# DBUS: Human Driving Behavior Understanding System

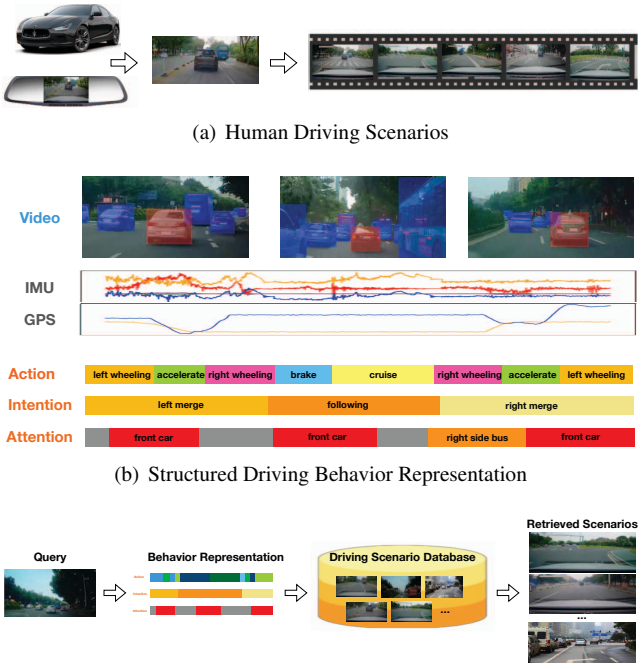
Max Guangyu Li[1,2], Bo Jiang[1], Zhengping Che[1], Xuefeng Shi[1], Mengyao Liu[1], Yiping Meng[1], Jieping Ye[1], Yan Liu[2] *
[1] Didi Chuxing [2]University of Southern California
{scottjiangbo, chezhengping, shixuefeng, liumengyao, mengyipingkitty, yejieping}@didiglobal.com
{guangyul, yanliu.cs}@usc.edu

## Abstract

*Human driving behavior understanding is a key ingredient for intelligent transportation systems. Either developing self-driving car drives like humans or building V2X systems to improve human driving experience, we need to understand how humans drive and interact with environments. Massive human driving data collected by top ride-sharing platforms and fleet management companies, offers the potential for in-depth understanding of human driving behavior. In this paper, we present DBUS, a real-time driving behavior understanding system which works with front-view videos, GPS/IMU signals collected from daily driving scenarios. Unlike previous work of driving behavior analysis, DBUS focuses on not only the recognition of basic driving actions but also the identification of driver's intentions and attentions. The analysis procedure is designed by mimicking the human intelligence for driving, powered with representation capability of deep neural networks as well as recent advances in visual perception, video temporal segmentation, attention mechanism, etc. Beyond systematic driving behavior analysis, DBUS also supports efficient behavior-based driving scenario search and retrieval, which is essential for practical application when working with large-scale human driving scenario dataset. We perform extensive evaluations of DBUS in term of inference accuracy of intentions, interpretability of inferred driver's attentions, as well as system efficiency. We also provide insightful intuitions as to why and how certain components work based on experience in the development of the system.*

## 1. Introduction

Understanding how human drives on the road is essential for the development of many intelligent transportation systems, including autonomous driving systems [16], Vehicle to Vehicle (V2V) and Vehicle to Infrastructure (V2I) systems [21], and driving safety monitoring systems [11, 20], just to name a few. For example, in an autonomous driving system, understanding human drivers' behavior is crucial for accurate prediction of surrounding vehicles' actions, proposing human-like planning and control strategies, as well as building a realistic simulator for the extensive autonomous driving virtual test. As for V2V and V2I systems, understanding driving behavior, especially how human drivers interact with each other and with transportation infrastructure, provides insight for an efficient and conve-




(a) Human Driving Scenarios

(b) Structured Driving Behavior Representation

(c) Behavior-based Driving Scenario Retrieval

Figure 1. Overview of *DBUS*: (a) Human driving scenarios collected via dash-cameras and GPS/IMU sensors. (b) Structured driving behavior representation (driving action, intention, attention) inferred from human driving data (videos, GPS/IMU signals). (c) Behavior-based driving scenarios retrieval via structured behavior representation.

*First two authors contributed equally to this work

nient design of such systems. In term of driving safety, driving behavior understanding enables real-time driving safety monitoring as well as constituting a comprehensive driving safety profile for each driver, which is of top priority for fleet management and ride-sharing platforms.

On the other hand, real-world traffic is a complicated multi-agent system in which multiple participants interacts with each other and with infrastructures. Moreover, each driver has his/her own driving style. Therefore, there is a huge diversity in daily driving scenarios and driving behaviors, which raises the need for extensive human driving data to build a comprehensive driving behavior understanding system. In recent years, the emerging of ride-sharing industry with multi million drivers, offers the potential for collecting vast volumes of human driving data. Equipped with on-vehicle sensors such as video cameras, GPS, inertial measurement unit (IMU), millions of drivers could share their driving data to power the in-depth analysis of human driving behavior.

However, mining large-scale human driving data for driving behavior understanding presents significant challenges in following aspects. First, human driving behavior is the outcome of the sophisticated human intelligent system, including perception of the current traffic scenario, reasoning surrounding traffic participants' intentions, paying special attention to certain objects, planning ego-trajectory and finally executing driving actions. Therefore, the focus of most existing work, e.g. detecting traffic participants [10, 7] or recognizing basic driving actions such as braking and steering [9], is only the first step. Toward a complete driving behavior understanding, we need to analyze human driving data at multiple levels, ranging from low-level driving action recognition all-the-way to high-level driver's intention and cause inference. On the other hand, human driving data is usually collected from multiple types of sensor simultaneously, including video camera, GPS, IMU etc., capturing different aspect of human driving behavior. Each type of data has its own properties and works with different methodologies. For example, IMU data is the best fit for low-level driving action recognition, while videos support driver's attention inference. Therefore, an integrated system, which not only mining each type of data properly but also fusion all information together, is required for comprehensive driving behavior understanding. Moreover, efficiency for such system is of crucial importance for practical deployment, especially when dealing with massive human driving data collected from millions of drivers.

In this work, we analyze the large-scale human driving data from one of the top ride-sharing companies and presents *DBUS*, **D**riving **B**ehavior **U**nderstanding **S**ystem. Through a comprehensive system design, *DBUS* tackles all the above challenges. Specifically, equipped with a structured representation, *DBUS* analyzes the driving behavior from three levels: basic driving action, driving intention, and driving attention which are formally defined in Section 3. Beyond structured driving behavior understanding, *DBUS* also supports behavior-based driving scenario search and retrieval, which cannot be achieved by conventional keywords or objects-based searches. Given a clip of human driving data, the system could find the most similar driving scenarios from millions of cases within a second.

*DBUS* has been deployed to power downstream applications within the ride-sharing company. Centered around a structured behavior representation produced from *DBUS*, a driver profiling system is built to summarize driving style of each driver with a special focus on driving safety. As for the driving scenario retrieval, a key application is for simulation test of autonomous driving system. Given a certain traffic scenario of interest, e.g. unprotected left turn with crossing pedestrians, relevant real-world driving scenarios would be retrieved automatically and then feed into autonomous driving simulator.

The main contributions of this work are summarized as follows:

- A structured representation of human driving behavior is designed with multiple levels of understanding, including basic driving action, driving intention and driving attention. This representation lays the foundation for comprehensive human driving behavior understanding.
- We develop INFER, a deep learning based model powered with attention mechanism, to jointly infer driver's intention and attention and provide interpretable causal reason behind driving behavior.
- We provides a total solution for mining human driving behavior through an integrated system *DBUS*, which supports multi-type sensors input, structured behavior analysis, as well as efficient driving scenario retrieval applicable to large-scale driving data collected from millions of drivers.

## 2. Related Work

**Driving Behavior Analysis** In general, research on driving behavior analysis can be classified according two perspectives: i) the target of the research, e.g. basic driving maneuver recognition, driving intention prediction, aggressive driving detection, driver identification, etc. ii) the data used for the analysis, e.g. GPS trajectories, IMU signals, CAN-bus data, dash-cam videos, etc [15] [4]. Early studies mainly focus on low-level kinetic behavior with CAN-bus data. For example, [17] asked drivers to perform basic driving maneuvers and investigate the pattern of recorded CAN-bus data. In the last decade, with the spread of smart phones,

Table 1. Categories of Structured Behavior Representation

| Behavior | # of Categories | Categories |
|---|---|---|
| Driving Action ($\mathbf{m}$) | 9 | { *left_accelerate*, *left_cruise*, *left_brake*, *straight_accelerate*, *straight_cruise*, *straight_brake*, *right_accelerate*, *right_cruise*, *right_brake* } |
| Driving Intention ($\mathbf{w}$) | 8 | { *following*, *left_turn*, *right_turn*, *left_lane_change*, *right_lane_change*, *left_merge*, *right_merge*, *U_turn* } |
| Driving Attention ($\mathbf{a}_{obj}$) | 8 | { car, bus, truck, person, bicycle, motorcycle, tricycle, traffic light } |

Table 2. Feature Definition and Notation of *DBUS*

| Notation | | | Type | Definition |
|---|---|---|---|---|
| Raw Data ($\mathcal{D}$) | Video ($\mathbf{V}$) | $\mathbf{v}$ | image | front-view video frames |
| | GPS/IMU ($\mathbf{S}$) | $vs._{speed}$ | $\mathbb{R}$ | vehicle speed in GPS signals |
| | | $vs._{acc}$ | $\mathbb{R}$ | forward accelerate in IMU signals |
| | | $vs._{yaw}$ | $\mathbb{R}$ | yaw angular velocity in IMU signals |
| Perception Result ($\mathcal{P}$) | Objects ($\mathbf{O}$) | $\mathbf{o}$ | mask | semantic mask of detected traffic participants and traffic lights |
| | Distance ($\mathbf{D}$) | $\mathbf{d}$ | $\mathbb{R}$ | distance between ego-vehicle and nearest front traffic participants |
| | Locations ($\mathbf{L}$) | $\mathbf{l}$ | category | vehicle's location on the road based on the lane perception results |
| Behavior Representation ($\mathcal{B}$) | Action ($\mathbf{M}$) | $\mathbf{m}$ | category | basic driving actions |
| | Intention ($\mathbf{W}$) | $\mathbf{w}$ | category | driving intention |
| | Attention ($\mathbf{A}$) | $\mathbf{a}_{mask}$ | mask | driving attention mask |
| | | $\mathbf{a}_{obj}$ | category | object category of driving attention |

large-scale GPS/IMU data collected from smart phone sensors are available for behavior analysis, which enables a series of tasks that cannot be achieved with CAN-bus data alone. To name a few, [23] tried to identify different drivers by analyzing their trajectories, [1] worked on safety risk analysis based on distractions such turning video, checking cell phones, smart phone sensors data have been coupled with CAN-bus data to isolate basic driving actions like accelerate, braking and turning events in [22]. Moreover, smart phone sensors have also beed used to detect aggressive driving [12] and drunk drivers [8].

**Driving Video Analysis** More recently, the collection of driving videos, especially front-view videos, become realistic with the emerge of dash-camera and other driving video recording systems. Given that humans are able to drive with vision information alone, driving videos carry massive information and offer the potential to take our understanding of human driving behavior to the next level. A few large-scale real-world human driving video datasets are published to encourage research activity. BDD dataset is proposed in [14] which includes video sequences, GPS/IMU data collected from multiple vehicles across three different cities in the US. Later, a subset of BDD dataset got human-annotated action descriptions and justifications, named BDD-X dataset. Recently, an attention dataset is generated from BDD dataset with human-annotated driver's attention at key frames, named BDD-A [24]. Another similar dataset, the HDD dataset [18], provides raw data collected not only from sensors including video, GPS/IMU but also from LiDAR and CAN together with human-annotated drivers' attentions and causal reason for driving actions.

Existing work with large-scale driving videos dataset mainly focus on behavior cloning towards end-to-end learning for a self-driving system. In general, a deep neural network based driving policy is learned with frame-action pairs in a supervised manner. For example, [2] present a model directly maps video frames to steering controls, while [25] use deep neural networks to predict vehicle's future motion given raw pixels and its current state. Although these end-to-end models have shown their potential of working with human driving video, lack of interpretability limits their contribution in term of in-depth understanding of human driving behavior. Recently, [13] bring explainability to end-to-end video-to-control models with a visual attention mechanism and an attention-based video-to-text model to produce textual explanations of human driving behavior.

## 3. Problem Formulation

Consider a driving behavior understanding task with time horizon $T$. The input is human driving data $\mathcal{D} = (\mathbf{V}, \mathbf{S})$ collected from on-vehicle sensors, denoted as a driving scenario, where $\mathbf{V} = \{\mathbf{v}_t\}_{t=1}^{T}$ refers frames from front-view camera videos, $\mathbf{S} = \{vs._t\}_{t=1}^{T}$ denotes GPS and IMU signals at every time step. The goal is to generate a three-level structured representation of human driving behavior $\mathcal{B} = (\mathbf{M}, \mathbf{W}, \mathbf{A})$ over the whole time horizon $T$ in a given driving scenario, including basic driving actions $\mathbf{M} = \{\mathbf{m}_t\}_{t=1}^{T}$, driver's intentions $\mathbf{W} = \{\mathbf{w}_t\}_{t=1}^{T}$ as well as driver's attentions $\mathbf{A} = \{\mathbf{a}_{mask}^t, \mathbf{a}_{obj}^t\}_{t=1}^{T}$. The proposed structured representation is exemplified with a driving scenario in Fig. 1.

At each time step, basic driving actions $\mathbf{m}_t$ and driver's intentions $\mathbf{w}_t$ belongs to one of pre-defined action and intention categories. As for driver's attention, $\mathbf{a}_{mask}^t$ is the mask over the video frame $\mathbf{v}_t$ with pixel-wise attention in-

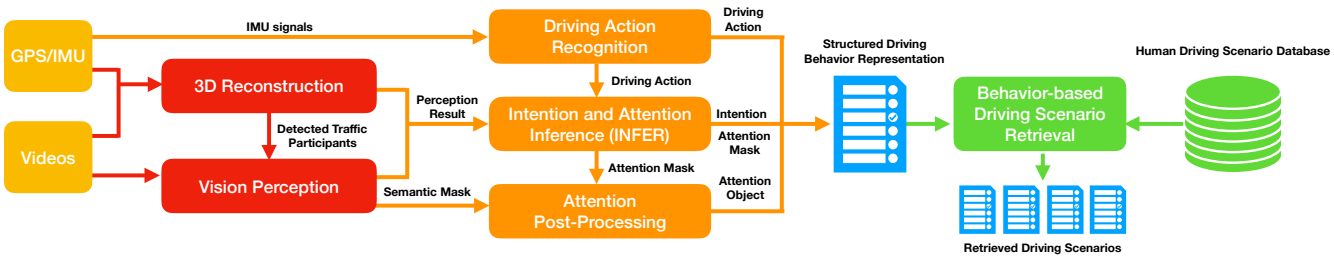



Figure 2. System architecture of *DBUS*

tensity between 0 and 1, while $\mathbf{a}_{obj}^t$ indicates which category of detected object the driver is focusing on. We consider 9 basic driving actions, 8 driver's intentions and 8 driver's attention object categories, summarized in Table 1.

On the application side, we consider driving scenario search and retrieval. Specifically, given a structured driving behavior representation $\mathcal{B} = (\mathbf{M}, \mathbf{W}, \mathbf{A})$ as the query input, one needs to retrieve the top K relevant driving scenarios $\{\mathcal{D}_k = (\mathbf{V}_k, \mathbf{S}_k)\}_{k=1}^K$ from a massive scenario database. Note that the query input could be manually designed representation or unstructured raw driving data.

## 4. System Architecture

In this section, we describe the architecture of *DBUS*. In general, *DBUS* could be divided into three modules as shown in Fig. 2:

- **Perception**: Vision perception and 3D reconstruction of driving scenarios, including detection, tracking and segmentation of traffic participants, traffic lights and signs in front-view driving videos, as well as distance estimation of traffic participants in driving scenarios.
- **Driving Behavior Analysis**: The core module of *DBUS* which processes perception results and GPS/IMU signals to generate the 3-level structured behavior representation.
- **Driving Scenario Retrieval**: Efficient behavior-based retrieval of relevant driving scenarios based on structured behavior representation.

Before diving into each module, we first provide notations and their definition in *DBUS*, summarized in Table 2. We now details the workflow of *DBUS*. First, human driving data $\mathcal{D} = (\mathbf{V}, \mathbf{S})$ including front-view camera videos $\mathbf{V}$ and GPS/IMU signals $\mathbf{S}$ feeds into the **Perception** module to generate perception results $\mathcal{P} = (\mathbf{O}, \mathbf{D}, \mathbf{L})$ where $\mathbf{O} = \{\mathbf{o}_t\}_{t=1}^T$ denotes semantic masks of detected traffic participants and traffic lights in video frames, $\mathbf{D} = \{\mathbf{d}_t\}_{t=1}^T$ denotes the distance between ego-vehicle and nearest traffic participants in the front, $\mathbf{L} = \{\mathbf{l}_t\}_{t=1}^T$ vehicle's relative location on the road based on the lane perception results. From this point on, *DBUS* discards raw videos $\mathbf{V}$ and takes $\mathcal{P} = (\mathbf{O}, \mathbf{D}, \mathbf{L})$ as the representation of full visual information. Next, perception results $\mathcal{P} = (\mathbf{O}, \mathbf{D}, \mathbf{L})$ together with GPS/IMU signals $\mathbf{S}$ feed into the **Driving Behavior Analysis** module to generate structured behavior representation $\mathcal{B} = (\mathbf{M}, \mathbf{W}, \mathbf{A})$. On the other hand, the **Driving Scenario Retrieval** takes certain behavior representation $\mathcal{B} = (\mathbf{M}, \mathbf{W}, \mathbf{A})$ and returns top K relevant driving scenarios $\{\mathcal{D}_k = (\mathbf{V}_k, \mathbf{S}_k)\}_{k=1}^K$.

Since **Perception** itself is a sophisticated independent sub-system with well-studied methodologies, we omit the methodology in **Perception** module and describe implementation details in Section 7.1. In following sections, we mainly focus on driving behavior analysis and driving scenario retrieval.

## 5. Driving Behavior Analysis

In term of the **Driving Behavior Analysis** module, the first challenge is how to predict structured representation properly, given that basic driving actions $\mathbf{M}$, driving intentions $\mathbf{W}$ and attentions $\mathbf{A}$ are highly correlated with each other. Generally speaking, a driver's behavior is switching between *active* and *passive* modes in daily driving scenarios. In the *active* mode, the driving intention would trigger all other behavior. For example, when a driver intends to turn right in an intersection, he/she would firstly slow down as approaching the intersection, pay attention to the traffic light and crossing traffic participants, then take a series of actions to complete the turn. In the *passive* mode, however, it is driver's attention that triggers driving actions and intentions. Imagining there is another vehicle cutting-in from right-side abruptly, which would first draw the driver's attention to the cut-in vehicle, then change his/her intention to left-lane-merge to avoid a collision, and finally trigger a series of driving actions such as breaking, left wheeling etc. Therefore, a key idea is to benefit from jointly infer all three types of driving behavior. On the other hand, the complexities of inferring three levels of driving behavior are different. Specifically, driving actions $\mathbf{M}$ have simple patterns and can be accurately inferred from just GPS/IMU signals $\mathbf{S}$, while driving intentions $\mathbf{W}$ and attentions $\mathbf{A}$ are relatively challenging to infer due to their own diversity and the complexity of vision representation, thus raises the need for sophisticated models.

Therefore, we develop the **Driving Behavior Analysis** module with two components: rule-based basic driving action inference, and deep learning based driving intention and attention inference.

### 5.1. Basic Driving Action Inference

We infer basic driving actions from on GPS/IMU signals $\mathbf{S}$ in a rule-based manner as follows. We focus on yaw angular velocity signal $\{vs._{yaw}^t\}_{t=1}^T$ and forward accelerate signal $\{vs._{acc}^t\}_{t=1}^T$ in GPS/IMU signals, and first smooth the noisy raw signals with hamming window. Then we classify each time step into one of three wheeling classes: left wheeling, right wheeling and going straight, by thresholding the yaw angular velocity $\{vs._{yaw}^t\}_{t=1}^T$ with pre-defined thresholds. Similarly, we also classify each time step into one of three accelerate classes: accelerate, brake and cruse at certain speed, by thresholding the forward accelerate $\{vs._{acc}^t\}_{t=1}^T$. By crossing wheeling and accelerate classes, we end up with classifying each time step into a driving action $\{\mathbf{m}_t\}_{t=1}^T$ where $\mathbf{m}_t$ belongs to one of 9 possible driving actions, summarized in Table 1.

### 5.2. Intention and Attention Inference

In this section, we introduce a deep neural network based model for driving intention and attention inference, denoted as INFER . INFER takes a set of features $\{\mathbf{o}_t, \mathbf{d}_t, \mathbf{m}_t, \mathbf{l}_t, vs._{yaw}^t, vs._{acc}^t, vs._{speed}^t\}_{t=1}^T$ as input, in which $\{\mathbf{o}_t\}_{t=1}^T$ denotes semantic masks of detected traffic participants and traffic lights, $\{\mathbf{d}_t\}_{t=1}^T$ denotes the distance between ego-vehicle and nearest traffic participants in the front, $\{\mathbf{l}_t\}_{t=1}^T$ denotes vehicle's relative location on the road, $\{\mathbf{m}\}_{t=1}^T$ refers to basic driving actions, and $\{vs._{yaw}^t, vs._{acc}^t, vs._{speed}^t\}_{t=1}^T$ represents yaw angular velocity, forward accelerate and vehicle speed in GPS/IMU signals. In a nutshell, INFER consists of two components: (1) An attention proposal network (APN) which takes above input and produces an attention mask $\{\mathbf{a}_{mask}^t\}_{t=1}^T$ representing driver's attention intensity over detected traffic participants and traffic lights. Note that the resolution of attention mask is lower than the resolution of semantic mask, to avoid generating extremely sparse attention mask. In practice, we find resolution ratio of 100 (each pixel of attention mask covers 100 pixels of semantic mask) reaches a good compromise between attention granularity and sparsity. (2) A intention inference network works with the same input feature set, except turning the semantic mask into the attention-weighted semantic mask by multiplying with the attention mask generated from APN. The intention inference network produces probabilities over all intention categories as intention output $\{\mathbf{w}_t\}_{t=1}^T$.

Fully-convolutional CNNs is adopted to embed semantic mask and attention-weighted semantic mask in two networks. Temporal information is extracted with LSTM or bi-LSTM depending on on-line or off-line mode of *DBUS*. In this paper, we consider the off-line mode in which data is collected beforehand and *DBUS* can leverage information from whole time horizon $T$. The structure of INFER is illustrated in Fig. 3. The training procedure has two stages: i) APN is first pre-trained with attention labels for a decent initialization. ii) The whole inference system (the intention inference network together with APN) is trained jointly in an end-to-end manner with both attention and intention labels.

After driving intention $\{\mathbf{w}_t\}_{t=1}^T$ and attention mask $\{\mathbf{a}_{mask}^t\}_{t=1}^T$ being inferred from the INFER model, a post-process step would match attention mask $\{\mathbf{a}_{mask}^t\}_{t=1}^T$ with semantic mask $(\{\mathbf{o}_t\}_{t=1}^T$ to find the category of the object with highest attention at each time step $\{\mathbf{a}_{obj}^t\}_{t=1}^T$. Specifically, we get the list of all detected traffic participants and traffic lights from semantic mask, and calculate the average attention for each detected object on its all pixels. Then find the object with the highest average attention intensity and output its category as $\mathbf{a}_{obj}^t$. In special, if no detection has the average attention value above a predefined threshold, we set $\mathbf{a}_{obj}^t$ as a special category to indicate no obvious attention exists in that frame.

In the end, driving intention $\mathbf{w}_t$, attention mask $\mathbf{a}_{mask}^t$, attention object category $\mathbf{a}_{obj}^t$ together with driving action $\mathbf{m}_t$ over whole time horizon $T$ would be the final output $\{\mathbf{m}_t, \mathbf{w}_t, \mathbf{a}_{obj}^t, \mathbf{a}_{mask}^t\}_{t=1}^T$ of the driving behavior analysis module.

## 6. Driving Scenario Retrieval

Given millions of driving scenarios and powerful behavior models in *DBUS*, fast retrieval methods is another key to efficiently conduct data analysis and driving behavior understanding. We design the **Driving Scenario Retrieval** module mainly for two types of retrieval task. First, we would like to get the most similar scenarios if a new driving scenario comes. Second, even if we do not have real data but can formulate our query in the form of the behavior representation $\mathcal{B}$, we can check if any scenario in *DBUS* is conceptually similar to what we seek. Searching in the original video and signal data directly is a straightforward but inferior solution in terms of both running time and result quality. Instead, we resort to behavior representations learnt from the data which provides a structured and compact depiction of the scenario embedded in a lower-dimensional space. Specifically, we take the flattened concatenation of $\{\mathbf{w}_t\}_{t=1}^T$ and $\{\mathbf{a}_{obj}^t\}_{t=1}^T$ as the feature vector of that drive scenario. This step produces a $T \times (D_b + D_a)$-dimensional vector for each scenario, where $T$ is the number of timestamps, $D_b$ is the number of different behaviors, and $D_a$ is the number of different object categories.

Given all the feature vectors, we apply ball tree for fast nearest neighbor search to retrieve the most similar scenar-

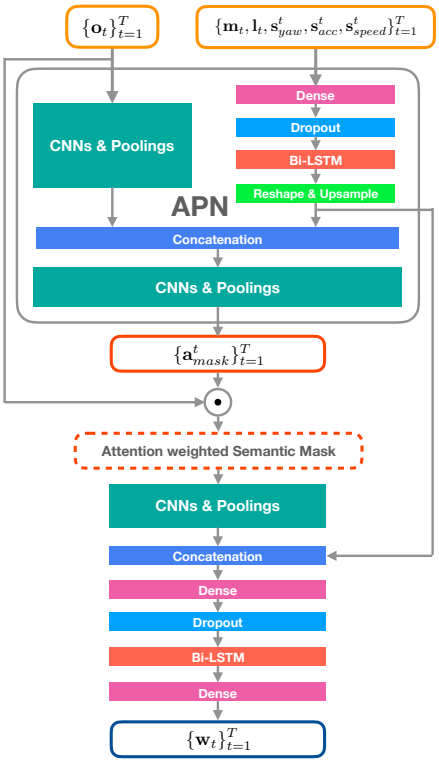



Figure 3. Model structure of INFER

ios given a query, which is also in the form of $\{\mathbf{w}_t, \mathbf{a}_{obj}^t\}_{t=1}^T$ and is either extracted from a query video or directly created by users. We take $L_1$ distance between two vectors as the distance metric. Since the original behavior and category vectors are both one-hot encoded, to speedup the tree building and the query steps, we instead convert them to categorical values and use Hamming distance as the metric. As shown in the experiments, this keeps the distance order and significantly reduced the time and space cost.

## 7. Experiments

In this work, to provide a comprehensive understanding of the effectiveness of the proposed *DBUS*, we conduct the experiments on a real-world large-scale human driving dataset from a top ride-sharing company to demonstrate that our *DBUS* can provide a total solution of mining human driving behavior with efficient and accurate analysis of human driving scenarios. Through the experiments, we answer the following questions: (1) How we utilize the learned multi-level representations from human driving scenarios? (2) How good is our proposed INFER model when provides driving intention and attention inference? (3) How we integrate multiple types of sensor data in *DBUS* to acquire a better perspective for analyzing human driving behaviors? In the remainder of this section, we illustrate the perception module, dataset, experimental design, quantitative results, etc., to answer the questions mentioned above.

### 7.1. Perception

In this section, we discuss some details of pre-trained **Perception** module of our *DBUS*, and show how we utilize the powerful perception models to get the perception results.

Table 3. Evaluations of traffic participants detection.

| | Car detection | Pedestrian detection |
|---|---|---|
| **AP** | 0.893 | 0.762 |
| **Recall** | 0.924 | 0.852 |

Table 4. Evaluations of traffic lights detection.

| | AP | Recall |
|---|---|---|
| **Traffic light detection** | 0.662 | 0.805 |

Table 5. Evaluations of lane detection.

| | Within 10 meters | Within 50 meters |
|---|---|---|
| **Precision** | 0.918 | 0.753 |
| **Recall** | 0.891 | 0.732 |

**Object detection** We leverage the power of **YOLOv3** model [19] for the traffic participants (e.g., car, person, etc.) detection and traffic lights detection. For the clarity, Table 3 only shows the performance of our pre-trained YOLOv3 model for car and pedestrian detections since these two parts play a crucial role in traffic participant detection. Furthermore, Table 4 shows the evaluations of traffic light detections. We report all the results in terms of average precision (AP) and recall only for the detected objects which have the intersection over union (IoU) larger than 0.5 with the true objects.

For each frame $\mathbf{v}$ of the video $\mathbf{V}$, we convert all its original traffic participants detections to a single-channel semantic mask and convert its traffic lights detection results to another single-channel semantic mask. Then, we concatenate these two masks to a 2-channel semantic mask. To reduce the computational complexity of *DBUS*, we resize the 2-channel semantic mask from original size $1920 \times 1080 \times 2$ to $160 \times 90 \times 2$ as the final objects $\mathbf{o}$ of frame $\mathbf{v}$.

**Lane detection** We trained **DeepLabv3** model [5] to detect lanes for each frame $\mathbf{v}$ of the video $\mathbf{V}$. Table 5 shows the evaluations of lane detection in terms of precision and recall. Note that, we only keep the detected lanes if they have the detection error less than 20cm.

Since the lane detection results are much more sparse than the traffic participants detections and traffic lights detections, we generate the vehicle location feature $\mathbf{l}$ (which shows in Table 2) for each frame $\mathbf{v}$ of video $\mathbf{V}$ according to the lane detection results instead of using them directly. The categorical feature $\mathbf{l}$ denotes the vehicle's location on the road, which represents one of the four categories: $\{$*in the middle of the road*, *on the left side of the road*, *on the right side of the road*, and *unknown location*$\}$.

**3D reconstruction** It can help us to rebuild the 3D coordinates of all detected traffic participants in video $\mathbf{V}$ of driving scenario $\mathcal{D}$. We make use of the reconstructed 3D coordinates of nearest front traffic participants to get the estimations of the following distance $\mathbf{d}$ of ego-vehicle for each frame $\mathbf{v}$. Since the accumulation of errors from video $\mathbf{V}$, signals $\mathbf{S}$, etc., within 50 meters, we have distance estimation error as $\pm 5$ meters.

## 7.2. Dataset

The human driving data $\mathcal{D}$ was collected by vehicle front-view cameras equipped with GPS/IMU sensors from millions of active service vehicles on a top ride-sharing platform. For each driving scenario $\mathcal{D}$, the length of raw video is around 25 to 30 seconds, the average FPS (frame per second) is 25, and the resolution of raw video is $1920 \times 1080$. Meanwhile, the GPS sensor samples latitude, longitude, bearing, speed, etc., per second (1Hz) and the IMU sensor retrieves the 3-axis forward accelerate signal $\{vs._{acc}\}_{t=1}^{T}$, 3-axis yaw angular velocity signal $\{vs._{yaw}\}_{t=1}^{T}$ with the retrieval frequency as 30Hz. For each human driving scenario $\mathcal{D} = (\mathbf{V}, \mathbf{S})$ and its labeled human driving behavior $\mathcal{B} = (\mathbf{M}, \mathbf{W}, \mathbf{A})$, $T$ equals to 25 and sampling rate is 5Hz. We acquired 2759 human driving scenarios from the 975 selected original driving scenarios.

## 7.3. Experimental Design

**Task and Evaluation Metrics** As we mentioned in previous sections, the human driving behavior understanding task takes driving data $\mathcal{D}$ as input and outputs the representation of human driving behavior $\mathcal{B}$. Considering the attention $\mathbf{A} = \{\mathbf{a}_{mask}^t, \mathbf{a}_{obj}^t\}_{t=1}^{T}$, we could obtain $\mathbf{a}_{obj}^t$ by mapping the $\mathbf{a}_{mask}^t$ with $\mathbf{o}_t$. Furthermore, since $\mathbf{M}$ is trivial and its generation process is heuristic, we mainly focus on inferring intention $\mathbf{W}$ and attention masks $\{\mathbf{a}_{mask}^t\}_{t=1}^{T}$ given $\mathcal{D}$. Specifically, in this work, the INFER model will take the collection of features as input and predict the corresponding intention and attention mask simultaneously. Due to the types of outputs of INFER model, we use *Mean Squared Error* (MSE) to evaluate the inferred attention mask $\{\mathbf{a}_{mask}^t\}_{t=1}^{T}$ and apply the *Categorical Accuracy* (ACC) to assess the prediction of intention $\mathbf{W}$.

**Baselines** We compare our INFER with the following baselines to show that our proposed INFER could provide accurate predictions of attention masks $\mathbf{A}$ and intention $\mathbf{W}$ at the same time.

- *SVM*, support vector machine [3]. It concatenates all the numerical features $\{\mathbf{d}_t, \mathbf{m}_t, \mathbf{l}_t, vs._{yaw}^t, vs._{acc}^t, vs._{speed}^t\}_{t=1}^{T}$ together for each time step as the input and ignores the semantic masks $\{\mathbf{o}_t\}_{t=1}^{T}$. Since the $\{\mathbf{a}_{mask}^t\}_{t=1}^{T}$ are inferred mainly based on $\{\mathbf{o}_t\}_{t=1}^{T}$, the *SVM* predicts the driving intentions only.
- *XGBoost* [6]. Similarly, it concatenates $\{\mathbf{d}_t, \mathbf{m}_t, \mathbf{l}_t, vs._{yaw}^t, vs._{acc}^t, vs._{speed}^t\}_{t=1}^{T}$ together for each time step as the input as well and ignores the semantic masks $\{\mathbf{o}_t\}_{t=1}^{T}$. The *XGBoost* predicts the driving intentions only.
- *INFER-NO-SM*, this model has the similar structure as the INFER does, but only takes $\{\mathbf{d}_t, \mathbf{m}_t, \mathbf{l}_t, vs._{yaw}^t, vs._{acc}^t, vs._{speed}^t\}_{t=1}^{T}$ as inputs **without** using semantic masks $\{\mathbf{o}_t\}_{t=1}^{T}$ of detected traffic participants and traffic lights. Conceivably, in order to generate a relatively reasonable outputs of $\{\mathbf{a}_{mask}^t\}_{t=1}^{T}$, it only infers driving intentions when it doesn't take $\{\mathbf{o}_t\}_{t=1}^{T}$ as the part of input.
- *INFER-ONLY-SM*, this model only takes semantic masks $\{\mathbf{o}_t\}_{t=1}^{T}$ as input **without** using $\{\mathbf{d}_t, \mathbf{m}_t, \mathbf{l}_t, vs._{yaw}^t, vs._{acc}^t, vs._{speed}^t\}_{t=1}^{T}$. Note that this model can still predict both attention masks $\{\mathbf{a}_{mask}^t\}_{t=1}^{T}$ and driving intentions $\{\mathbf{w}_t\}_{t=1}^{T}$.
- *INFER-NO-ATTN*, this model takes all the features as the inputs but does not utilize and infer the attention mask $\{\mathbf{a}_{mask}^t\}_{t=1}^{T}$ during the training process. Note that this model can purely output intention predictions.

## 7.4. Quantitative Results

Table 6. Results of attention prediction and intention prediction

| | **MSE** (Attention masks) | **ACC** (Intentions) |
|---|---|---|
| **SVM** | - | 0.193 |
| **XGBoost** | - | 0.258 |
| **INFER-NO-SM** | - | 0.276 |
| **INFER-ONLY-SM** | 0.032 | 0.693 |
| **INFER-NO-ATTN** | - | 0.628 |
| **INFER** | **0.025** | **0.772** |

Table 6 shows the attention mask prediction and intention prediction results in terms of *mean squared error* (MSE) and *Categorical Accuracy* (ACC) respectively. Our proposed INFER model outperforms all baselines. The results demonstrate that semantic masks and other features all play the crucial roles for attention mask predictions and intention predictions, especially for semantic masks. Additionally, we could see that attention can help our model to improve the performance of the intention inference.

Fig. 4 shows the confusion matrix based on the INFER's predictions of intentions on the held-out test set. For a better view, we normalize the rows accordingly. Under most circumstances, the INFER can infer the intentions accurately. However, since the intention *U_turn* occurs less frequently, the accuracy of its inference is slightly worse compared with other intentions.

## 7.5. Evaluations on Driving Scenario Retrieval

We evaluate our **Driving Scenario Retrieval** in terms of the running time in practice. We compare the proposed

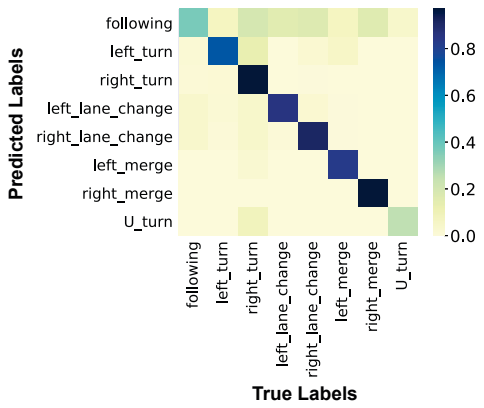



Figure 4. Confusion matrix for intention predictions on the test set using INFER.

Table 7. Timing on **Driving Scenario Retrieval** with different data size. Time unit: second.

| | | Brute-force Search | | Ball-Tree | |
|---|---|---|---|---|---|
| Distance Metric | | $L_1$ | Hamming | $L_1$ | Hamming |
| $N = 2759$ | Tree-built Time | – | – | 0.122 | 0.006 |
| | Search Time | 0.188 | 0.189 | 0.118 | 0.110 |
| $N = 10000$ | Tree-built Time | – | – | 0.643 | 0.025 |
| | Search Time | 0.181 | 0.181 | 0.103 | 0.102 |
| $N = 50000$ | Tree-built Time | – | – | 16.96 | 0.564 |
| | Search Time | 0.730 | 0.618 | 0.173 | 0.118 |
| $N = 285766$ | Tree-built Time | – | – | 656.7 | 6.168 |
| | Search Time | 0.993 | 0.719 | 0.403 | 0.288 |

Ball-tree structure with Hamming distance with brute-force search and/or $L_1$ distance. We vary $N$ from 2759, which is the number of our filtered labeled samples, to 285766, in which all scenario clips are included, to test the performance and the potential of these methods. For each $N$, we randomly choose ten queries and test these queries in different algorithm and metric settings, and report the average time. As shown in Table 7, using ball-tree is more efficient than brute-force search, and taking Hamming distance significantly reduces time to build the ball-tree and further improves search speed. Especially when all 285766 samples are included, the proposed method can still finish one query within 0.3 seconds. For ball-tree evaluations, the time is evaluated with ten parallel threads for tree building.

## 8. Case Study

Some concrete examples are shown in Fig. 5. These examples demonstrate *DBUS* can accurately infer drivers' attention in various driving scenarios, including intersection crossing, lane-following and highway cruising. As for the top-row scenario, the driver mainly focuses on the cluster of a bicyclist and a black car close to ego-vehicle in the front. A retrieval process for a scenario similar to the top-row scenario is illustrated in Fig. 6.

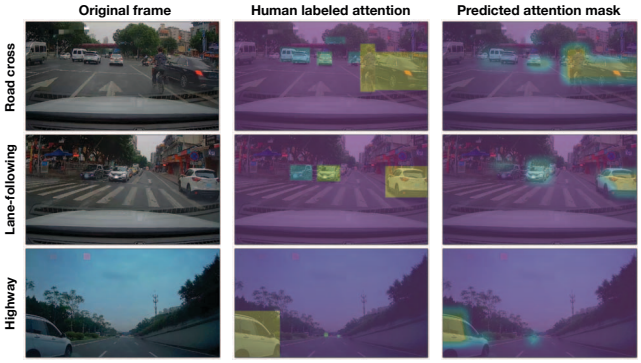



Figure 5. Examples of driver's attention inferred by *DBUS*. **Top row:** the ego-vehicle is approaching an intersection when a bicyclist is in the near front and a black car tends to cut-in. **Middle row:** the ego-vehicle is going to enter a single lane road with a white SUV parked illegally on the roadside. **Bottom row:** the ego-vehicle is passed by speeding SUV from the left side in a highway.

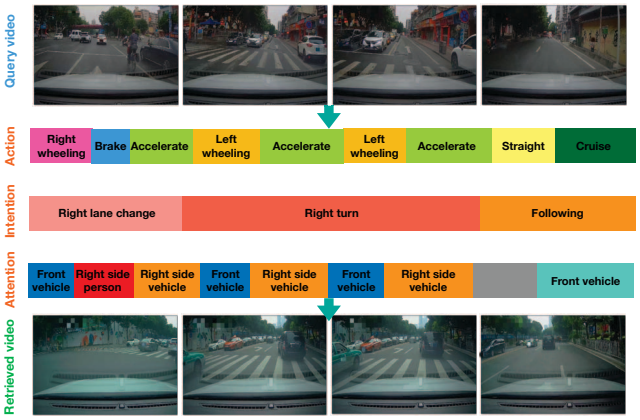



Figure 6. An example of retrieved driving scenario with query and its corresponding structured behavior representation.

## 9. Conclusion

In this paper, we presented *DBUS*, an integrated system for human driving behavior understanding. Built upon a novel structured behavior representation, powerful vision perception and behavior analysis modules, *DBUS* enables an in-depth understanding of human driving behaviors in term of basic driving actions, driving intentions and driving attentions, take a step forward towards intelligent transportation systems. *DBUS* also supports efficient behavior-based driving scenario search and retrieval, which is essential for practical application with driving scenarios collected from millions of drivers. Experiments on daily driving scenarios demonstrate its ability to recognize a driving intention and infer driving attention, as well as efficiently retrieve relevant driving scenario from a large-scale database. The system has been deployed to analysis daily human driving data collected in a top ride-sharing company and powers several downstream applications.

## References


[1] Luis M Bergasa, Daniel Almería, Javier Almazán, J Javier Yebes, and Roberto Arroyo. Drivesafe: An app for alerting inattentive drivers and scoring driving behaviors. In *Intelligent Vehicles Symposium Proceedings, 2014 IEEE*, pages 240–245. IEEE, 2014.

[2] Mariusz Bojarski, Davide Del Testa, Daniel Dworakowski, Bernhard Firner, Beat Flepp, Prasoon Goyal, Lawrence D Jackel, Mathew Monfort, Urs Muller, Jiakai Zhang, et al. End to end learning for self-driving cars. *arXiv preprint arXiv:1604.07316*, 2016.

[3] Chih-Chung Chang and Chih-Jen Lin. Libsvm: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2(3):27, 2011.

[4] Zhengping Che, Guangyu Li, Tracy Li, Bo Jiang, Xuefeng Shi, Xinsheng Zhang, Ying Lu, Guobin Wu, Yan Liu, and Jieping Ye. $D^2$-city: A large-scale dashcam video dataset of diverse traffic scenarios. *arXiv preprint arXiv:1904.01975*, 2019.

[5] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

[6] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794. ACM, 2016.

[7] Aleksa Ćorović, Velibor Ilić, Siniša Durić, Malisa Marijan, and Bogdan Pavković. The real-time detection of traffic participants using yolo algorithm. In *2018 26th Telecommunications Forum (TELFOR)*, pages 1–4. IEEE, 2018.

[8] Jiangpeng Dai, Jin Teng, Xiaole Bai, Zhaohui Shen, and Dong Xuan. Mobile phone based drunk driving detection. In *Pervasive Computing Technologies for Healthcare (PervasiveHealth), 2010 4th International Conference on-NO PERMISSIONS*, pages 1–8. IEEE, 2010.

[9] AB Ellison and S GREAVES. Driver characteristics and speeding behaviour. In *Australasian Transport Research Forum (ATRF), 33rd, 2010, Canberra, ACT, Australia*, 2010.

[10] Dariu M Gavrila. Pedestrian detection from a moving vehicle. In *European conference on computer vision*, pages 37–49. Springer, 2000.

[11] William J Horrey, Mary F Lesch, Marvin J Dainoff, Michelle M Robertson, and Y Ian Noy. On-board safety monitoring systems for driving: review, knowledge gaps, and framework. *Journal of safety research*, 43(1):49–58, 2012.

[12] Derick A Johnson and Mohan M Trivedi. Driving style recognition using a smartphone as a sensor platform. In *Intelligent Transportation Systems (ITSC), 2011 14th International IEEE Conference on*, pages 1609–1615. IEEE, 2011.

[13] Jinkyu Kim, Anna Rohrbach, Trevor Darrell, John Canny, and Zeynep Akata. Textual explanations for self-driving vehicles. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 563–578, 2018.

[14] Vashisht Madhavan and Trevor Darrell. *The BDD-Nexar Collective: A Large-Scale, Crowsourced, Dataset of Driving Scenes*. PhD thesis, Master's thesis, EECS Department, University of California, Berkeley, 2017.

[15] Clara Marina Martinez, Mira Heucke, Fei-Yue Wang, Bo Gao, and Dongpu Cao. Driving style recognition for intelligent vehicle control and advanced driver assistance: A survey. *IEEE Transactions on Intelligent Transportation Systems*, 19(3):666–676, 2018.

[16] Natasha Merat, A Hamish Jamson, Frank CH Lai, Michael Daly, and Oliver MJ Carsten. Transition to manual: Driver behaviour when resuming control from a highly automated vehicle. *Transportation research part F: traffic psychology and behaviour*, 27:274–282, 2014.

[17] Nuria Oliver and Alex P Pentland. Driver behavior recognition and prediction in a smartcar. In *PROC SPIE INT SOC OPT ENG*, volume 4023, pages 280–290. Citeseer, 2000.

[18] Vasili Ramanishka, Yi-Ting Chen, Teruhisa Misu, and Kate Saenko. Toward driving scene understanding: A dataset for learning driver behavior and causal reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7699–7707, 2018.

[19] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

[20] Heikki Summala. Traffic psychology theories: Towards understanding driving behaviour and safety factors. In *International Conference of Traffic and Transport Psychology*, 2005.

[21] Alireza Talebpour, Hani S Mahmassani, and Fabián E Bustamante. Modeling driver behavior in a connected environment: Integrated microscopic simulation of traffic and mobile wireless telecommunication systems. *Transportation Research Record: Journal of the Transportation Research Board*, (2560):75–86, 2016.

[22] Minh Van Ly, Sujitha Martin, and Mohan M Trivedi. Driver classification and driving style recognition using inertial sensors. In *Intelligent Vehicles Symposium (IV), 2013 IEEE*, pages 1040–1045. IEEE, 2013.

[23] Pengyang Wang, Yanjie Fu, Jiawei Zhang, Pengfei Wang, Yu Zheng, and Charu Aggarwal. You are how you drive: Peer and temporal-aware representation learning for driving behavior analysis. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2457–2466. ACM, 2018.

[24] Ye Xia, Danqing Zhang, Alexei Pozdnoukhov, Ken Nakayama, Karl Zipser, and David Whitney. Predicting driver attention in critical situations. *CoRR*, abs/1711.06406, 2018.

[25] Huazhe Xu, Yang Gao, Fisher Yu, and Trevor Darrell. End-to-end learning of driving models from large-scale video datasets. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2174–2182, 2017.