Pedestrian Crossing Intention Prediction at Red-Light Using Pose Estimation

Shile Zhang, Mohamed Abdel-Aty, Member, IEEE, Yina Wu, and Ou Zheng

Abstract—Pedestrians’ red-light crossing can present a threat to traffic safety. Among all the existing work related to pedestrian’s red-light crossing, there are few studies using trajectory data in time sequence. This paper uses pose estimation (keypoint detection) to generate pedestrians’ variables from CCTV videos. Four machine learning models are used to predict pedestrians’ crossing intention at intersections’ red-light. The best model achieves an accuracy of 0.920 and AUC value of 0.849, with data from three intersections. Different prediction horizons (up to 4 sec) are used. With longer prediction horizons, the sample size gets smaller, which partially leads to worse model performance. However, the performance with prediction horizon up to 2 sec is still good (AUC value as 0.841). It is found that keypoint variables such as the angles between ankle and knee (left side) and elbow and shoulder (right side) are important. This model can be further implemented in the Infrastructure-to-Vehicle (I2V) applications and thus prevent accidents due to pedestrians’ red-light crossing by issuing warnings to drivers.

Index Terms—Pedestrian crossing intention, red-light crossing, pose estimation, artificial intelligence (AI).

I. INTRODUCTION

PEDESTRIANS are regarded as the most vulnerable road users. According to the World Health Organization (WHO), 1.35 million fatalities were caused due to road crash annually. Among the total fatalities, 23% were pedestrians’ fatalities [1]. In the U.S., the number of pedestrians’ death increased by 54% during the 10-year period from 2010 to 2019, from 4,280 deaths in 2010 to 6,590 deaths in 2019 [2]. For pedestrian-related crashes, pedestrian’s unexpected crossing behavior such as suddenly walking out from the designated crosswalk/sidewalk can be one of the causal factors, especially the red-light crossings at signalized intersections. Based on NHTSA Fatality Analysis and Reporting System (FARS), pedestrians’ red-light crossings can cause hundreds of fatalities annually. And the number of fatalities has been growing in recent years [3]. Numerous studies have been carried out on the prediction of pedestrian’s crossing intention, however, there are few studies about pedestrian’s crossing intention during the red-light signal phase at signalized intersection, which is a special case that can be more critical.

This paper uses CCTV (closed-circuit television) videos with pose estimation (keypoint detection) technique to extract the key landmarks on pedestrians’ bodies. While CCTVs are not the ideal cameras for advanced computer vision application such as pose estimation, substantial benefits would be made with them to improve pedestrian safety. Besides, they are cost-efficient due to the wide coverage. Four machine learning models, Support Vector Machine (SVM), Random Forest (RF), Gradient Boosting (GBM), and Extreme Gradient Boosting (XGBoost) based on the generated variables are used. The dependent variable is divided into three classes, standing, walking (normal) (i.e., starting to cross during pedestrian signal phase), and walking (red-light) (starting to cross during red-light signal phase). The best model achieves an AUC value of 0.849. Different prediction horizons are taken into consideration as well.

Compared with traditional studies, this paper uses pedestrians’ trajectory data to predict pedestrian’s red-light crossing intention. This is an application of Artificial Intelligence (AI) in transportation safety. With the development of Infrastructure-to-Vehicle (I2V) technologies, the established model can be used to warn drivers of unexpected crossing pedestrians. It can also be used for signal timing optimization at signalized intersections.

II. LITERATURE REVIEW

A. Pedestrians’ Crossing Intention Prediction

Pedestrians’ crossing intention prediction was typically conducted in the same context with trajectory prediction. Among all the sensors, Wi-Fi and Bluetooth were usually used for indoor localization, while camera and LiDAR were used more in the road environment [4]–[6]. The related studies are summarized in Table I. It can be found that most studies used cameras to predict pedestrians’ crossing intention or trajectories [4], [5], [7]. From the perspective of modeling methods, three types of methods were mainly used in the literature, including parametric models such as Kalman Filter (KF) and Gaussian Process Dynamical Models (GPDMs), machine learning models such as SVM, and deep learning models such as long short-term memory (LSTM) [4], [8]–[10]. From the perspective of the predicting objectives, the output data were trajectories or crossing/non-crossing intentions [11]–[14]. It was found that pedestrians could change their motions abruptly, or could stop at any time. Quintero, et al. [15] used GPDMs and naïve-Bayes classifiers to predict pedestrians’
TABLE I

<table>
<thead>
<tr>
<th>Title</th>
<th>Sensor</th>
<th>Method</th>
<th>Objective/output</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bonnin, et al. [17], 2014</td>
<td>Camera</td>
<td>Context model tree</td>
<td>Crossing intention (crossing/not crossing)</td>
</tr>
<tr>
<td>Kooij, et al. [18], 2014</td>
<td>Camera</td>
<td>Neural network</td>
<td>Trajectory</td>
</tr>
<tr>
<td>Ferguson, et al. [19], 2015</td>
<td>Lidar</td>
<td>Gaussian process mixture model</td>
<td>Crossing intention (crossing/not crossing), trajectory</td>
</tr>
<tr>
<td>Völz, et al. [20], 2015</td>
<td>Lidar</td>
<td>Machine learning</td>
<td>Crossing intention (crossing/not crossing)</td>
</tr>
<tr>
<td>Goldhammer, et al. [21], 2015</td>
<td>Camera</td>
<td>Neural network</td>
<td>Trajectory</td>
</tr>
<tr>
<td>Quintero, et al. [15], 2015</td>
<td>Camera</td>
<td>Gaussian Process Dynamical Models (GPDM)</td>
<td>Crossing intention (crossing/not crossing), trajectory</td>
</tr>
<tr>
<td>Hashimoto, et al. [22], 2015</td>
<td>Camera</td>
<td>Dynamic Bayesian Network (DBN)</td>
<td>Crossing intention (crossing/not crossing)</td>
</tr>
<tr>
<td>Bock, et al. [23], 2017</td>
<td>Camera</td>
<td>Neural network</td>
<td>Trajectory</td>
</tr>
<tr>
<td>Rehder, et al. [24], 2018</td>
<td>Camera</td>
<td>Neural network (LSTM)</td>
<td>Trajectory, goal prediction</td>
</tr>
<tr>
<td>Saleh, et al. [12], 2018</td>
<td>Camera</td>
<td>Neural network (LSTM)</td>
<td>Behavior (bending in/starting/crossing/ stopping)</td>
</tr>
<tr>
<td>Míguez, et al. [13], 2019</td>
<td>Camera</td>
<td>Gaussian process dynamical models (GPDMs)</td>
<td>Behavior (walking/standing/starting/ stopping)</td>
</tr>
<tr>
<td>Abughaliheh and Alawneh [25], 2020</td>
<td>Camera</td>
<td>Neural network</td>
<td>Moving direction, distance to vehicle</td>
</tr>
<tr>
<td>Goldhammer, et al. [26], 2020</td>
<td>Camera</td>
<td>KF, machine learning</td>
<td>Trajectory, behavior (waiting/starting/moving/ stopping)</td>
</tr>
</tbody>
</table>

It should be noted that it’s usually complicated to define pedestrians’ crossing intention. Most of the traditional studies defined pedestrians’ crossing intention as binary categories, crossing/not crossing. To better define crossing intentions, some studies classified the pedestrians’ intention into several categories such as walking, standing, starting, stopping, etc. [11]–[14], [27]. Hariyono and Jo [11] used observers’ ratings to label the levels of pedestrians’ intention. In most of the cases, the pedestrians were labeled with certain categories such as crossing(1)/not crossing(0). Other categories between 0 and 1 were caused by some of the pedestrians’ behaviors, such as turning heads to watch for vehicles. Rasouli, et al. [28] collected a data set labeling pedestrians’ behaviors across various countries under different lighting conditions. Most of the behavioral patterns found are the sequences of “standing, looking, and crossing”, or “moving, looking, and crossing”.

B. Human Pose Estimation

Traditional studies learned pedestrians’ trajectories for predicting future states. However, it was found that merely trajectories of pedestrians and vehicles were not sufficient [29], [30]. Body languages such as leg movements or turning of body were indispensable among all the factors used for predicting pedestrians’ crossing intention. And there were controversial conclusions about whether pedestrian’s gaze or head orientation were important [31]–[33].

The development of pose estimation (keypoint detection) could better help recognize pedestrians’ states [34]. The pose estimation technique were used to detect the key points on human body. Pavllo, et al. [35] first applied a convolutional neural network on keypoint data generated from video. Luvizon, et al. [36] used pose estimation to conduct activity recognition. [37] used videos to recognize drivers that were distracted by phones while driving. Face detection, hand detection, as well as pose estimation of the upper body were used. Moreover, pose estimation offered a robust and effective way to estimate pedestrians’ crossing intention. Ghori, et al. [33] used a long short-term memory (LSTM) model to predict pedestrians’ and bicyclists’ crossing intention. A Bayesian inference function was used to predict the probabilities of five categories of behaviors (crossing, stopping, starting, etc.). Konrad, et al. [38] used a sequence of poses to extract variables such as lengths, angles, rotation rates, and linear accelerations formed by pedestrians’ joints. The kinematic variables of pedestrians were found to be reliable and accurate enough compared with an inertial measurement unit (IMU).
C. Pedestrians’ Red-Light Crossing Behavior

To investigate pedestrians’ red-light crossing intention, behavioral models such as the theory of planned behavior (TPB) model and statistical models were used [22], [39]–[41]. It was found that pedestrians’ characteristics, such as age, gender, grouping behavior, pedestrian volume, and safety awareness were significant factors [42], [43]. More emphasis should be placed on integrating pedestrians’ characteristics into the analysis. Besides, the pedestrians’ red-light crossing intention increased with longer waiting time, especially during the last few seconds before crossing [42], [44].

Compared with traditional traffic lights, countdown displays can significantly improve pedestrians’ signal compliance [45]. However, countdown displays are often installed near schools or busy intersections. As the studied locations in this paper are mostly located on major arterials in suburban area, the traditional traffic signals are installed. Pedestrians who cross at red-light have potential conflicts with the high-speed vehicular traffic. Besides, the push-button operation at intersections can help to separate the vehicular traffic and pedestrians, thus improving pedestrian safety [46], [47].

Despite the existing work on modeling of pedestrians’ red-light crossing intention, few work used sequential data [48]. With the development of Infrastructure-to-Vehicle (I2V) technology, the prediction of the pedestrians’ red-light crossing intention can be integrated with other communication media to warn drivers.

This study presents the prediction of pedestrians’ red-light crossing intention. Pose estimation is used to generate pedestrians’ variables from videos. Four machine learning models, Support Vector Machine (SVM), Random Forest (RF), Gradient Boosting (GBM), and Extreme Gradient Boosting (XGBoost), are used to predict pedestrians’ red-light crossing intentions from 1 sec up to 4 sec ahead. The best model achieves a recall value of 0.757 on the walking (red-light) class and an overall AUC value of 0.849. The model performance is still good when the model is used for predicting pedestrians’ red-light crossing intentions 2 sec ahead, with the AUC value as 0.841. This work can be applied in the I2V environment to better warn drivers.

III. DATA COLLECTION

The videos used in this study are from three signalized intersections located in Seminole County, Florida. All the videos are collected using CCTV (closed-circuit television) cameras during 8:00-19:00 on five sunny workdays in October and November 2019. All the intersections are four-lane by two-lane intersections to ensure the performance of the pose estimation model. The detailed information is listed in Table II. A total of 150-hour of videos are processed with 182 pedestrians collected as valid samples. The pedestrians’ trajectories before crossing (in waiting zone) are extracted. Another data source is ATSPM (Automated Traffic Signal Performance Measures) signal timing data to label pedestrians who cross at the red-light [49].

<table>
<thead>
<tr>
<th>Intersection</th>
<th>Road width (major/minor)</th>
<th>Vehicle volume (daily, major/minor road)</th>
<th>Vehicle approach speed (major road)</th>
</tr>
</thead>
<tbody>
<tr>
<td>US 17-92@3rd St</td>
<td>61 ft/20 ft</td>
<td>240/239519</td>
<td>28 mph</td>
</tr>
<tr>
<td>US 17-92@13th St</td>
<td>62 ft/36 ft</td>
<td>24251/4769</td>
<td>27 mph</td>
</tr>
<tr>
<td>SR 46@Park Dr</td>
<td>63 ft/39 ft</td>
<td>9959/5864</td>
<td>32 mph</td>
</tr>
</tbody>
</table>

Fig. 1. Pedestrian keypoint detection and transformation.
and their corresponding projections on the world plane [53]. A linear least squares method is used [54], [55]. As shown in Equation (2), the image coordinates \((u_t, v_t)\) and the world coordinates \((X_t, Y_t)\) (GPS coordinates in decimal degrees) are used to form matrix \(A\). Each pair of points forms two rows of matrix \(A\). Singular value decomposition (SVD) method is used to derive the solution by minimizing the value of \(\|Ah\|\) with \(h_0 = 1\). After obtaining \(h\) matrix and the inverse matrix of \(h\), all generated image coordinates can be transformed to world coordinates. The walking speed is calculated using haversine formula, which is the distance traveled by a pedestrian between two timestamps \(t_1\) and \(t_2\) (Equation (3)).

\[
\begin{pmatrix}
u \\
v \\
\end{pmatrix} = \begin{pmatrix}
X & Y \\
1 & 1 \\
\end{pmatrix} \begin{pmatrix}
h_1 & h_4 & h_7 \\
h_2 & h_5 & h_8 \\
h_3 & h_6 & h_9 \\
\end{pmatrix} \begin{pmatrix}
X \\
Y \\
1 \\
\end{pmatrix}
\]

(1)

\[
A \ast h = \begin{pmatrix}
0 & 0 & 0 & -X_1 & -Y_1 & v_1X_1 & v_1Y_1 & v_1 \\
X_1 & Y_1 & 1 & 0 & 0 & -u_1X_1 & -u_1Y_1 & -u_1 \\
0 & 0 & 0 & -X_2 & -Y_2 & v_2X_2 & v_2Y_2 & v_2 \\
X_2 & Y_2 & 1 & 0 & 0 & -u_2X_2 & -u_2Y_2 & -u_2 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
\end{pmatrix} = 0
\]

(2)

\[
\text{Walking speed} = \frac{\text{haversine}(X_{t1}, Y_{t1}), (X_{t2}, Y_{t2}))}{(t_2 - t_1)}
\]

(3)

B. Input Variables Overview

Using pose estimation, the angles between some of the key joints are generated. Besides, pedestrians’ walking directions, waiting time (time elapsed after the pedestrian reaches the waiting zone), walking speed, and whether pedestrian presses the pushbutton (to activate pedestrian signal phase), are also used as input variables. Some external variables are also included. The hourly temperature data are from National Oceanic Atmospheric Administration (NOAA). Total vehicle volume and right-turn vehicle volume at the current signal cycle, and green time of the vehicle signal phase on pedestrian’s conflicting direction are extracted from ATSPM. An overview of all input variables is listed in Table III.

C. Pedestrians’ Crossing Intention Labeling

Previous studies found that pedestrians’ red-light intention increased when waiting time increased [42], [44]. Thus, the last few moments are an important research target when a pedestrian approaches the road, stops at the curb (waiting zone), and finally starts crossing at red-light. Fig. 2 (a) shows a sequence of video frames. The time-to-cross has been previously used in the related work as the time difference between each frame and the frame when the pedestrian starts crossing [33], [56]. Time-to-cross equals zero means that the pedestrian starts to cross. As the time-to-cross get closer to zero (shown in Fig. 2 (b)), the pedestrian behaves more and more impatiently while looking around and watching for approaching traffic. Meanwhile, his crossing intention becomes clearer over time.

On average, the time intervals pedestrians spent on observing the surrounding environment are between 1 sec and 2 sec, which are around 1.32 sec for adults and 1.45 sec for the elderly and children [28]. This time interval is important for the decision-making of the crossing/not-crossing behavior. So, the last 1 sec to 2 sec before crossing can be important for the prediction of pedestrian’s red-light crossing behavior.

In this study, the dependent variable is pedestrians’ crossing intention. The labeling procedure is shown in Fig. 3. The CCTV videos are first processed using pose estimation and object tracking techniques. The frame rate of CCTV videos is 30 frames per second (fps). The samples in every 0.5 sec are then smoothed and aggregated into one sample to remove noise. The samples in the waiting zones are labeled with three classes, standing, walking normally (for pedestrians who stand still, and the other two classes are from the video frames when the pedestrian stands still, and the other two classes are from the video frames when the pedestrians start to cross (last 1 sec - 2 sec before time-to-cross=0). The labels are validated through manual checks to ensure accuracy.

For prediction purpose, we suppose the driver will yield to pedestrians after capturing the pedestrians’ crossing intentions after the reaction time 1 sec [57]. In this case, vehicles travel at 20 mph will have a stopping distance of 40 ft. The dependent variable is shifted 1 sec ahead of time (Fig. 4). This is regarded as the prediction horizon. The generated data set is later split into training set and test set for further modeling.
TABLE III
INPUT VARIABLE OVERVIEW

<table>
<thead>
<tr>
<th>Description</th>
<th>Mean</th>
<th>Standard deviation</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Walking direction</td>
<td>3.154</td>
<td>1.659</td>
<td>0.079</td>
<td>6.259</td>
<td>Rad</td>
</tr>
<tr>
<td>Walking speed</td>
<td>1.418</td>
<td>1.042</td>
<td>0.026</td>
<td>3.091</td>
<td>Ft/s</td>
</tr>
<tr>
<td>Pushing button</td>
<td>0.725</td>
<td>0.446</td>
<td>0.000</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>Waiting time</td>
<td>10.485</td>
<td>10.001</td>
<td>0.500</td>
<td>55.667</td>
<td>Sec</td>
</tr>
<tr>
<td>Angle α (ear &amp; eye, left)</td>
<td>1.362</td>
<td>0.792</td>
<td>7.5e-04</td>
<td>3.135</td>
<td>Rad</td>
</tr>
<tr>
<td>Angle α (ear &amp; eye, right)</td>
<td>0.521</td>
<td>0.270</td>
<td>2e-04</td>
<td>3.141</td>
<td>Rad</td>
</tr>
<tr>
<td>Angle α (nose &amp; eye, left)</td>
<td>1.374</td>
<td>0.870</td>
<td>3e-04</td>
<td>3.134</td>
<td>Rad</td>
</tr>
<tr>
<td>Angle α (nose &amp; eye, right)</td>
<td>0.397</td>
<td>0.214</td>
<td>9.7e-05</td>
<td>3.091</td>
<td>Rad</td>
</tr>
<tr>
<td>Angle α (elbow &amp; shoulder, left)</td>
<td>0.583</td>
<td>0.340</td>
<td>1e-03</td>
<td>3.116</td>
<td>Rad</td>
</tr>
<tr>
<td>Angle α (elbow &amp; shoulder, right)</td>
<td>0.692</td>
<td>0.458</td>
<td>1e-03</td>
<td>3.140</td>
<td>Rad</td>
</tr>
<tr>
<td>Angle α (wrist &amp; elbow, left)</td>
<td>0.649</td>
<td>0.406</td>
<td>4e-04</td>
<td>3.139</td>
<td>Rad</td>
</tr>
<tr>
<td>Angle α (wrist &amp; elbow, right)</td>
<td>0.629</td>
<td>0.487</td>
<td>1e-04</td>
<td>3.129</td>
<td>Rad</td>
</tr>
<tr>
<td>Angle α (ankle &amp; knee, left)</td>
<td>0.801</td>
<td>0.592</td>
<td>8e-04</td>
<td>3.132</td>
<td>Rad</td>
</tr>
<tr>
<td>Angle α (ankle &amp; knee, right)</td>
<td>0.756</td>
<td>0.576</td>
<td>7.49e-05</td>
<td>3.122</td>
<td>Rad</td>
</tr>
<tr>
<td>Vehicle volume (current cycle)</td>
<td>72.000</td>
<td>16.102</td>
<td>20.000</td>
<td>120.000</td>
<td></td>
</tr>
<tr>
<td>Vehicle green time (current cycle)</td>
<td>45.208</td>
<td>21.575</td>
<td>0.007</td>
<td>84.256</td>
<td>Sec</td>
</tr>
<tr>
<td>Vehicle counts (right-turn, current cycle)</td>
<td>5.000</td>
<td>3.060</td>
<td>0.000</td>
<td>11.000</td>
<td></td>
</tr>
<tr>
<td>Temperature</td>
<td>83.847</td>
<td>3.337</td>
<td>68.000</td>
<td>89.000</td>
<td>Fahrenheit</td>
</tr>
</tbody>
</table>

**IV. EXPERIMENT AND RESULTS**

Four machine learning models, SVM, RF, GBM, and XGBoost are established to predict pedestrians’ red-light crossing intentions. The models’ hyper-parameters are tuned to reach the best performance.

**A. Evaluating Metrics**

The evaluating metrics, such as precision, recall, F1 score, accuracy, and AUC are illustrated as below.

1. **Precision**: the proportion of correctly classified samples among classified positive samples, as shown in Equation (4).

2. **Recall**: or sensitivity, the proportion of correctly classified samples among actual positive samples, as shown in Equation (5).

3. **F1 score**: a weighted average of precision and recall, as shown in Equation (6).

4. **Accuracy**: the proportion of correctly classified samples among all the samples, as shown in Equation (7).

5. **AUC**: area under the ROC curve. The ROC (Receiver Operating Characteristic) curve is used as a comprehensive metric to evaluate the model’s performance. This curve plots two parameters, recall and false alarm rate (FAR), at different classification thresholds. The AUC value, which ranges from 0.5 to 1, is the area under the ROC curve. For an imbalanced data set, the AUC value is more reliable than accuracy.

\[
\text{Precision} = \frac{TP}{TP + FP} \tag{4}
\]

\[
\text{Recall} = \frac{TP}{TP + FN} \tag{5}
\]

\[
F1\text{score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{6}
\]

\[
\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN} \tag{7}
\]

**B. Experiment Results**

Among all the 182 pedestrians collected from CCTV data, 61 pedestrians start to cross the road during the red-light...
signals. With the sampling time window as 0.5 sec, there are 2,375 data samples collected, with the number of samples between the three classes is 1,725: 407: 243. Eighty percent of the samples are used as the training data set, and twenty percent of the samples are used as the test data set. Synthetic Minority Over-Sampling Technique (SMOTE) is used to balance the numbers of samples in three classes in the training data set, to make all three categories balanced [58], [59].

In this study, the dependent variable is divided into three classes, standing, walking (normal), and walking (red-light). The last class is the most critical class. So, the model’s performance of this class should be put more emphasis on. Meanwhile, the average value of metrics over three classes, which is usually called macro average value, is also calculated. The modeling results of the four models with prediction horizon as 1 sec are listed in Table IV.

The best model is determined to be RF. The recall value for walking (red-light) class is 0.757, which means the model can recognize 75.7% of the samples (video frames) in which the pedestrians start walking at red-light. Meanwhile, the precision value is 0.800. It also achieves the best performance over three classes compared with the other models. Overall, RF achieved an accuracy of 0.920 and an AUC value of 0.849 over the test data set.

Confusion matrix is usually used to check the overall performance of the model, and identify the specific errors affecting each class. The confusion matrix of the RF model on the test data set is shown in Table V. Most of the samples in each class are classified correctly, denoting the model has a good performance.

The variable importance plot with the top fifteen important variables is shown in Fig. 5. It can be found that walking speed, waiting time, green time (vehicle signal phase), pushing button behavior play important roles for predicting pedestrians’ red-light crossing intention. Besides, the angles between knee and ankle (on the left side) also play an important role. Facial variables are also found to be important, such as the angle between left ear and left eye. This may be related to head orientation. As the data are collected in Florida, where extremely hot weather is usually present at noon, the temperature is also an influencing factor.

Given higher speed limits, there is a need to use longer prediction horizon to build the model. Thus, the other values, 2 sec, 3 sec, and 4 sec, are also taken into consideration. The experiment results are shown in Table VI. When the prediction horizon is 2 sec, the model still maintains an AUC value of 0.841. With the prediction horizon increases to up to 4 sec, the sample size keeps shrinking. So, the model’s performance on walking (red-light) class (the most minority class) gets worse, resulting in low values of the evaluating metrics. Besides, the macro average values of evaluating metrics show an overall tendency of decreasing. For an imbalanced data set, the AUC can better reflect model performance than
the model still shows a fairly good performance over all target classes. Overall, the AUC value decreases as prediction horizon increases. When the prediction horizon is 2 sec the model’s performance is still good, with the recall value as 0.623 (on walking (red-light) class and AUC value as 0.841.

A more generic model with the dependent variable labeled as standing/walking is also established for comparison. It can be found that the model’s performance gets improved, with the AUC value as 0.889. If the signal timing data is not available, then this model can be used instead for warning approaching vehicles, especially the right-turn vehicles.

The limitation of this study is that data from merely three intersections are used with similar geometric design (four-lane by two-lane intersections). But as the CCTV cameras are installed at different locations with different angles, the model successfully deals with the heterogeneity of the generated data set. The study sheds light on the application of pose estimation for studying pedestrian safety. The variables automatically generated from pose estimation can better predict pedestrians’ red-light crossing intention compared with only mobility variables such as position and speed [48]. Future work can be conducted to implement the proposed model in field test.

V. CONCLUSION AND DISCUSSION

This paper uses video data to predict pedestrians’ red-light crossing intentions at the signalized intersections with pose estimation and various machine learning models. The highlights of the study mainly include:

1) The pose estimation technique is used to capture the variables of the pedestrians’ bodies, such as angles formed by some of the key joints (wrist, elbows, etc.) and facial landmarks (nose, eyes, and ears) over time.

2) Upon labeling the dependent variable, pedestrians’ red-light crossing intention, the study takes into consideration both mobility (standing/walking) and pedestrians’ red-light crossings.

3) Four machine learning models are used to predict the pedestrians’ red-light crossing intentions with multiple prediction horizons. The best model achieves an AUC value of 0.849.

Through the established models, there are a few points to be marked on pedestrians’ crossing intention prediction. The walking speed is the top important variable to reflect pedestrians’ crossing intentions. The other variables such as button pushing and waiting time may be related to red-light violations. The leg movement denoting by the angle between knee and ankle is an important variable. Compared with the body part, the facial landmarks also reveal early signs of starting walking, which can be related to head orientation.

With respect to different prediction horizons, though the evaluating metrics on walking (red-light) class fluctuates, the model still shows a fairly good performance over all target classes. Overall, the AUC value decreases as prediction horizon increases. When the prediction horizon is 2 sec the model’s performance is still good, with the recall value as 0.623 (on walking (red-light) class and AUC value as 0.841.

A more generic model with the dependent variable labeled as standing/walking is also established for comparison. It can be found that the model’s performance gets improved, with the AUC value as 0.889. If the signal timing data is not available, then this model can be used instead for warning approaching vehicles, especially the right-turn vehicles.

The limitation of this study is that data from merely three intersections are used with similar geometric design (four-lane by two-lane intersections). But as the CCTV cameras are installed at different locations with different angles, the model successfully deals with the heterogeneity of the generated data set. The study sheds light on the application of pose estimation for studying pedestrian safety. The variables automatically generated from pose estimation can better predict pedestrians’ red-light crossing intention compared with only mobility variables such as position and speed [48]. Future work can be conducted to implement the proposed model in field test.

V. CONCLUSION AND DISCUSSION

This paper uses video data to predict pedestrians’ red-light crossing intentions at the signalized intersections with pose estimation and various machine learning models. The highlights of the study mainly include:

1) The pose estimation technique is used to capture the variables of the pedestrians’ bodies, such as angles formed by some of the key joints (wrist, elbows, etc.) and facial landmarks (nose, eyes, and ears) over time.

2) Upon labeling the dependent variable, pedestrians’ red-light crossing intention, the study takes into consideration both mobility (standing/walking) and pedestrians’ red-light crossings.

3) Four machine learning models are used to predict the pedestrians’ red-light crossing intentions with multiple prediction horizons. The best model achieves an AUC value of 0.849.

Through the established models, there are a few points to be marked on pedestrians’ crossing intention prediction. The walking speed is the top important variable to reflect pedestrians’ crossing intentions. The other variables such as button pushing and waiting time may be related to red-light violations. The leg movement denoting by the angle between knee and ankle is an important variable. Compared with the body part, the facial landmarks also reveal early signs of starting walking, which can be related to head orientation.

With respect to different prediction horizons, though the evaluating metrics on walking (red-light) class fluctuate, the model still shows a fairly good performance over all target classes. Overall, the AUC value decreases as prediction horizon increases. When the prediction horizon is 2 sec the model’s performance is still good, with the recall value as 0.623 (on walking (red-light) class and AUC value as 0.841.

A more generic model with the dependent variable labeled as standing/walking is also established for comparison. It can be found that the model’s performance gets improved, with the AUC value as 0.889. If the signal timing data is not available, then this model can be used instead for warning approaching vehicles, especially the right-turn vehicles.

The limitation of this study is that data from merely three intersections are used with similar geometric design (four-lane by two-lane intersections). But as the CCTV cameras are installed at different locations with different angles, the model successfully deals with the heterogeneity of the generated data set. The study sheds light on the application of pose estimation for studying pedestrian safety. The variables automatically generated from pose estimation can better predict pedestrians’ red-light crossing intention compared with only mobility variables such as position and speed [48]. Future work can be conducted to implement the proposed model in field test.

V. CONCLUSION AND DISCUSSION

This paper uses video data to predict pedestrians’ red-light crossing intentions at the signalized intersections with pose estimation and various machine learning models. The highlights of the study mainly include:

1) The pose estimation technique is used to capture the variables of the pedestrians’ bodies, such as angles formed by some of the key joints (wrist, elbows, etc.) and facial landmarks (nose, eyes, and ears) over time.

2) Upon labeling the dependent variable, pedestrians’ red-light crossing intention, the study takes into consideration both mobility (standing/walking) and pedestrians’ red-light crossings.

3) Four machine learning models are used to predict the pedestrians’ red-light crossing intentions with multiple prediction horizons. The best model achieves an AUC value of 0.849.

Through the established models, there are a few points to be marked on pedestrians’ crossing intention prediction. The walking speed is the top important variable to reflect pedestrians’ crossing intentions. The other variables such as button pushing and waiting time may be related to red-light violations. The leg movement denoting by the angle between knee and ankle is an important variable. Compared with the body part, the facial landmarks also reveal early signs of starting walking, which can be related to head orientation.

With respect to different prediction horizons, though the evaluating metrics on walking (red-light) class fluctuate, the model still shows a fairly good performance over all target classes. Overall, the AUC value decreases as prediction horizon increases. When the prediction horizon is 2 sec the model’s performance is still good, with the recall value as 0.623 (on walking (red-light) class and AUC value as 0.841.

A more generic model with the dependent variable labeled as standing/walking is also established for comparison. It can be found that the model’s performance gets improved, with the AUC value as 0.889. If the signal timing data is not available, then this model can be used instead for warning approaching vehicles, especially the right-turn vehicles.

The limitation of this study is that data from merely three intersections are used with similar geometric design (four-lane by two-lane intersections). But as the CCTV cameras are installed at different locations with different angles, the model successfully deals with the heterogeneity of the generated data set. The study sheds light on the application of pose estimation for studying pedestrian safety. The variables automatically generated from pose estimation can better predict pedestrians’ red-light crossing intention compared with only mobility variables such as position and speed [48]. Future work can be conducted to implement the proposed model in field test.


Shile Zhang is currently pursuing the Ph.D. degree with the University of Central Florida (UCF). She has worked on several funded projects in the context of connected vehicle and traffic safety. She has several publications in journals, including Accident Analysis and Prevention and Transportation Research Record. Her research interests include pedestrian safety, surrogate safety measures (SSMs), and computer vision applications.

Mohamed Abdel-Aty (Member, IEEE) is currently a Pegasus Professor and the Chair of the Department of Civil, Environmental, and Construction Engineering, University of Central Florida (UCF), Orlando, FL, USA. He has managed more than 70 research projects of $20. He has published more than 650 articles, 350 in journals (As of October 2020, Google Scholar citations are 19 800 H-Index 74, Science direct papers downloaded more than 360 000 times). His main expertise and research interests include ITS, traffic safety, simulation, CAV, and active traffic management. He received nine best paper awards from ASCE, TRB, and WCTR.

Ou Zheng received the master’s degree in transportation (smart cities) and computer vision. He has strong engineering professional with his master’s degree. He is currently a Computer Vision Research Engineer with the Department of Civil, Environmental, and Construction Engineering, University of Central Florida (UCF). He is also an experienced Algorithm Engineer with a demonstrated working experience from traffic safety-related research. He has been instrumental in many major projects with the UCF SST Team, including the USDOT award-winning safety system, UCF SST A.R.C.I.S computer vision system, and FDOT ATTAIN smartphone applications in connected vehicles. He is skilled in Python, C++, Distributed Systems, Deep Learning, and Computer Vision.

Yina Wu received the Ph.D. degree in civil engineering (transportation) from the University of Central Florida (UCF). She is currently a Research Associate Professor with the Department of Civil, Environmental, and Construction Engineering, UCF. She is the PI or Co-PI for six projects that are related to traffic safety. She has published 20 peer-reviewed articles related to real-time prediction, connected and autonomous vehicles (CAVs), human/behavioral factors, computer vision technologies, and other aspects of traffic safety in top journals, including Transportation Research Part C, Accident Analysis and Prevention, IEEE, Journal of Intelligent Transportation Systems, and Transportation Research Record. Her research interests include driver behavior, surrogate safety measure, computer vision, crash risk analysis, simulation, human factors, and connected vehicles. She serves as a Committee Member for the TRB road weather committee.