Proposal: Recognizing Pedestrian Behavior in Transit Crossings by Using Pose Estimation

(XCS229II)

Sebastián Barrios Slight

sbarrios93@gmail.com

I. General Problem

Autonomous Vehicles (AVs) have been at the center of the news as the next (or current?) big thing in the tech world. AVs are poised to transformed our roads in safer places for everyone: passengers, pedestrians and every other road participant. However, reaching a safety standard where everyone feels confident with AVs has proven difficult, as the underlying algorithms and networks have to deal with implicit road rules, unexpected actions by road participants and a myriad of edge cases that NNs have not been exposed to before.

One of the topics that needs further work is the one that has a new level of complexity – the one related to the estimation of pedestrian intentions. This complexity comes from the inherent randomness in pedestrian behavior as individuals, and because pedestrian roads do not have many implicit rules (such as bike lanes and others).

The papers discussed in this Literature Review all approach this topic in novel ways, and they all shared the same dataset, providing an interesting view at the different ways researchers and practitioners are looking to solve this task.

II. Concise Summary

A. *Are they going to Cross? A Benchmark Dataset and Baseline for Pedestrian Crosswalk Behavior*

The main contribution of Rasouli et al. was the introduction of the Joint Attention in Autonomous Driving (JAAD) dataset. In addition to bounding boxes for pedestrian detection, the dataset also includes pedestrian behavior and contextual annotations for each video clip. The paper aims to prove that incorporating the latter information improves prediction over the case of only using bounding boxes.

The dataset introduced by Rasouli et al. includes 346 videos between 5 and 15 seconds of duration at 30 frames per second. The data was mainly obtained in Europe (82.7%) of the total number of video clips, and the rest (17.3%) was sourced in North America. All videos were recorded with monocular cameras positioned inside the car below the rearview mirror. These 346 comprise a total of ~82,000 frames, close to 2,200 pedestrians, and nearly 337,000 total bounding boxes.

Even though these videos already give researchers a new rich dataset to perform machine learning research tasks, the uniqueness of JAAD comes from the inclusion of behavioral tags for the labeled pedestrians (654 unique pedestrians out of the ~2,200 samples) and the contextual information of the scene. More specifically, the behavioral dataset includes attributes like the state of the pedestrians before crossing, the way the pedestrian becomes aware of the vehicle action, and the response of the pedestrian to the action of the car (the ego-vehicle which recorded the scene). The dataset also includes behavioral tags for the driver (of the ego-vehicle), which include the current state: moving slow or fast, and the response: slow down or speed up. Information about the demographic of the pedestrians is also included (age, gender, size of group).

On the other hand, contextual tags include information about the scene where the video was recorded. The information consists of the configuration of the road: whether it is a street (and the number of lanes) or if it is a parking lot or a garage. It also contains information on traffic signals (stop or pedestrian), zebra crossings, and traffic lights.

Finally, the dataset also includes occlusion tags for the pedestrians: partial occlusion (between 25% and 75% visible) and complete occlusion (less than 25% visible).

While using this dataset, researchers focused on determining whether pedestrians were moving (walking) or standing and whether they performed any attention action (i.e., looking at the oncoming car). In terms of prediction, Rasouli et al. reduced the problem into an image classification problem as they were only interested in distinguishing between 4 types of actions (walking, standing, looking towards the traffic, and not looking.

Furthermore, to compensate for not having annotations for body parts (i.e., a fitted skeleton for each pedestrian), the training and testing were done by cropping the top third and bottom half from images with minor occlusions ($\geq$ 75% visible).

Researchers found that using the AlexNet architecture over the cropped images could achieve ~80% levels in average precision for classifying whether pedestrians were looking or walking. After this step, researchers obtained the 10-15 frames previous to the pedestrian's decision to cross or not, and by applying the previous classification, they reached nearly 40% average precision on the crossing / not crossing prediction. The notable works come from the addition of contextual data to aid the action data. By using these two categories of labels, researchers obtained an average precision of ~63%.

B. Intention Recognition of Pedestrians and Cyclists by 2D Pose Estimation

This paper focuses on anticipating the intentions of pedestrians and cyclists - what the researchers call "Vulnerable Road Users"(VRUs). They work on pedestrian intention estimation is based on the JAAD dataset published by Rasouli et al.[?], while the cyclists estimation is done with data collected by researchers for this experiment (I will not focus on the cyclists side of the research as it is out of scope of my project). The researchers seek to propose a better method to estimate VRUs intentions while relying only on monocular 2D-type frames - in contrast with more complex approaches that involve the use of stereo information, optical flow, or ego-motion compensation.

This work by Fang et al. is the one that most closely resembles the current methodology I am proposing for this project. In their experiment, the process involves detecting and tracking pedestrians, fitting a skeleton to each pedestrian, and then applying a cross/no-cross classifier that relies on the fitted skeleton features.

For training the C/NC classifier, Fang et al. used only samples with pedestrians which bounding boxes were at least 60 pixels wide and had no occlusions. Because they were tracking the pedestrians across the frames, these requirements had to stay valid for more than T frames. For tracks longer than T frames, they used a sliding window approach to obtain different training samples. The selected value for T = 14.

The researchers showed that using the skeleton-fitting approach could outperform models based on CNN approaches (based on AlexNet such as [?]), using values of T=14 and T=1), and obtain an accuracy of ~80%. On testing, the model was able to obtain a prediction after tracking a pedestrian over eight frames (around 250ms in the JAAD dataset).

C. Predicting Intentions of Pedestrians from 2D Skeletal Pose Sequences with a Representation-Focused Multi-Branch Deep Learning Network

In this paper, researchers took a more complex approach to estimating pedestrians' intentions and finally sought to propose a new network (Skeleton-based Pedestrian Intention network, or SPI-Net) for pedestrians' discrete intentions. The data also comes from the JAAD dataset, and it is transformed in a similar way to the work done by Fang et al. by fitting a skeleton representation to each pedestrian, this time by using the Cascaded Pyramid Network (CPN) algorithm. With this method, researchers are able to obtain 14 different key points from each skeleton.

On the network architecture, Gesnouin et al. designed it as a two-branch system. One branch focuses on Euclidean distances between the key points of the skeleton and their evolution over time, while the other one tracks the position of the key points over time in a Cartesian plane.

More specifically, the pedestrian pose sequence was obtained by using a sliding window to maintain a fixed size of a 3D tensor, where the shape is represented by (T, K, d) and where $(T = 300, K = 14, d = 2)$, for T representing a fixed number of frames, K representing the different key points of the skeleton and d representing the coordinates of each key point.

Researchers then trained an auto-encoder for the Euclidean distance branch (the Geometric Features branch) and concatenated the encoder part of the auto-encoder with the CNN network trained for the Cartesian plane branch (the Cartesian Features branch) deprived of the output layer.

With the method introduced on SPI-Net together with CPN, the experiment reached a total accuracy of 94.4% on the JAAD dataset.

D. Pedestrian Motion State Estimation from 2D Pose

This paper presents a straightforward method to estimate the intention and motion of pedestrians and other VRUs while using the same dataset as the other experiments (the JAAD dataset). The researchers proposed a method comprising three main steps: i) Using an existing pose estimation software, they obtain the 2D pose of the pedestrian (in a skeleton-type format); ii) with the key points of the skeleton obtained, they extract high-level features, static *micro motion* features, and dynamic *micro motion* features; iii) by using a seq2seq network, a model is trained to estimate probabilities of different motions states for each pedestrian using a softmax classifier.

Researchers refer to *micro motion* features as the features related to the movement of the joints (or key points) of the pedestrian-related to the torso. Static micro motions include the normalized position of key points, the normalized distance between the different key points, and the direction of different joints. As the paper notes *Dynamic micro motion features are differential calculations of static micro motion features in the time domain, mainly consisting of dynamic Euclidean distance features and dynamic angle features.*

Li et al. then divided the micro-motion features into four distinct groups: position features, distance features, static angle features, and dynamic angle features. Each of these features is then encoded separately in embedding layers and then concatenated together. For modeling the sequence data, researchers used gated recurrent units (GRUs), controlling the info that the model received at each time step.

The success of this experiment was measured against the original JAAD dataset paper [?],and the researchers note that by using their proposed method, they were able to improve the precision of the predictions by 11.6%.

III. Compare and Contrast

The most notable commonality between all these papers is the use of the same dataset to answer the question about whether a pedestrian (or a set of pedestrians) intends to cross a street or intersect with the ego-vehicle. By sharing the use of the JAAD dataset, it is relatively straightforward to compare the best approach, or set of approaches, for this task.

Because the nature of the task is the one with recognizing an intention by estimating the current pose of a pedestrian, we are dealing with a set of assumptions that still need to be validated, but interestingly, all the experiments here described have taken this assumption as given. As Gesnouin et al. describe it in their paper: *Most of those approaches are currently based on the assumption that one can link the position of a pedestrian's joints previous to his action to an intention. Consequently, the problem of a pedestrian discrete intention prediction is, therefore, dealt with as an action recognition task prior to the action, and the action label becomes the intention.*

Although the dataset is the same one for all the experiments, and it has already been noted that the main difference is the one related to the network architecture and algorithms used (which I will mention shortly), it is interesting to note the way how each of these experiments treats and manipulates the data, as well as how they selected which data is actually fit for use and which is not (mainly in the context of occlusion and whether the problem of not having 100% visibility of the pedestrian renders the frame unusable). Rasouli et al.[?] took a simple approach and cropped each bounding box in two different areas to obtain a focused image on the head area and another one on the area below the torso while accepting occlusions of up to 25%. On the other hand, the remaining papers all used some algorithm to fit a skeleton to the pedestrian, obtaining the position on the frame of the joints and other key points(between 14-16 key points). Most of them also selected only the frames where there was no occlusion at all.

Finally, on the topic of algorithms, this is where we see a range of different approaches taken to achieve the same tasks. While Rasouli et al. treated the problem as an image classification problem, Fang et al. decided to create a dual approach using the skeleton features plus a CNN architecture. On the other hand, Li et al. decided to fit the skeleton data but treated these features as two different subsets (static and dynamic micro-motions) while leveraging the use of RNNs, embeddings, and gated recurrent units. Finally, a more complex approach came from the work of Gesnouin et al., where an encoding layer was extracted from a trained auto-encoder and then merged with a CNN network stripped of its classification layer to obtain estimations of the intention of the pedestrians.

There is clear merit to each of these approaches, with the simplest ones more apt (hardware- and economic- constraints-wise) to be deployed on a large scale to AVs and other types of assisted driving systems.

IV. Future Work

The preceding papers have made several contributions, and most importantly, a new dataset published by Rasouli et al. gave these researchers rich information to pursue these pedestrian intention estimation tasks. Although while most of the experiments proposed novel ways of designing a network optimized for this task, only Rasouli et al. proposed utilizing contextual data to improve the prediction algorithm. To the best of my knowledge, all the other experiments focused on the video clips but did not take advantage of much of the behavioral and contextual attributes that the dataset contains. In this sense, it would be interesting to understand what kind of achievements could we see when some of the best algorithms presented in this literature review ,such as the one by Gesnouin et al., would include contextual data (already knowing that the paper for the original dataset increased their precision in 20 percentage points by including this information.

Furthermore, most of the datasets discarded frames and pedestrians that were at some level occluded (only Rasouli included occluded pedestrians up to 25%). Even though it makes sense from a training perspective, in real life scenarios, AVs will need to be able to make predictions over heavily occluded VRUs, as the occlusion is no excuse not to be able to perform safety maneuvers to keep the passengers of the ego-vehicle and other road participants safe.

Finally, Rasouli et al.mentioned in their paper that further work needs to be performed as dynamic factors such as velocity changes, changes in the state of the vehicle, and pedestrian's sequences of actions were not taken into consideration in their research or dataset. Also, more work needs to be performed where the demographic context of the road participants, group behavior, and weather conditions are supplied in future models.