Cross or Not: Pedestrian Behavior Prediction

David Chu dc788@cornell.edu Siyuan Hu sh2442@cornell.edu Eric Jing epj32@cornell.edu

Abstract

We aim to predict whether a pedestrian is about to cross. Previous work done in this area combined pedestrian features, global environmental cues (such as the existence of traffic lights within the video frame), and ground-truth labels about the location and the pedestrian to predict if a pedestrian will cross. Our contribution is the incorporation of local environmental cues into the prediction estimate. Motivated by the belief that objects in a pedestrian's immediate surroundings influence their decision, we improve upon the accuracy of pedestrian crossing prediction by appending those detections to the parameters for estimation.

1. Introduction

Self-driving vehicles can provide many conveniences to our every day lives. Recent years have seen a surge in research in robotics, computer vision, and machine learning in order to equip the first generation of self-driving vehicles with the ability to drive on its own and react to any abrupt changes. [8]

In order for an autonomous vehicle to maneuver itself safely, it must be able to avoid collisions while arriving at its destination within at a reasonable velocity. Out of context, a pedestrian's behavior appears unpredictable; the machine cannot tell, based on a pedestrian's appearance alone, whether he or she is about to cross. This creates a dangerous scenario; if a self driving car sees a pedestrian and does not know if he or she will cross, should the car stop?

Many factors play into the decision of a pedestrian to cross; one of the most significant influences, according to [8], is environmental context. In other words, a pedestrian's surroundings impacts whether or not he or she will cross.

We aim to improve the accuracy of predictions for pedestrian-crossing events. We will do so by incorporating the presence of objects in a pedestrian's vicinity into the predictor.

2. Related Works

Many attempts have been made in recent years to improve the accuracy for pedestrian prediction: given footage from a dashcam, predict whether he or she will cross.

2.1. Context-based Pedestrian Path Prediction

Before the neural network approach became popular, a common context-based approach called Dynamic Bayesian Network was used to predict pedestrian path trajectory [5]. The model incorporates the pedestrian situational awareness, situation criticality, and spatial layout of the environment, as latent states on top of a Switching Linear Dynamical System (SLDS) to anticipate changes in the pedestrian dynamics. Using computer vision, situational awareness is assessed by head orientation, situation criticality by the distance between vehicle and pedestrian at the expected point of closest approach, and spatial layout by the distance of the pedestrian to the curbside. In their experiments using stereo vision data obtained from a vehicle, they demonstrate that the proposed approach results in more accurate path prediction than only SLDS, at a one-second time horizon.

2.2. Odometry-based Prediction

The task can be modified to predicting the future location of a pedestrian [1]. By changing the binary cross-or-not task to a continuous prediction of location, the authors of [1] can create heatmaps of predicted locations and increase the available training data since annotation is unnecessary. However, since the vehicle's movement also influences the pedestrian's relative position in frame, the authors had to provide odometry (steering angle and velocity) of the vehicle in previous frames, and predict future odometry in parallel. Errors emerged when the car stop abruptly, and past and future odometry become uncorrelated.

A self-driving car determines its own odometry. A prediction of location in 3D space would avoid translation in the camera frame when the car moves, but 3D datasets are limited. The neural network in [1] may have learned where a pedestrian would be based on the driver's reaction, reflected in odometry. Therefore, we opted not to consider methods presented in [1]. This eliminates the possibility of unsupervised learning and predicting the exact position of a pedestrian, since that is influenced by the driver's movement. An ideal dataset would instead provide labels for whether a pedestrian will cross, independent of position within the camera frame.

2.3. JAAD

Joint Attention on Autonomous Driving [4] is a dataset that contains timelines detailing when each pedestrian is crossing the road within each video. A baseline has been created in [3], combining ground truth annotations in the dataset with pedestrian features extracted with AlexNet to obtain 62.73 ± 13.16 percent accuracy. The authors include ground-truth labels such as *looking*, *walking*, and global *street parameters*, such as the width of the road, into the input of the model. Our model differs in that it does not require any ground-truth annotations for prediction. We aim to incorporate local environmental cues and surpass the baseline accuracy.

3. Datasets

Our model needs to extract feature descriptors of each pedestrian and the objects in each pedestrian's vicinity. We surmised that the presence of select objects closeby to the pedestrian would influence his or her decision to cross. We analyzed cues that may be detected on roads for which annotation were provided in existing datasets. This included road surfaces (such as lane markings), vehicles, and miscellaneous objects that may appear in a traffic scene (such as traffic cones, poles).

The objects we deemed influential to a pedestrian's decision to cross are vehicles, traffic lights, traffic signs, roads, sidewalks, and crosswalks.

3.1. Traffic lights and vehicles

Microsoft's COCO dataset [9] contains segmentations for everyday objects, including traffic lights and multiple vehicle classes. We combined the following into one "vehicle" class: *car, motorcycle, airplane, bus, train, truck, boat*.

3.2. Traffic signs

American Stop signs is also a category included within COCO. However, the majority of JAAD [4] videos were sampled from non-American countries; out of 357 video clips, 291 were from Ukraine, 55 from Canada, 6 from Germany, and 5 from the United States. This meant that any model trained on American stop signs from COCO would not be able to detect signs within videos from JAAD, which differed based on the country. Instead, we used Cityscapes [7], a popular dataset for automonous vehicles that includes traffic signs collected in Germany.

3.3. Roads and sidewalks

Both Cityscapes and Berkeley Deep Drive (BDD) dataset provide masks for roads and sidewalks that are used to train our model. Cityscapes contains high quality pixel-level annotations of 5,000 frames and 20,000 weakly annotated frames. BDD has over 10,000 diverse images with pixel-level and rich instance-level annotations.

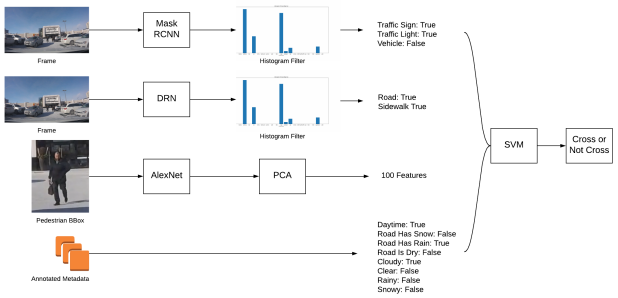

Figure 1. Pipeline of networks combining AlexNet features, Mask R-CNN segmentations, DRN segmentations, and ground truth annotations into SVM input.

3.4. Crosswalks

Mapillary Vistas is a dataset that comprises 20,000 images, out of which 18,000 are used for training and the remaining 2,000 for validation. They provide pixel-wise polygon annotations for 66 object classes, including crosswalks. We planned on training a detector for crosswalks; however, only 500 images out of 18,000 training images contained crosswalks, too few to create a reliable model.

3.5. JAAD

JAAD consists of around 300 annotated clips of dashcam footage. Annotations include per-frame labels of the current scene, such as weather, the presence of traffic signs, and ambient locations such as parking lots. There are 88,000 pedestrian bounding boxes, each one annotated with age, clothing, and carried items. Each pedestrian is uniquely identified throughout a single video clip. The dataset also includes timelines for other pedestrian actions, such as *looking* or *slowing down*. Actions are labeled with start and end frames and can overlap each other.

4. Models

Our model is a pipeline (see Figure 1) of CNNs that detect the presence of local environmental cues and pedestrian attributes in parallel, before coalescing these features as inputs to a SVM that outputs a binary prediction of whether the pedestrian will cross.

4.1. Mask R-CNN

Mask R-CNN [6] was designed to create instance segmentations of objects within an image. Our model is based off Facebook Research's at `https://github.com/facebookresearch/maskrcnn-benchmark`. To

Figure 2. Annotated frame of `video_0067.mp4` from JAAD. Mask R-CNN model was trained on Cityscapes.

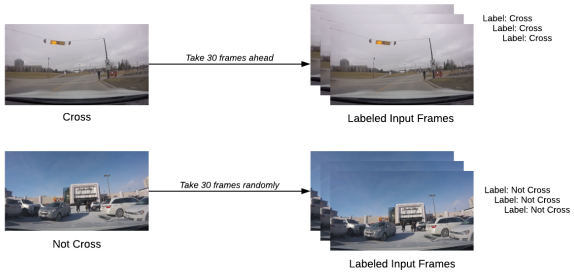

Figure 3. Selection of 30 frames before crossing/non-crossing event.

detect traffic lights and vehicles, we used their model, which was based on ResNet-101, pretrained on COCO.

We trained a separate model, also based on ResNet-101, pretrained on ImageNet, to detect traffic signs. In order to train the model, we first converted the formatting files from Cityscapes format to COCO format, creating bounding boxes and calculating areas of the segmentation masks provided by Cityscapes. Since Cityscapes did not provide segmentation masks for its testing set, we had to create our own 80-10-10 dataset split. The neural net obtained 73.6% average precision on the test data after training. See Figure 2 for an example of an annotated image.

The code is available here: `https://github.com/davidchuyaya/maskrcnn-benchmark`.

4.2. Diluted Residual Networks

Dilated Residual Networks (DRNs) are shown in [2] to have outperformed their non-dilated counterparts in image classification without increasing the models depth or complexity. We used the predefined architecture DRN-D-22 to detect large objects such as roads and sidewalks.

The code is available here: `https://github.com/kelvinhu9988/pedestrian-prediction`.

4.3. AlexNet

To reproduce the global scene cue detector in [3], we implemented a FCN AlexNet as a baseline. The network took in 1920×1080 video frames and outputs a vector, each entry encoding the presence of scene cues. The modification involved replacing the fully-connected layers of the original AlexNet architecture with a global max-pooling layer, followed by the output layer with a linear activation. This allowed AlexNet to work on input images of any size, which is vital to detecting small scene cues in a high-resolution video frame.

The ground-truth images and scene tags were split into 90-10 train-test partitions. The model converged in 5 epochs, and it achieved performance comparable to the baseline networks used in the original paper. It struggled with detecting small elements, such as traffic signs and traffic lights. This could be because this implementation, like [3], processes entire frames at once, forgoing anything more sophisticated like scanning through sub-images. One solution would be to have the second-to-last convolutional layer of the network output a tensor with the same number of channels as the number of classes to detect, and then only max-pooling without adding another fully-connected output layer. In this way, the network would still be fully convolutional, but scene detection is unaffected by the position or size of features.

4.4. Feature transformation

To integrate JAAD into our data pipeline, we processed the data into a format more suited for training predictors for pedestrian crossing events. For every pedestrian that crossed a street in a video clip, 30 frames were chosen such that all of them were at least 30 frames before the crossing starts (see Figure 3). The pedestrian's bounding box image was evaluated by ResNet50 with weights pretrained on ImageNet, producing a 2048-dimensional feature vector. To reduce the dimensions, PCA was performed on these feature vectors to produce 100-dimensional vectors.

4.5. Hyperparameters

For each pedestrian, we defined "vicinity" as a square centered at the pedestrian's bounding box's center, with edges of length $2 \times height$. The square is trimmed off at the edges of the image. The edge length was chosen assuming that all pedestrians were taller than they were wide. We chose a square sized proportionally to the pedestrian since those closer to the dashcam will appear larger, and the items in their vicinity will appear farther away. In contrast, those farther away from the dashcam will have objects in their vicinity appear closer. A square too large would capture objects inconsequential to the pedestrian, while one too small might fail to include any local cues.

Any instance segmentation whose number of pixels captured within the square exceeded 1% would have its binary label set to 1; otherwise it would be set to 0. This threshold was chosen to eliminate noisy segmentations. Again, a

Method	Accuracy
JAAD paper baseline	$62.73 \pm 13.16\%$
AlexNet + attributes	$69.85 \pm 5.92\%$
AlexNet + DRN + Mask R-CNN + attributes	$70.06 \pm 6.72\%$
AlexNet	$53.98 \pm 6.03\%$
AlexNet + DRN + Mask R-CNN	**61.83** $\pm 5.86\%$

Table 1. Evaulation results. *Attributes* refer to ground truth annotations about the pedestrian, global environmental cues, and whether the pedestrian is standing still or walking.

threshold that is too large would ignore smaller local cues, while a threshold that is too small would include misclassified objects. Therefore we decided to make the threshold proportional to the size of the square.

5. Evaluation and discussion

Our pipeline begins with an image containing a pedestrian and his or her bounding box. We then use AlexNet to extract pedestrian features, Mask R-CNN to find traffic lights, signs, and vehicles, and DRN to find roads and sidewalks. We then process the segmentations and preserve only those within the vicinity of the pedestrian and account for more than the threshold% of pixels. We feed our input, consisting of 100 dimensions from AlexNet, 3 from Mask R-CNN (binary labels), and 2 from DRN, into an SVM with L1 loss.

We sampled from 310 pedestrians, 196 of which crossed and 114 did not. This is based on [3], which sampled from 315 pedestrians, 234 of which crossed and 81 did not. Slight numerical adjustments were required since we did not want the model to be biased towards "crossing."

Due to the low number of videos available from JAAD, we determined that a SVM would be a suitable classifier. For both PCA and the SVM, we set `random_state=0` so that our results would be replicatable. Table 1 illustrates our results after averaging on 5-fold cross validation, with the exception of `JAAD paper baseline` from [3], where evaluation methods are unclear. Since we did not wish to train and validate on the same videos, each fold contains roughly the same number of pedestrians, with no pedestrians from the same video in any 2 folds.

`JAAD paper baseline` theoretically uses the pedestrian features detected by AlexNet and ground truth attributes, similar to our `AlexNet + attributes` benchmark, but we obtain a higher accuracy. This may be caused by multiple factors; the paper never clearly explains how it reduces its AlexNet features and ground truth attributes into inputs for the SVM, only stating that it obtained a 121 dimension input vector. The paper also does not elaborate on its evaluation methodology.

We found through the ablation study that our local environmental cues, when added alongside ground truth annotations, offering a statistically insignificant 0.26% increase in accuracy. However, ground truth annotations do not exist in the real world, so the more important contributions lie in predictions based on the given image without annotations.

When given only the image, we find that the detection of local environmental cues increases accuracy by 7.85%. This appears to confirm our hypothesis that a pedestrian's decision to cross is influenced by the presence of local objects.

5.1. Discussion

Due to our limited sample size, there are hyperparameters (such as the size of a pedestrian's "vicinity") that we decided not to tune, fearing overfitting.

Only 8 traffic lights were ever detected in the vicinity of a pedestrian, causing 3 of the 5 SVMs to give it no weight.

We did not have time to compute the contribution of each environmental cue to the accuracy.

Further work can be done, given a larger dataset, on analysis of what environmental objects are the most influential to a pedestrian. We used binary labels to signal presence of an object in the scene, but a better label would be the exact location, size, or relative position to the pedestrian or his/her gaze.

6. Conclusion

The presence of traffic signs, traffic lights, vehicles, roads, and sidewalks can be used to improve pedestrian prediction in the absence of ground truth annotations.

References

[1] B. S. Apratim Bhattacharyya, Mario Fritz. Long-term on-board prediction of people in traffic scenes under uncertainty, 2016.

[2] T. F. Fisher Yu, Vladlen Koltun. Dilated residual networks, 2017.

[3] A. R. Iuliia Kotseruba and J. K. Tsotsos. Are they going to cross? a benchmark dataset and baseline for pedestrian crosswalk behavior, 2017.

[4] A. R. Iuliia Kotseruba and J. K. Tsotsos. Joint attention in autonomous driving (jaad), 2017.

[5] N. S. F. F. D. M. G. Julian Francisco, Pieter Kooij. Context-based pedestrian path prediction, 2014.

[6] P. D. R. G. Kaiming He, Georgia Gkioxari. Mask r-cnn, 2018.

[7] S. R. T. R. M. E. R. B. U. F. S. R. B. S. Marius Cordts, Mohamed Omran. The cityscapes dataset for semantic urban scene understanding, 2016.

[8] A. Rasouli and J. K. Tsotsos. Joint attention in driver-pedestrian interaction: from theory to practice, 2018.

[9] S. B. L. B. R. G. J. H. P. P. D. R. C. L. Z. P. D. Tsung-Yi Lin, Michael Maire. Microsoft coco: Common objects in context, 2015.