Are They Going to Cross? A Benchmark Dataset and Baseline for Pedestrian Crosswalk Behavior

Amir Rasouli, Iuliia Kotseruba and John K. Tsotsos
York University, Toronto, Ontario, Canada
{aras,yulia_k,tsotsos}@eecs.yorku.ca

Abstract

Designing autonomous vehicles suitable for urban environments remains an unresolved problem. One of the major dilemmas faced by autonomous cars is how to understand the intention of other road users and communicate with them. The existing datasets do not provide the necessary means for such higher level analysis of traffic scenes. With this in mind, we introduce a novel dataset which in addition to providing the bounding box information for pedestrian detection, also includes the behavioral and contextual annotations for the scenes. This allows combining visual and semantic information for better understanding of pedestrians' intentions in various traffic scenarios. We establish baseline approaches for analyzing the data and show that combining visual and contextual information can improve prediction of pedestrian intention at the point of crossing by at least 20%.

1. Introduction

Visual perception and scene understanding are key components in autonomous driving. Tasks such as road detection and following [24], pedestrian [42] and car detection [40] have been extensively researched in the past decades and shown to be essential in designing robust and safe systems.

In typical traffic scenarios, there are numerous factors that influence the behavior of road users. Aside from official rules that govern traffic flow, which can be accommodated by observing signs, signals or position of the other road users, traffic participants often engage in some form of non-verbal communication to resolve ambiguous situations, such as yielding, taking the right of way, or even crossing the street [39].

The importance of such interactions between road users is evident in the recent report on Google's self-driving car [1] indicating that 90% of the failures occur in busy streets out of which 10% is due to the incorrect prediction of traffic participants behavior. Among the problems reported in recent years are the failure to estimate other vehicle's movement [33] or inability to respond to unexpected behaviors of other drivers [19]. This is a concern for pedestrian safety. For example, a recent study [32] shows that pedestrians, in particular the ones at crosswalks, have the highest risk of being the victim of perceptual discrepancies.

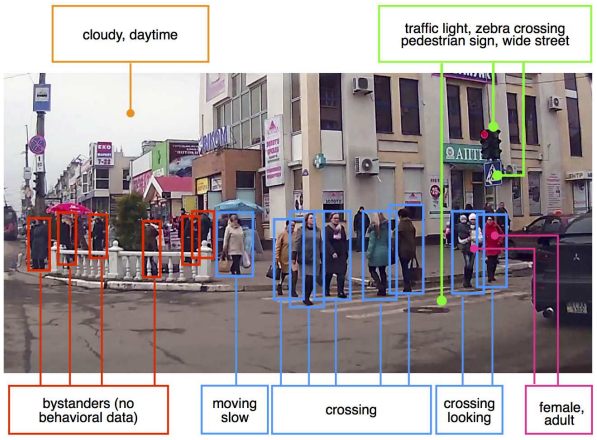


Figure 1: Example of annotations provided in the dataset: bounding boxes for all pedestrians, behavioral labels, gender and age for pedestrians crossing or intending to cross, contextual tags (weather, time of the day, street structure)

The common approach to remedy the problem of predicting road users behavior is to employ dynamic factors, such as trajectory [3] or velocity [30], or the expected final goal of pedestrians [18, 31]. More recently researchers also looked at other behavioral cues such as pedestrian head orientation to measure the level of awareness at the point of crossing [21, 35]. These studies, however, are limited in scope and look at very few contextual elements to predict the behavior of pedestrians. In practice, there are other factors, in addition to spatiotemporal ones, that can influence the crossing behavior of a pedestrian including the structure of the crosswalk (e.g. sign, delineation) [38], environmental factors (e.g. weather condition, visibility) or individual characteristics of pedestrians (e.g. demographics) [39].

To provide a testbed and facilitate further studies in the field of pedestrian behavior analysis, we make the following contributions: 1) We introduce a novel dataset called Joint Attention in Autonomous Driving (JAAD)[1] that enables behavioral analysis of pedestrians at the point of crossing. Compared to existing large-scale datasets such as KITTI [13] and Caltech pedestrian dataset [8], in addition to ground truth for all pedestrians in the scene and occlusion information, our dataset contains behavioral tags describing actions of pedestrians intending to cross. Each frame also includes contextual information such as the type of the crosswalk, traffic signs and road type as well as weather conditions and time of the day. 2) We examine baseline approaches using Convolutional Neural Networks (CNNs) to detect and analyze the context of the scenes as well as pedestrians' behavioral cues. 3) Finally, we combine the best performing baseline approaches using a linear classifier to evaluate how well the crossing behavior of pedestrians can be predicted.

2. Why context matters

In the computer vision community human activity recognition has been extensively studied [26, 23, 17, 6, 41]. These works either focus on recognizing the basic activity of a subject such as walking, running or jumping [41] or understanding the underlying semantics of a certain activity, e.g. standing in a queue [6], performing an offensive action in a sports game [17, 23] or grasping certain objects [26]. To resolve the latter problem, the context in which the activity is perceived is also considered. The relevance of a contextual element to understanding a certain activity may vary depending on the task. For instance, in a sports scene to analyze a certain game strategy the current state of all subjects in the scene [23] as well as the spatiotemporal context that resulted in the perceived behavior [17] are important.

Understanding pedestrian activities and predicting their behavior in traffic scenes is no exception and requires higher level reasoning by taking into account various contextual elements. Here the context can either be target specific such as demographics (e.g. gender and age) or culture specific attributes [39], spatiotemporal (e.g. trajectory or velocity) [25] or environmental context (e.g. signs, delineation) [38].

A number of recent works have attempted to capture some of the above contextual elements to predict pedestrian behavior. For instance, trajectory [3] or head orientation of pedestrians [21] are used to predict whether they will cross. In these works, however, the physical context is known (streets with no designated crossing) and is not linked to the perceived behavior (e.g. the pedestrian is looking towards the approaching car). The pedestrian behavior is also limited to looking and non-looking actions.

[1]The JAAD dataset is available at http://data.nvision2.eecs.yorku.ca/JAAD_dataset/

Dataset	# ped. samples	# frames	occ. labels	temporal corr.	video sequence	behavioral data	context data	weather variation
INRIA[4]	1.8k	2.5k						
Daimler[10]	72k	28.5k			x			
Caltech[8]	347k	250k	x	x	x			
KITTI [13]	12k	80k	x		x			
MPD [16]	86.2k	95k	x	x	x			
Our Dataset	337k	82k	x	x	x	x	x	x

Table 1: Comparison of our dataset with existing large-scale pedestrian datasets. The last six columns indicate the properties of each dataset.

There are a number of large-scale datasets publicly available that can be potentially used for pedestrian behavior understanding [13, 8, 10]. However, the majority of these datasets are designed for the purpose of pedestrian detection and only provide bounding boxes for pedestrians. Few exceptions, such as KITTI [13], also provide optical flow and stereo information for mapping and localization.

Some small-scale datasets are also available that are generated for pedestrian path prediction [34], [21]. The one of interest is from Daimler [21], which, in addition to dynamic factors, also takes into account context in the form of the degree of pedestrians' head orientation. This dataset, however, only contains 58 sequences that are collected from 4 pedestrian participants who were instructed to perform certain behaviors during the recording (i.e. the dataset is not naturalistic). This feature limits the scope of the dataset as "in the wild" much more varied behaviors may occur during the course of crossing. In addition, in the Daimler dataset, context only refers to the pedestrians head orientation, the curbside location and distance and trajectory of the pedestrians. Apart from head orientation there are no other behavioral tags reflecting the actions of the driver or the pedestrian. In terms of environmental context, the videos are only recorded in streets without any traffic signals or zebra crossing and no ground truth information is available for characteristics of the scene (both for the pedestrians and environments).

We are introducing a novel dataset to facilitate the study of traffic scene analysis and pedestrians' behavior understanding. Since most of the interactions with pedestrians, from the perspective of an autonomous car, are at the point of crossing, in our dataset we particularly focus on various crossing scenarios. To the best of our knowledge, this is the first publicly available large-scale dataset that combines pedestrian detection with behavioral and contextual information. Table 1 summarizes the differences between our dataset and some of the state-of-the-art datasets.

Camera Model	resolution	FOV	# clips
Garmin GDR-35	1920x1080	110^o	276
GoPro Hero+	1920x1080	170^o	60
Highscreen BB Connect	1280x720	100^o	10

Table 2: The properties of the cameras used to collect the data and corresponding number of video clips captured with each one.

3. The dataset

3.1. Data collection

There are over 300 video clips in our dataset ranging from 5 to 15 seconds in duration. The data is collected in North America (60 clips) and Europe (286 clips) using three different high-resolution monocular cameras (see Table 2). The cameras were positioned inside the car below the rear view mirror. The frame rate of the videos is 30 fps.

3.2. Ground truth

Our dataset comes with three types of ground truth annotations: bounding boxes for detection and tracking of pedestrians, behavioral tags indicating the state of the pedestrians and scene annotations listing the environmental contextual elements (shown in Figure 1).

Bounding boxes: the dataset contains approximately 82k frames and 2.2k unique pedestrian samples comprising a total of 337k bounding boxes. The bounding boxes and annotations are done using Piotr's annotation toolbox [7]. We used 3 types of ids: *pedestrian* to identify the pedestrians with behavioral tags (i.e. the ones demonstrating the intention of crossing or located close to the curb), *ped* for all other individual pedestrians in the scene (bystanders) and *people* for groups of pedestrians, where individual people are hard to distinguish. All identified pedestrians are assigned a unique id such as *pedestrian1, pedestrian2, ped1, ped2, people1, people2*, etc. This form of annotation enables tracking of pedestrians throughout a sequence.

Occlusion information is provided in the form of tags for each bounding box: partial occlusion (between 25 and 75% visible) and full occlusion (less than 25% visible).

Behavioral tags: there are 654 unique pedestrian samples (out of 2.2k samples) with behavioral tags in the dataset. The behavioral data is created using the BORIS software [12]. The behavioral information captures the type and duration of pedestrians' actions and also the actions of the driver (i.e. the car with the recording camera). The pedestrians' actions are categorized into 3 groups: *Precondition*- this refers to the state of the pedestrian prior to crossing and can be either standing, moving slow or fast. *Attention*- the way the pedestrian becomes aware of the approaching vehicle. These actions, depending on their duration, are looking ($> 1s$) or glancing ($\leq 1s$). *Response*- this includes the behaviors that pedestrians exhibit in response to the action of the approaching vehicle, namely, stop, clear path, slow down, speed up, hand gesture and nod.

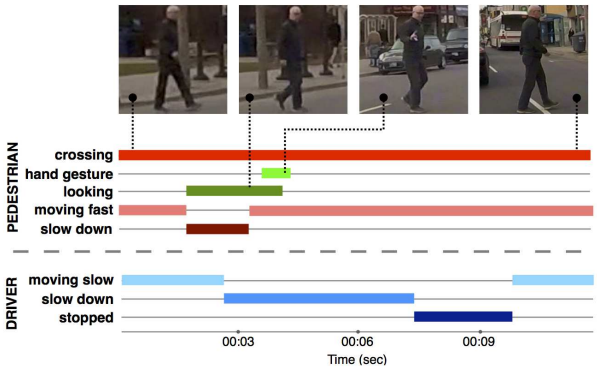


Figure 2: Behavioral labels with timestamps to represent the sequence of the events that took place during the crossing and observe the correspondence between the actions of the driver and the pedestrian.

In addition to individual behavioral information, complementary tags are also included that reflect the demographics of the pedestrians such as gender, age (adult, child or elderly), and the size of the group each individual pedestrian is associated with.

The driver's behavioral tags capture the state of the approaching vehicle which can be one of the following: moving slow or fast (*current state*) and slow down or speed up (*response*). Figure 2 shows an example of behavioral annotation with selected corresponding frames.

Contextual tags: each frame is assigned a contextual tag that describes the scene. There are four types of contextual information: *Configuration*- includes the number of lanes or whether it is a parking lot/garage. *Traffic signals*- refers to the presence of zebra crossing, pedestrian sign, stop sign or traffic light in the scene. *Weather*- includes sunny, cloudy, snowy or rainy. *Time of day*- this tag crudely indicates the lighting conditions and can be day, afternoon or nighttime.

3.3. Properties

Visibility- The majority of the dataset is collected during daytime, only a few videos are recorded at night. In some videos there is strong sun glare making it particularly difficult to observe the scene properly. Weather is another factor that varies significantly and changes from clear sky to heavy rain or snow. Figure 3 shows some examples of visibility variation and how challenging detecting pedestrians and analyzing their actions can be in some cases.

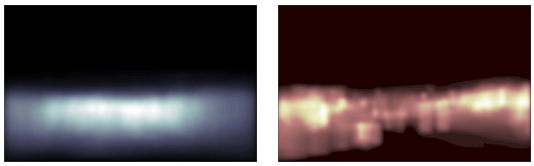
Position distribution- Figure 4 shows the distribution of pedestrians positions in the scene. As one would expect, the position of the pedestrians with behavioral data is more concentrated in the center of the images since most of them are crossing the street. Other pedestrian samples (bystanders) are more uniformly distributed on either side of the frame.

Scale-Figure 5 shows the scale changes of data for pedes-

sun glare night

snow cloudy

Figure 3: Samples of visibility changes in our dataset. These examples show how weather/lighting conditions can make the analysis of pedestrians behavior challenging.



with behavioral data without behavioral data

Figure 4: The distribution of bounding boxes around pedestrians in the scenes with and without behavioral data. The pedestrians with behavioral data appear mainly in the center of the images because they are crossing.

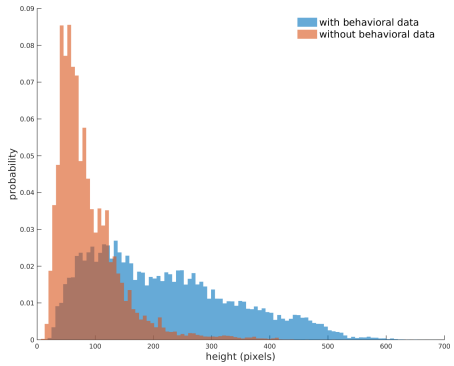


Figure 5: The scale variation of pedestrians height (in pixels) in the scenes. The pedestrians with behavioral data are generally larger in scale and closer to the camera with an average distance of 6-20m. The pedestrians without behavioral data (bystanders) are farther away from the vehicle (20-45m).

trians with and without behavioral data in terms of the height of bounding boxes. The average distance to pedestrians from the recording vehicle was between 6 to 20 meters for the ones with behavioral data and 20 to 45 m for the rest.

4. Baseline methods for scene analysis

In order to understand pedestrians' intentions to cross, we take into account two elements: static environmental context and behavior of the pedestrians. For our baseline approaches, we first identify how these elements can be extracted from the visual data using the textual annotations. Next, we examine how combining static contextual information with behavioral data can affect the prediction accuracy of pedestrian crossing.

Our data samples only have labels that list the elements presented in the scene and the behavior exhibited by the pedestrians. There is no bounding box information available for contextual data or for the position of the pedestrians body parts. Thus, our problem is an instance of weakly-supervised learning where we try, by providing sufficient amount of data, to identify the attributes of the image. This is particularly challenging for context understanding since there is a large amount of background clutter that can affect detection of certain elements.

Convolutional Neural Networks (CNNs) are commonly used to deal with weakly-supervised data and showed state-of-the-art performance in various applications [36, 29, 2]. In our experiments we apply the widely used AlexNet architecture [22] as our baseline and evaluate different variants of this architecture against our data.

4.0.1 Contextual elements

Attribute recognition in weakly-supervised data is an active topic of research [36, 43, 37]. There are two ways to recognize attributes in a weakly-supervised manner. The first is to explicitly classify attributes in the scenes [36]. The second is to implicitly infer the attributes by classifying the scenes and then identifying the attributes from the shared features [43]. Since our data mainly contains inner city street scenes, we use the former method to directly learn the attributes.

We selected seven classes that best represent the structure of a scene relevant to crossing:

Street width-*narrow* ($<$ 2 lanes) or *wide* ($\geq$ 3 lanes). The wider the street, the longer it takes the pedestrian to cross, therefore he/she would behave more conservatively.

Traffic signal- *pedestrian crossing sign*, *zebra crossing*, *stop sign* and *traffic light*. Lack of street signal does not constrain the approaching vehicles to slow down or stop, therefore pedestrians will be more conservative to cross in the presence of a vehicle. In addition, different signs may have a different level of strength in controlling the traffic. For instance, a pedestrian sign indicates yielding to pedestrians (in some instances the driver might not yield), whereas stop sign or red traffic light prohibits the driver to go any further without stopping. Most of the locations in the dataset are without the stop signs and traffic signals.

Location-*parking lot* (indicates whether the scene is in a regular street or a parking area). The driving speed in

parking lot areas is significantly slower than in streets (on average 20 km/h). Therefore, pedestrians may have a higher level of confidence to cross.

As the baseline model, we first train randomly initialized AlexNet end-to-end on our dataset. Next, we examine the option of transferred learning by fine-tuning the network on our data. For this purpose we use pre-trained AlexNet on two large image datasets, ImageNet [5] and places, and both datasets combined [44].

One drawback of AlexNet is that it accepts images of size 227x227 pixels as input. This limits its ability to learn smaller objects in large scenes. One remedy to this problem is to convert AlexNet to a fully convolutional network (FCN) allowing learning of images with larger dimensions. To do so, we transform the last three fully connected layers (*fc6-fc8*) to convolution layers. We follow a similar approach as in [28] and use a global max-pooling layer at the output to generate the final class scores.

Depending on where the attributes appear in the scene, they might be captured in small or large scale images. For this reason, we also consider multi-scale learning of the scenes using two approaches. The first is the Spatial Pyramid Pooling (SPP) [15] technique which allows the max-pooling of the features from the last convolutional layer (*conv5*) at different scales. Next, we also combine two different scale models, AlexNet and FCN, and take max prediction for each class from either network.

4.0.2 Pedestrian behavior estimation

We use behavioral annotation to determine pedestrians' actions from the still frames. In particular, we focus on determining pedestrians' gait (walking/standing) and the presence of attention (looking), since moving towards the road or standing at the curb and examining the traffic are strong indicators of crossing intention [14]. Some works in the context of pedestrian safety explicitly estimate body and head orientation [11], detect early signs of crossing intention [20] or focus on path estimation [21]. Since our dataset focuses on pedestrians that are about to cross or are moving/standing sufficiently close to the road we are only interested in distinguishing between four actions, namely, walking, standing, looking (towards the traffic) and not looking. Thus, to establish a baseline we reduce this problem to image classification.

We train separate models for gait and attention estimation. In each case we train a randomly initalized AlexNet end-to-end on cropped images of pedestrians from our dataset (with minor occlusions up to 25% allowed) and then try transfer learning by fine-tuning an AlexNet pre-trained on ImageNet [27].

We also verify whether using a full body pose improves classification compared to analyzing only the cropped image of the head for orientation or lower body for gait estimation. Since our dataset does not have annotations for body parts, we simply crop the top third and the bottom half of the bounding box for attention and gait estimation respectively.

5. Evaluation and discussion

In this section we describe the experimental evaluation of our classification problem for both context and pedestrian behavior. We report on the results in terms of Average Precision (AP) for each class and mean of AP for the overall performance of each method.

5.0.1 Environment analysis

For context detection experiments we divided our total dataset into three sets of train, test and validation each containing 50%, 10% and 40% of the data respectively. In all of our experiments, the following parameters were fixed: we used Stochastic Gradient Descent (SGD) learning method with constant step learning update with γ set to 0.1. Momentum, μ, and weight decay, ω, were set to 0.9, and 0.0005 respectively.

In the original AlexNet method, the softmax loss function is used to perform classification. A particular feature of using softmax is that it normalizes the final prediction to sum to 1. This means one class eventually gets the most importance among the other classes. In our case, since multiple objects may be present in each scene, we require individual confidence estimation for each class. As a result, we choose the sigmoid cross entropy loss function instead.

We used the default parameters for the base case end-to-end training of AlexNet with the learning rate of 0.01 and batch size of 256. For fine-tuning the pre-trained models on AlexNet, we reduced the learning rate to 0.001 and the local rate of the remaining layers by a factor of 10 times. We found that in practice reducing the learning rate of convolutional layers instead of fixing their learning results in a better performance, on average up to 5%.

Fine-tuning the FCN and SPP models is similar. We set the weights the same way as regular AlexNet models. We chose the image size of 540x540 pixels for input with batch size of 32. Here a lower learning rate (0.000625) was used to accommodate the learning on a smaller batch of samples. The SPP model was evaluated with pyramid height of 2 and 3. It should also be noted that in the SPP models the *fc6* layers were learned from scratch due to the change in the dimensionality of their inputs.

The results of the experiments are summarized in Table 3. Overall, combining AlexNet with its fully convolutional counterpart achieved the best results. This is primarily due to the fact that AlexNet and FCN have complementary performance in detection of objects in different scales. Such a multi-scale detection performance, however, was not achieved using the SPP models. Overall, the performance of the SPP models was even inferior comparing to those of

Method								mAP %
	29040	17460	12817	20754	1862	3372	6576	
AlexNet	63.42	71.95	50.85	73.73	5.40	4.63	53.46	46.20
AlexNet-places	84.61	88.50	74.46	92.10	20.75	29.33	77.77	66.79
AlexNet-hybrid	83.82	91.23	78.74	92.05	21.07	29.68	74.83	67.35
AlexNet-imagenet	84.94	91.94	77.19	92.32	17.34	**42.45**	78.21	69.20
FCN-imagenet	86.82	91.02	83.20	90.75	19.62	35.13	**80.85**	69.63
SPP-imagenet-p2	83.48	91.14	77.93	89.10	**22.82**	22.24	70.74	65.35
SPP-imagenet-p3	85.80	91.76	79.42	91.49	21.36	24.21	80.34	67.77
AlexNet + FCN	**87.33**	**93.65**	**84.02**	**92.94**	17.62	42.25	79.38	**71.03**

Table 3: The performance of weakly-supervised classification models for environmental contextual elements. The results from the left to right refer to classes *narrow*, *wide*, *pedestrian crossing sign*, *zebra crossing*, *stop sign*, *traffic light*, *parking lot* and mean AP of all the classes. The first line of numerical values below the class symbols indicate the number of instances of each element in the training data.

single scale models (with exception of stop sign detection).

5.0.2 Action

Our dataset contains a total of 88K unoccluded pedestrian samples, however, their distribution between classes is uneven. For instance, there are 14K samples for people standing and 17K for people looking. In order to have balanced training data, we use all of the samples for the least represented classes and randomly select an equivalent number of negative samples. We use 60% of the data for training, 10% for validation and 30% for testing. For all experiments we used SGD learning and kept the default parameters of AlexNet unchanged: γ set to 0.1, momentum, μ, set to 0.9 and and weight decay, ω of 0.0005. For end-to-end training we set the learning rate to 0.001 and reduced it by a factor of 10 for fine-tuning.

Average precision for each model is shown in Table 4. We found that fine-tuning AlexNet gave the best results, possibly due to the fact that the network was able to leverage the existing representation for people during the training. Also using only the top and bottom part of the bounding box for attention and gait classification offered slight improvement over using the whole image. Figure 6 shows samples from the dataset with predicted action labels. Overall, all learned models were successful at recognizing typical cases such as people looking straight at the camera and profile view of people walking. In many misclassified looking samples the pedestrian's face is partially obscured by clothes or sunglasses. The biggest issue for distinguishing walking pedestrians from the ones standing is that some of the key frames from these actions may look very similar and can be disambiguated only by taking into account temporal context.

5.0.3 Crossing or not?

Finally, we examine the contribution of gait/attention and environmental context (width of the road, pedestrian crossing, signs, traffic lights, etc.) to determining pedestrians' intentions of crossing. We select from the dataset 315 instances of pedestrians approaching or standing at the curb with the following ground truth: walking/standing, looking/not looking, street parameters and whether they cross or not. For each scenario, we select 10-15 frames and corresponding bounding boxes of the pedestrians preceding their decision to cross or not to cross (omitting frames with heavy occlusion). Overall, there are 81 non-crossing and 234 crossing scenarios with the total number of 3324 frames.

Method	walking	looking
AlexNet-full	78.34	67.45
AlexNet-cropped	74.23	74.98
AlexNet-imagenet-full	80.45	75.23
AlexNet-imagenet-cropped	**83.45**	**80.23**

Table 4: The average precision (AP%) of the classification results for pedestrians' walking and looking actions.

Figure 6: Example predictions of pedestrians actions for looking/not looking (top) and walking/standing (bottom). The captions in green and red indicate correct and wrong predictions respectively.

Method	AP
Action	39.24 $\pm$16.23
Action + context	**62.73 $\pm$ 13.16**

Table 5: Prediction accuracy (%) of pedestrians' crossing. Adding the context information significantly improves the prediction results.

To get corresponding visual features for gait and attention we convert *fc6*, *fc7* and *fc8* layers of AlexNet to fully-convolutional layers and use the output of *fc8* as the compact representation of the scene and pedestrian action (such features proved useful for generic recognition tasks [9]). Given that the size of the input image is 227x227 pixels, each CNN for gait and attention produces a 4D feature vector. For environmental elements classification, we use the output of the last layer of the fully-convolutional model (FCN-imagenet). Since each input image is resized to 540x540 pixels, the resulting feature vector for scene is 121D. Next, we train a linear SVM model to classify instances of crossing or not crossing based on gait/attention information with and without the context. We perform a 10-fold cross-validation and report the mean accuracy and 95% confidence interval of the accuracy estimate in Table 5.

Classification using only the attention/gait information can correctly predict approximately 40% of crossing behavior observed, however, adding context, such as the width of the street and the presence of the designated crossing, improves the classification by 20%.

It should be noted that environmental contextual elements, e.g. zebra crossing or pedestrian sign, are not the only contributing factors in predicting pedestrian crossing. There are also dynamic factors, such as reaction from the driver (slowing down or maintaining the speed), which also have a significant effect. For example, pedestrians are more likely to take the right of way at the designated crossing assuming that the driver would yield. On the other hand, pedestrians at non-designated locations usually wait for the explicit driver's reaction or large enough gap between vehicles before making a decision to cross.

6. Conclusion

We introduced a novel dataset to facilitate research on traffic scene understanding, in particular, pedestrian behavioral analysis. Our dataset combines localization information of pedestrians in the scenes with their behavioral data and contextual information allowing for higher-level reasoning about their actions.

We showed that relying only on low-level behavioral information such as looking or walking does not suffice for reliable prediction of pedestrian crossing action. However, combining behavioral clues with contextual elements such as scene structure can significantly increase crossing prediction accuracy.

While environmental contextual information improves understanding of crossing behavior, it is still not sufficient for robust and reliable behavioral prediction in practical traffic scenarios. There are also dynamic factors such as velocity changes, changes in the state of the vehicle and pedestrians' sequences of actions that should be taken into consideration. Furthermore, the effects of demographic context (e.g. age and gender), group behavior and ambient weather conditions need to be investigated.

References

[1] Google Self-Driving Car Testing Report on Disengagements of Autonomous Mode. Online, Dec. 2015. Accessed: 2017-03-05.

[2] R. Cabral, F. De la Torre, J. P. Costeira, and A. Bernardino. Matrix completion for weakly-supervised multi-label image classification. *PAMI*, 37(1):121–135, 2015.

[3] W. Choi and S. Savarese. Understanding collective activitiesof people from videos. *PAMI*, 36(6):1242–1257, 2014.

[4] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, pages 886–893, 2005.

[5] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009.

[6] Z. Deng, A. Vahdat, H. Hu, and G. Mori. Structure inference machines: Recurrent neural networks for analyzing relations in group activity recognition. In *CVPR*, pages 4772–4781, 2016.

[7] P. Dollár. Piotr's Computer Vision Matlab Toolbox (PMT). https://github.com/pdollar/toolbox.

[8] P. Dollár, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: A benchmark. In *CVPR*, pages 304–311, 2009.

[9] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. *arXiv preprint arXiv:1310.1531*, 2013.

[10] M. Enzweiler and D. M. Gavrila. Monocular pedestrian detection: Survey and experiments. *PAMI*, 31(12):2179–2195, 2009.

[11] F. Flohr, M. Dumitru-Guzu, J. F. Kooij, and D. M. Gavrila. A probabilistic framework for joint pedestrian head and body orientation estimation. *IEEE Transactions on Intelligent Transportation Systems (ITSC)*, 16(4):1872–1882, 2015.

[12] O. Friard and M. Gamba. Boris: a free, versatile open-source event-logging software for video/audio coding and live observations. *Methods in Ecology and Evolution*, 7(11):1325–1330, 2016.

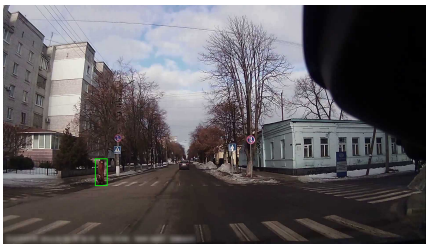
[13] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013.

[14] N. Guéguen, S. Meineri, and C. Eyssartier. A pedestrians stare and drivers stopping behavior: A field experiment at the pedestrian crossing. *Safety science*, 75:87–89, 2015.

[15] K. He, X. Zhang, S. Ren, and J. Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. In *ECCV*, pages 346–361, 2014.

[16] S. Hwang, J. Park, N. Kim, Y. Choi, and I. So Kweon. Multispectral pedestrian detection: Benchmark dataset and baseline. In *CVPR*, pages 1037–1045, 2015.

[17] M. S. Ibrahim, S. Muralidharan, Z. Deng, A. Vahdat, and G. Mori. A hierarchical deep temporal model for group activity recognition. In *CVPR*, pages 1971–1980, 2016.



context: looking, standing, narrow, zebra crossing ped. sign
prediction: crossing

context: looking, walking, narrow, ped. sign
prediction: crossing

context: looking, standing, wide, zebra crossing
prediction: not crossing

Figure 7: Example predictions of pedestrians actions for looking/not looking (top) and walking/standing (bottom). The captions and bounding boxes in green and red indicate correct and wrong predictions respectively.

[18] V. Karasev, A. Ayvaci, B. Heisele, and S. Soatto. Intent-aware long-term prediction of pedestrian motion. In *ICRA*, pages 2543–2549, 2016.

[19] W. Knight. Driverless Cars Are Further Away Than You Think. Online, Oct. 2013. Accessed: 2017-03-05.

[20] S. Koehler, M. Goldhammer, S. Bauer, S. Zecha, K. Doll, U. Brunsmann, and K. Dietmayer. Stationary detection of the pedestrian's intention at intersections. *IEEE Intelligent Transportation Systems Magazine*, 5(4):87–99, 2013.

[21] J. F. P. Kooij, N. Schneider, F. Flohr, and D. M. Gavrila. Context-based pedestrian path prediction. In *ECCV*, pages 618–633, 2014.

[22] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, pages 1097–1105, 2012.

[23] T. Lan, L. Sigal, and G. Mori. Social roles in hierarchical models for human activity recognition. In *CVPR*, pages 1354–1361, 2012.

[24] D. Levi, N. Garnett, E. Fetaya, and I. Herzlyia. Stixelnet: A deep convolutional network for obstacle detection and road segmentation. In *BMVC*, pages 109–1, 2015.

[25] N. Lubbe and J. Davidsson. Drivers comfort boundaries in pedestrian crossings: A study in driver braking characteristics as a function of pedestrian walking speed. *Safety Science*, 75:100–106, 2015.

[26] M. Ma, H. Fan, and K. M. Kitani. Going deeper into first-person activity recognition. In *CVPR*, pages 1894–1903, 2016.

[27] M. Oquab, L. Bottou, I. Laptev, and J. Sivic. Learning and transferring mid-level image representations using convolutional neural networks. In *CVPR*, pages 1717–1724, 2014.

[28] M. Oquab, L. Bottou, I. Laptev, J. Sivic, et al. Weakly supervised object recognition with convolutional neural networks. In *NIPS*, 2014.

[29] G. Papandreou, L.-C. Chen, K. P. Murphy, and A. L. Yuille. Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation. In *ICCV*, pages 1742–1750, 2015.

[30] S. Pellegrini, A. Ess, K. Schindler, and L. Van Gool. You'll never walk alone: Modeling social behavior for multi-target tracking. In *ICCV*, pages 261–268, 2009.

[31] E. Rehder and H. Kloeden. Goal-directed pedestrian prediction. In *ICCVW*, pages 50–58, 2015.

[32] Z. Ren, X. Jiang, and W. Wang. Analysis of the influence of pedestrians eye contact on drivers comfort boundary during the crossing conflict. *Procedia Engineering*, 137:399–406, 2016.

[33] M. Richtel and C. Dougherty. Googles driverless cars run into problem: Cars with drivers. *New York Times*, 1, 2015.

[34] N. Schneider and D. M. Gavrila. Pedestrian path prediction with recursive bayesian filters: A comparative study. In *German Conference on Pattern Recognition*, pages 174–183, 2013.

[35] A. T. Schulz and R. Stiefelhagen. Pedestrian intention recognition using latent-dynamic conditional random fields. In *Intelligent Vehicles Symposium (IV)*, pages 622–627, 2015.

[36] S. Shankar, V. K. Garg, and R. Cipolla. Deep-carving: Discovering visual attributes by carving deep neural nets. In *CVPR*, pages 3403–3412, 2015.

[37] J. M. Susskind, A. K. Anderson, G. E. Hinton, and J. R. Movellan. *Generating facial expressions with deep belief nets*. 2008.

[38] A. Tom and M.-A. Granié. Gender differences in pedestrian rule compliance and visual search at signalized and unsignalized crossroads. *Accident Analysis & Prevention*, 43(5):1794–1801, 2011.

[39] I. Wolf. The interaction between humans and autonomous agents. In *Autonomous Driving*, pages 103–124. 2016.

[40] T. Wu, B. Li, and S.-C. Zhu. Learning and-or model to represent context and occlusion for car detection and viewpoint estimation. *PAMI*, 38(9):1829–1843, 2016.

[41] L. Xia and J. Aggarwal. Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera. In *CVPR*, pages 2834–2841, 2013.

[42] L. Zhang, L. Lin, X. Liang, and K. He. Is faster r-cnn doing well for pedestrian detection? In *ECCV*, pages 443–457, 2016.

[43] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Object detectors emerge in deep scene cnns. *arXiv preprint arXiv:1412.6856*, 2014.

[44] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning deep features for scene recognition using places database. In *NIPS*, pages 487–495, 2014.