Survey Paper on Existing Technologies for Pedestrian Detection and Tracking

Ashrey Gupta

Department of Computer Science and Engineering, MSU

CSE 891 – Advance Topics in Autonomous Vehicles

Dr. Josh Siegel

***Abstract***

This survey paper will present an insight into existing technologies and research used to detect pedestrians and track them on crosswalks and intersections. This type of technology is crucial to avoiding pedestrian accidents as 75% of pedestrian accident takes place close to urban intersections. With the help of extensive research and development of a highly accurate detection and tracking system, vehicles will be able to communicate with road infrastructure and avoid these accidents. Present algorithms used for this technology are neural network modelling, deep-sort multi tracking algorithm, logistic regression for pedestrian behavior analysis and lidar detection for multiple pedestrian detection. Finally, a novel approach for infrastructure to vehicle(V2I) communication for this system will be presented.

1. **Introduction**

One-fifth of total pedestrian accidents occurred at intersections in 2019 which is a high rate considering intersections are supervised traffic systems and vehicles are driven governed by set of rules set by laws local to the country. With the advent of vehicles with ADAS technology and self-driving capabilities, the use of V2I technology these vehicles can obtain current pedestrian data at crosswalks and act according to it. With live pedestrian detection, vehicles can slow down and even come to a stop if a danger is detected or a pedestrian acts irrationally(jaywalking). With the help of cameras at intersections, videos can be used to build multiple models using Convolutional Neural Networks and other machine learning based models to track pedestrians and monitor their behavior. The most used parameters are pedestrian speed, vehicle speed, pedestrian path changes, relative distance from pedestrian to vehicle, and time spent at curb.

1. **Detection**

Most pedestrian detection systems use Convolutional Neural Network (CNN) to classify and detect pedestrians. The current detection system is a two-stage process [3]. The first step is to classify and identify regions including crosswalk and intersections. The second step is to identify pedestrians and make bounding boxes to track them.

1. *Yolo5*

The yolo frame consists of the backbone or trunk, neck and the head [2]. The backbone is used to extract image features. The image is transmitted to the head by the neck through upsampling where synthetically generated data points are concatenated to balance the data set.. After transmission, the output is predicted, and the bounding box is generated.

These Yolo V5 models are generally more accurate than traditional machine learning models, however, they cannot run on robot controllers in real time. Lidar sensors which have high accuracy and measurement speed have been getting popular. However, since they are not standard sensors, the use can be expensive since they require high-end processors.

1. *Density Maps*

One of the other methods used for pedestrian detection is through 3-D reconstruction of the intersection. This takes place with the help of two CMOS cameras for video input and an IR illuminator for night vision (see figure 2). The 3-D points are obtained using calibration and transformation between stereo platform (camera rig) and the ground plane (the intersection)(see figure 1). These 3-D points are then filtered relative to the ground plane to detect pedestrians with points inside 0.2m < Y < 2m removed using a 3-D filtering procedure. A temporal density map is obtained through these projected points as shown in Figure 1. Next step is to apply object segmentation on temporal density map using a region growing algorithm. The background is also removed in this process. The resulting objects obtained on the dark plane are then further segmented into pedestrian or vehicle based on their orientation, size and speed. An occlusion reasoning algorithm is applied to detect shape of objects that are big in size which may correspond to multiple pedestrians.

A picture containing shape

Description automatically generated A picture containing sport

Description automatically generated

Figure : Object segmentation on dark plane  
Source: Image adapted from [7]

Figure 2: Plane transformation



Figure 3: CMOS Cameras(L), IR Illuminator(R)  
Source: Image adapted from [7]

1. **Pedestrian Behavior**

The next important component of pedestrian detection and tracking is pedestrian behavior modelling. The JAAD dataset [5] has shown that pedestrian tracking is not enough for AV’s to accurately predict the trajectory of pedestrians. The M.L. models that are trained for tracking pedestrians can become more accurate if pedestrian behavior is given weight in models as well. The JAAD dataset found that pedestrians communicate intention of crossing by looking in direction of approaching vehicles in 90% of the cases [6]. This minor stat could be very important as V2I communication can be used to signal AV’s if pedestrians have intention of crossing when they do not have the right of way.

A. JAAD Dataset

Going in depth into the dataset, it emphasizes driver’s and pedestrians’ behavior when they are about to cross the road. It contains 346 HD clips, approximately 5-10 seconds long, that are recorded in North America and Europe with an on-board camera at 30 FPS. Various types of annotations for the video and the 2793 pedestrians that are present in the video. These include bounding boxes for pedestrians, behavior tags like walking, standing, crossing etc. The dataset also includes pedestrians’ crossing direction and the frames where pedestrians take the decision to cross and where they cross.

B. Behavior Detection and Prediction

While different kinds of techniques can be used for pedestrian behavior modelling, some sort of pedestrian data is unquestionably required for accurately predicting pedestrian movement at crosswalks based on their behavior.

a) *Using JAAD dataset [4]*

The prediction model uses YOLOv4 object detection model [8] to create bounding boxes around pedestrians. The model is trained on COCO MS dataset [8]. The image obtained from bounding boxes is cropped to contain pedestrian’s head.

The two cropped images are passed through to two CNN classifiers to predict whether the pedestrian is walking or standing and if the pedestrian is looking towards the vehicle or not. The CNN models are trained using Walking/standing and Looking/N-looking data extracted from JAAD dataset. The training and validation is split into 0.8 and 0.2 respectively to guarantee no overlapping between the data chosen. Finally, spatial dependency is obtained by combining the output of the two CNN classifiers. Spatial dependency is defined as the relationship between two variables measures in a particular space. In this case spatial dependency is used to measure the relationship between state of the pedestrian and determine whether the pedestrian is standing/walking or looking/n-looking (not looking).

The next phase combines the results from the two CNN classifiers to predict whether the pedestrian is not crossing, about to cross or crossing. The labels are defined as ‘N-crossing’, ‘A-crossing’ and ‘Crossing’ respectively. The prediction is based on obtaining temporal dependency which is further obtained from relating the spatial information obtained to frames before or after that frame. The temporal relationship is obtained by training data-driven models like GRU, RNN and LSTM on data from the JAAD dataset. For the A-crossing and N-crossing labels, a decision frame is chosen as the starting frame for training while a separate crossing frame is chosen for label “Crossing” as shown in Figure 3.

Timeline

Description automatically generated

Figure 4: Decision Frames between scenarios

For the earlier two CNN models, where the intentions were to detect the state of pedestrians, the result was as follows. For walking/standing model, 92% f1 score or accuracy was achieved for “walking” while 96% f1 score was achieved for “standing”. For looking/n-looking model, f1 score of 87% was obtained for “Looking” and 94% was obtained for “N-looking”. As shown in figure 4, some spikes were obtained for the model results for “standing” as the pedestrian was perceived to be standing in the middle of walking. Therefore, a threshold was applied to ensure prediction stability. The same threshold was applied for “looking/n-looking” model as well.

For the prediction phase where the model was trained to determine the crossing state of pedestrian, the following results were achieved. As depicted by figure 5, pedestrian is predicted to cross 2.33 seconds or about 70 frames before they actually cross.

Chart, box and whisker chart

Description automatically generated

Figure 5: Prediction result for pedestrian state  
Source: Adapted from [4]

Chart, histogram

Description automatically generated

Figure 6: Crossing Prediction for Pedestrian Source: Adapted from [4]

The results using JAAD dataset are accurate as they work well for pedestrian behavior prediction. Similarly, other prediction models, if they use tagged pedestrian data like the JAAD dataset, predictions can become more accurate.

b) *Using Probabilistic models*

The following probabilistic model used acceptance gaps and time-to-collision as variables for determining pedestrian behavior [7]. Acceptance gaps are defined as gaps, whether distance or time, to oncoming vehicles that pedestrians deem safe before they cross [9]. On the other hand, time-to-collision is defined as the time for collision between the pedestrian and oncoming vehicle.

The data of acceptance gaps is distributed vs the time-to-collision and a log-normal distribution is fitted to data (figure 5).

Chart

Description automatically generated

Figure : Accepted gap distribution lognormal fitted Source: Adapted from [7]

In order to model these decisions, binary logistical regression is used. The world state w ∈ {0,1} follows the Bernoulli distribution where Pr(w = 1) = λ for accepting the gap and Pr(w = 0) = 1 - λ for rejecting the gap. According to the model, the probability of the accepting the gaps is contingent to eq (1):

Pr(w|x, β) = *Bern*[sig(β Tx)] = 1/(1+e-β^Tx)

Where x corresponds to data like Time-to-Collision, Distance, Vehicle speed etc. For the above equation, β Tx is a linear combination of inputs that will be used to model λ. It uses a sigmoid function to ensure a positive λ value under 1.

*Results*

According to the model, time-to-collision made the most contribution to the regression model, achieving a predictive quality of 88.97 % correct classifications. This follows he results achieved by previous studies that a strong correlation is present between time-to-collision and road-crossing probability. [10]

The values for x were changed for other variables. The predictive quality of the overall model was 93.10% classifications. It had 4 % more accurate classification than with just one variable – time-to-collision.

**4.** **Conclusion and Future Opportunities**

In this survey paper, we covered a variety of pedestrian detection and pedestrian behavior modelling techniques. While all techniques are accurate, the question is how these models can be used efficiently with the current infrastructure present. While detection, tracking and behavior modelling are interdependent on each other, the use of historical pedestrian data is imperative to produce more accurate results along with high-speed communication with vehicle through the infrastructure and vice-versa.

V2I communication will play an important role in this process as this technology will be used to communicate between the vehicle and a road-side unit [11]. V2I communication can also broadcast or spread information more effectively as they can transmit select information depending on pedestrian traffic, vehicular traffic, time of the day to control overcrowding of information.

V2I can also allow more data to be collected directly from other vehicles that have passed through the intersection. In the first step, data is passed to a road-side infrastructure unit. Then it is passed on to other vehicles coming through intersection if set conditions like time passed or speed of the vehicle are met. This can allow more historical data to be collected so that pedestrians can be more accurately detected.

**References**

[1] Y. Sun, Y. Yan, J. Zhao and C. Cai, "Research on Vision-based pedestrian detection and tracking algorithm," *2022 IEEE International Conference on Mechatronics and Automation (ICMA)*, Guilin, Guangxi, China, 2022, pp. 1021-1027, doi: 10.1109/ICMA54519.2022.9855902.

[2] F. Xudong, G. Xiaofeng, K. Ping, L. Xianglong and Z. Yalou, "Pedestrian Detection and Tracking with Deep Mutual Learning," *2021 18th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP)*, Chengdu, China, 2021, pp. 217-220, doi: 10.1109/ICCWAMTIP53232.2021.9674099.

[3] K. S. Wickramasinghe and G. U. Ganegoda, "Pedestrian Detection, Tracking, Counting, Waiting Time Calculation and Trajectory Detection for Pedestrian Crossings Traffic light systems," *2020 20th International Conference on Advances in ICT for Emerging Regions (ICTer)*, Colombo, Sri Lanka, 2020, pp. 172-177, doi: 10.1109/ICTer51097.2020.9325479.

[4] M. A. Manzour, O. M. Shehata and E. I. Morgan, "Pedestrian Crossing Intention Prediction via Spatio-Temporal Visual Data in Urban Traffic Scenes," *2021 IEEE 94th Vehicular Technology Conference (VTC2021-Fall)*, Norman, OK, USA, 2021, pp. 1-5, doi: 10.1109/VTC2021-Fall52928.2021.9625360.

[4] A. Rasouli, I. Kotseruba, and J. 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, 2017, pp. 206-213, doi: 10.1109/ICCVW.2017.29.

[5] A. Rasouli, I. Kotseruba, and J. K. Tsotsos, "Agreeing to cross: How drivers and pedestrians communicate," in 2017 IEEE Intelligent Vehicles Symposium (IV), 2017, pp. 264-269, doi: 10.1109/IVS.2017.7995703.

[6] A. Rasouli, I. Kotseruba, and J. K. Tsotsos, "The JAAD dataset: A large-scale video dataset for object detection, tracking, and understanding in realistic traffic videos," in 2017 IEEE International Conference on Computer Vision Workshop (ICCVW), 2017, pp. 853-861, doi: 10.1109/ICCVW.2017.106.

[7] R. Quintero Mínguez, M.-A. Sotelo, I. Parra, D. Fernández-Llorca, and C. Fernández, "Stereo-based Pedestrian Detection in Crosswalks for Pedestrian Behavioural Modelling Assessment," in *ICINCO 2014 - Proceedings of the 11th International Conference on Informatics in Control, Automation and Robotics*, vol. 2, pp. 102-109, 2014, doi: 10.5220/0005055401020109.

[8]A. Bochkovskiy, C. Wang, and H. Liao, "YOLOv4: Optimal Speed and Accuracy of Object Detection," in IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2020, pp. 1-8, doi: 10.1109/CVPRW50498.2020.00134.

[9] J. Zhao, J. O. Malenje, Y. Tang, and Y. Han, "Gap acceptance probability model for pedestrians at unsignalized mid-block crosswalks based on logistic regression," Accid Anal Prev, vol. 129, pp. 76-83, Aug. 2019, doi: 10.1016/j.aap.2019.05.012.

[10] S. Schmidt and B. Färber, "Pedestrians at the kerb—Recognising the action intentions of humans," Transportation Research Part F: Traffic Psychology and Behaviour, vol. 12, no. 4, pp. 300-310, Jul. 2009, doi: 10.1016/j.trf.2009.02.003

[11] W. He et al., "Overview of V2V and V2I Wireless Communication for Cooperative Vehicle Infrastructure Systems," 2019 IEEE 4th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC), Chengdu, China, 2019, pp. 127-134, doi: 10.1109/IAEAC47372.2019.8997786.