Vehicle-Pedestrian Dynamic Interaction through Tractography of Relative Movements and Articulated Pedestrian Pose Estimation

Rifat Mueid
Indiana University-Purdue University Indianapolis
Indianapolis, IN 46202
rmueid@iupui.edu

Lauren Christopher
Indiana University-Purdue University Indianapolis
Indianapolis, IN 46202
lauchris@iupui.edu

Renran Tian
Indiana University-Purdue University Indianapolis
Indianapolis, IN 46202
rtian@iupui.edu

***Abstract*—To design robust Pre-Collision Systems (PCS) we must develop new techniques that will allow a better understanding of the vehicle-pedestrian dynamic relationship, and which can predict pedestrian future movements. This paper focuses on the potential-conflict situations where a collision may happen if no avoidance action is taken from driver or pedestrian. We have used 1000 15-second videos to find vehicle-pedestrian relative dynamic trajectories and pose of pedestrians. Adaptive structural local appearance model and particle filter methods have been implemented to track the pedestrians. We have obtained accurate tractography results for over 82% of the videos. For pose estimation, we have used flexible mixture model for capturing co-occurrence between pedestrian body segments. Based on existing single-frame human pose estimation model, we have implemented Kalman filtering with other new techniques to make stable stick-figure videos of the pedestrian dynamic motion. These tractography and pose estimation data were used as features to train a neural network for classifying 'potential conflict' and 'no potential conflict' situations. The training of the network achieved 91.2% true label accuracy, and 8.8% false level accuracy. Finally, the trained network was used to assess the probability of collision over time for the 15 seconds videos which generates a spike when there is a 'potential conflict' situation. The paper enables new analysis on potential-conflict pedestrian cases with 2D tractography data and stick-figure pose representation of pedestrians, which provides significant insight on the vehicle-pedestrian dynamics that are critical for safe autonomous driving and transportation safety innovations.**

Keywords—Pre-Collision System, Pose estimation, Transportation Safety

I. INTRODUCTION

Vehicle (driver)-pedestrian interaction is a very important aspect for transportation safety, especially as driving becomes more autonomous. Current systems operate using Crash Imminent Braking (CIB) where the brakes are only applied at the last minute to avoid collisions. As driving becomes more autonomous, earlier actions by the vehicle must be developed in a more comprehensive way to avoid getting into the CIB situations. The dynamic behavior of the pedestrian in traffic can indicate whether the pedestrian is aware or unaware of the oncoming vehicle. Pedestrians also have a negotiation strategy for crossing traffic which is dynamic in nature, and the pedestrian pose (waving vehicle ahead, running, starting and stopping) can indicate to the vehicle important information.

We have used machine learning techniques to analyze existing IUPUI TASI 110-Car naturalistic driving video to classify and understand the dynamic vehicle-pedestrian interactions. These videos represent the view of 110 drivers around the Indianapolis metropolis area for one whole year. We have implemented visual tracker model [1] [2] to recognize pedestrians and track them. Finally, we have computed depth and lateral position of the pedestrians with respect to the vehicle. Next we have employed the flexible mixture-of-parts method [3] to estimate human pose and improved the base single frame pose estimation remarkably. Later these data were used to classify between 'potential conflict' and 'no potential conflict' situations and to generate instantaneous danger for 'potential conflict' videos over time. The results of this research can be used for developing autonomous driving rules, or for autonomous vehicle testing.

II. TRACTOGRAPHY

A. Tracking

Precise tractography data largely depends on the accuracy of the tracking. This tracking is important for two reasons: 1) The pedestrian must be tracked well in each frame to produce an accurate graph, and 2) The dimension of the tracking box around the pedestrian is later used for calculating depth and lateral

This research is supported by Samsung Global Research Outreach (GRO) Program.

This is the author's manuscript of the article published in final edited form as:
Mueid, R., Christopher, L., & Tian, R. (2016). Vehicle-pedestrian dynamic interaction through tractography of relative movements and articulated pedestrian pose estimation. In 2016 IEEE Applied Imagery Pattern Recognition Workshop (AIPR) (pp. 1–6). https://doi.org/10.1109/AIPR.2016.8010592

position. In this case we have used the adaptive structural local appearance model and particle filter methods described in [1] and [2]. From this base, we have done extensive experiments with these methods and modified several parameters so the code is best customized for pedestrians.

Pedestrians were already extracted using (HOG-based) pattern recognition techniques from the TASI 100 car naturalistic video dataset. After detection of a single pedestrian (verified manually and best pedestrian image frame chosen manually), these videos were organized into 5 seconds and 15 second videos centered in time on the detected pedestrian single frame. A database of manually identified features was made, and we used one of these features as the starting point for this research: potential conflict. Potential conflict is defined as the direction and current speed of vehicle or pedestrian would cross at a point in the future. Because we had no crashes, this implied that either the vehicle or the pedestrian changed speed or trajectory to avoid the collision. We have used these 15 second videos with potential conflict cases as our analysis starting point. Each video in the database has a log that contains some data analyst information, including position boxes of pedestrians and bicyclists in a frame and the reference frame number (out of the 15 seconds of frames). In our research, we used this log file for the starting frame and position box to further track the pedestrian in both directions (forward and back in time) from this center frame. Target template from the reference frame is obtained using the log information as shown in Figure 1. Then we apply an affine transformation to make a customized template size.

Figure 1. Target template from the reference frame

The particle filter provides an estimate of posterior distribution of a random variable related to a Markov chain. In visual tracking applications, it is an important tool for estimating the target tracked in the next frame without knowing the concrete observation probability. We generate 600 particles and employ a Gaussian distribution to model the state transition distribution. An example of 10 particle windows is shown in Figure 2.

Figure 2. 10 particle windows

In figure 2, by applying an affine transformation using the state information of the pedestrian as parameters, crop the region of interest (ROI) from the image and normalize it to be the same size as the target template. Similarity between target candidates and the target template is calculated to find out the most similar one. The weights of particles are updated based on the calculated similarity results.

B. Focus of Expansion

The Focus of expansion (FoE) is an important parameter to compute the relative distance between the pedestrian and vehicle, and the height parameter of the FoE strongly effects the depth calculation in the tractography as discussed in our previous work [2]. We have developed a process to automate the FoE calculation. We calculate this improved focus of expansion using the following process:

(1). Take 30 frames (a 1-second clip) of a sequence where the vehicle is moving in the videos, then average them, forming a single image. This produces a smear of the video, centered at the FoE.

(2). Apply a Hough transform to find lines in the image which will converge to the FoE, along the smeared video. Some lines are then eliminated based on orientation angle, as the FOE is expected only in the center of the image, and must pass through a center ROI. These remaining lines are extended in both directions and are shown in Figure 3.

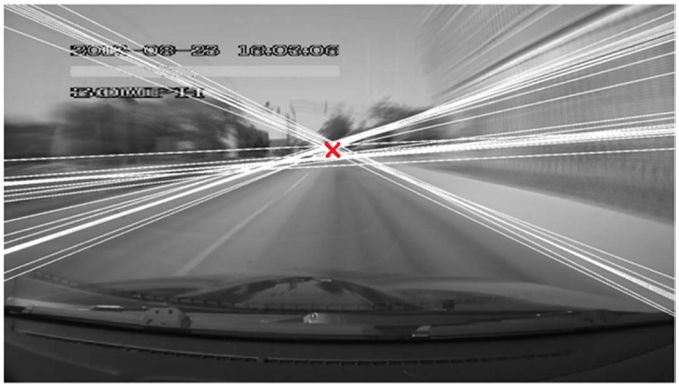

Figure 3. FoE

(3). Calculate the FoE from the intersection of lines using the center of mass of the crossing points of the lines as the expected FOE, as shown in Figure 3 (small red "x" in FOE center).

We have applied several conditions and imposed restrictions to find out the accurate FoE. For more accuracy, we have calculated FoE for each second and selected the best one (highest number of lines, with low variance for the crossing points) for the total video.

C. *Tractography Data Extraction*

Understanding of the vehicle-pedestrian dynamic position is very important. Therefore, we need to plot the relative distance of pedestrian from vehicle over time. We have developed a prototype of this using computer vision techniques combined with 2D to 3D projection geometries. These techniques discussed above are used to track pedestrian-vehicle spatial interaction. The time series of the tracks can be collected together and visualized as a tractographs as shown in Figure 4.

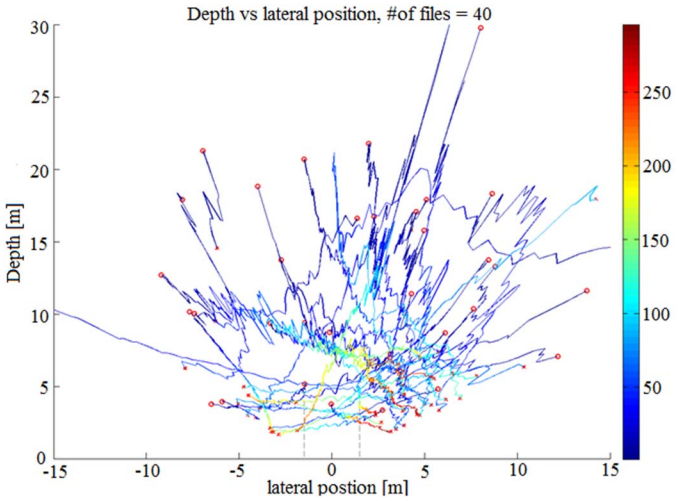

Figure 4. Tractography

In this figure, the instantaneous positions of 40 randomly chosen pedestrian videos have been overlaid into a single plot. The graph is relative position of the pedestrian with respect to the vehicle position, where the vehicle is centered anchored between the dashed lines at the center-bottom of the graph. The start point (first appearance of the pedestrian in time) is denoted by red 'o' and end point (disappearance point) is denoted by red 'x'. The direction of the movement with time is denoted by the color of the trace as shown in the sidebar of the figure. The sidebar scale represents frame number at 30 frames per second. So from the color of a trace at the end point we can understand how much time that trace denotes. The features from the relative positions and dynamic behaviors developed from this tractography can also be used to inform the semantic human-vehicle interaction. Later in the data analysis, large variations may be smoothed with filtering or outliers eliminated to produce smoother track. Also, we know from our previous work [2], that the position data accuracy reduces with distance from the car, so the best region of interest for accurate data will be in a half circle in front of the car. Typical scenarios can then be gleaned from this data that are useful for autonomous driving control or for testing such vehicle systems.

From the database of our TASI 100-car study, human generated vehicle-pedestrian motion has tagged this pedestrian motion into four different broad categories: 1) pedestrian crossing from right to left (of the vehicle), 2) crossing left to right, 3) pedestrian walking towards the vehicle, and 4) pedestrian walking in the same direction as the vehicle. These four scenarios are shown in figure 5-8, collecting 15 cases of each scenario.

Tractography can give us significant insight of vehicle-pedestrian negotiation. For example, when pedestrian walking in the opposite direction of the vehicle the relative distance decreases very fast. We can see that for some cases in Figure 7 relative distance (depth) has decreased a lot in just a small period of time which is visualized by a little change in color. This we can glean from the color change in the trace. This understating can be an important factor for calculating risk factor, time to collision and other parameters which are essential for autonomous driving.

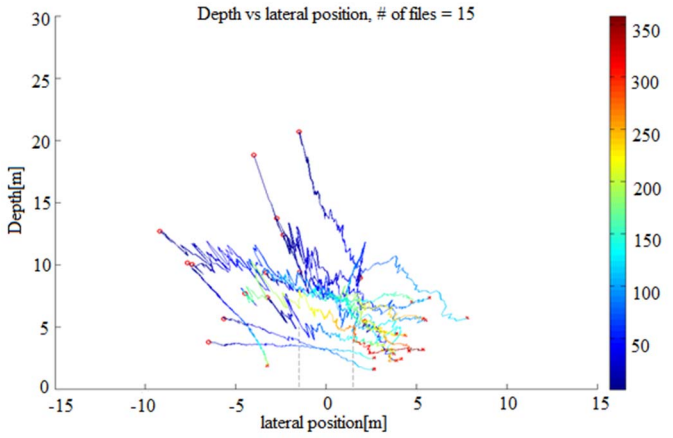

Figure 5. Pedestrians crossing from left to right

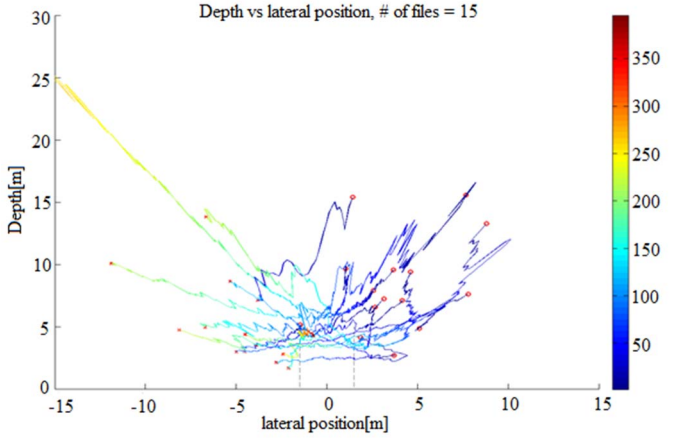

Figure 6. Pedestrian crossing from right to left

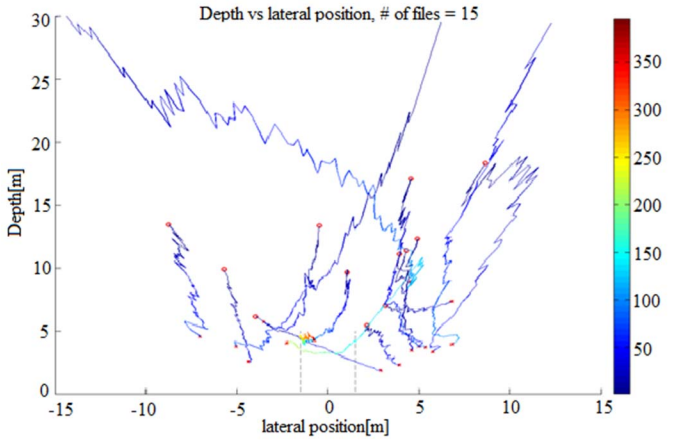

Figure 7. Pedestrians walking towards the vehicle

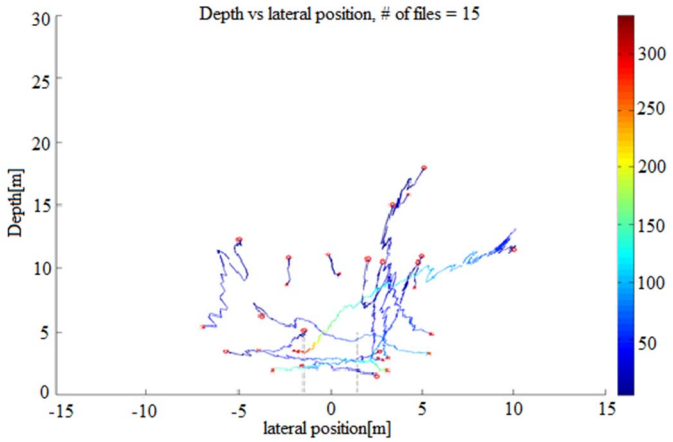

Figure 8. Pedestrians walking with the vehicle

III. POSE ESTIMATION

We have used a method for human pose estimation in static images based on a representation of part models described in [3]. The method does not use articulated limb parts, but rather captures how the templates of each part orients with each other. A general, flexible mixture model is used for capturing co-occurrence relations between segments. Then, standard spring models are augmented that encode spatial relations. It has been shown in [3] that such relations can capture notions of underlying local structure. The model can be effectively optimized with greedy algorithm when co-occurrence and spatial relations are tree-structured.

Frames of a video are cropped using the pedestrian position information from the tracking. The cropped images are little larger than the box size. In our case, we will provide these cropped images from tracking as input rather than the whole image to reduce computational time.

For image pre-processing initially we used histogram equalization and adaptive histogram equalization. Though adaptive histogram equalization performed better than histogram equalization we experimented with different sharpening filters for more accuracy. Sharpening filter with an unsharp masking performed better than adaptive histogram equalization. However, adaptive histogram equalization along with the sharpening filter performed almost as good as the sharpening filter alone. So, we used the sharpening filter alone and also both sharpening filter and adaptive histogram equalization interchangeably to boost the high frequency components of the images.. This is a very important finding from our current research. The original paper [3] used very good high-contrast, high resolution human figure pictures. Our data has varying resolutions and contrast, due to the natural light variation across the day, and the distance to the pedestrian.

Then following from the work in [3], a feature pyramid is created for each image considering all limb parts of the human. Corresponding confidence scores are also calculated for each limb part. Then a greedy algorithm is implemented to select the best combination among the parts and corresponding confidence score of the whole human pose is also calculated.

To further improve the result we have used Kalman filter to estimate the future location of the different body parts that has been used to create the estimated stick figures. It improved the result significantly, especially with the frames where there was

Figure 9. Pose Estimation

no detection of pedestrians. We could predict the stick figure joint locations of the non-detected frame with the Kalman filter. Also the filter reduces the weight of wrong detections among frames remarkably.

After the use of Kalman filter, a moving average smoothing filter was to smooth the changes of positions of the stick figures. The final output is stable and smooth stick figures of the pedestrian poses.

Using this, we have been able to reduce frame to frame pixel offset by 86% compared with the previous single frame model.

Ten consecutive output frames for pedestrian pose estimation for three videos are shown in Figure 9. The color of the line represents the match (blue-left leg, red-right leg, yellow-torso, green-head, magenta-right arm, cyan-left arm).

IV. Statistical Analysis

A. Neural Network Training

The 34 examples of the 5-second video sequences used in this part were human-labelled with "potential conflict" if at some point in the video, the path of the pedestrian and the path of the automobile would cross. Since none of the vehicles in our study encountered a true crash, we are using the potential conflict as a training set for the statistical analysis. We also obtained 34 vehicle-pedestrian videos that were labelled "no potential conflict". For example, a pedestrian on a sidewalk parallel to the motion of the vehicle is considered "no potential conflict". This test data was used for neural network training.

First the Principle Component Analysis (PCA) was done on the input data (tractography and pose) similarly as it was done for feature selection, and then a 2-layer neural network was trained with this input, the network can be seen in Table 1.

Table 1. Confusion Matrix

		Target Class	
Output Class	30 44.1%	2 2.9%	93.8% 6.3%
	4 5.9%	32 47.1%	88.9% 11.1%
	88.2% 11.8%	94.1% 5.9%	91.2% 8.8%

B. Danger Assesment

Since any real-time system will not have the advantage of the "future" time of the whole interaction between the pedestrian and the vehicle, the data must be labelled over time, so this task is to run short segments of time through the trained network to identify the key features that indicate potential conflict outcomes.

As from the tractography data we can extrapolate that there is a chance of collision if the path of the vehicle and pedestrian crosses, it can be used for used for continuous danger assessment over time.

The neural network that we have trained to classify between potential conflict and no potential conflict situations was used to assess the risk of collision over time for the 15 seconds potential conflict videos, our main database for this project.

Figure 10 shows probability of 'potential conflict' over 15 seconds of a video.

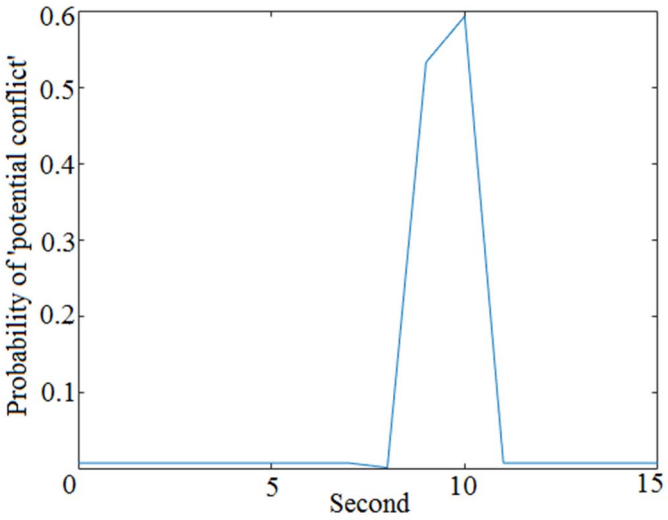

Figure 10. Time series of danger assessment of a 15 second video

From the graph we can see that there is chance of 'potential conflict' above the threshold of 0.5 value around 9-11 seconds. The accuracy of the graph can be verified manually by watching the video if really there is a 'potential conflict' situation in the video in similar time as the spike of the graph.

With a random selection of videos, we have checked that for 90% cases the trained network can produce a spike when there is visibly a 'potential conflict' situation.

V. Conclusion

The long-term research goal is to enable advances in computer vision, robotics, vehicle safety, and consumer electronics by capturing human semantic information from naturalistic driving movies. So far, there has been no dynamic testing of vehicle-pedestrian interaction, only time to collision (TTC) with automatic braking (CIB) has been tested. The research from this study will inform the future test scenarios and provide more advanced concepts for autonomous driving systems about the probable vehicle-pedestrian interactions.

As pedestrian-vehicle interaction is better understood, systems can be created to reduce confusion and wrong decisions from the two parties, improve traffic efficiencies, and prevent injuries or fatalities. In addition, the false alarms (false positives) from these autonomous or semi-autonomous driving

systems can be reduced, if normal pedestrian behavior is understood.

ACKNOWLEDGMENT

We would like to thank SAMSUNG Global Research Outreach (GRO) Program for supporting this project.

REFERENCES

[1] X. Jia, H. Lu, and M. Yang, "Visual Tracking via Adaptive Structural Local Appearance Model", IEEE Conference on Computer Vision and Pattern Recognition, Providence, June, 2012.

[2] C. Liu, R. Fujishiro, L. Christopher, J. Zheng, "Vehicle-Bicyclist Dynamic Position Extracted from Naturalistic Driving Videos", IEEE Transactions on Intelligent Transportation Systems, 2016.

[3] Y. Yang, D. Ramanan, "Articulated Human Detection with Flexible Mixtures-of-Parts", IEEE Transactions on Pattern Recognition and Machine Intelligence, 2013, 35(12), pp. 2878 – 2890.