A MULTI-CAMERA DEEP NEURAL NETWORK FOR DETECTING ELEVATED ALERTNESS IN DRIVERS

John Gideon[1], Simon Stent[2], and Luke Fletcher[2]

Department of Computer Science and Engineering, University of Michigan[1] and Toyota Research Institute[2]

COMPUTER SCIENCE & ENGINEERING
UNIVERSITY OF MICHIGAN

TOYOTA RESEARCH INSTITUTE

Motivation

Path to autonomous cars will require driver-facing technology

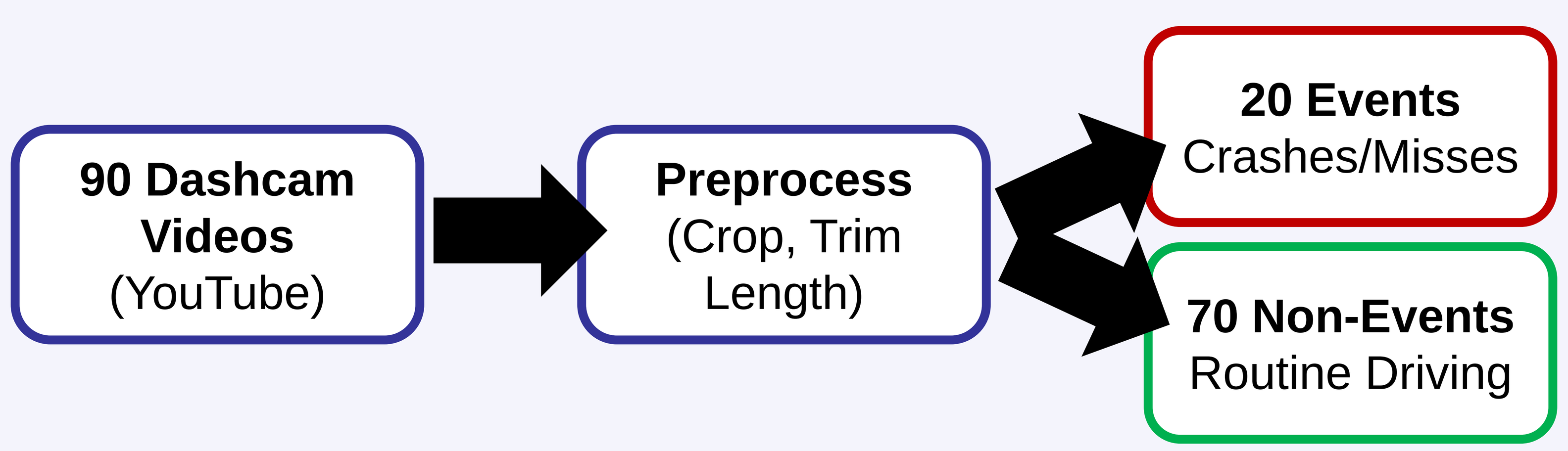
- Detect drowsiness, attention, and cognitive load
- Needed for intermediate (level-2 and 3) stages of autonomous driving

There is a need for a driver-facing affect dataset

- Many driver facing datasets without annotated affect
- Many affect datasets only consider conversations

TRIAD Dataset Collection

The Toyota Research Institute Affective Driving (TRIAD) dataset



Participants: 25 people
Task: Monitor level-2 vehicle and **react when needed**
Input: Force feedback steering wheel and pedals
Cameras: Left, center, and right view
monochromatic - 968x728 pixels - 15hz
Face Alignment: Cut to 8 seconds and use OpenFace
90 clips x 8 sec. x 25 participants x 3 cameras = **15 hours**

Driving Simulator Setup

Proctor View **Participant View**

TRIAD Dataset Annotation

Annotators: 5 people
Ground Truth: Watch each reaction clip and continuously rate between 0 and 1
Threshold: Binarize the rating using a 0.25 cutoff
Folds: Divided based on order seen by participants

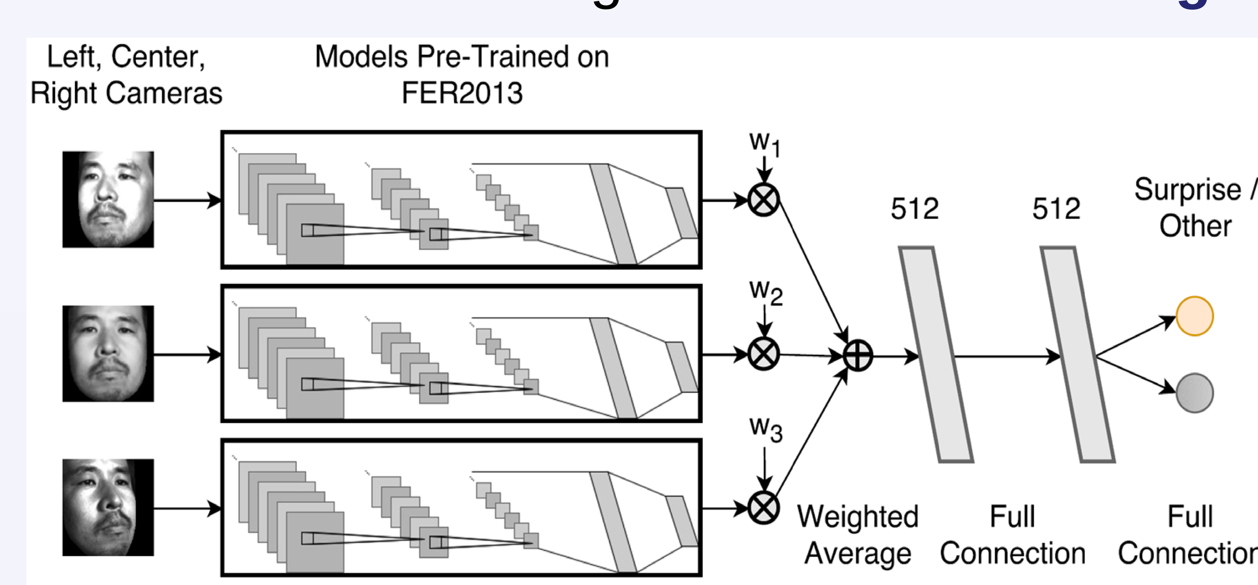
Baseline Methodology

Pretraining

- FER2013 model and dataset
- Specialize for surprise vs. other

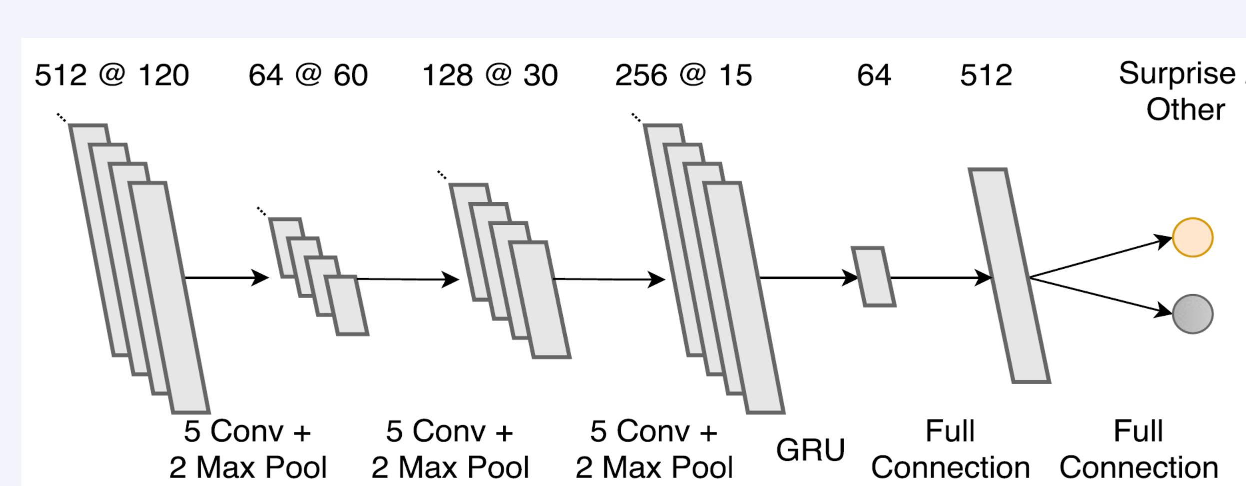
Frame-Level Modeling

- Each camera shares weights (**FER2013**)
- Weighted mean based on camera frame validity
- Also modeled **camera confidence** using 512 dense
- Used confidence with weighted **mean/max merge**



Temporal Modeling

- Uses final frame-level 512 dimensional representation
- Clip surprising if any frame-level label above threshold



Main Results

- **Metric:** Unweighted average recall (**UAR**)
- **Training:** Use single camera or different methods of combining cameras
- **Testing:** Use one, two, or all three cameras

	Cameras Used at Test Time		
Training Method	**One**	**Two**	**Three**
Left Camera	0.766	-	-
Center Camera	0.720	-	-
Right Camera	0.792	-	-
Merge by Valid	0.757	0.815	0.854
Merge by Mean	0.773	0.819	0.823
Merge by Max	**0.862**	**0.866**	**0.897**

- Merge methods show performance increase by **dealing with obstructions**
- Max merging method shows best performance by capturing **most salient features** from each camera

Conclusions

- Captured the **TRIAD dataset** combining affect annotation with simulated driving conditions (multiple, sometimes obstructed, cameras)
- Demonstrated the ability of a multi-camera system to detect driver surprise, **even when missing data**

Future work will explore how to generalize this system for the detection of anomalous events in naturalistic driving datasets

Bibliography

[1] Edmund Wascher et al., "Driver state examination – Treading new paths," *Accident Analysis & Prevention*, vol. 91, pp. 157–165, 2016.
[2] SAE On-Road Automated Vehicle Standards Committee et al., "Taxonomy and definitions for terms related to on-road motor vehicle automated driving systems," *SAE Standard J3016*, pp. 01–16, 2014.
[3] Ian J Goodfellow at al., "Challenges in representation learning: A report on three machine learning contests," in *International Conference on Neural Information Processing*. Springer, 2013, pp. 117–124.

Project Page: JohnGideon.me/projects/TRIAD