DEEP PASSENGER STATE MONITORING USING VIEWPOINT WARPING

IAN TU, ABHIR BHALERAO, NATHAN GRIFFITHS, MAURICIO MUÑOZ, THOMAS POPHAM, AND ALEX MOUZAKITIS

DEPARTMENT OF COMPUTER SCIENCE, UNIVERSITY OF WARWICK, COVENTRY, UK
JAGUAR LAND ROVER, ENGINEERING CENTRE, COVENTRY, UK

CONTACT DETAILS: I.TU@WARWICK.AC.UK | ABHIR.BHALERAO@WARWICK.AC.UK | NATHAN.GRIFFITHS@WARWICK.AC.UK
AMUNOZD1@JAGUARLANDROVER.COM / TPOPHAM@JAGUARLANDROVER.COM / AMOUZAK1@JAGUARLANDROVER.COM

INTRODUCTION

With the advent of autonomous vehicles and smart vehicles systems, all the occupants inside a vehicle have become relevant.

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, it could adjust for a smoother ride to not disturb or wake you.

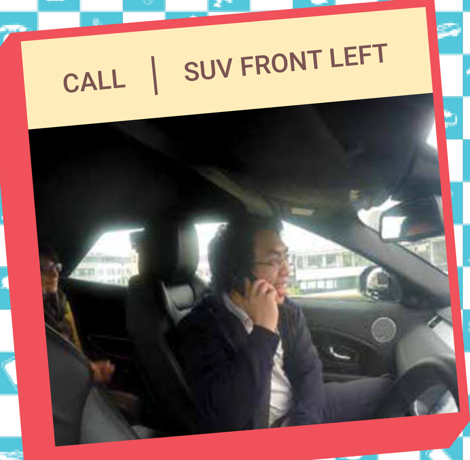
Or if the vehicle knows you are holding a drink, it can avoid or warn you of any incoming bumps to prevent you from spilling your drink.

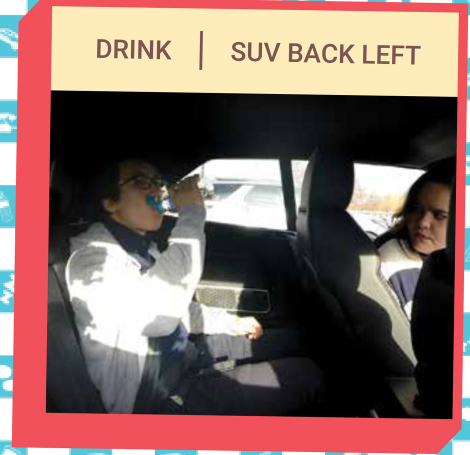
We propose a deep learning method to monitor and classify passenger state using data captured from in-vehicle cameras.

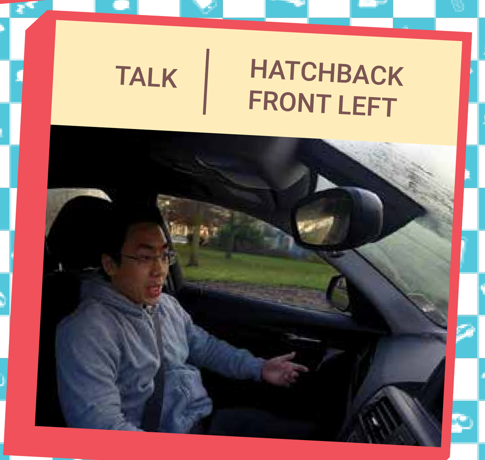
We conducted validation experiments on data acquired from two vehicles: an **SUV** and a **Hatchback**.

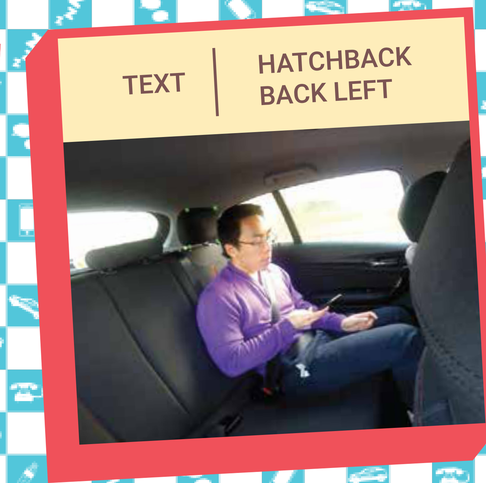
In these datasets, participants were asked to do typical in-vehicle actions whilst being filmed.

- There were **25** unique people in the **SUV** dataset.
- There were **8** different people in the **Hatchback** dataset.
- There were **5** different **viewpoints**.
- The **states** present in the dataset were phone **calling**, **drinking**, **resting**, **talking** and **texting**.



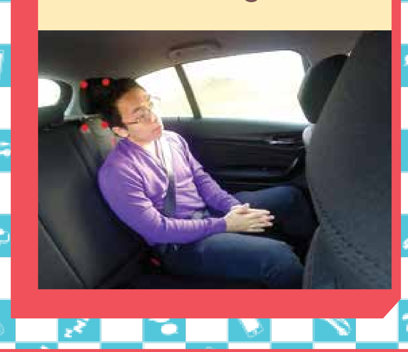




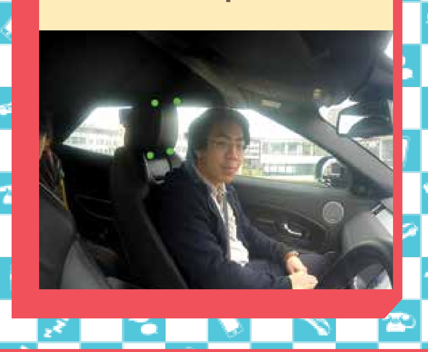


METHOD

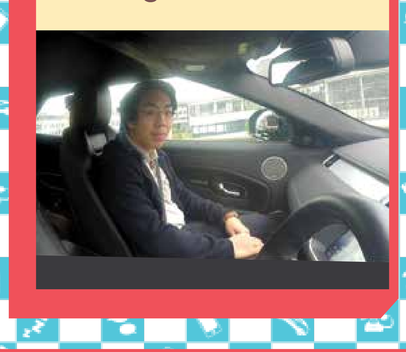
1. Label Target



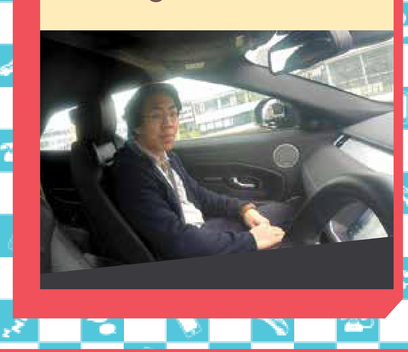
2. Label Input



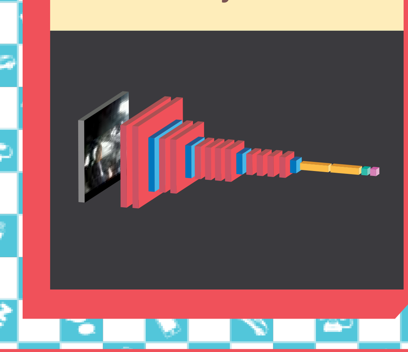
3. Align



4. Augment



5. Classify



The proposed method has two stages: image alignment and classification using a convolutional neural network.

The image alignment stage uses homography, in this phase a single viewpoint is chosen to which all the other images are mapped to and this mapping is calculated by marking corresponding points from two example images.

The second stage uses a pre-trained CNN model to predict state. The training samples were augmented with small viewpoint variations.

Image Alignment

A homography is a projective transformation from one plane to another and can be defined as the algebraic linear mapping $h : \mathbb{R}^2 \mapsto \mathbb{R}^2$ is a homography if and only if there exist a non-singular 3×3 matrix $\mathbf{H}$ such that for any point in $\mathbb{R}^2$, represented by a homogeneous coordinate x, $h(\mathrm{x}) = \mathrm{H}\mathrm{x}$ [1]. We can express this as:

$$\begin{pmatrix} u \\ v \\ 1 \end{pmatrix} \sim \begin{pmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{pmatrix} \begin{pmatrix} x \\ y \\ 1 \end{pmatrix}$$

This mapping can be solved using the Direct Linear Transform (DLT) algorithm.

The homography matrix $\mathbf{h}$ is scale-invariant, so there are only 8 unknowns to solve for. As a result, 4 pairs of non-colinear points are required, with each pair of source and target points providing two equations. The homographic matrix $\mathbf{h}$ can be found by solving, $A_i\mathbf{h} = \mathbf{0}$, with SVD, where

$$A_i = \begin{pmatrix} x_i & y_i & 1 & 0 & 0 & 0 & -u_i x_i & -u_i y_i & -u_i \\ 0 & 0 & 0 & x_i & y_i & 1 & -v_i x_i & -v_i y_i & -v_i \end{pmatrix}.$$

Viewpoint Augmentation

For data augmentation, we apply image warping to our training data set by randomising using homography given knowledge of the intrinsic camera matrix of the target viewpoint. The camera projection without lens distortion is modelled as by the perfect pin-hole camera geometry such that 3D world coordinate points $\mathbf{X}$ project to the camera plane as $\mathbf{x}$ through the product of the intrinsic and extrinsic camera matrices, K and $(R|T)$,

$$\begin{pmatrix} \mathbf{x} \\ 1 \end{pmatrix} \sim K \begin{pmatrix} R & T \\ \mathbf{0}^T & 1 \end{pmatrix} \begin{pmatrix} \mathbf{X} \\ 1 \end{pmatrix}$$

To synthesise small viewpoint changes around the principal axis of the camera i.e. $T = \mathbf{0}$, we induce random motions of the principal axis and rotations around this axis. The randomly chosen R matrix is then substituted into the above equation to generate the 3×3 random homography matrix, $H = K(R|0)$.

Classification

The CNN model architecture is based on VGG19 [2] which has shown to generalise well on a variety of datasets.

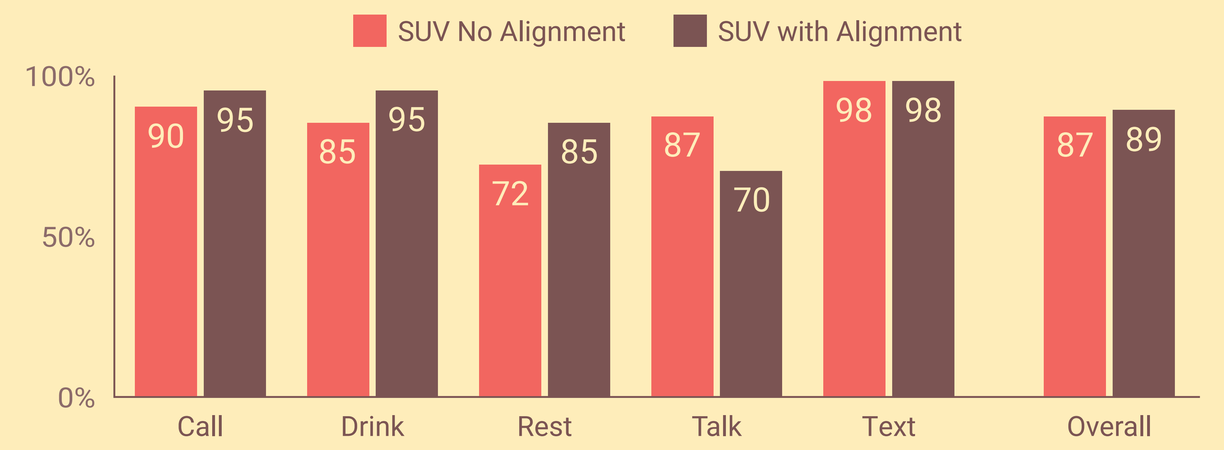
Transfer learning was used, this is where a network weights are initialised with weights from another network trained on a different dataset. Commonly, ImageNet weights are used to pre-initialise a network.

The input to the network was a 224 × 224 RGB image, with the output being one of the 5 states: calling, drinking, resting, talking or texting.

RESULTS – BETWEEN VEHICLES

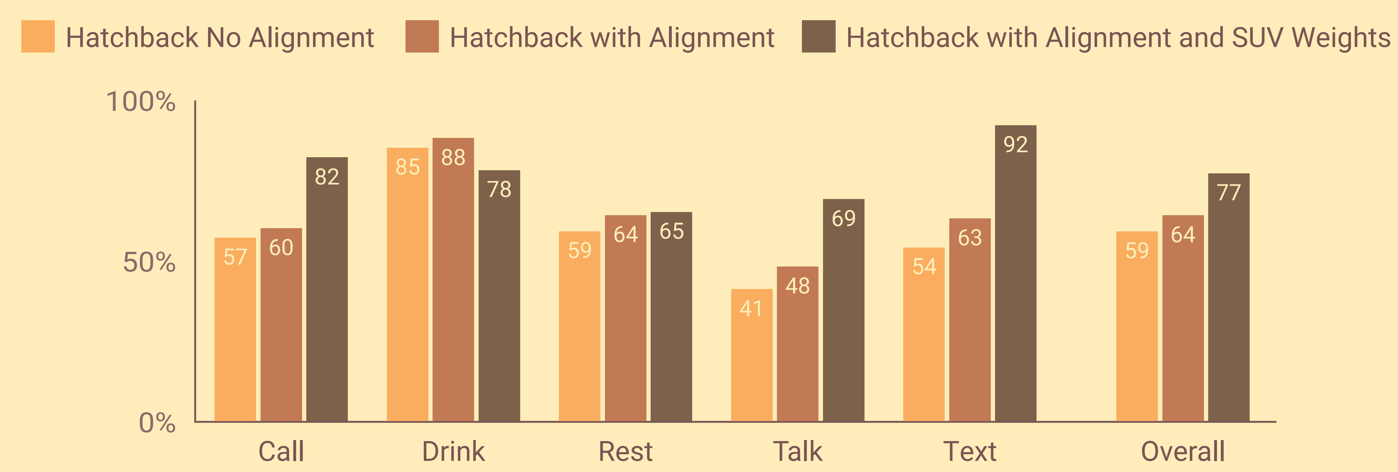
SUV Dataset

The alignment process on the SUV dataset increases the overall accuracy by 2%.



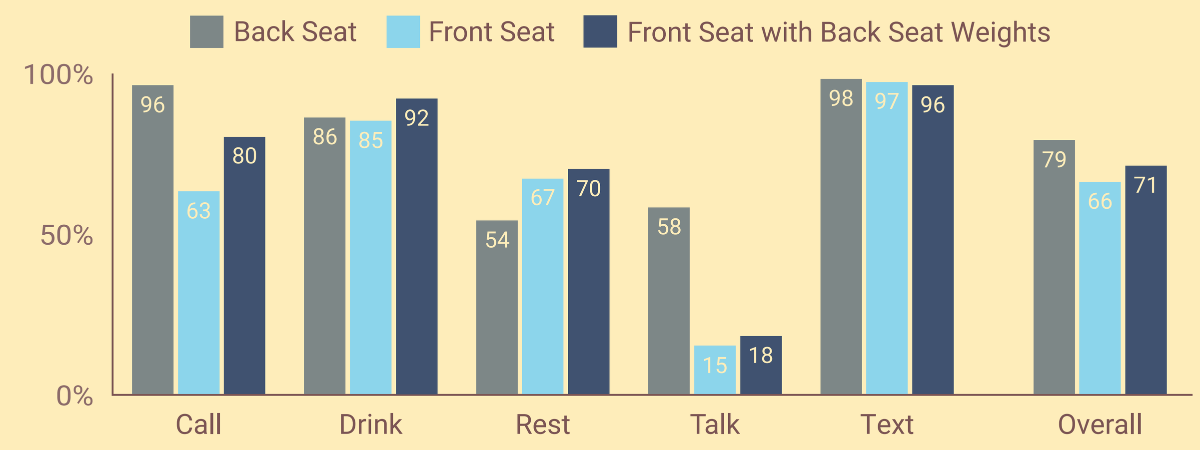
Hatchback Dataset

The dataset with alignment increases overall accuracy by 5%, with accuracy increased on all states. Transfer learning alongside the alignment process improves overall accuracy by a further 13%.



RESULTS – BETWEEN SEATS

The two datasets are combined and separated by seats. Alignment and transfer learning on the front seat dataset gives a 5% increase in overall accuracy and increases the accuracy of the most states.



CONCLUSIONS

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

- The viewpoint normalisation and augmentation allows the trained model to be re-trained with additional data to work between vehicle types and between seat positions.
- This approach allows data to be re-purposed from driver monitoring to use in occupant state classification.
- Viewpoint augmentation helps the learnt model become more robust to small viewpoint changes and enables the model to generalise better.

WARWICK
THE UNIVERSITY OF WARWICK

REFERENCES
[1] Chuan, Z., Da Long, T., Feng, Z., Li, D.Z.: A planar homography estimation method for camera calibration. In: Proc. of IEEE Comp. Int. in Robotics and Automation. vol. 1, pp. 424–429. IEEE (2003)
[2] Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)

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.

EPSRC
Pioneering research and skills

JAGUAR

LAND ROVER