown in Figure 4.

4. Conclusion

Accurate and early prediction of the intention of a pedestrian helps an autonomous vehicle to take safe navigation steps. This is crucial for the acceptance of autonomous vehicles and their coexistence with humans. The proposed novel 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 80%. Computing the human pose explicitly and superimposing on the image boosts the intention detection accuracy to over 84%. Our experiments show that 3D Convolution networks can learn the pose and surrounding information well and can determine the intention with reliable accuracy.

In future work, pedestrian intention estimation can benefit from using additional information such as ego vehicle speed, map information including pedestrian crossings, traffic lights, etc.

References

[1] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017. 4322

[2] Zhijie Fang and Antonio M. López. Is the pedestrian going to cross? answering by 2d pose estimation. In *2018 IEEE Intelligent Vehicles Symposium, IV 2018, Changshu, Suzhou, China, June 26-30, 2018*, pages 1271–1276. IEEE, 2018. 4321

[3] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6546–6555, 2018. 4322

[4] S. Ji, W. Xu, M. Yang, and K. Yu. 3d convolutional neural networks for human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):221–231, 2013. 4322

[5] Will Kay, João Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset. *CoRR*, abs/1705.06950, 2017. 4322

[6] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 4322

[7] Bingbin Liu, Ehsan Adeli, Zhangjie Cao, Kuan-Hui Lee, Abhijeet Shenoi, Adrien Gaidon, and Juan Carlos Niebles. Spatiotemporal relationship reasoning for pedestrian intent prediction, 2020. 4323

[8] Satyajit Neogi, Michael Hoy, Kang Dang, Hang Yu, and Justin Dauwels. Context model for pedestrian intention prediction using factored latent-dynamic conditional random fields. *CoRR*, abs/1907.11881, 2019. 4321

[9] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. 4322

[10] Francesco Piccoli, Rajarathnam Balakrishnan, Maria Jesus Perez, Moraldeepsingh Sachdeo, Carlos Nunez, Matthew Tang, Kajsa Andreasson, Kalle Bjurek, Ria Dass Raj, Ebba Davidsson, Colin Eriksson, Victor Hagman, Jonas Sjöberg, Ying Li, L. Srikar Muppirisetty, and Sohini Roychowdhury. Fussi-net: Fusion of spatio-temporal skeletons for intention prediction network. *CoRR*, abs/2005.07796, 2020. 4321, 4323

[11] Adithya Ranga, Filippo Giruzzi, Jagdish Bhanushali, Émilie Wirbel, Patrick Pérez, Tuan-Hung Vu, and Xavier Perrotton. Vrunet: Multi-task learning model for intent prediction of vulnerable road users. *CoRR*, abs/2007.05397, 2020. 4321

[12] Amir Rasouli, Iuliia Kotseruba, Toni Kunic, and John K. Tsotsos. Pie: A large-scale dataset and models for pedestrian intention estimation and trajectory prediction. In *International Conference on Computer Vision (ICCV)*, 2019. 4321

[13] Amir Rasouli, Iuliia Kotseruba, and John K Tsotsos. Agreeing to cross: How drivers and pedestrians communicate. In *IEEE Intelligent Vehicles Symposium (IV)*, pages 264–269, 2017. 4321, 4322

[14] Amir Rasouli, Iuliia Kotseruba, and John K Tsotsos. Are they going to cross? a benchmark dataset and baseline for pedestrian crosswalk behavior. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 206–213, 2017. 4321, 4322, 4323

[15] Amir Rasouli, Iuliia Kotseruba, and John K Tsotsos. Pedestrian action anticipation using contextual feature fusion in stacked rnns. In *BMVC*, 2019. 4322

[16] Khaled Saleh, Mohammed Hossny, and Saeid Nahavandi. Real-time intent prediction of pedestrians for autonomous ground vehicles via spatio-temporal densenet. In *International Conference on Robotics and Automation, ICRA 2019, Montreal, QC, Canada, May 20-24, 2019*, pages 9704–9710. IEEE, 2019. 4321, 4323

[17] Nicolas Schneider and Dariu M. Gavrila. Pedestrian path prediction with recursive bayesian filters: A comparative study. In Joachim Weickert, Matthias Hein, and Bernt Schiele, editors, *Pattern Recognition - 35th German Conference, GCPR 2013, Saarbrücken, Germany, September 3-6, 2013. Proceedings*, volume 8142 of *Lecture Notes in Computer Science*, pages 174–183. Springer, 2013. 4321

[18] Dimitrios Varytimidis, Fernando Alonso-Fernandez, Boris Durán, and Cristofer Englund. Action and intention recognition of pedestrians in urban traffic. In Gabriella Sanniti di Baja, Luigi Gallo, Kokou Yétongnon, Albert Dipanda, Modesto Castrillón Santana, and Richard Chbeir, editors, *14th International Conference on Signal-Image Technology & Internet-Based Systems, SITIS 2018, Las Palmas de Gran Canaria, Spain, November 26-29, 2018*, pages 676–682. IEEE, 2018. 4321