Exploring the Situational Awareness of Humans inside Autonomous Vehicles

Akshay Rangesh, Nachiket Deo, Kevan Yuen, Kirill Pirozhenko, Pujitha Gunaratne, Heishiro Toyoda and Mohan M. Trivedi

Institution: UCSD LISA, Toyota Collaborative Safety Research Center (CSRC)
2018 21st International Conference on Intelligent Transportation Systems (ITSC)

Presenter: Yuting Jiang ~ May 19th, 2020
Background: Humans are lazy animals....

Humans are not particularly good at tasks that require vigilance and sustained attention over long periods of time......

_Irony of automation:_ “the more advanced a control system is, the more crucial may be the contribution of the human operator”
System must decide if the driver is ready to take over control!

Even in a highly automated system!

But how?
Objectives

- Explore the circumstances where vehicles are highly automated, humans only drive intermittently.
- Explore a new set of features that have not been produced or analyzed in this context before.
Brief Introduction

- The paper proposes a pipeline that reliably generates these feature of interest, and carry out the analysis of these features on a large dataset comprising of naturalistic driving data in a variety of conditions, captured using a unique testbed.
What do we need exactly?

- Sensors: RGB cameras from different angles, depth cameras, infrared sensors...
- Algorithms:
  - Face and gaze analysis (OpenPose, GazeNet)
  - Hand analysis (HandyNet)
  - Pose Analysis (OpenPose)
  - Foot analysis (SqueezeNet)

![Diagram showing flow of sensors and algorithms]

- Coarse localization:
  - On brake pedal
  - On gas pedal
  - Hovering over brake pedal
  - Hovering over gas pedal
  - Foot in camera view
  - Foot outside camera view

- Fine localization:
  - Distance to gas pedal
  - Distance to break pedal

Legend:
- Forward
- Right
- Left
- Display
- Rearview
- Speedometer

- Blink frequency
- PERCLOS70
- Eye closure duration
- Mouth opening
What do we want?

- Observable situational awareness (OSA)
Any Example?
Looking at the Driver/Rider in Autonomous Vehicles to Predict Take-Over Readiness

Author: Nachiket Deo, Mohan M. Trivedi
Institution: UCSD LISA
Published on IEEE Transactions on Intelligent Vehicles, 2019
First, some backgrounds....

Continuous estimation of driver’s take-over readiness is critical for safe and timely transfer of control for conditionally autonomous vehicles.
While sophisticated computer vision algorithms have been proposed for driver activity analysis, relatively few works have addressed the problem of mapping driver activity to take-over readiness. ready? not ready?
First, some backgrounds....

While electroencephalogram (EEG) sensors allow for the most faithful representation of the driver’s brain activity, they are too intrusive to be viable in commercial vehicles. Another approach used in recent studies is to define takeover readiness based on take-over time and take-over quality in experimental trials with take-over requests issued to drivers performing secondary activities. However, the nature of the task restricts such experiments to simulator settings.
Objectives

- Build a naturalistic driving dataset from a conditionally autonomous vehicle
- Continuously estimate the take-over readiness of drivers using vision (non-intrusive) sensors
Methodology - Overview
Methodology - Details
A 2 hour 10 min dataset of drivers behind the wheel of a commercially available conditionally autonomous vehicle. The first study evaluating take-over readiness of drivers using a naturalistic driving dataset from a conditionally autonomous vehicle.

20 clips were rated by all 5 raters. This subset is referred as the common set, used to analyze the agreement across raters and to normalize for rater bias. Remaining 240 clips were rated by 2 different raters. referred as the expansion set. Both are used for training and testing.
Dataset

Testbed: LISA-T, built on top of a Tesla Model S. The testbed is equipped with a suite of non-intrusive sensors monitoring the driver’s state. Four high resolution cameras observe the driver’s face, hands, body pose and feet, for driver’s gaze activity, hand activity, objects held and body posture. A depth sensor observing the driver allows for analysis of the 3-D distances of the driver’s hands from the vehicle controls. Infrared sensors on its brake and gas pedals to measure the distance of the driver’s foot from each pedal. All sensors are synchronized and record at a frame rate of 30 frames per second.
Dataset

The entire dataset involves naturalistic drives with 11 drivers, with the car operating in autonomous mode on Californian multi-lane freeways. 7 drivers were male, 4 were female. 5 drivers were in the age group of 20 to 30, 3 in the age group of 30 to 40, 1 in the age group of 40 to 50, and 2 drivers were over 50. All drivers were familiarized with the Tesla autopilot functionality prior to data collection.

TABLE 1: Secondary activities in collected dataset

<table>
<thead>
<tr>
<th>Secondary task based on SHRP2 NEST nomenclature [56]</th>
<th>Present in collected data?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Talking/Singing to Self</td>
<td>✓</td>
</tr>
<tr>
<td>Talking/Singing to Passenger(s)</td>
<td>✓</td>
</tr>
<tr>
<td>Dancing</td>
<td>×</td>
</tr>
<tr>
<td>Reading</td>
<td>✓</td>
</tr>
<tr>
<td>Writing</td>
<td>×</td>
</tr>
<tr>
<td>Holding object</td>
<td>✓</td>
</tr>
<tr>
<td>Manipulating object</td>
<td>✓</td>
</tr>
<tr>
<td>Reaching for object</td>
<td>✓</td>
</tr>
<tr>
<td>Talking/listening on handheld phone</td>
<td>✓</td>
</tr>
<tr>
<td>Adjusting steering wheel buttons</td>
<td>✓</td>
</tr>
<tr>
<td>Adjusting/monitoring center stack controls</td>
<td>✓</td>
</tr>
<tr>
<td>Adjusting/monitoring other devices</td>
<td>✓</td>
</tr>
<tr>
<td>Looking for object internal to vehicle</td>
<td>✓</td>
</tr>
<tr>
<td>Looking at object external to vehicle</td>
<td>✓</td>
</tr>
<tr>
<td>Eating/drinking</td>
<td>✓</td>
</tr>
<tr>
<td>Grooming (combing hair, removing glasses)</td>
<td>✓</td>
</tr>
</tbody>
</table>
Definition of Ground Truth

Human ratings for take-over readiness:

Subjective ratings were collected from human raters viewing the sensor feed. The experiments show a high consistency in the trend of the ratings across raters, rather than their absolute value. The dataset also normalize for rater bias using a percentile based approach. The mean value of the normalized ratings, averaged across raters, is then treated as the ground truth for the models. This is termed:

Observable Readiness Index (ORI).
Definition of Ground Truth

Intraclass correlation coefficients (ICCs) are used to evaluate inter-rater agreement. Human ratings are modeled as a two-way random-effect model without interaction, assuming n observations and k raters. Under this model, the rating $x_{ij}$ assigned by rater j to clip i can be expanded as:

$$x_{ij} = \mu + r_i + c_j + e_{ij}$$

where, $\mu$ is the global average rating, $r_i$’s are the deviations based on the content of the rated clips, and $c_j$’s are the deviations due to rater bias. The $r_i$’s and $c_j$’s are independent, with mean 0 and variance $\sigma_r^2$ and $\sigma_c^2$ respectively. And finally, $e_{ij}$ is the normally distributed measurement error with zero mean and variance $\sigma_e^2$.

$$ICC(C, 1) = \frac{\sigma_r^2}{\sigma_r^2 + \sigma_c^2 + \sigma_e^2}$$

$$ICC(A, 1) = \frac{\sigma_r^2}{\sigma_r^2 + \sigma_c^2 + \sigma_e^2}$$

$$ICC(A, k) = \frac{\sigma_r^2}{\sigma_r^2 + \sigma_c^2 + \sigma_e^2_k}$$
Use a percentile-based approach on the common set for normalization of the ratings:
1. Pool and sort ratings provided by each rater on the common set to obtain rater specific look-up tables.
2. Pool and sort ratings of all raters to obtain a combined look-up table.
3. Find the percentile range of the assigned value in the rater’s lookup table.
4. Replace it with the average of all values in that percentile range in the combined look-up table.
This percentile-based lookup can be applied to the entire dataset, including the expansion set.

< 0.5, 0.5 < ICCs < 0.75, 0.75 < ICCs < 0.9, > 0.9 are indicative of poor, moderate, good, and excellent reliability, respectively.
Model Used For Estimating ORI
Model Used For Estimating ORI

Gaze Analysis:

**Inputs:** frames from the face camera

**Outputs:** frame-wise probabilities given by a Softmax layer for each gaze zone out of 9 in \{forward, left shoulder, left mirror, lap, speedometer, infotainment unit, rear-view mirror, right mirror, right shoulder\}


Hand Analysis (Camera-based):

**Inputs:** a cropped image near the localized wrist joints

**Outputs:** a hand-activity class. 4 for the left hand: \{on lap, in air (gesturing), hovering wheel, on wheel\}. 6 for the right hand: \{on lap, in air (gesturing), hovering wheel, on wheel, interacting with infotainment unit, on cup-holder\}


Model Used For Estimating ORI

Hand Analysis (Depth-based)

**Inputs:** each frame of the depth sensor

**Outputs:** the locations of the driver’s hands, a 5-dimensional feature vector: distance to the wheel, and a 4-dimensional output of the hand object classifier include 4 classes: {no-object, cell-phone/tablet, beverage/food, other item}.


Pose Analysis:

**Inputs:** each frame of the pose camera

**Outputs:** a 20-dimensional feature vector includes the x and y coordinates of the 10 body key-points.

Model Used For Estimating ORI

Foot Analysis:
IR sensors give the distance of the driver's foot to the gas and brake pedal for each frame
Model Used For Estimating ORI

Proposed LSTM model:

(a) Vanilla LSTM  (b) LSTM with key-frame weighting
Results & Analysis

260 video clips are split into 3 sets, 172 for training, 32 for validation, 56 for testing.

<table>
<thead>
<tr>
<th>Features used</th>
<th>Frame SVM</th>
<th>Vanilla LSTM</th>
<th>LSTM keyframe wts</th>
<th>Bi-LSTM keyframe wts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gaze</td>
<td>0.779</td>
<td>0.621</td>
<td>0.578</td>
<td>0.581</td>
</tr>
<tr>
<td>Hand</td>
<td>0.639</td>
<td>0.571</td>
<td>0.573</td>
<td>0.572</td>
</tr>
<tr>
<td>Pose</td>
<td>0.836</td>
<td>0.855</td>
<td>0.831</td>
<td>0.823</td>
</tr>
<tr>
<td>Foot</td>
<td>0.986</td>
<td>1.018</td>
<td>1.043</td>
<td>1.001</td>
</tr>
<tr>
<td>Gaze + Hand</td>
<td>0.611</td>
<td>0.470</td>
<td>0.463</td>
<td>0.457</td>
</tr>
<tr>
<td>Gaze + Pose</td>
<td>0.602</td>
<td>0.468</td>
<td>0.463</td>
<td>0.456</td>
</tr>
<tr>
<td>Gaze + Foot</td>
<td>0.699</td>
<td>0.467</td>
<td>0.456</td>
<td>0.449</td>
</tr>
<tr>
<td>Gaze + Hand + Pose</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gaze + Hand + Foot</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## Results & Analysis

### TABLE IV: Average inference times for components of our model

<table>
<thead>
<tr>
<th>Component</th>
<th>Run time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Face detector</td>
<td>62 ms</td>
</tr>
<tr>
<td>Gaze analysis</td>
<td>6 ms</td>
</tr>
<tr>
<td>Hand analysis (Camera based)</td>
<td>25 ms</td>
</tr>
<tr>
<td>Hand analysis (Depth based)</td>
<td>55 ms</td>
</tr>
<tr>
<td>Pose analysis</td>
<td>45 ms</td>
</tr>
<tr>
<td>Foot analysis</td>
<td>-</td>
</tr>
<tr>
<td>LSTM for estimating ORI</td>
<td>3 ms</td>
</tr>
</tbody>
</table>

Input: 30Hz  
Output: 15Hz
Results & Analysis
Results & Analysis
Advantages & Disadvantages

Advantages:

- Proposed an approach to characterize the observable take-over readiness of drivers in autonomous vehicles, map the driver activities directly to take-over readiness.

- Achieved a MAE of 0.449 on the 5 points scale.

- Key-frame detecting and weighting that leads to smoother and more reliable prediction.

Disadvantages:

- Proposed metric (ORI) is based on subjective rating.

- Doubts on foot analysis.
Key Takeaways

- Feed high dimensional inputs to the CNNs and get feature vectors, use these vectors as the input of next level
- Define ground-truth based on subjective ratings
- Use bidirectional LSTM to learn the frame-wise weight
- Identify key-frames, apply key-frame weighting to get smooth prediction
Quiz

For paper 1: What method(s) can you think of to eliminate the bias of raters?

For paper 2: According to Table 3, we can see that using pose feature brings little improvement to the MAE of the predicted ORI, in some cases, does not improve. What reasons can you think of that may cause this fact?
Q&A