Cognitive Load Estimation in the Wild

Lex Fridman[1] **Bryan Reimer**[1] **Bruce Mehler**[1] **William T. Freeman**[1,2]
[1]Massachusetts Institute of Technology
[2]Google Research

ABSTRACT

Cognitive load has been shown, over hundreds of validated studies, to be an important variable for understanding human performance. However, establishing practical, non-contact approaches for automated estimation of cognitive load under real-world conditions is far from a solved problem. Toward the goal of designing such a system, we propose two novel vision-based methods for cognitive load estimation, and evaluate them on a large-scale dataset collected under real-world driving conditions. Cognitive load is defined by which of 3 levels of a validated reference task the observed subject was performing. On this 3-class problem, our best proposed method of using 3D convolutional neural networks achieves 86.1% accuracy at predicting task-induced cognitive load in a sample of 92 subjects from video alone. This work uses the driving context as a training and evaluation dataset, but the trained network is not constrained to the driving environment as it requires no calibration and makes no assumptions about the subject's visual appearance, activity, head pose, scale, and perspective.

INTRODUCTION

Any time a study of human behavior seeks to leverage measurements of the mental aspect of human performance, the at once obvious and complicated question arises: how do we measure the state of the human mind? Cognitive load is one category of measurements that falls within this challenge. Over three decades of research in various disciplines [4] has shown cognitive load to be an important variable impacting human performance on a variety of tasks including puzzle solving, scuba diving, public speaking, education, fighter aircraft operation, and driving. The breadth and depth of the published work in this field also highlights the difficulty of identifying useful measures of cognitive load that do not interfere with the behavior of interest or otherwise influence the state of the individual being measured. Various physiological measures have been shown to be sensitive to changes in cognitive load; however, establishing practical, non-contact approaches that do not unduly constrain continuous monitoring is far from a solved problem.

Figure 1: Illustrative example of real-time cognitive load estimation during active conversation between driver and passenger. Videos of real-time cognitive load estimation in various contexts (including outside the driving context) are available on https://hcai.mit.edu/cognitive. This visualization shows (1) the video of the driver's face, (2) the 10 recent snapshots of the eye region, (3) the 30Hz cognitive load estimation plot, (4) the video of the cabin, and (5) the estimated class of cognitive load.

CHI 2018, April 21–26, 2018, Montréal, QC, Canada.

ACM ISBN 978-1-4503-5620-6/18/04...$15.00
http://dx.doi.org/10.1145/3173574.3174226

Video-based metrics that assess various characteristics of the physiological reactivity and movement of the eye in response to varying cognitive load have been studied for some time using various measures derived from eye and pupil tracking technologies. However, the bottom line is that these approaches to cognitive load estimation are generally difficult even in the lab, under controlled lighting conditions and where subject movement can be minimized [31]. Many of the video-based eye metrics used and validated in the lab become virtually impossible to detect "in the wild" in an accurate and robust way using established sensor technology.

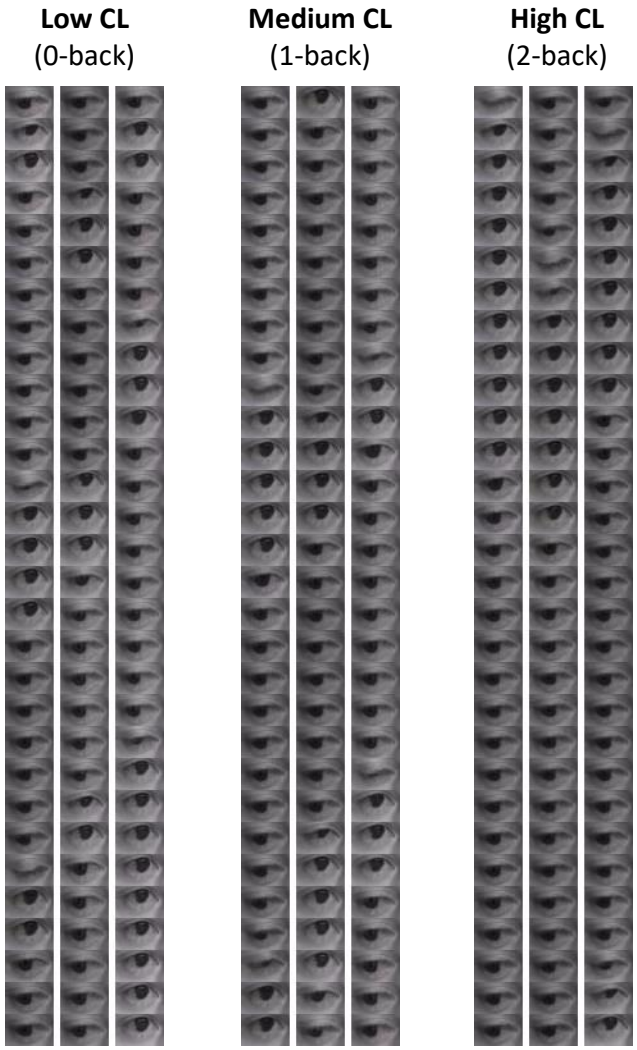


Figure 2: Three sample 90-frame 6-second video clips of a frontalized eye region. Each clip is selected from a task associated with one of 3 levels of cognitive load. Each of these video clips serve as input to both the HMM and 3D-CNN frameworks.

As a contextual grounding for our proposed approach to this problem, we review which eye-based physiological metrics have been shown to be effective predictors of cognitive load, and describe the computer vision challenges that arise when attempting to accurately estimate those metrics in outdoor environments (see the "Related Work" section). We then propose two approaches (see the "Cognitive Load from Eye Movement" section) that capture the temporal dynamics of eye movement and eye blinks in order to estimate the cognitive load of drivers engaged in three tasks of varying cognitive difficulty while driving: 0-back (low difficulty), 1-back (medium difficulty), and 2-back (high difficulty) tasks [13, 14]. Differences in the loading of these tasks have been behaviorally validated using physiological measurements (e.g. heart rate, skin conductance, and pupil diameter), self-report ratings, and detection response tasks. Furthermore, the tasks have been used as anchor points in standards development [2], as well as being compared against various in-vehicle tasks to represent a meaningful partition of cognitive load into distinctive levels [19, 20].

To the best of our knowledge, the dataset (see the "Driver Cognitive Load Dataset" section) used for evaluation is the largest on-road driver-facing video dataset of its kind, including subjects engaged in real highway driving while performing the aforementioned n-back cognitive load tasks. This dataset is unique both in the number of subjects, availability of ground truth, and the fact that is captured not in the simulator but on-road [4]. The last point is one that is worth emphasizing, because most of the work with driver cognitive load has been done in the controlled conditions of an indoor driving simulator.

The focus of our work is to develop cognitive load estimation algorithms that successfully operate in the on-road driving environment where the computer vision based detection task is difficult and the time to make a decision that ensures the driver's safety is short. We envision that robust estimation of levels of cognitive load can be integrated into an intelligent vehicle safety system both for (1) assistive technology such as future advanced driver assistance systems (ADAS) and (2) semi-autonomous vehicles that use driver state in optimizing transfer of control decisions and motion planning.

The main contributions of our work can be summarized in the following way:

1. **Novel Approach:** We propose two methods for extracting the discriminative signal in eye movement dynamics for predicting cognitive load. In the domain of data-driven approaches that are open to public validation, the use of eye movement for predicting cognitive load is novel.
2. **Open Source Implementation:** One of the key missing elements in research on cognitive load is an easily-accessible tool for detecting cognitive load in raw video of a person's face. We provide the source code and tutorial for running the code at https://hcai.mit.edu/cognitive.

RELATED WORK

For over three decades, researchers in applied psychology have looked to study human performance through measuring various aspects of cognitive load [15, 25]. Objective measurement techniques fall into two categories: (1) looking for decrements in performance measures in response to potentially cognitively loading task conditions [1] and (2) changes in physiological measures known to be responsive to

increased workload [7]. The former set of approaches measure performance of a subject on quantifiable aspect of a well-defined task. The latter set of approaches measure the physiological response produced by the subject's body through sensors such as those that monitor the electric activity of the heart (ECG), of the brain (EEG), of the skin (electrodermal activity / EDA), and through visually-identifiable metrics such as movement of the head and eyes.

The goal for our proposed cognitive load estimation system is three fold; it ought to: (1) be non-intrusive, (2) be robust to variable "in the wild" conditions, and (3) be capable of producing an accurate classification of cognition load given only a few seconds of measurement data. The non-intrusive requirement eliminates the ability to use classical ECG, EEG, and EDA recording methods. The real-world robustness and time-critical requirements eliminate many of the other options as discussed below. The open question is what metrics do provide enough discriminative signal for a non-intrusive, real-time system to effectively estimate cognitive load in the wild? In this paper, we consider one of the most promising candidate metrics that combines pupil and eyelid dynamics, and evaluate its performance on a real on-road dataset.

The term used to refer to cognitive load (CL) varies in literature depending on application context and publication venue. For the driving context, "cognitive workload", "driver workload", and "workload" are all typically used to refer to the same general concept. We consistently use the term "cognitive load" throughout this paper in discussion of related work even if the cited paper used different terminology.

An extensive meta-analysis of which eye-based metrics correlate well with cognitive load was published in 2016 and should be consulted for a detailed view of prior studies [4]. Most of the over 100 studies considered in this meta-analysis were conducted in the controlled condition of an indoor laboratory. And still, the key takeaway from work is that the impact of cognitive load on eye movement, blink rate, pupil diameter, and other eye based metrics is multi-dimensional in a number of latent variables that are difficult to account for, making its estimation (even in the lab) very challenging. Nevertheless, this prior work motivates our paper and the propose supervised-learning approach that leverages data without the need to explicitly account for the multitude of variables that impact cognitive load especially in the real-world on-road driving environment. Several cognitive load estimation methods have been proposed in recent years [3, 31], but to the best of our knowledge none have been proposed and validated in outdoor, on-road setting.

Driver eye movements have been linked to variations in cognitive load [18]. While differences in experimental approaches, data and analysis methods make direct comparisons between studies difficult, a prevailing trend across the literature suggests that gaze concentration, a narrowing of a drivers search space around the center of the roadway, occurs with increased levels of cognitive load. Some work [21] suggests that a plateau may exist in the narrowing of gaze at higher levels of demand. The overall concentration effect, often confounded with "visual tunneling", results in a reduced sensitivity across the entire visual field including the central concentrated areas [18]. As such, drivers response to threats presented across the visual field are diminished, conceptually reducing reaction time in safety critical situations. Direct comparison was made in [27] between several prevailing methodologies for computing changes in gaze dispersion (the point at which a drivers gaze measured through an eye tracker intersects a vertical plane ahead). This comparison showed that eye movements in the horizontal plane showed greatest sensitivity to changes in cognitive demand. Vertical changes show less sensitivity.

DRIVER COGNITIVE LOAD DATASET

Cognitive Load Task

The version of the n-back task considered in this paper presents subjects with single digit numbers auditorially which they need to hold in memory and repeat back verbally, either immediately (0-back), after another number has been presented (1-back), or after two additional numbers have been presented (2-back). Each of these three levels thus places an incrementally greater demand on working memory to carry out the task. The numbers are presented as a random ordering of the digits 0-9 with a typical spacing of 2.25 seconds between numbers. Single 10 item stimulus sets were employed with subjects considered in this analysis, resulting in task periods of approximately 30 seconds in duration.

In addition to the objectively defined increase in demand on working memory across the three levels of the task, objective physiological measures of workload (heart rate and skin conductance) have been shown to increase in an ordered fashion across the task levels, as do self-report ratings of workload [13]. Numerous research groups have employed this form of the n-back task as a structured method for imposing defined levels of cognitive load including ISO associated standards research [2, 16] and in work carried out for the National Highway and Transportation Safety Administration [17]. The three levels of the n-back task have been found to effectively bracket a range of real secondary tasks carried out while driving such as adjusting the radio or entering an address into a navigation system [19].

On-Road Data Collection

Data for the evaluation of methods proposed in this paper was drawn from two on-road studies that included the n-back cognitive load reference task [13, 20]. Subjects were trained on the n-back in the lab and given additional practice while parked in the study vehicle prior to going on-road. Data collection occurred on a multilane, divided interstate highway, and a minimum of 30 minutes of adaptation to driving was provided prior to subjects engaging in the n-back or other experimental tasks. The ordering of each of the difficulty levels of the n-back task was randomized across the analysis sample.

The study vehicle was instrumented with a customized data acquisition system for time synchronized recording of vehicle information from the CAN bus, a medical grade physiological monitoring unit for recording EKG, EDA and other signals, a FaceLAB eye tracking system, a microphone, and a series of

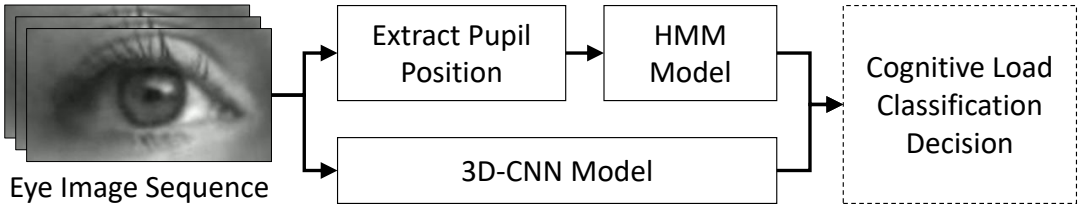


Figure 3: A high-level diagram of the cognitive load estimation task as presented in this paper. The input is a sequence of 90 eye region images from a 6 second video sequences. The output is a cognitive load level classification decision. The HMM approach requires explicit feature extraction prior to classification. The 3D-CNN approach is end-to-end in that it performs both the spatial and temporal feature extraction implicitly.

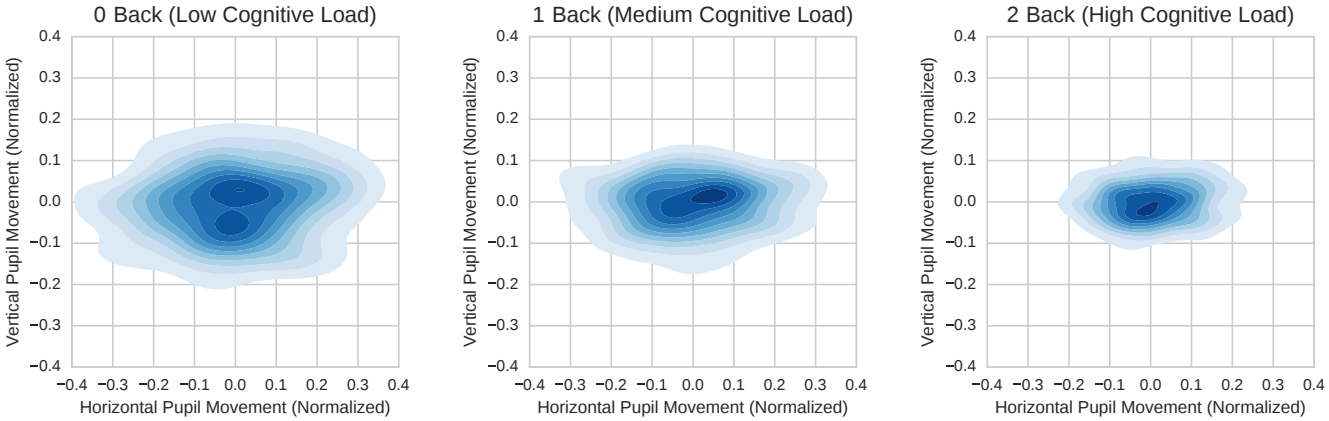


Figure 4: Kernel density estimation (KDE) visualization of the relative pupil movement for each of the 3 cognitive load levels. The axes are normalized by the "intraoccular distance" defined as the distance between the estimated landmark positions of the two corners of the eye.

cameras for tracking various aspects of driver behavior and the surrounding driving environment. The camera used for the analysis in this report was positioned to capture a subject's whole head and upper torso with sufficient margins to keep the face in view as the driver showed normal forward oriented movements while driving; images were recorded in black and white using a 30 fps capture rate and a 640x480 image size.

The video clips associated with each of the 3 cognitive load levels were extracted from the raw on-road footage and were annotated for the computer vision tasks described in the next section. For each frame in the dataset, these annotations include: (1) the bounding box of the driver's face, (2) 43 face landmarks, (3) visibility state of the pupil, (4) 25 eyelid landmarks, and (5) 14 iris and pupil landmarks when the pupil is annotated as visible.

COGNITIVE LOAD FROM EYE MOVEMENT

The high level architecture of the cognitive load estimation system proposed in this paper is shown in Fig. 3. The input is a video clip of the eye and the output is a cognitive load classification decision as to the level of cognitive load the person in the video clip is under.

The duration of the video clip used for classification is fixed to 6 seconds and is downsampled from 30 fps to 15fps. The result is a sequence of 90 grayscale eye region images. We refer to this 6 second period of data as a "classification epoch". For the 3D-CNN approach it includes the raw images. For the HMM approach it includes the extracted pupil position and blink state.

As described in the "Related Work" section, prior work in analysis of gaze patterns has shown some correlation between dispersion of gaze and cognitive load. In order to motivate the estimation task detailed below, we first perform a similar type of analysis on the patterns of gaze in the dataset we use for evaluation. Fig. 4 shows the kernel density estimation (KDE) visualization of normalized pupil position movement for each of the 3 cognitive load levels. The 3 KDE functions show a decreased dispersion of gaze as the level cognitive load increases. In particular, the change in horizontal dispersion is greater than the change in vertical dispersion (matching results of prior studies [27]). In other words, in aggregate, gaze does seem to provide some signal for discriminating between levels of cognitive load. The question we answer in the following subsections is how we can pull that signal out for predictive purposes based on only a 6 second video clip of a driver's face.

Preprocessing Pipeline

The initial input to the cognitive load estimation system is a 6 second video clip of a driver's face taken from a longer video where the driver was performing either 0-back, 1-back, 2-back secondary tasks while driving. This video clip is first downsampled in time from 30fps to 15fps by removing either other frame. The result is 90 temporally-ordered images of a driver's face.

The preprocessing operations shown in Fig. 5 are repeated on each frame without placing constraints on temporal consistency. First step is face detection. For this task we use a Histogram of Oriented Gradients (HOG) combined with a linear SVM classifier, an image pyramid, and sliding window detection scheme implemented in the DLIB C++ library [10]. The performance of this detector has lower FAR than the widely-used default Haar-feature-based face detector available in OpenCV [12] and thus is more appropriate for our application. Face alignment in the preprocessing pipeline is performed on a 43-point facial landmark that includes features of the eyes, eyebrows, nose and mouth as shown in Fig. 5. The active appearance model (AAM) algorithm for aligning the 43-point shape to the image data uses a cascade of regressors as described in [9]. The characteristics of this algorithm most important to driver gaze localization is it has proven to be robust to partial occlusion and self-occlusion.

Both eye regions are extracted from the image of the face using the localized fiducial points for the eyes. We only choose one of the eye regions for the input to the cognitive load estimation model. Specifically, we choose the eye region that is closer to the camera in estimated world coordinates. This is done by mapping the face aligned features to a generic 3d model of a head. The resulting 3D-2D point correspondence is used to compute the orientation of the head with OpenCV's SolvePnP solution of the PnP problem [23]. Once the 43 fiducial points have been localized, and the eye region has been selected, we use the face frontalization algorithm in [8] to synthesize a frontal view of the driver's face. This is done for the purpose of frontalizing the eye region but in practice full face frontalization has the indirect effect of producing a more robust eye region frontalization than if the synthesis is performed on eye-aligned landmarks alone.

The same AAM optimization as done for face alignment is performed for 25 points on the eye lids of the selected eye. The aligned points and the raw image is loaded into a standard 2D CNN (with 3 convolutional layers and 2 fully connected layers) to predict the visibility state of the pupil as it relates to the occlusion caused by the blinking action. Finally, if the eye is deemed to sufficiently open for the pupil to be visible, the AAM process is repeated one last time with 39 points that includes 14 extra points localizing the iris and the pupil.

Steps 4, 5, and 6 in the preprocessing pipeline (see Fig. 5) serve as the feature extraction step for the HMM cognitive load estimation approach described next. However, it also allows for a higher accuracy re-alignment of the eye region image provided as input to the network in the 3D-CNN approach. In practice, this re-alignment resulted in a small reduction of classification performance. We hypothesize that imperfect alignment of the eye region serves as a data augmentation technique for the training set allowing for the resulting model to generalize more effectively.

Pupil Trajectories with HMMs

The result of the preprocessing pipeline is an estimate of the pupil position. The pixel position of the pupil is normalized by by the magnitude of the line segment between the two corners of the eye. To determine the normalized position, the midpoint of the "intraoccular" line segment is used as the origin, the x-axis is made parallel to it, and the y-axis is made perpendicular to it. When the pupil is not visible the last known position is assigned or if no prior position was determined, a position of $(0,0)$ is assigned.

For the purpose of modeling cognitive load as a set of Hidden Markov Models (HMMs), each 6-second classification epoch is defined as a sequence of 90 normalized pupil positions. A bivariate continuous Hidden Markov Model (see [11]) is used for this purpose. An HMM is constructed and trained for each of the 3 cognitive load classes. The number of hidden states in each HMM is set to 8, which does not correspond to any directly identifiable states in the cognitive load context. Instead, this parameter was programmatically determined to maximize classification performance. One HMM is constructed for each of the 3 classes in the classification problem. The HMM model parameters are learned using the GHMM implementation of the Baum-Welch algorithm [22].

The result of the training process are three HMM models. Each model can be use to provide a log-likelihood of an observed sequence. The HMM-based classifier then takes a 90-observation sequence, computes the log-likelihood from each of the 3 HMM models, and returns the class associated with the maximum log-likelihood.

Raw Eye Region Video with 3D-CNNs

In contrast with the HMM approach in the "Pupil Trajectories with HMMs" section that performs "late temporal fusion" after the feature extraction step, the three-dimensional convolutional neural network (3D-CNN) approach performs "early temporal fusion" by aggregating temporal dynamics information in conjunction with the spatial convolution on the raw grayscale image data of the eye region. The architecture for the network used is shown in Fig. 6.

The input to the network is a temporally-stacked sequence of grayscale images. Unlike prior methods we do not explicitly provide dense optical flow of the eye region as input to the network [28]. Instead, we structure the network in a way that allows it to learn the salient motion both in terms of pupil movement and eyelid movement. See the "Results" section for discussion of the implicit learning of both spatial and temporal characteristics of eye region dynamics. Each image is converted to grayscale and resized to 64x64. As described above, a classification epoch include 90 of images. Therefore, the input to the 3D-CNN network is $1 \times 90 \times 64 \times 64$ which includes 1 grayscale channel, 90 temporally ordered images, 64 pixels in height, and 64 pixels in width.

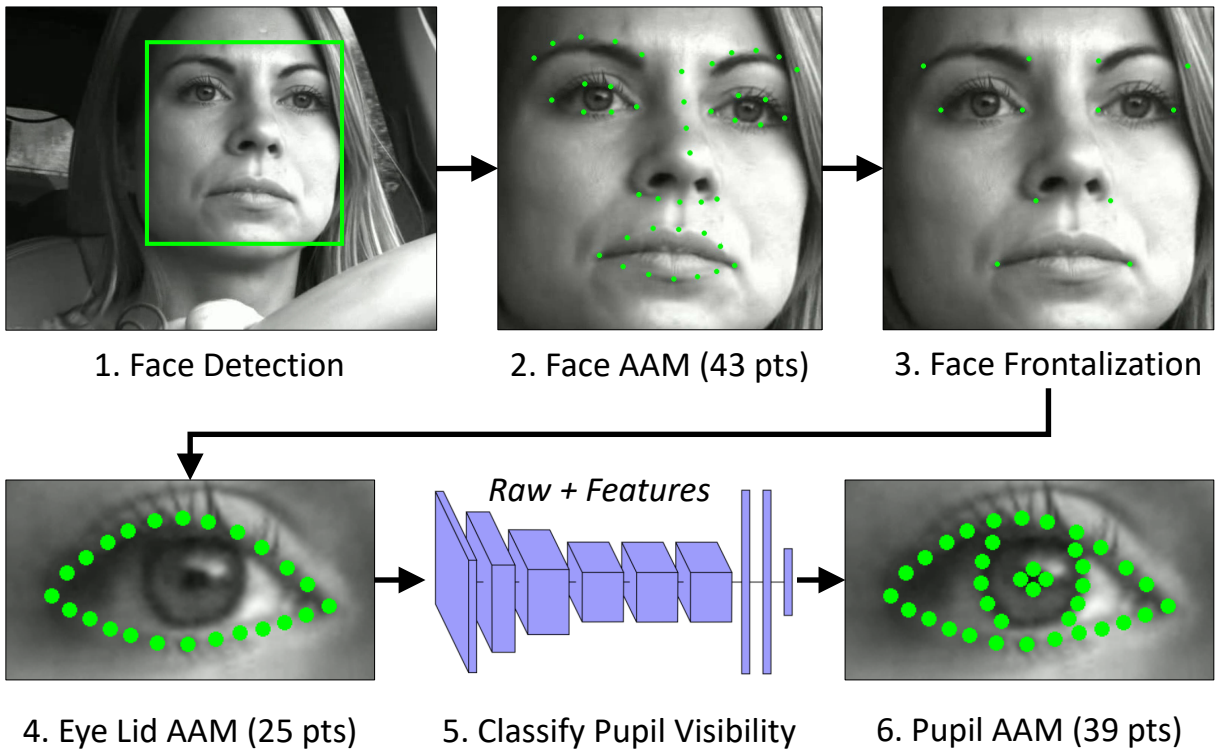


Figure 5: Image preprocessing steps that prepare the data for the two classification approaches. The steps go from the original image of the driver's head in arbitrary pose to the aligned, frontalized eye region image. Features extracted as part of steps 4, 5, and 6 are used as input only by the HMM approach. These latter steps are optimally used by the 3D-CNN approach to ensure proper alignment of the raw grayscale input to network. See the "Preprocessing Pipeline" section for details on these steps and the "Raw Eye Region Video with 3D-CNNs" section on how they are optionally leveraged by the the 3D-CNN approach.

The network (shown in Fig. 6) is 7 convolutional layers and 2 fully connected layers. Convolutional layers are followed by max-pooling layers. The softmax layer at the end produces the 3-class prediction. Based on the exploration of convolutional kernels in [24, 26] we use a kernel size of $3 \times 3 \times 3$ with temporal dimension size of 3 and spatial dimensions of size 3 as well. 128 filters are used at each convolutional layer with stride $1 \times 1 \times 1$. All pooling layers are sized $2 \times 2 \times 2$ with stride $2 \times 2 \times 2$. Appropriate padding is used such that the size of the image is maintained through the convolutional layers.

RESULTS

Evaluation of both approaches was performed using 10 randomly selected training-testing splits of a dataset of 92 subjects. An 80-20 split across subjects was used which corresponded to 74 subjects in the training set and 18 subjects in the testing set. In any one instance of cross-validation, no subject appeared in both the training and the testing set.

The HMM models were trained using a GHMM implementation of the Baum-Welch algorithm [22]. The 3D-CNN models were training using a TensorFlow implementation of stochastic gradient descent with mini-batches of 100 video clips per subject, and a total of 80 training epochs. The results achieved by both methods are shown as confusion matricies in Fig. 7. These are average over the 10 cross-validation folds. For the 3-class cognitive load estimation problem as defined in this paper, the HMM approach achieves an average accuracy of 77.7% and the 3D-CNN approach achieve 86.1%.

Both the HMM and 3D-CNN approaches perform two tasks: (1) extract pupil position and blink state and (2) track changes in those variables over time. The HMM approach does both explicitly, while the 3D-CNN approach does both implicitly (end-to-end). To confirm the latter, we investigated what it is that the 3D-CNN network learns by using a deconvolutional network [30] as detailed in [29] for the visualization task to probe each of layers in the network. For the initial frames, the 3D-CNN learns the spatial characteristics of the eye region. For the remainder of the frames, it switches to activating on motion of both the pupil and the eyelids in the remainder of the frames. In other words, it performs the explicit pupil detection task of the HMM approach implicitly in an end-to-end way. This is a promising observation, because pupil detection in visible light has been repeatedly shown to be very difficult under a range of lighting variations and vehicles vibrations present in the on-road driving context [5, 6].

Given the size of the dataset, and the inherent complexity of cognitive load estimation as a task, especially in a real-world environment, the results are impressive for both HMM and

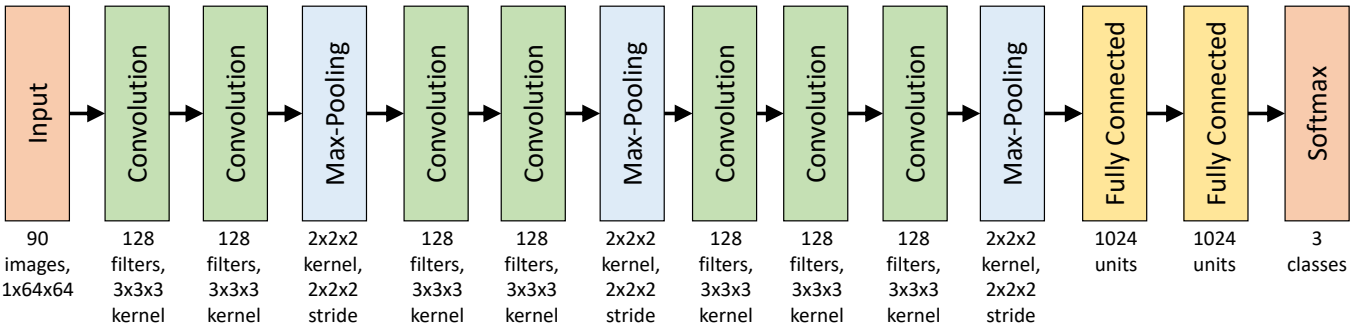
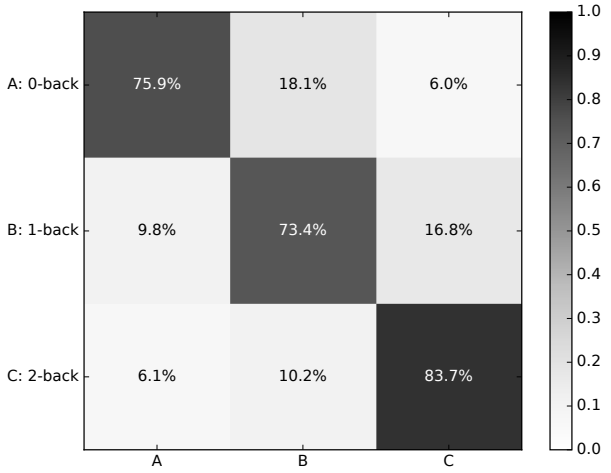
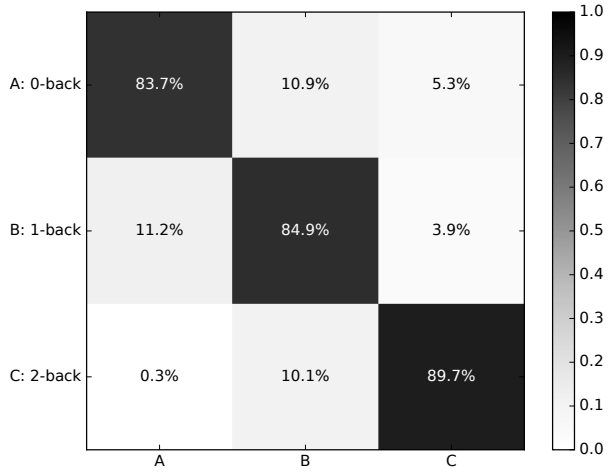


Figure 6: The 3D-CNN architecture with 90 stacked $1 \times 64 \times 64$ images as input and 3 class prediction as output. Each of the convolutional layers has 128 filters with $3 \times 3 \times 3$ kernels of equal size in temporal and spatial dimensions. The 2 fully connected layers have 1024 output units.



(a) HMM approach. Average accuracy: 77.7%



(b) 3D-CNN approach. Average accuracy: 86.1%

Figure 7: Confusion matrices for the two cognitive load estimation approaches proposed in this paper. The results are averaged over 10 random training-testing splits where the group of subjects in the training set was always distinct from the group of subjects in the testing set.

3D-CNN approaches. It remains an open question of how well this approach generalizes beyond the context of a driver engaging in a secondary task (n-backs in this case) while driving. Nevertheless, as the results indicate, the metric of eye region dynamics as captured through visible light and processed through modern computer vision approaches is a promising one for the general cognitive load estimation task.

CONCLUSION

Cognitive load estimation in the wild is an important and challenging problem. We propose two computer vision based approaches for addressing this problem. The first approach uses HMM models. The second approach uses a 3D-CNN model. Both are based on temporal dynamics of the eye over a period of 6 seconds as captured by 90 visible light video frames. The HMM method tracks explicitly-extracted pupil positions over time, while the 3D-CNN method operates end-to-end on the raw grayscale eye region image sequences. On a dataset of 92 subjects, the HMM approach achieves 77.7% average accuracy and the 3D-CNN approach achieves 86.1%. The source code for the implementation of both approaches is made publicly available.

Acknowledgments

This work was in part supported by the Toyota Class Action Settlement Safety Research and Education Program. The views and conclusions being expressed are those of the authors, and have not been sponsored, approved, or endorsed by Toyota or plaintiffs class counsel.

REFERENCES

1. Linda S Angell, J Auflick, PA Austria, Dev S Kochhar, Louis Tijerina, W Biever, T Diptiman, J Hogsett, and S Kiger. 2006. *Driver Workload Metrics*. Technical Report.

2. Marie-Pierre Bruyas, Laëtitia Dumont, and France Bron. 2013. Sensitivity of Detection Response Task (DRT) to the driving demand and task difficulty. In *Proceedings of the Seventh International Driving Symposium on Human*

Factors in Driver Assessment, Training, and Vehicle Design. 64–70.

3. Siyuan Chen and Julien Epps. 2013. Automatic classification of eye activity for cognitive load measurement with emotion interference. *Computer methods and programs in biomedicine* 110, 2 (2013), 111–124.
4. Melissa Patricia Coral. 2016. *Analyzing Cognitive Workload Through Eye-related Measurements: A Meta-Analysis*. Ph.D. Dissertation. Wright State University.
5. Lex Fridman, Daniel E. Brown, William Angell, Irman Abdic, Bryan Reimer, and Hae Young Noh. 2016a. Automated Synchronization of Driving Data Using Vibration and Steering Events. *Pattern Recognition Letters* (2016).
6. Lex Fridman, Joonbum Lee, Bryan Reimer, and Trent Victor. 2016b. Owl and Lizard: Patterns of Head Pose and Eye Pose in Driver Gaze Classification. *IET Computer Vision* (2016).
7. Eija Haapalainen, SeungJun Kim, Jodi F Forlizzi, and Anind K Dey. 2010. Psycho-physiological measures for assessing cognitive load. In *Proceedings of the 12th ACM international conference on Ubiquitous computing*. ACM, 301–310.
8. Tal Hassner, Shai Harel, Eran Paz, and Roee Enbar. 2015. Effective face frontalization in unconstrained images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4295–4304.
9. Vahid Kazemi and Josephine Sullivan. 2014. One millisecond face alignment with an ensemble of regression trees. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*. IEEE, 1867–1874.
10. Davis E. King. 2009. Dlib-ml: A Machine Learning Toolkit. *Journal of Machine Learning Research* 10 (2009), 1755–1758.
11. Sergey Kirshner. 2005. *Modeling of multivariate time series using hidden Markov models*. Ph.D. Dissertation. UNIVERSITY OF CALIFORNIA, IRVINE.
12. Rainer Lienhart and Jochen Maydt. 2002. An extended set of haar-like features for rapid object detection. In *Image Processing. 2002. Proceedings. 2002 International Conference on*, Vol. 1. IEEE, I–900.
13. Bruce Mehler, Bryan Reimer, and Joseph F Coughlin. 2012. Sensitivity of physiological measures for detecting systematic variations in cognitive demand from a working memory task: an on-road study across three age groups. *Human factors* 54, 3 (2012), 396–412.
14. Bruce Mehler, Bryan Reimer, and JA Dusek. 2011. MIT AgeLab delayed digit recall task (n-back). *Cambridge, MA: Massachusetts Institute of Technology* (2011).
15. Fred Paas, Juhani E Tuovinen, Huib Tabbers, and Pascal WM Van Gerven. 2003. Cognitive load measurement as a means to advance cognitive load theory. *Educational psychologist* 38, 1 (2003), 63–71.
16. Bastian Pfleging, Drea K Fekety, Albrecht Schmidt, and Andrew L Kun. 2016. A Model Relating Pupil Diameter to Mental Workload and Lighting Conditions. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 5776–5788.
17. Thomas A Ranney, GH Baldwin, Larry A Smith, Elizabeth N Mazzae, and Russell S Pierce. 2014. *Detection Response Task (DRT) Evaluation for Driver Distraction Measurement Application*. Technical Report.
18. Miguel A Recarte and Luis M Nunes. 2003. Mental workload while driving: effects on visual search, discrimination, and decision making. *Journal of experimental psychology: Applied* 9, 2 (2003), 119.
19. Bryan Reimer, Bruce Mehler, J Dobres, and JF Coughlin. 2013. The effects of a production level "voice-command" interface on driver behavior: summary findings on reported workload, physiology, visual attention, and driving performance. (2013).
20. Bryan Reimer, Bruce Mehler, Jonathan Dobres, Hale McAnulty, Alea Mehler, Daniel Munger, and Adrian Rumpold. 2014. Effects of an'Expert Mode'Voice Command System on Task Performance, Glance Behavior & Driver Physiology. In *Proceedings of the 6th international conference on automotive user interfaces and interactive vehicular applications*. ACM, 1–9.
21. Bryan Reimer, Bruce Mehler, Ying Wang, and Joseph F Coughlin. 2012. A field study on the impact of variations in short-term memory demands on drivers' visual attention and driving performance across three age groups. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 54, 3 (2012), 454–468.
22. Alexander Schliep, Benjamin Georgi, Wasinee Rungsarityotin, I Costa, and A Schonhuth. 2004. The general hidden markov model library: Analyzing systems with unobservable states. *Proceedings of the Heinz-Billing-Price* 2004 (2004), 121–135.
23. Gerald Schweighofer and Axel Pinz. 2008. Globally Optimal O (n) Solution to the PnP Problem for General Camera Models.. In *BMVC*. 1–10.
24. Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).
25. John Sweller, Paul Ayres, and Slava Kalyuga. 2011. Measuring cognitive load. In *Cognitive load theory*. Springer, 71–85.
26. Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning spatiotemporal features with 3d convolutional networks. In *2015 IEEE International Conference on Computer Vision (ICCV)*. IEEE, 4489–4497.

27. Ying Wang, Bryan Reimer, Jonathan Dobres, and Bruce Mehler. 2014. The sensitivity of different methodologies for characterizing drivers' gaze concentration under increased cognitive demand. *Transportation research part F: traffic psychology and behaviour* 26 (2014), 227–237.

28. Joe Yue-Hei Ng, Matthew Hausknecht, Sudheendra Vijayanarasimhan, Oriol Vinyals, Rajat Monga, and George Toderici. 2015. Beyond short snippets: Deep networks for video classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4694–4702.

29. Matthew D Zeiler and Rob Fergus. 2014. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision*. Springer, 818–833.

30. Matthew D Zeiler, Graham W Taylor, and Rob Fergus. 2011. Adaptive deconvolutional networks for mid and high level feature learning. In *2011 International Conference on Computer Vision*. IEEE, 2018–2025.

31. Yilu Zhang, Yuri Owechko, and Jing Zhang. 2004. Driver cognitive workload estimation: A data-driven perspective. In *Intelligent Transportation Systems, 2004. Proceedings. The 7th International IEEE Conference on*. IEEE, 642–647.