DUAL VIEWPOINT PASSENGER STATE CLASSIFICATION USING 3D CNNS

IAN TU[1]
ABHIR BHALERAO[1]
NATHAN GRIFFITHS[1]
MAURICIO MUÑOZ DELGADO[2]
ALASDAIR THOMASON[2]
THOMAS POPHAM[3]
ALEX MOUZAKITIS[2]

[1]DEPARTMENT OF COMPUTER SCIENCE, UNIVERSITY OF WARWICK, COVENTRY, UK
[2]JAGUAR LAND ROVER, ENGINEERING CENTRE, COVENTRY, UK
[3]SCHOOL OF ENGINEERING, UNIVERSITY OF WARWICK, COVENTRY, UK

CONTACT EMAIL:
I.TU@WARWICK.AC.UK

INTRODUCTION

With the advent of smart vehicles systems and autonomous vehicles, everyone inside a vehicle will become relevant in the future.

For car manufacturers, being able to monitor and predict occupant state can help maximise the experience of a vehicle journey.

For example, if the vehicle knows you are asleep, then it could adjust for a smoother ride to not disturb you.

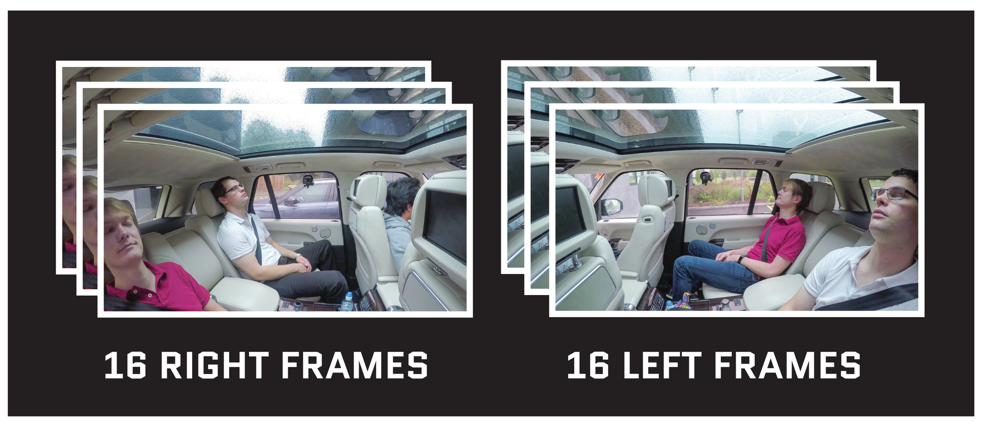
We propose a **deep learning** method to monitor and classify **passenger state** using video data captured from **dual in-vehicle cameras**.

Data was captured inside a large SUV with passengers performing various common in-vehicle actions.

The video dataset contains:

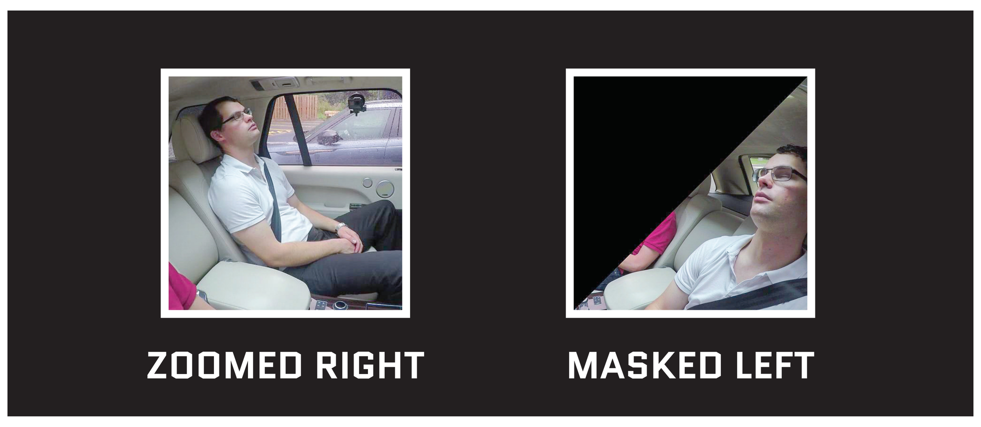
- 13 unique people.
- 7 different actions.
- 2 different viewpoints.

METHOD - THE DUAL VIEWPOINT PIPELINE



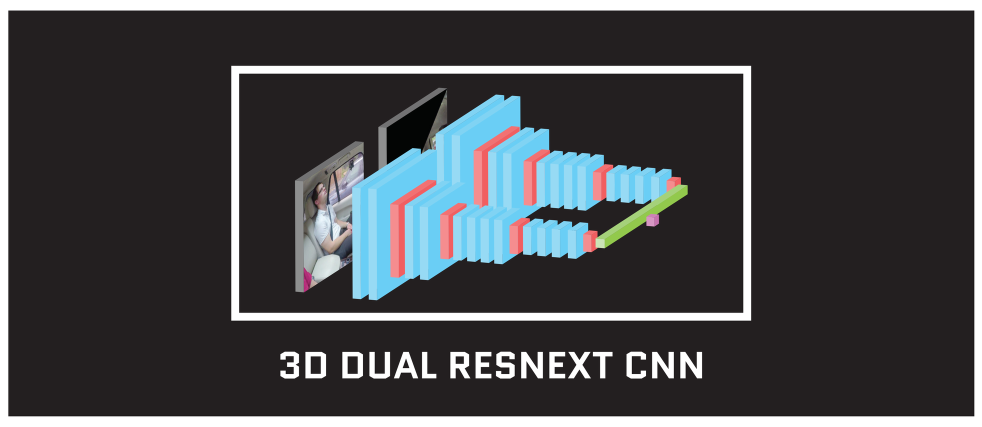
1. SETUP AND INPUT

The videos were filmed in a full-sized SUV while stationary. Two GoPro Hero 5 cameras were placed at the top corner of each backseat window using suction cups. These cameras captured video data of the backseat passengers at 4K 30fps. Then our method takes 16 consecutive frames from each viewpoint to be preprocessed.



2. PREPROCESSING

Any uneccessary information is removed from the original images. The right viewpoint images were square cropped and zoomed into the subject. The left viewpoint images were square cropped and masked so only the relevant subject remains. The output of this is then subsampled to give 16 RGB frames at 112 x 112 resolution for each viewpoint.



3. CLASSIFICATION

To classify the passenger state from the images a convolutional neural network (CNN) was used. In order to process multiple images, or video data, a 3D CNN was required, our single view model was based on the 3D ResNeXt architecture in [1], the dual view model used a fusion of two 3D ResNeXt models. The final output of this model is the state.

RESULTS – BEST

The best model was the **dual view model** using the **vehicle weights** with an overall accuracy of **75%**.

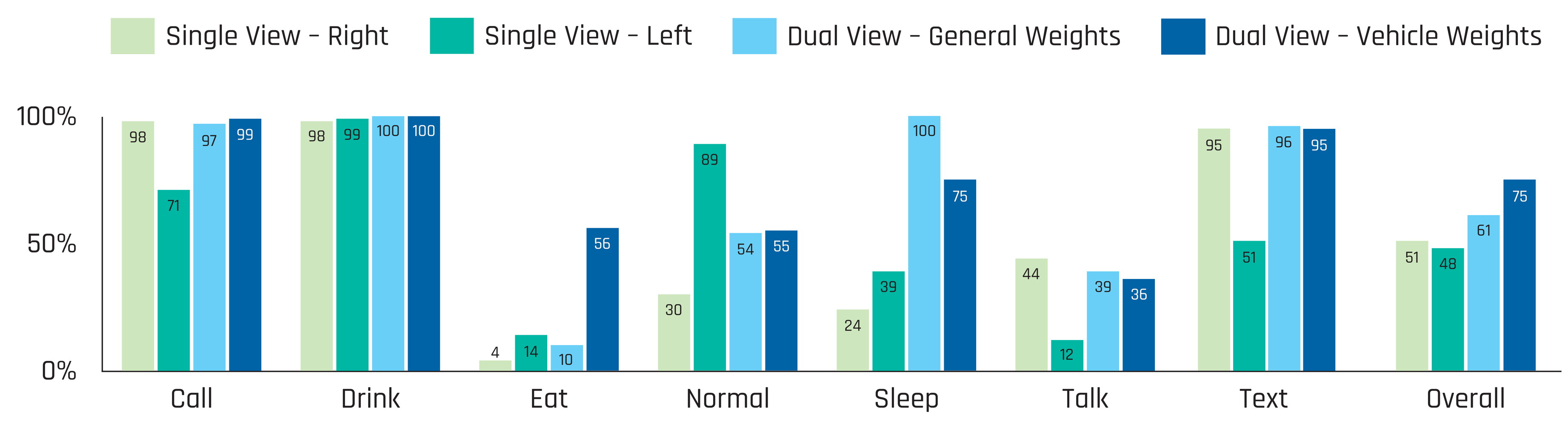
The confusion matrix below shows for each class how much was correctly classified and what it was also being misclassified as. For example, eating was misclassified as drinking 42% of the time.

Predicted State

True State	Call	Drink	Eat	Norm	Sleep	Talk	
Call	99%	0%	0%	0%	0%	0%	0%
Drink	0%	100%	0%	0%	0%	0%	0%
Eat	0%	42%	56%	2%	0%	0%	1%
Norm	10%	0%	0%	55%	0%	23%	12%
Sleep	0%	0%	0%	25%	75%	0%	0%
Talk	14%	0%	0%	33%	0%	36%	17%
Text	0%	1%	1%	2%	0%	0%	95%

RESULTS – SINGLE VS DUAL VIEW

Single viewpoint models for the **right** and **left** side only achieved around **51%** and **48%** in overall accuracy, respectively. Meanwhile, the **dual viewpoint model** which used **general weights** in the training process achieved an overall accuracy of **61%**. Furthermore, the **dual viewpoint model** which used **vehicle weights**, meaning it used the weights of the right and left single viewpoint models in the training process, achieved an overall accuracy score of **75%**, a 20% increase compared to single viewpoint models.



CONCLUSIONS

We demonstrated that the proposed method has a number of benefits:

- 3D CNNs can be used for video vehicle occupant monitoring.
- The use of dual viewpoints aids the model in overcoming occlusions and perceiving more detail.
- Transfer learning can be used between vehicles to improve performance further.

Future work will involve more subjects and include evaluating performance in moving vehicles.

ACKNOWLEDGEMENT

This work was supported by Jaguar Land Rover and the UK-EPSRC grant EP/N012380/1 as part of the jointly funded Towards Autonomy: Smart and Connected Control (TASCC) Programme. We wish to thank all who volunteered to take part in the data collection.

REFERENCES

[1] Hara, K., Kataoka, H., & Satoh, Y. (2017). Can Spatiotemporal 3D CNNs Retrace the History of 2D CNNs and ImageNet?. arXiv preprint arXiv:1711.09577.

WARWICK THE UNIVERSITY OF WARWICK

EPSRC Pioneering research and skills

JAGUAR

LAND ROVER

IV2018