Multi-Input Fusion for Practical Pedestrian Intention Prediction

Ankur Singh[1,2], Upendra Suddamalla[1]
[1]Moovita Pte. Ltd. [2]Indian Institute of Technology Kanpur
ankuriit@iitk.ac.in, upendr@moovita.com

Abstract

Pedestrians are the most vulnerable road users and are at a high risk of fatal accidents. Accurate pedestrian detection and effectively analyzing their intentions to cross the road are critical for autonomous vehicles and ADAS solutions to safely navigate public roads. Faster and precise estimation of pedestrian intention helps in adopting safe driving behavior. Visual pose and motion are two important cues that have been previously employed to determine pedestrian intention. However, motion patterns can give erroneous results for short-term video sequences and are thus prone to mistakes. In this work, we propose an intention prediction network that utilizes pedestrian bounding boxes, pose, bounding box coordinates, and takes advantage of global context along with the local setting. This network implicitly learns pedestrians' motion cues and location information to differentiate between a crossing and a non-crossing pedestrian. We experiment with different combinations of input features and propose multiple efficient models in terms of accuracy and inference speeds. Our best-performing model shows around 85% accuracy on the JAAD dataset.

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[1], Joint Attention for Autonomous Driving (JAAD)[2, 3], Pedestrian Intention Estimation (PIE)[4].

Predominantly trajectory-based approaches have been used to predict pedestrian intentions. Methods like [5], [6] rely on the past trajectory of the pedestrian to predict their future locations. Though motion analysis is a key feature for estimating the future course of the pedestrian, it may be inconsistent for small changes in the pedestrians' actions and are often subject to errors. To overcome this problem, recent pedestrian intention estimation techniques have adopted bounding boxes[3, 7, 8], pose[9, 10], semantic segmentation maps[11, 12] as their input. However, these techniques focus on specific information and are usually prone to failures in certain scenarios. To obtain a more generalized solution that is also robust, there is a need to utilize information, global as well as local, from various sources.

In this work, we propose an approach that uses pedestrian bounding boxes, pose information, both with global and local context along with bounding box coordinates for pedestrian intention prediction. We perform a thorough analysis of our proposed approach. We experiment with different inputs to determine the best possible input combination for the task of pedestrian intention prediction and compare our method with state-of-the-art techniques. 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. We extend our experiments to evaluate the impact of observation length on model performance and inference speed. We further study the behavior of our model during the beginning of a crossing event to find out any latency in model prediction.

[1] WHO Global status report on road safety 2018

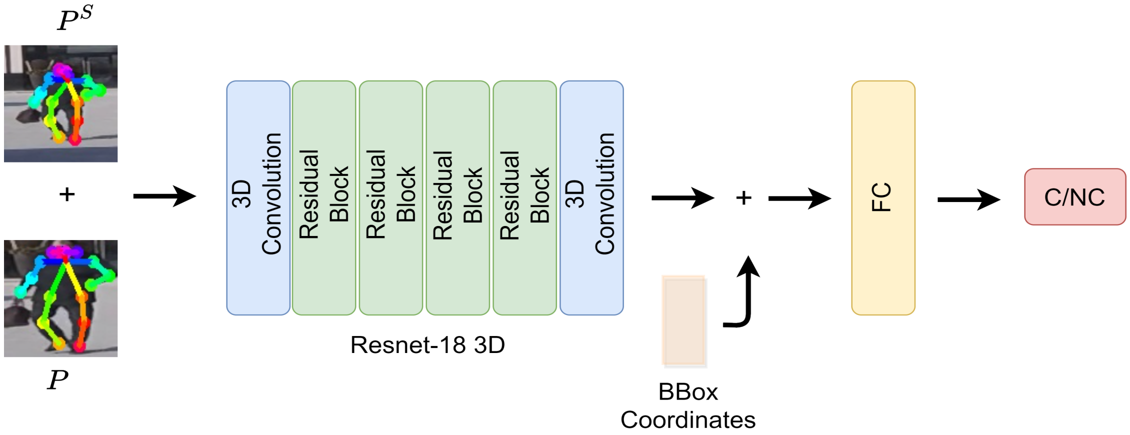

Figure 1. The architecture of our best performing model: Here P^s(pose having surrounding information) is concatenated with P(pose having local context only). The concatenated input is fed to a pretrained Resnet-18 3D. The features extracted from the last convolutional layer of Resnet 3D are then concatenated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

We set the key objectives of this work as follows:

1. Predicting the crossing intention as soon as the crossing event begins.
2. Evaluating the significance of different combinations of pose, bounding box and surrounding information.
3. Discussing the practical aspects of the proposed method - in terms of accuracy vs computation time.

2. Related Work

Action Prediction: Action recognition has been widely studied; however, in many scenarios such as autonomous driving systems, criminal activities, etc., we do not have the luxury to wait for the action to end. Hence action prediction becomes a more viable option in such cases. Ryoo et al. [13] proposed one of the first works in action prediction that used a bag of words approach. For accident prediction, Zeng et al. [14] take an agent-centric approach to anticipate accidents through soft-attention RNNs. [15] extends accident risk assessment to autonomous driving by employing a ConvLSTM model on camera images and driving commands. Apart from accident risk estimation, several works [16], [17] are focused on understanding the intent of different agents on the road. Next, we will look at previous approaches related to the particular topic of pedestrian crossing intention estimation.

Pedestrian crossing intention estimation: Previous works in pedestrian crossing intention estimation have relied on different input features and network architectures. [3] was one of the first works to report its results on JAAD. It utilized a single frame of pedestrian and traffic scene information to predict the crossing intention. However, a better technique to predict crossing intentions is through sequence analysis by incorporating multiple frames. [9] uses 14 frames of pedestrian pose data, which is later fed to SVM/Random Forests for classification. [7] employs a spatio-temporal Densenet for classification based on sequences of pedestrian bounding boxes. Piccoli et al. [10] follow a similar technique; however, they additionally use pose features as input. [18] incorporates a Transformer for classifying pedestrian intentions based on bounding box features.

Recently feature fusion or feature concatenation has been explored for pedestrian crossing estimation. [19] is an early work that uses multiple modalities, including bounding box, pose, ego-vehicle speed, and then performs the classification step through stacked RNNs. Yang et al. [20] use similar features, however, with a Spatio-temporal attention module in their network architecture. In [21], the authors utilize a RubiksNet [22] along with a transformer to extract features followed by a classification network. [11] incorporates depth maps, semantic segmentation maps, optical flow output, and bounding boxes for feature extraction.

Graph-based approaches are also being adopted to solve the intention estimation problem. [23] uses 2D human pose and Graph Convolutional Networks as a solution. [24] employs a Graph-based network to model interactions between different agents in the scene such as pedestrians, ego-vehicle, other vehicles, etc.

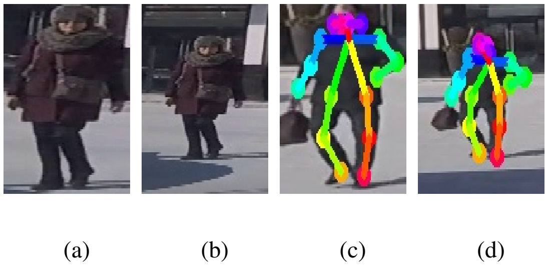
(a) (b) (c) (d)

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

.

3. Proposed Approach

We define the problem of pedestrian intention estimation as a binary classification task, the two classes being crossing (C) and not crossing (NC). 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 $X = \{x_1, x_2, ..., x_n\}$ 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 $B^s = \{b^s_1, b^s_2, ..., b^s_n\}$, pose $P = \{p_1, p_2, ..., p_n\}$, pose with surrounding information $P^s = \{p^s_1, p^s_2, ..., p^s_n\}$ and the bounding box coordinates $C = \{c_1, c_2, ..., c_n\}$.

3.1. 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 (Figure 2(a)) and resize it to 100×100. 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[25] 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.

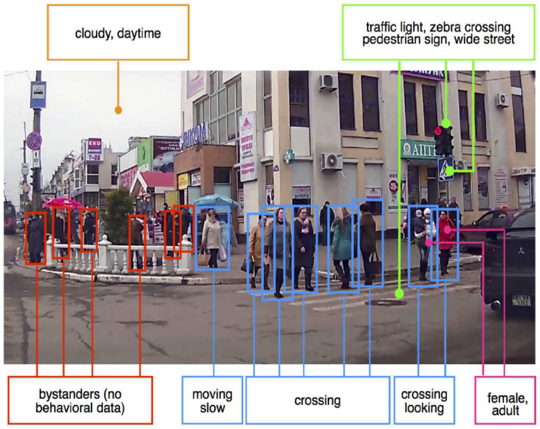
cloudy, daytime
traffic light, zebra crossing pedestrian sign, wide street
bystanders (no behavioral data)
moving slow
crossing
crossing looking
female, adult

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

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 [19], 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.

3.2. Classification

Owing to the success of 3D-CNNs[26] in video classification tasks in the recent past, we use a 3D Resnet-18[27] pre-trained on Kinetics-400[28] 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.

4. Experiments and Results

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

4.1. Dataset

We use the Joint Attention in Autonomous Driving (JAAD) Dataset[2, 3] 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.

4.2. 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[29] 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[30] deep learning framework.

4.3. Evaluation Technique

We train on the first 250 videos and evaluate on the remaining 96 videos from JAAD. Since most of the previous works on pedestrian intention estimation utilize 16 frames of temporal information, therefore for a fair comparison, even we observe sequences of 0.53 seconds(16 frames) before making a prediction. In later sections, we also look at the effect of different observation lengths on intention estimation performance. 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. For the test set, we have 39155 observations, of which 20041 are crossing and 19114 are not crossing.

4.4. Comparison of different input combinations

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
$\boldsymbol{P^s, P, C}$	**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.

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%.

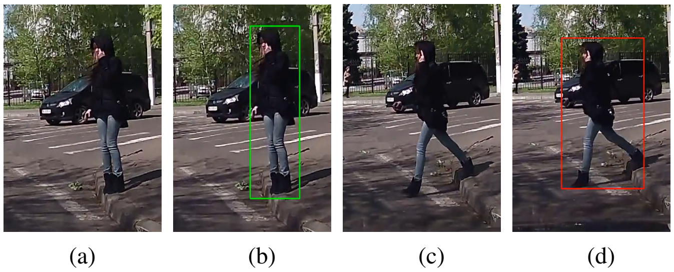

Figure 4. Results of intention prediction (a) Pedestrian is standing on the curb (b) The green bounding box around the pedestrian generated by our network shows that the pedestrian is not crossing the road (c) The pedestrian is intending to cross the road (d) The red bounding box around the pedestrian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

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.

4.5. Comparison with prior works

Method	Obs. Length	Pred. Horizon	Acc.
ATGC[3]	1 (0.03s)	next frame	63
Fussi-Net[10]	16 (0.533s)	next frame	75.6
STIP [24]	30 (1s)	next frame	76.98
Ours	16 (0.533s)	next frame	**84.9**

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

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[3] 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[7] for classification. STIP[24] uses a graph-based network to interact with different objects in the surrounding and achieves a prediction accuracy of 76.98%. 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.

For future experiments, we only focus on bounding box features for inputs as bounding box coordinates have been provided in JAAD annotations. Since pose key points are not present in the available annotations, estimating pose separately from a pre-trained model can be prone to errors and time-consuming for real-time inferences.

4.6. Effect of Observation lengths on performance

In this section, we indicate the effect of the number of past frames on crossing intention performance. We show a comparison between history lengths of 1 frame, 8 frames, and 16 frames. In Table 3, we consider B^s, C features as input, and for Table 4, we use B as input features. Both the experiments show similar trends.

The accuracy of 1 frame experiment is the lowest, as expected. This is because a single frame is unable to provide any temporal information, and the prediction is solely based on the spatial information available from the frame. A history length of 8 frames performs better than 16 frames in both our experiments. One possible explanation for this could be the enhanced performance of 8 frames during the transition state of the pedestrian. We define transition as the change of pedestrian's crossing intention from not crossing (NC) to crossing (C) or vice versa. We suspect that utilizing a larger number of past frames would delay intention prediction during transition. To enforce this belief, we perform quantitative analysis and qualitative analysis comparing transition and non-transition accuracies for different observation lengths. Section 4.7 presents the results of the aforementioned experiment.

Input	Obs. Length	Acc.	F1	Precision	Recall
B^s, C	1 (0.03s)	78.60	79.43	78.09	80.77
B^s, C	**8** (0.26s)	**84.63**	**85.08**	**84.59**	**85.56**
B^s, C	16 (0.53s)	82.54	82.41	80.12	84.71

Table 3. Effect of different observation lengths on performance using B^s, C features as input

4.7. Transition state analysis

In this section, we discuss the results of different experiments during transition state of the pedestrian. A transition state is defined as the change of pedestrians' intention. To calculate transition accuracy, we compare the predictions of our models with the ground truths over 16 frames after the state change. Let f^{th} frame be the transition frame; then we calculate the transition accuracy over the frames $f+1$, $f+2$, $f+3 \ldots$ $f+16$. Table 5 presents our results for the experiment mentioned above. The term T-C indicates the behavior change from not crossing to crossing. In comparison, T-NC suggests a change from crossing to not crossing.

Input	Obs. Length	Acc.	F1	Precision	Recall
B	1 (0.03s)	76.40	77.09	76.16	78.02
B	**8** (0.26s)	**81.05**	**81.66**	**83.06**	**80.27**
B	16 (0.53s)	79.8	80.15	80.53	79.77

Table 4. Effect of different observation lengths on performance using B features as input

Input	Obs. Length	T-C Accuracy	T-NC Accuracy
B^s, C	**1** (0.03s)	**73.71**	**45.16**
B^s, C	8 (0.26s)	71.46	38.54
B^s, C	16 (0.53s)	61.02	36.93

Table 5. Effect of different observation lengths on Transition accuracies using B^s, C features as input

During a transition, the temporal information available to the network consists of frames f, $f-1$, $f-2$, $\ldots$, $f-(N-1)$. Here N is the number of frames used for observation. Of all the frames utilized for observation, only the f^{th} frame belongs to the same class as frame $f+1$, $f+2$, $\ldots$. Whereas frame $f-1$, $f-2$, $f-(N-1)$ belong to the opposite class. As N gets larger, the temporal data has an adverse effect on the prediction accuracy during transition. We suspect a lower N to show better results during transition. This is clearly shown in Table 5, where an observation length of 1 gives the best T-C accuracy as well as T-NC accuracy. As expected, a history length of 16 gives the lowest numbers. This is because during transition, a 16 frames sequence consists of a majority of frames of the other class, which plays an opposing role and results in wrong predictions.

Qualitative analysis of model performance during transition has been presented in Figure 5. The first row represents the ground truth, while the second, third, and fourth row show the outputs of 1-frame, 8-frame and 16-frame model respectively. The transition frame in the figure is indicated by f. The qualitative results present a similar picture as the quantitative results. The 1-frame model outputs follow the ground truths during transition, and there is no delay in prediction. However, as we move to the outputs of the 8-frame model, we observe a delay of 3 frames—the results of the 16-frame model show the highest latency where we see a lag of 7 frames.

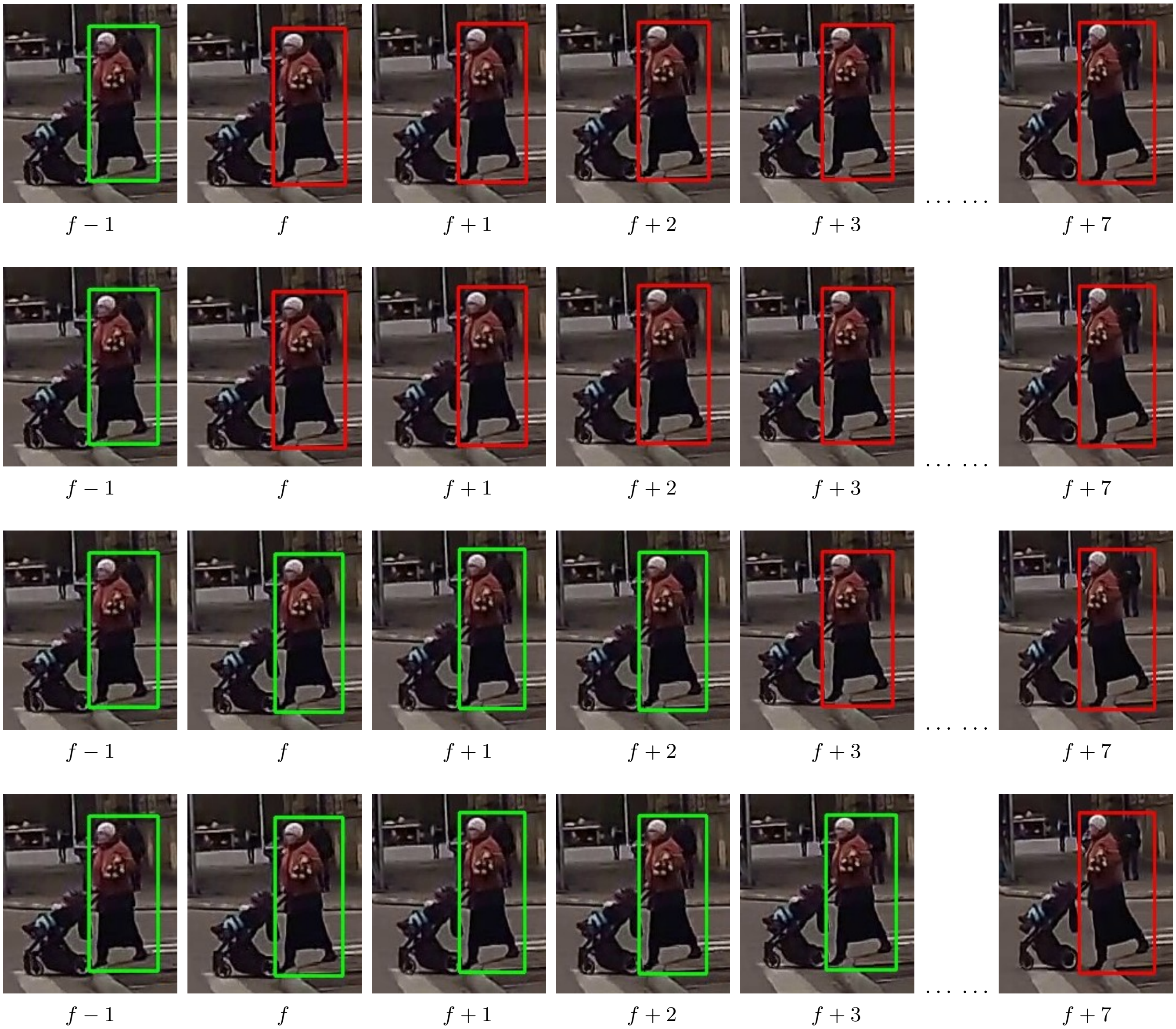

Figure 5. Qualitative Analysis of model performance during transition: The first row represents the ground truths. The second, third, and fourth row provide the outputs of 1-frame, 8-frame and 16-frame model respectively. Here f represents the frame when the intention changes from not-crossing to crossing. The 1-frame model does not show any delay, whereas we observe a delay of 3 frames and 7 frames in the predictions of the 8-frame and 16-frame models respectively.

4.8. Effect of Sampling

In this section, we briefly discuss the effect of sampling on the performance of our model. We uniformly sample frames from a history length of 8 frames and utilize them to train our model. Hence for a 4 frame history, every alternate frame is sampled, while for a 2 frame history, every fourth frame is sampled. The accuracy results, along with the inference speed of the model, is provided in Table 6. Since there can be multiple pedestrians in a single frame, therefore to calculate the inference rate, we measure the pedestrians processed by the network in one second. We call this pedestrians per second (PPS).

Input	Obs. Length	PPS	Accuracy
B^s, C	**8** (0.26s)	97	**84.63**
B^s, C	4 (0.13s)	125	84.14
B^s, C	**2** (0.06s)	**141**	82.01

Table 6. Sampling from 8 frames history using B^s, C features

The results show that the accuracy of the 8 frames model is the highest. There is a slight decrease in performance while utilizing a 4 frames model; however, it has a much higher PPS of 125 as compared to 97 of the former. The accuracy of the 2 frames model is around 2% lower and has a slightly higher PPS than the 4-frame model. The high PPS and a satisfactory accuracy of the 4 frames model indicate that it might be an optimal solution for the task of pedestrian crossing intention which requires accurate as well as quick predictions.

5. 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 on the public road. The proposed novel method shows that using an implicit pose from the appearance and surrounding information is simple, straightforward, requires less computation, and gives high accuracy of over 84%. Computing the human pose explicitly and superimposing on the image boosts the intention detection accuracy further by a small amount. Our study on the effect of observation length shows that using the data from a quarter of a second (250msec) is faster and achieves better accuracy compared to a commonly used 16 frames (530msec) duration or a single frame (33msec). 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] 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.

[2] 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.

[3] 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.

[4] 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.

[5] Julian Francisco Pieter Kooij, Nicolas Schneider, Fabian Flohr, and Dariu M Gavrila. Context-based pedestrian path prediction. In *European Conference on Computer Vision*, pages 618–633. Springer, 2014.

[6] Alexandre Alahi, Kratarth Goel, Vignesh Ramanathan, Alexandre Robicquet, Li Fei-Fei, and Silvio Savarese. Social lstm: Human trajectory prediction in crowded spaces. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 961–971, 2016.

[7] 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.

[8] 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.

[9] 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.

[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.

[11] 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.

[12] 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.

[13] Michael S Ryoo. Human activity prediction: Early recognition of ongoing activities from streaming videos. In *2011 International Conference on Computer Vision*, pages 1036–1043. IEEE, 2011.

[14] Kuo-Hao Zeng, Shih-Han Chou, Fu-Hsiang Chan, Juan Carlos Niebles, and Min Sun. Agent-centric risk assessment: Accident anticipation and risky region localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2222–2230, 2017.

[15] Mark Strickland, Georgios Fainekos, and Heni Ben Amor. Deep predictive models for collision risk assessment in autonomous driving. In *2018 IEEE International Confer-*

ence on Robotics and Automation (ICRA), pages 4685–4692. IEEE, 2018.

[16] Sergio Casas, Wenjie Luo, and Raquel Urtasun. Intentnet: Learning to predict intention from raw sensor data. In *Conference on Robot Learning*, pages 947–956. PMLR, 2018.

[17] Pratik Gujjar and Richard Vaughan. Classifying pedestrian actions in advance using predicted video of urban driving scenes. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 2097–2103. IEEE, 2019.

[18] Lina Achaji, Julien Moreau, Thibault Fouqueray, Francois Aioun, and François Charpillet. Is attention to bounding boxes all you need for pedestrian action prediction? *arXiv preprint arXiv:2107.08031*, 2021.

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

[20] Dongfang Yang, Haolin Zhang, Ekim Yurtsever, Keith Redmill, and Ümit Özgüner. Predicting pedestrian crossing intention with feature fusion and spatio-temporal attention. *arXiv preprint arXiv:2104.05485*, 2021.

[21] J Lorenzo, I Parra, and MA Sotelo. Intformer: Predicting pedestrian intention with the aid of the transformer architecture. *arXiv preprint arXiv:2105.08647*, 2021.

[22] Linxi Fan, Shyamal Buch, Guanzhi Wang, Ryan Cao, Yuke Zhu, Juan Carlos Niebles, and Li Fei-Fei. Rubiksnet: Learnable 3d-shift for efficient video action recognition. In *European Conference on Computer Vision*, pages 505–521. Springer, 2020.

[23] Pablo Rodrigo Gantier Cadena, Ming Yang, Yeqiang Qian, and Chunxiang Wang. Pedestrian graph: Pedestrian crossing prediction based on 2d pose estimation and graph convolutional networks. In *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*, pages 2000–2005. IEEE, 2019.

[24] 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.

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

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] 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.