Toward Driving Scene Understanding: A Dataset for Learning Driver Behavior and Causal Reasoning

Vasili Ramanishka[1] vram@bu.edu
Yi-Ting Chen[2] ychen@honda-ri.com
Teruhisa Misu[2] tmisu@honda-ri.com
Kate Saenko[1] saenko@bu.edu

[1]Boston University, [2]Honda Research Institute, USA

Abstract

Driving Scene understanding is a key ingredient for intelligent transportation systems. To achieve systems that can operate in a complex physical and social environment, they need to understand and learn how humans drive and interact with traffic scenes. We present the Honda Research Institute Driving Dataset (HDD), a challenging dataset to enable research on learning driver behavior in real-life environments. The dataset includes 104 hours of real human driving in the San Francisco Bay Area collected using an instrumented vehicle equipped with different sensors. We provide a detailed analysis of HDD with a comparison to other driving datasets. A novel annotation methodology is introduced to enable research on driver behavior understanding from untrimmed data sequences. As the first step, baseline algorithms for driver behavior detection are trained and tested to demonstrate the feasibility of the proposed task.

1. Introduction

Driving involves different levels of scene understanding and decision making, ranging from detection and tracking of traffic participants, localization, scene recognition, risk assessment based on prediction and causal reasoning, to interaction. The performance of visual scene recognition tasks has been significantly boosted by recent advances of deep learning algorithms [27, 9, 35, 8], and an increasing number of benchmark datasets [6, 21]. However, to achieve an intelligent transportation system, we need a higher level understanding.

Different vision-based datasets for autonomous driving [7, 12, 22, 28, 4, 34, 23] have been introduced and push forward the development of core algorithmic components. In core computer vision tasks, we have witnessed significant advances in object detection and semantic segmentation because of large scale annotated datasets [6, 7, 4]. Additionally, the Oxford RobotCar Dataset [22] addresses the challenges of robust localization and mapping under significantly different weather and lighting conditions. However, these datasets do not address many of the challenges in the higher level driving scene understanding. We believe detecting traffic participants and parsing scenes into the corresponding semantic categories is only the first step. Toward a complete driving scene understanding, we need to understand the interactions between human driver behaviors and the corresponding traffic scene situations [29].

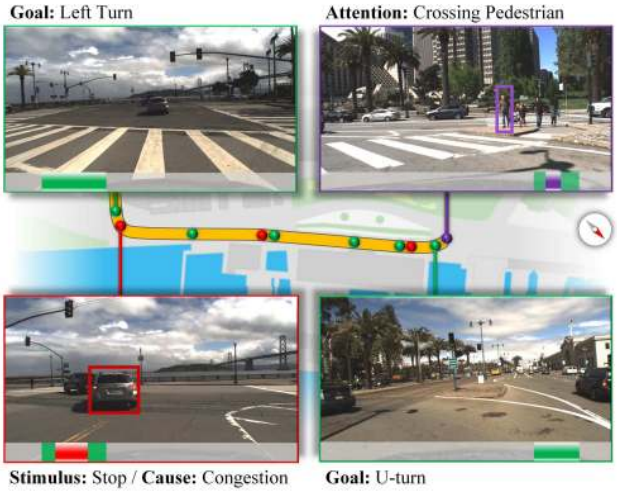


Figure 1: An example illustrating different driver behaviors in traffic scenes. The yellow trajectory indicates GPS positions of our instrumented vehicle. The driver performs actions and reasons about the scenes. To understand driver behavior, we define a 4-layer annotation scheme: **Goal-oriented action**, **Stimulus-driven action**, **Cause** and **Attention**. In **Cause** and **Attention**, we use bounding boxes to indicate when the traffic participant causes a *stop* or is attended by the driver. Best viewed in color.

To achieve the goal, we design and collect HDD[1] with the explicit goal of learning how humans perform actions and interact with traffic participants. We collected 104 hours of real human driving in the San Francisco Bay Area

[1]The dataset will be made available at https://usa.honda-ri.com/HDD

using an instrumented vehicle. The recording consists of 137 sessions, and each session represents a navigation task performed by a driver. Further details about the dataset will be discussed in Section 3.

In each session, we decompose the corresponding navigation task into multiple predefined driver behaviors. Figure 1 illustrates the decomposition of a navigation task. The yellow trajectory indicates GPS positions of our instrumented vehicle. A 4-layer annotation scheme is introduced to describe driver behaviors. The first layer is **Goal-oriented action**, colored green. In this example, the driver is making a *left turn* as shown in the upper left image of Figure 1. The second layer is **Stimulus-driven action**, colored red and is shown in the lower left image. In this example, the driver makes a *stop* because of *stopped car*. The *stop* action corresponds to **Stimulus-driven action** layer and *congestion* belongs to **Cause**, which is designed to indicate the reason the vehicle makes a *stop*. The red bounding box localizes the **Cause**.

While driving, human drivers are aware of surrounding traffic participants. We define the fourth layer called **Attention**, colored purple. In this example, the purple bounding box is used to indicate the traffic participant attended by the driver. Note that our annotation scheme is able to describe multiple scenarios happening simultaneously. In Figure 1, two different scenarios are illustrated. First, the driver intends to make a left turn but stops because of congestion. Second, the driver is making a U-turn while paying attention to a crossing pedestrian. A detailed description of our annotation methodology is presented in Section 3.1.

With the multimodal data and annotations, our dataset enables the following challenging and unique research directions and applications for intelligent transportation systems. First, detecting unevenly (but naturally) distributed driver behaviors in untrimmed videos is a challenging research problem. Second, interactions between drivers and traffic participants can be explored from cause and effect labels. Third, a multi-task learning framework for learning driving control can be explored. With the predefined driver behavior labels, the annotations can be used as an auxiliary task (i.e., classification of behavior labels) to improve the prediction of future driver actions. Fourth, a multimodal fusion for driver behavior detection can be studied.

Learning how humans drive and interact with traffic scenes is a step toward intelligent transportation systems. In this paper, we start from a driver-centric view to describe driver behaviors. However, driving involves other aspects. Particularly, it involves predicting traffic participants' intentions and reasoning overall traffic situations for motion planning and decision making, which is not discussed in this work. Toward the goal of developing intelligent driving systems, a scalable approach for constructing a dataset is the next milestone.

2. Related work

We review a range of datasets and highlight the uniqueness and the relationship between the current datasets and the proposed dataset.

Driving Scene Datasets. The emergence of driving scene datasets has accelerated the progress of visual scene recognition for autonomous driving. KITTI [7] provides a suite of sensors including cameras, LiDAR and GPS/INS. They launch different benchmarks (e.g., object detection, scene flow and 3D visual odometry) to push forward the algorithmic developments in these areas.

Cordts et al. [4] proposed a large scale road scene dataset, Cityscapes Dataset, with 5000 images with fine pixel-level semantic labeling. It enables research in category-level and instance-level semantic segmentation [35, 8] in driving scenes that KITTI dataset does not address. For long-term localization and mapping, the Oxford Robotcar dataset [22] presents a huge data collection collected under a variety of weather and lighting conditions over a year.

Our dataset is complementary to [7] and [22] since we focus on learning driver behavior under various traffic situations. A joint effort of ours and these existing datasets can lead to intelligent transportation systems.

Recently, learning a vision-based driving model [3, 12, 28, 34, 13] for autonomous driving has attracted a lot of attention. Chen et al. [3] used the driving game TORCS to obtain training data for learning a driving model by defining different affordance indicators. Jain et al., [12] proposed a dataset and algorithms to anticipate driver maneuvers. Santana and Hotz [28] presented a dataset with 7.25 hours of highway driving data to support research in this task. Earlier developments are constrained by limited amount of real-world driving data or simulated environment data [3]. With these limitations in mind, the BDD-Nexar dataset [34, 23], which includes video sequences, GPS and IMU, was proposed and adopted a crowdsourcing approach to collect data from multiple vehicles across three different cities in the US.

The proposed dataset provides additional annotations to describe common driver behaviors in driving scenes while existing datasets only consider turn, go straight, and lane change. Moreover, CAN signals are captured to provide driver behaviors under different scenarios, especially interactions with traffic participants.

Recently, Xu et al. [34] proposed an end-to-end FCN-LSTM network for this task. They considered 4 discrete actions in learning a driving model. The definition of four actions is based on CAN signals with heuristics. Instead, we provide an explicit definition of driver behaviors as shown in Figure 6a. Multitask learning frameworks for learning a driving model by introducing auxiliary tasks (i.e., classifica-

Dataset	Purpose	Sensor types	Hours	Areas
Princeton DeepDriving [3]	Vision-based control	Driving game TORCS	4.5	Driving game TORCS
KITTI [7]	Semantic understanding & vision-based control [19]	Camera, LiDAR, GPS, and IMU	1.4	Suburban, urban and highway
BDD-Nexar [23]	Vision-based control & semantic understanding & representation learning using videos	Camera, GPS and IMU	1000	Suburban, urban and highway
Udacity [33]	Steering angle prediction & image-based localization	Camera, LiDAR, GPS, IMU, and CAN	8	Urban and highway
comma.ai [28]	Driving simulator	Camera, GPS, IMU, and CAN	7.25	Highway
Brain4Car [12]	Driver Behavior Anticipation	Camera, GPS, and speed logger	N/A (1180 mi)	Suburban, urban and highway
Ours	Driver behavior & causal reasoning	Camera, LiDAR, GPS, IMU and CAN	104	Suburban, urban and highway

Table 1: Comparison of driving scene datasets

tion of current driver behavior and prediction of multisensor values) can be designed. A similar idea is proposed in [34] in that they introduced semantic segmentation as a side task. Integrating a behavior classification task can make the models explainable to humans, and can allow the use of common sense in traffic scenes from human priors.

A detailed comparison of different driving scene datasets is shown in Table 1.

Human Activity Understanding Datasets. Human activity understanding plays an important role in achieving intelligent systems. Different datasets have been proposed [10, 31, 15, 25] to address the limitations in earlier works. Note that our data can enable research in learning driver behaviors as mentioned in the introduction. Particularly, recognizing a **Goal-oriented action** is an *egocentric activity recognition* problem. The Stanford-ECM dataset [25] is related to our dataset in the following two aspects. First, they define egocentric activity classes for humans as in our **Goal-oriented** classes for drivers. Second, they provide egocentric videos and signals from a wearable sensor for jointly learning activity recognition and energy expenditure while we provide multisensor recordings from an instrumented vehicle for learning driver behavior. In addition to learning egocentric activities, we also annotate how traffic participants interact with drivers.

Datasets for research on pedestrian behaviors are released [18, 26]. Kooij et al., [18] proposed a dataset that annotates a pedestrian with the intention to cross the street under different scenarios. Rasouli et al., [26] provides a dataset (Joint Attention in Autonomous Driving) with annotations for studying pedestrian crosswalk behaviors. Understanding interactions between the driver and pedestrian are important for decision making. Robust pedestrian behavior modeling is also necessary [17, 16].

Visual Reasoning Datasets. Visual question answering (VQA) is a challenging topic in artificial intelligence. A VQA agent should be able to reason and answer questions from visual input. An increasing number of datasets [1, 14] and algorithms [2, 11] have been proposed recent years. Specifically, CLEVR dataset [14] is presented to enable the community to build a strong intelligent agent instead of solving VQA without reasoning. In our work, we hope to enable the community to develop systems that can understand traffic scene context, perform reasoning and make decisions.

3. Honda Research Institute Driving Dataset

3.1. Data Collection Platform

The data was collected using an instrumented vehicle equipped with the following sensors (their layout is shown in Figure 2):

(i) 3 x Point Grey Grasshopper 3 video camera, resolution: 1920 1200 pixels, frame rate: 30Hz, field of view (FOV): 80 degrees x 1 (center) and 90 degrees x 2 (left and right).

(ii) 1 x Velodyne HDL-64E S2 3D LiDAR sensor, spin rate: 10 Hz, number of laser channel: 64, range: 100 m, horizontal FOV: 360 degrees, vertical FOV: 26.9 degrees.

(iii) 1 x GeneSys Eletronik GmbH Automotive Dynamic Motion Analyzer with DGPS outputs gyros, accelerometers and GPS signals at 120 Hz.

(iv) a Vehicle Controller Area Network (CAN) that provides various signals from around the vehicle. We recorded throttle angle, brake pressure, steering angle, yaw rate and speed at 100 Hz.

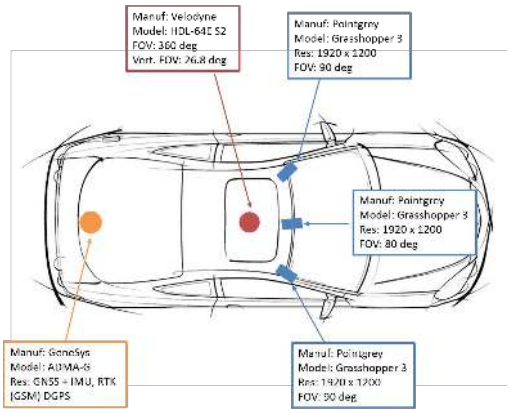


Figure 2: Sensor layout of the instrumented vehicle.

Figure 3: The figure shows GPS traces of the HDD dataset. We split the dataset into training and testing according to the vehicle's geolocation. The blue and red color traces denote the training and testing sets, respectively.

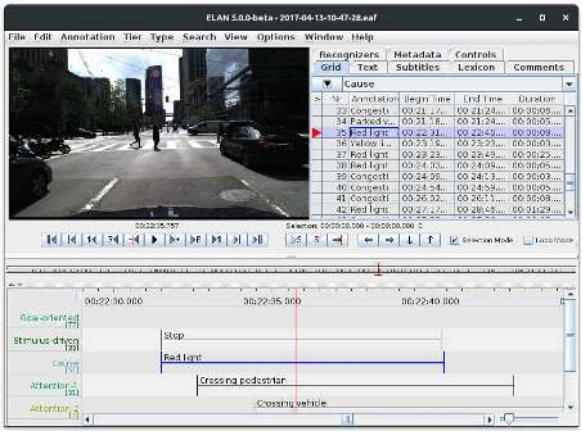


Figure 4: Annotation interface. We use the open source software toolkit ELAN to annotate different driver behaviors and causal relationships.

All sensors on the vehicle were logged using a PC running Ubuntu Linux 14.04 with two eight-core Intel i5-6600K 3.5 GHz Quad-Core processors, 16 GB DDR3 memory, and a RAID 0 array of four 2TB SSDs, for a total capacity of 8 TB. The sensor data are synchronized and timestamped using ROS[2] and a customized hardware and software designed for multimodal data analysis.

For our dataset, we are interested in having a diverse set of traffic scenes with driver behaviors. The current data collection spans from February 2017 to October 2017. We drove within the San Francisco Bay Area including on urban, suburban and highway roads, as shown in Figure 3. The total size of the post-processed dataset is around 150 GB and 104 video hours. The video is converted to a resolution of 1280×720 at 30 fps.

3.2. Annotation Methodology

It is challenging to define driver behavior classes since it involves cognitive processes and vehicle-driver interaction. It is especially challenging to identify an exact segment of driver behavior from data we collected, in particular from video sequences. In our annotation processes, we make the best effort in annotating different driver behaviors with a mixture of objective criteria and subjective judgment.

Our annotation methodology is motivated by human factor and cognitive science. Michon [24] proposed three classes of driving processes: **operational processes** that correspond to the manipulation of the vehicle, **tactical processes** that are the interactions between the vehicle, traffic participants and environment, and **strategic processes** for higher level reasoning, planning and decision making.

With these definitions in mind, we propose a 4-layer representation to describe driver behavior, i.e., **Goal-oriented action**, **Stimulus-driven action**, **Cause** and **Attention**, which encapsulate driver behavior and causal reasoning. A complete list of labels in the 4-layer representation for describing driver behavior can be found in Figure 6a.

Goal-oriented action involves the driver's manipulation of the vehicle in a navigation task such as *right turn*, *left turn*, *branch* and *merge*. While operating the vehicle, the driver can make a *stop* or *deviate* due to traffic participants or obstacles. *Stop* and *deviate* are categorized as **Stimulus-driven action**. When the driver performs a *stop* or a *deviate* action, there is a reason for it. We define the third layer **Cause** to explain the reason for these actions. For example, a *stopped car* in front of us is an immediate cause for a stop as in Figure 1. Finally, the fourth layer **Attention** is intro-

[2]http://www.ros.org/

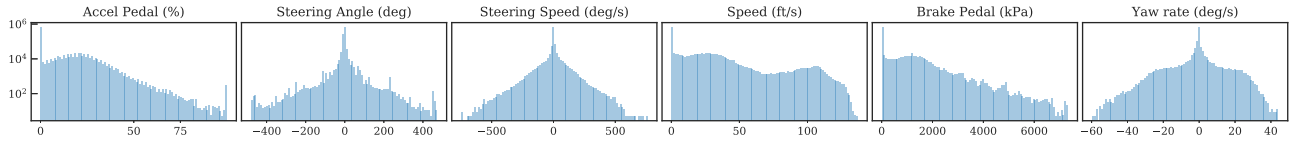


Figure 5: Histograms of sensor measurements for the dataset.

duced to localize the traffic participants that are attended by drivers. For example, a *pedestrian near ego lane* may be attended by drivers since the pedestrian might perform certain actions that would affect driver behavior.

Based on the aforementioned definition, we work with experienced external annotators on this task. The annotators use an open source software package ELAN[3] to label videos as shown in Figure 4. To ensure the consistency in annotations, we conduct the following quality control strategy. Given a driving session, it is first annotated by 2 independent human annotators. Then, a third annotator merges the three annotations with his/her own judgment into a single annotation. Finally, we have an internal expert annotator to review and obtain the final version.

To analyze the annotation consistency, we compare the annotation quality of the third external annotator to the internal expert annotator on 10 sessions. Based on the same annotation procedures, we found a 98% (driver behavior label) agreement between the third external annotator and the internal expert annotator. However, the start time and end time of a driver behavior is not trivial to assess since it involves a subjective judgment. A systematic evaluation of action localization consistency is needed and requires a further investigation.

3.3. Dataset Statistics

In the current release, we have a total of 104 video hours, which are annotated with the proposed 4-layer structure. Within 104 video hours, we have 137 sessions corresponding to different navigation tasks. The average duration of each session is 45 minutes. The statistics of session duration can be found in Figure 6b. Figure 6a shows the number of instances of each behavior. A highly imbalanced label distribution can be observed from the figure.

4. Multimodal Fusion for Driver Behavior Detection

Our goal is to detect driver behaviors which occur during driving sessions by predicting a probability distribution over the list of our predefined behavior classes at every given point of time. As the first step, we focus on the detection of **Goal-oriented** and **Cause** layers in our experiments. To detect driver behaviors, we design an algorithm to learn a representation of driving state which encodes the necessary history of past measurements and effectively translates them into probability distributions. Long-Short Term Memory (LSTM) networks were shown to be successful in many temporal modeling tasks, including activity detection. We thus employ an LSTM as the backbone architecture for our model.

In addition to the input video stream, our model has access to an auxiliary signal which provides complimentary information about the vehicle dynamics. This auxiliary signal includes measurements from the following CAN bus sensors: car speed, accelerator and braking pedal positions, yaw rate, steering wheel angle, and the rotation speed of the steering wheel, illustrated in Figure 5. This makes the task and approach different from the standard activity detection setup where models usually have access only to a single modality. The proposed baseline architecture of driver behavior detection is shown in Figure 6c.

Our application domain dictates the necessity to design a model which can be used in real-time in a streaming manner. Thus, we constrain our model to the case where predictions are made solely based on the current input and previous observations.

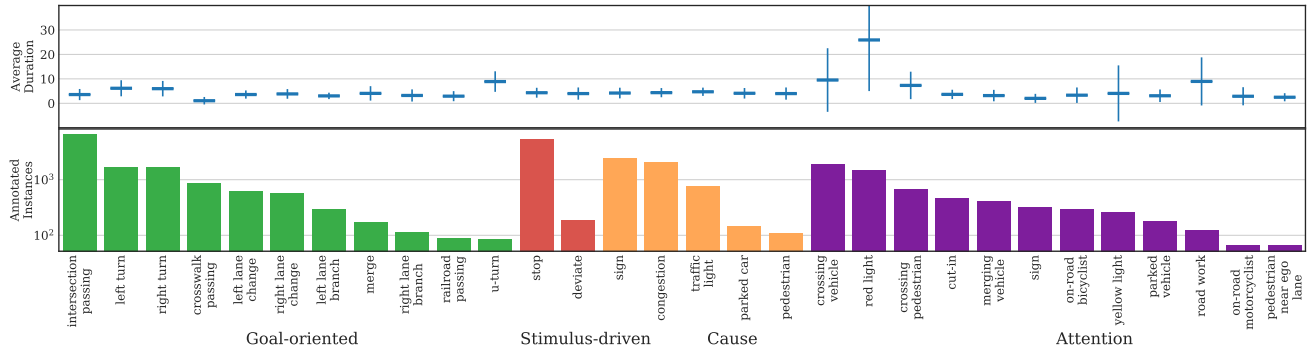
5. Experiments

To assess the quality of the visual representations learned by our model on our challenging dataset, we perform rigorous experiments as well as comparisons to several baselines. First of all, we split the dataset based on the geolocation data, thus, minimizing spatial overlap of train and test routes. This way we avoid testing on the very same locations as those used to train the model. Fig 3 shows the geolocation measurements in the training (in blue) and test (in red) splits.

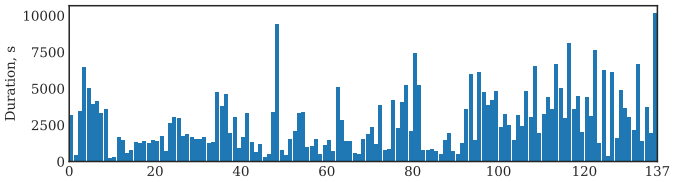
For baseline models we sample input frames from video streams and values from CAN bus sensors at 3 Hz. We found this to be a reasonable trade-off between modeling complexity and precision. We employed the *Conv2d_7b_1x1* layer of InceptionResnet-V2 [32] pretrained on ImageNet [5] to get feature representations for each frame.

Raw sensor values are passed through a fully-connected layer before concatenation with visual features. In turn, visual features are represented by spatial grid of CNN activations. Additional 1x1 convolution is applied to reduce dimensionality of *Conv2d_7b_1x1* from $8 \times 8 \times 1536$ to $8 \times 8 \times 20$ before flattening it and concatenating with sensor

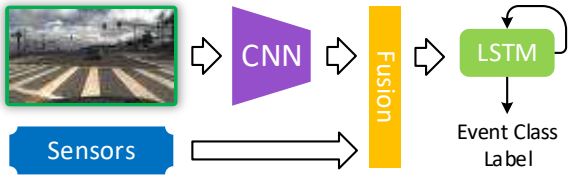
[3]https://tla.mpi.nl/tools/tla-tools/elan/



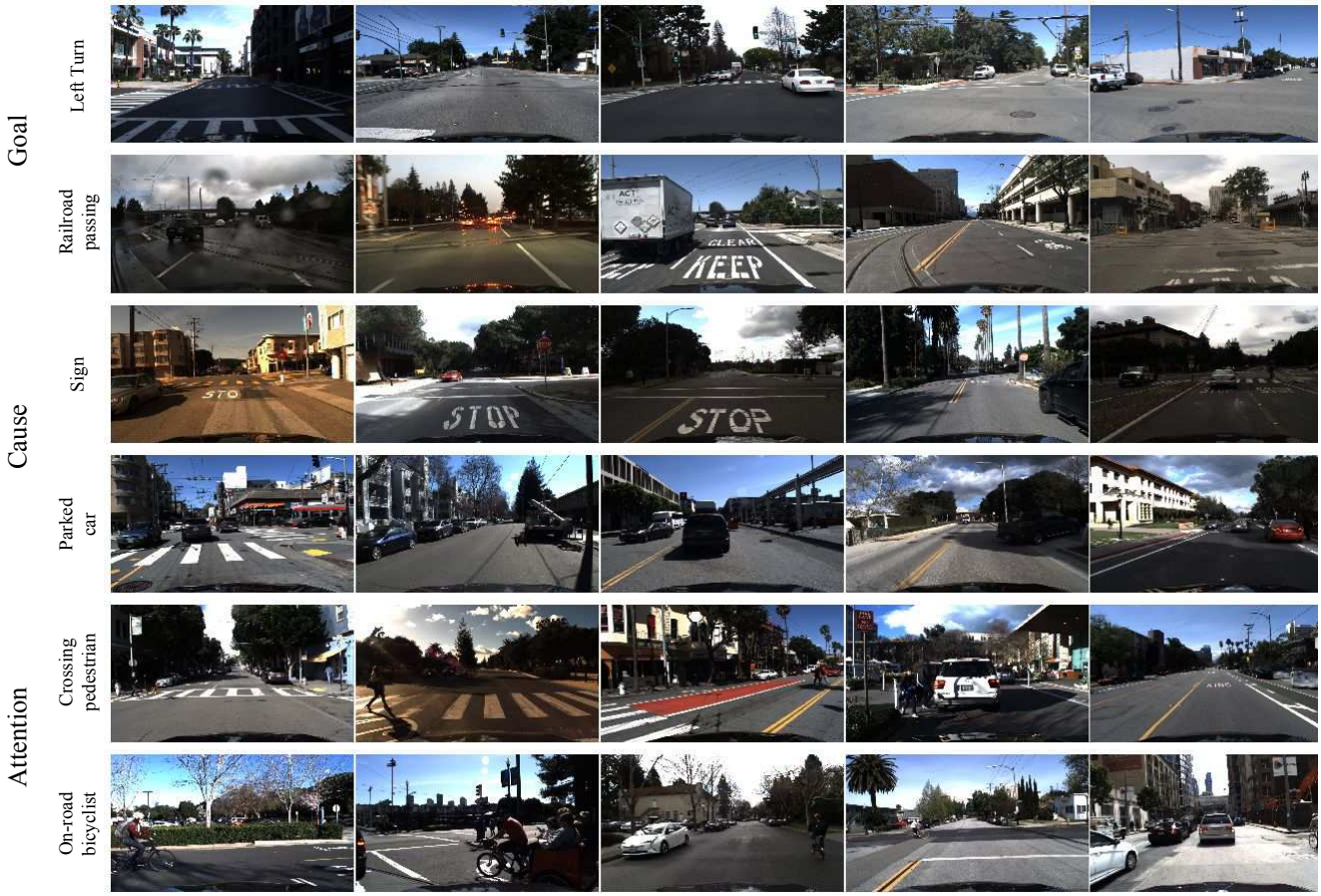
(a) Driver behavior statistics (annotated instances and instance duration) in different layers.



(b) A total of 137 annotated sessions with a 45-minute average duration.



(c) A baseline architecture of driver behavior detection.



(d) Examples of video frames from our dataset which belong to driver behaviors designated on the left.

Table 2: Detection results on test set for **Goal-oriented action** using per-frame average precision, *Random* illustrates the portion of frames which belongs to the given action. *Sensors* model uses LSTM to encode the stream of 6-dimensional vectors from CAN bus. Features from the last convolutional layer of InceptionResnet-V2 on RGB frames are provided as an input to LSTM in *CNN* models.

Models	right turn	left turn	intersection passing	railroad passing	left lane branch	right lane change	left lane change	right lane branch	crosswalk passing	merge	u-turn	mAP
Random	2.24	2.61	6.58	0.07	0.15	0.46	0.42	0.07	0.21	0.13	0.18	1.19
CNN pool	30.11	31.88	56.49	3.96	2.02	1.35	1.43	0.15	8.71	7.13	4.89	13.46
Sensors	74.27	66.25	36.41	0.07	8.03	13.39	26.17	0.20	0.30	3.59	33.57	23.84
CNN conv	54.43	57.79	65.74	2.56	25.76	26.11	27.84	1.77	16.08	4.86	13.65	26.96
CNN+Sensors	77.47	76.16	76.79	3.36	25.47	23.08	41.97	1.06	11.87	4.94	17.61	32.71

Table 3: Detection results on test split for **Cause layer**. Behaviors from this layer are immediate causes for either *stop* or *deviate* actions.

Models	sign	congestion	traffic light	pedestrian	parked car	mAP
Random	2.49	2.73	1.22	0.20	0.15	1.36
CNN+Sensors	46.83	39.72	45.31	2.15	7.24	28.25

features. The necessity to preserve the spatial resolution is illustrated by Table 2 where 'CNN conv' demonstrates the substantial advantage in detection of turns. LSTM hidden state size is set to 2000 in all experiments. During training, we formed batches of sequence segments by sequentially iterating over driving sessions. The last LSTM hidden state from the previous batch is used to initialize the LSTM hidden state on the next step. Training is performed using truncated backpropagation through time.

For Goal and Cause layers, we trained separate LSTMs using batches of size 40 with each sequence length set to 90 samples. We confirmed that the larger batch size improves convergence. We set dropout *keep* probability on the input and output of the LSTM to 0.9. Taking into account the data imbalance between foreground and background frames, and also the imbalance of behavior classes themselves, we used the recently proposed technique for modifying cross-entropy loss to deal with class imbalance [20]. This modification was originally applied to the task of object detection where negative region proposals also dominate.

Our evaluation strategy is inspired from the activity detection literature [30] where each frame is evaluated for the correct activity label. Specifically, Shou *et al.* [30] treated the per-frame labeling task as a retrieval problem and computed Average Precision (AP) for each activity class by first ranking all frames according to their confidence scores. Following this procedure, we compute the AP for individual driver behavior classes as well as the mean AP (mAP) over all behavior classes. The test split we obtained via the geospatial selection procedure described above includes 37 driving sessions, which contain a total of approximately 274,000 frames sampled at 3 FPS for the mean average precision evaluation.

6. Results and Discussion

Goal-oriented layer Table 2 provides the APs for 11 Goal-oriented Actions (starting from 'right turn' to 'u-turn') for our model and its ablated versions. The last column provides the mAP value for all the methods. First, we provide a description of the baselines used in this experiment. The first baseline ('Random') simply assigns random behavior labels to each frame and serves as the lower bound on model performance. It also illustrates the portion of frames in the test data for which every class label is assigned. The next one ('CNN pool') encodes each frame by extracting convolutional features using an InceptionResnet-V2 network and pooling them spatially to a fixed-length vector. These pooled representations of frames are sequentially fed to the LSTM to predict the behavior label. The third baseline ('Sensors') uses only the sensor data as input to the LSTM. The next method ('CNN conv') is a variant of the second method: instead of spatially pooling CNN feature encodings, we used a small convnet to reduce the dimensionality of the CNN encodings of the frames before passing them through the LSTM. Finally, the 'CNN+Sensors' method adds sensor data to the 'CNN conv' method.

We can see that the performance of 'CNN pool' is quite low. This can be attributed to the fact that information is lost by the spatial pooling operation. 'CNN conv' replaces pooling by a learnable conv layer and significantly increases mAP. Sensor measurements (brake, steering wheel, etc.) alone result in slightly better AP for simple actions like left/right turns where the information about steering wheel position can be sufficient in most of the cases. When it comes to actions like lane changes, visual information used in 'CNN conv' allows for proper scene interpretation, thus, improving over 'Sensor' model. It is clear, that only sensor information is not sufficient for driver behavior detection, especially in an imbalanced scenario. The visual data and data from sensors are complementary to each other in this respect and thus their fusion gives the best results, as shown in the last row of the table.

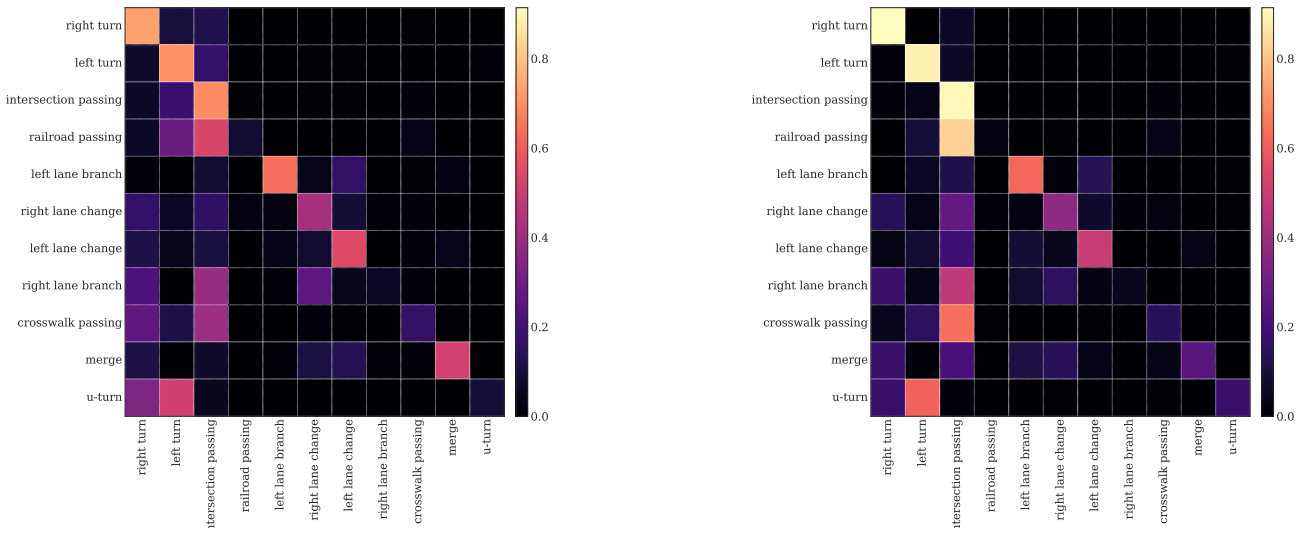


Figure 7: Confusion matrices for *Goal-oriented* driver behavior classes using 'CNN conv' model (left) and 'CNN+Sensors' (right). For better view we omit background class from visualization and normalize rows accordingly. Note the clear disambiguation between the detection of turns and 'intersection passing' after adding the input data from sensors. Intersections usually have crosswalks, signs and road markings which influence detection performance for other classes.

It is interesting to note that the 'blind model' (without camera input) is able to successfully guess 'intersection passing' because most of them are happening in a very specific pattern: 'deceleration/wait/acceleration'. A 'railroad passing' is surprisingly hard for the CNN model because this behavior type includes not only railroad crossing in the designated locations which have discriminative visual features but also tram rails crossing. The confusion of behavior classes with a 'background' class remains the most frequent source of errors for all layers.

Cause Layer Table 3 represents the detection results of causes for stimulus-driven actions. We observe a better detection performance for *sign*, *congestion* and *traffic light*. The corresponding motion pattern should be similar, i.e., deceleration. On the other hand, the vehicle dynamics for *pedestrian* and *parked car* are very different from the rest. For *pedestrian*, the vehicle usually makes a stop action for pedestrians while making turns. For *parked car*, the vehicle deviates from the original trajectory to avoid a collision. We hypothesize that the weak performance of the proposed model is due to the following two reasons. First, the two classes are underrepresented in the dataset as shown in Figure 6a. A better approach to deal with an imbalanced distribution is necessary. The same is true for rare **Goal-oriented** actions. Second, the motion modeling for a short duration of cause (e.g., pedestrian) may not be captured in the baseline model (similar to railroad passing). The motion pattern of a *deviate* action may not be modeled effectively to detect *parked car*. This would benefit from better motion modeling using optical flow features. We leave this for future work.

In the current version of the dataset, we have only four causes for a stop action, namely, *sign*, *congestion*, *traffic light*, *pedestrian*, and one cause *parked car* for a deviate action. Because detection of these immediate causes directly implies detection and discrimination of their respective actions we do not provide separate results for **Stimulus-driven** layer.

7. Conclusion

In this paper, we introduce the Honda Research Institute Driving Dataset, which aims to stimulate the community to propose novel algorithms to capture the driver behavior. To enable this, we propose a novel annotation methodology that decomposes driver behaviors into a 4-layer representation, i.e., **Goal-oriented**, **Stimulus-driven**, **Cause** and **Attention**. A variety of baselines for detecting driver behaviors in untrimmed videos were proposed and tested on this dataset. Our preliminary results show that this task is challenging for standard activity recognition methods based on RGB frames. Although adding sensor data improves accuracy, we need better representations, temporal modeling and training strategy to achieve reasonable performance in driver behavior detection before exploring the actual relationship between behaviors in different layers, i.e., the relationship between a driver and traffic situations.

8. Acknowledgements

This work is supported in part by the DARPA XAI program and Honda Research Institute USA.

References

[1] A. Agrawal, J. Lu, S. Antol, M. Mitchell, C. L. Zitnick, D. Batra, and D. Parikh. VQA: Visual Question Answering. In *ICCV*, 2015. 3

[2] J. Andreas, M. Rohrbach, T. Darrell, and D. Klein. Learning to Compose Neural Networks for Question Answering. In *CVPR*, 2016. 3

[3] C. Chen, A. Seff, A. Kornhauser, and J. Xiao. DeepDriving: Learning Affordance for Direct Perception in Autonomous Driving. In *ICCV*, 2015. 2, 3

[4] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The Cityscapes Dataset for Semantic Urban Scene Understanding. In *CVPR*, 2016. 1, 2

[5] J. Deng, W. Dong, R. Socher, L. jia Li, K. Li, and L. Fei-fei. ImageNet: A large-scale Hierarchical Image Database. In *CVPR*, 2009. 5

[6] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR*, 2009. 1

[7] A. Geiger, P. Lenz, and R. Urtasun. Are We Ready for Autonomous Driving? The KITTI Vision Benchmark Suite. In *CVPR*, 2012. 1, 2, 3

[8] K. He, G. Gkioxari, P. Dollar, and R. Girshick. Mask R-CNN. In *ICCV*, 2017. 1, 2

[9] K. He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition. In *CVPR*, 2016. 1

[10] F. C. Heilbron, V. Escorcia, B. Ghanem, and J. C. Niebles. ActivityNet: A Large-Scale Video Benchmark for Human Activity Understanding. In *CVPR*, 2015. 3

[11] R. Hu, J. Andreas, M. Rohrbach, T. Darrell, and K. Saenko. Learning to Reason: End-To-End Module Networks for Visual Question Answering. In *ICCV*, 2017. 3

[12] A. Jain, H. S. Koppula, B. Raghavan, S. Soh, and A. Saxena. Car that Knows Before You Do: Anticipating Maneuvers via Learning Temporal Driving Models. In *ICCV*, 2015. 1, 2, 3

[13] D. Jayaraman and K. Grauman. Learning Image Representations Tied to Egomotion from Unlabeled Video. *International Journal of Computer Vision*, 125(1):136–161, 2017. 2

[14] J. Johnson, B. Hariharan, L. van der Maaten, L. Fei-Fei, C. L. Zitnick, and R. Girshick. CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning. In *CVPR*, 2017. 3

[15] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, and A. Zisserman. The Kinetics Human Action Video Dataset. *CoRR*, abs/1705.06950, 2017. 3

[16] C. G. Keller and D. M. Gavrila. Will the Pedestrian Cross? A Study on Pedestrian Path Prediction. *IEEE Transactions on Intelligent Transportation Systems*, 15(2):494–506, 2014. 3

[17] K. Kitani, B. Ziebart, J. Bagnell, and M. Hebert. Activity forecasting. In *ECCV*, 2012. 3

[18] J. Kooij, N. Schneider, F. Flohr, and D. Gavrila. Context-based Pedestrian Path Prediction. In *ECCV*, 2014. 3

[19] N. Lee, W. Choi, P. Vernaza, C. B. Choy, P. H. S. Torr, and M. Chandraker. DESIRE: Distant Future Prediction in Dynamic Scenes with Interacting Agents. In *CVPR*, 2017. 3

[20] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar. Focal Loss for Dense Object Detection. In *ICCV*, 2017. 7

[21] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollr, and C. L. Zitnick. Microsoft COCO: Common Objects in Context. In *ECCV*, 2009. 1

[22] W. Maddern, G. Pascoe, C. Linegar, and P. Newman. 1 Year, 1000km: The Oxford RobotCar Dataset. *The International Journal of Robotics Research*, 36(1):3–15, 2017. 1, 2

[23] V. Madhavan and T. Darrell. The BDD-Nexar Collective: A Large-Scale, Crowsourced, Dataset of Driving Scenes. Master's thesis, EECS Department, University of California, Berkeley, May 2017. 1, 2, 3

[24] J. A. Michon. A Critical View of Driver Behavior Models: What Do We Know, What Should We Do? *Human Behavior and Traffic Safety*, pages 485–520, 1985. 4

[25] K. Nakamura, S. Yeung, A. Alahi, and L. Fei-Fei. Jointly Learning Energy Expenditures and Activities using Egocentric Multimodal Signals. In *CVPR*, 2017. 3

[26] A. Rasouli, I. Kotseruba, and J. K. Tsotsos. Are They Going to Cross? A Benchmark Dataset and Baseline for Pedestrian Crosswalk Behavior. In *ICCVW*, 2017. 3

[27] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *NIPS*, 2015. 1

[28] E. Santana and G. Hotz. Learning a Driving Simulator. *CoRR*, abs/1608.01230, 2016. 1, 2, 3

[29] M. Schmidt, U. Hofmann, and M. Bouzouraa. A Novel Goal Oriented Concept for Situation Representation for ADAS and Automated Driving. In *ITSC*, 2014. 1

[30] Z. Shou, J. Chan, A. Zareian, K. Miyazawa, and S. Chang. CDC: Convolutional-De-Convolutional Networks for Precise Temporal Action Localization in Untrimmed Videos. In *CVPR*, 2017. 7

[31] G. A. Sigurdsson, G. Varol, X. Wang, A. Farhadi, I. Laptev, and A. Gupta. Hollywood in Homes: Crowdsourcing Data Collection for Activity Understanding. In *ECCV*, 2016. 3

[32] C. Szegedy, S. Ioffe, and V. Vanhoucke. Inception-v4, inception-resnet and the impact of residual connections on learning. *CoRR*, abs/1602.07261, 2016. 5

[33] Udacity. Public driving dataset. https://www.udacity.com/self-driving-car, 2017. 3

[34] H. Xu, Y. Gao, F. Yu, and T. Darrell. End-To-End Learning of Driving Models From Large-Scale Video Datasets. In *CVPR*, 2016. 1, 2, 3

[35] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia. Pyramid Scene Parsing Network. In *CVPR*, 2017. 1, 2