Toward Real-time Assessment of Workload: A Bayesian Inference Approach

Ruikun Luo[1*], Yifan Wang[1*], Yifan Weng[1], Victor Paul[2], Mark J. Brudnak[2], Paramsothy Jayakumar[2], Matt Reed[1], Jeffrey L. Stein[1], Tulga Ersal[1], and X. Jessie Yang[1]

[1]University of Michigan, Ann Arbor, MI, US
[2]U.S. Army Ground Vehicles System Center, Warren, MI, US
[*]These authors contributed equally to this work

Workload management is of critical concern in teleoperation of unmanned vehicles, because high workload can lead to sub-optimal task performance and can harm human operators' long-term well-being. In the present study, we conducted a human-in-the-loop experiment, where the human operator teleoperated a simulated High Mobility Multipurpose Wheeled Vehicle (HMMWV) and performed a secondary visual search task. We measured participants' gaze trajectory and pupil size, based on which their workload level was estimated. We proposed and tested a Bayesian inference (BI) model for assessing workload in real time. Results show that the BI model can achieve an encouraging 0.69 F_1 score, 0.70 precision, and 0.69 recall.

INTRODUCTION

Teleoperation has been used in a wide variety of applications, such as urban search and rescue (USAR) and border patrol (Burke et al., 2004; Girard et al., 2004). Teleoperation allows human operators to access difficult or hazardous areas. However, it can impose high workload on the operator, leading to sub-optimal task performance and even task failures (Lu et al., 2019).

Workload can be measured offline or online. Offline retrospective measures are used after a human operator completes a task, usually via a questionnaire. In contrast, online real-time measures of workload are assessed while the operator is performing the task and therefore could be used in the design of adaptive automation. Both performance measures and physiological measures can provide online assessment of workload. For performance measures, a variety of secondary tasks have been employed, such as the mental arithmetic and the auditory n-back memory task. However, performance measures are not applicable if the secondary task performance is ambiguous or is not available immediately. Physiological measures rely on changes in human physiological signals. Common measures include heart rate related measures (Backs et al., 2003), electroencephalogram (EEG) (Liu et al., 2017b), eye-related measures (Lu et al., 2019; Di Nocera et al., 2007), Galvanic Skin Response (GSR) (Xu, 2014) and near infrared spectroscopy (NIRS) (Liu et al., 2017b; Ayaz et al., 2012).

Among all the physiological measures, some could be intrusive and the some could be easily affected by body movements (Chen et al., 2015). Therefore, with the development of advanced eye-tracking technology, research effort has been spent on using eye-related measurements to assess operators' workload, including pupil diameter (Recarte and Nunes, 2003), gaze distribution (Reimer, 2009) , gaze trajectory (Wang et al., 2014; Fridman et al., 2018), blink rate (Coral, 2016) and so on.

To assess workload online using physiological data, previous studies largely adopted statistical methods to show the relationships between certain physiological signals and workload. Recently, researchers started to apply machine learning techniques to classify mental workload into different levels. Using a decision tree, Zhang et al. (2004) classified drivers' workload into 2 levels by analyzing a 30-second time window of the pupil diameter and driving data. Solovey et al. (2014) examined the impact of the size of the time window (10, 15, 20, 25, and 30 seconds) on workload estimation accuracy and found that the accuracy tends to increase with increased window size. A recent work of Fridman et al. (2018) proposed a deep neural network to analyze a 6-second window of eye videos and classified operators' workload into 3 categories. In the present study, we developed a new method, which required only 4-second physiological data as input. With this method, we obtained an encouraging result and were able to assess the operator's workload in nearly real time.

In addition, Solovey et al. (2014) compared different classification algorithms including Decision Tree, Logistic Regression, Multilayer Perceptron, Naive Bayes, and Nearest Neighbor. They found that no one algorithm fits all - certain algorithms provide higher estimation accuracy when analyzing certain physiological measurements. Compared to previous work, which used a single computational model for analyzing a single type (Fridman et al., 2018; Chen and Epps, 2013) or multiple types of physiological measurements (Hogervorst et al., 2014), the present study focused on how to leverage different computational models for different types of physiological measurements. We measured human operators' gaze points and pupil sizes and put forward a Bayesian inference model for assessing operators' workload.

Our main contributions are:

- We proposed a Bayesian inference model that leverages

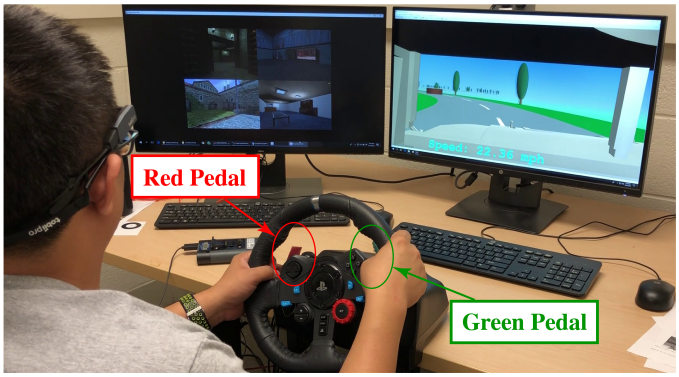

Figure 1: Experiment setup. Front screen is for driving task. Side screen is for visual search task.

Figure 2: Illustration of the visual search task (left). The participant is shown four images at a time and is required to detect potential threats. Illustration of a potential threat (right).

different computational models for different types of physiological measurements, namely, Hidden Markov Model (HMM) for analyzing gaze trajectory and Support Vector Machine (SVM) for pupil size.

- Our proposed model can estimate human operators' workload based on data from a 4-second time window.

METHOD

Participants

A total of 20 students participated in the experiment. Data of 8 participants were discarded due to equipment malfunction. The remaining 12 participants were on average 22.7 years old (SD = 2.6) and had an average of 4.5 years of driving experience (SD = 2.2). All participants had normal or corrected-to-normal vision.

Simulation Testbed

In the experiment, participants were asked to teleoperate a simulated High Mobility Multipurpose Wheeled Vehicle (HMMWV) and to perform a secondary visual search task (See Figure 1). The HMMWV has a high center of gravity, making it difficult to turn the vehicle. In the teleoperation task, the participants interacted with an autonomous navigation algorithm in shared-control mode via a Logitech G29 driving force racing wheel. The autonomy only performed lane keeping at a fixed driving speed of 10 m/s (around 22 mph). Each participant drove on four different tracks, each with a length of 1000 m. There were 6 obstacles along each track. The participants were asked to drive the HMMWV as close to the center lane as possible with the aid of the autonomy while avoiding the obstacles by themselves.

In the visual search task, the participants received image feeds and were asked to identify potential threats in the images (See Figure 2). The participants were informed that the image feeds may contain potential threats and they should report to their commander as soon as possible once they spotted the potential threats. Participants reported "danger" by clicking the red pedal at the back of the steering wheel. Otherwise, the participant reported "clear" by clicking the green pedal. As the steering wheel can only rotate from -90 degree to 90 degree, the participant would not need to cross their hands and could always keep the hands on the steering wheel. The potential threat will appear in only one of the four images. The visual search task utilized a combined pace design: If the participant responded within 8 seconds, there would be a gap until the display of the next set of four images; if the participant responded after 8 seconds, the next set of images would be displayed immediately. There was an auditory alert every 3 seconds to remind the participant of the secondary task. Figure 3 illustrates the design of the visual search task, where R_t denotes the human operator's response time, $A_t = 3$ s is the alert time, $a = 8$ s is a parameter that limits the participant's pace on the visual search task and W_t is the gap between the display of the current set of images and the display of the next set of images. We define $W_t = \max(a - R_t, 0)$. Thus if the human operator's response time R_t is smaller than $a = 8$ s, a blank image will be shown for $a - R_t$ seconds; if the human operator's response time R_t is larger than $a = 8$ s, the next set of images will appear immediately.

The autonomy used in this project implemented the nonlinear model predictive control approach (NMPC) (Febbo et al., 2017; Liu et al., 2017a). It leverages a 3 degrees-of-freedom vehicle model for its embedded model. It receives the vehicle's state information such as position, lateral speed, steering angle, yaw angle and yaw rate as inputs, and outputs the steering angle series for the vehicle as a result of minimizing a cost function. The cost function comprises three different terms: deviation from the center line of the track, a penalty on vehicle tire lift-off, and steering effort regularization. Also, constraints are imposed to ensure that the vehicle is operated within its dynamic limits. By optimizing this cost function, the autonomy can achieve a solution that minimizes the deviation from the center line with least control effort while ensuring the vehicle does not experience tire-lift off. The update rate for the autonomy is less than 500 ms, which is considered as a real-time update rate.

Experimental Design

The experiment used a within-subject design. Each participant drove on four different tracks, each with 1000 m length. There were six different obstacles on the track, with varying

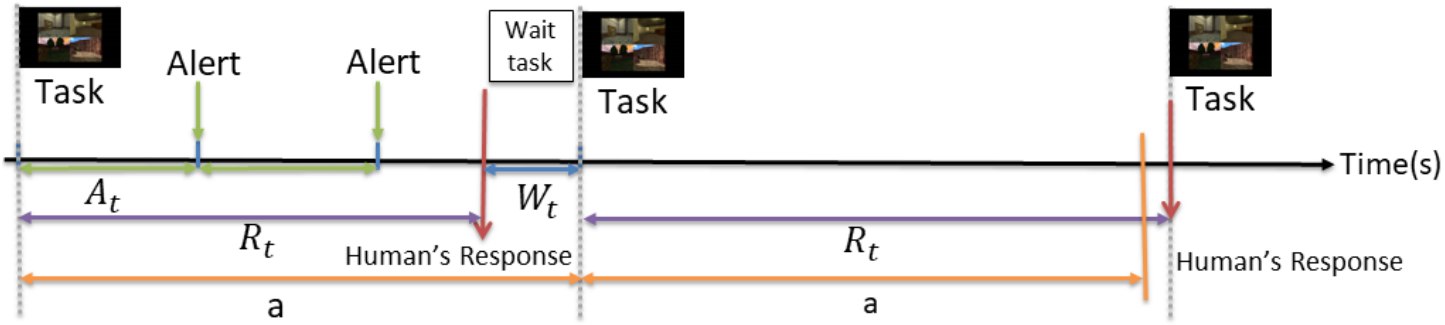

Figure 3: Illustration of the design of the secondary visual search task.

obstacle sizes (1-, 3-, 5-meter diameter) and visible distances (25- and 80-meter). The visible distance indicates how far away the participant can see the obstacle. As the vehicle's speed is 10 m/s, participants had 2.5 s or 8 s to perform the obstacle avoidance. The presentation order of the 6 obstacles followed a 6×6 Latin square to eliminate potential order effects.

Procedure

The participants provided informed consent and filled in a demographic survey prior to the experiment. The participants' baseline pupil sizes were then collected by asking them to look at a white wall twice, each for 30 seconds. Four training trials were provided to them before the real experiment: (1) driving on a track without obstacles; (2) driving on a track with obstacles; (3) performing the secondary visual search task; (4) performing the primary and the secondary tasks on a track with obstacles. The participants were asked to report the perceived difficulty after avoiding each obstacle during the training and real experiment. A debriefing survey was taken at the end of the experiment.

During the experiment, the participants wore the Tobii Pro Glasses 2 to gather their eye-related data, i.e. gaze points and pupil sizes.

Preparation of Data

During the experiment, the participants' gaze points, pupil sizes and perceived difficulty of each obstacle were collected. The results revealed a significant effect of visible distance on perceived difficulty ($F(1, 11) = 101.928$, $p < .001$). Therefore we considered the event of avoiding obstacles with a 25-meter visible distance as imposing high workload on human operators and the event of avoiding obstacles with a 80-meter visible distance as imposing low workload. The sampling rate for the Tobii Pro Glasses 2 is 50 Hz. In the case of data dropout, we re-sampled the data to 50 Hz sampling rate. We used data of 4 seconds in the middle of each obstacle avoidance event and had 288 data points (12 participants × 4 tracks × 6 obstacles).

WORKLOAD INFERENCE

In this section, we first discuss how to use HMM and SVM to classify workload into different levels (high and low workload in our experiment). Then we introduce our proposed Bayesian inference method to leverage these two models.

HMM for Gaze Trajectory

Hidden Markov model (HMM) has been used to model gaze trajectory to estimate workload (Fridman et al., 2018). Let $X_H = \{x_H^1, x_H^2, ..., x_H^T\}$ represent a gaze trajectory captured from the eye tracker, where x_H^t represents the gaze point (location of where the human is looking at relative to the external world coordinate) at time t.

HMM contains a set of hidden states y, observations x, observation model $p(x|y)$ and state transition probabilities $p(y_i|y_j)$. To model the gaze trajectory using HMM, we defined the hidden states as centers of the gaze points and the observation model is a multivariate normal distribution over the centers. The number of hidden states was determined by Bayesian Information Criterion (BIC) (Calinon and Billard, 2005; Schwarz et al., 1978). We trained two HMMs, one for the high workload and one for the low workload. The parameters of HMMs were learned by the Expectation Maximization algorithm. We used the open source implementations from Rozo et al. (2016); Calinon (2016).

Given a gaze trajectory X_H, we computed the likelihood $p(X_H|H)$ via the forward algorithm, where H represented different HMMs for the high workload and low workload. To estimate the workload of X_H, we found the HMM with the maximum likelihood, i.e. $\arg\max_H p(X_H|H)$.

SVM for Pupil Size Changes

Support-vector machine (SVM) has been used to classify human operators' workload using the changes of the pupil size as features (Hogervorst et al., 2014). SVM is a supervised learning algorithm that aims to find the optimal hyperplane that separates data points into different clusters. Wu et al. (2004) showed how to estimate probabilities for multi-class classification by pairwise coupling, i.e. given a data point X_S, the proposed algorithm can estimate $p(S|X_S)$, where S is a different class label. We used the LIBSVM package (Chang and Lin, 2011) in our implementations.

Given a sequence of pupil sizes $D = \{D_1, ..., D_T\}$, we first computed the change of the pupil size $\hat{D}$ based on each participant's baseline pupil size D_B as $\hat{D} = \{D_i - D_B | i = 1, ..., T\}$. Sim-

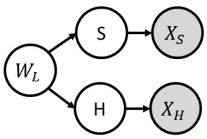

Figure 4: Bayesian inference model to combine HMM and SVM.

ilar to Solovey et al. (2014), we used a 0.4-second time window (20 time steps) and a 0.2-second overlap (10 time steps) to extract the feature vector from $\hat{D}$ as $X_S = \{\frac{\sum_{i=1}^{20}\hat{D}_{t+i}}{20} | t = 0, 10, 20, \ldots\}$. To estimate the workload of X_S, we found the S that maximized the posteriori, i.e. $\arg\max_S p(S|X_S)$. We used the linear kernel SVM and regularization parameter $C = 1$ in this paper.

Bayesian Inference

In order to combine the HMM and SVM models, we employed a Bayesian inference approach. Figure 4 shows the proposed probabilistic graphical model for the Bayesian inference, where W_L is the human's workload, S is the workload estimation by SVM, H is the workload estimation by HMM, X_S is the feature vector of the pupil size changes, X_H is the gaze trajectory. The shaded circles represent the observed data and other circles represent the hidden states. The maximum posteriori estimation of workload is to compute $\arg\max_{W_L} p(W_L|X_H, X_S)$. Given the probabilistic graphical model, we had the following equations based on the Bayes' rule and the law of total probability:

$$\begin{aligned} & p(W_L|X_H, X_S) \\ \propto\ & p(X_H, X_S|W_L)p(W_L) \\ =\ & p(W_L)\sum_{H,S} p(X_H, X_S, H, S|W_L) \\ \propto\ & p(W_L)\sum_{H,S} p(H|W_L)p(S|W_L)p(X_H|H)\frac{p(S|X_S)}{p(S)} \end{aligned} \quad (1)$$

where $p(W_L)$ is the prior of the human operators' workload (0.5 for both high and low workload in our case), $p(H|W_L)$, $p(S|W_L)$ are the prior knowledge of how the HMM and SVM model works independently (we use empirical results of HMM and SVM performance to approximate), $p(X_H|H)$, $p(S|X_S)$ are the likelihood and posteriori output of HMM and SVM, and $p(S)$ is the prior knowledge of SVM (we use empirical results of SVM to approximate). As $p(X_H|H)$ is the probability density of the gaze trajectory, the longer the trajectory is, the smaller this value is. In order to eliminate the influence of the length of the trajectory, one can use geometric mean of the probability density of a trajectory (Luo et al., 2018). Thus we used the following equation to estimate the human's workload, where N is the length of the gaze trajectory (200 in our experiment).

$$\arg\max_{W_L} p(W_L)\sum_{H,S} p(H|W_L)p(S|W_L)\sqrt[N]{p(X_H|H)}\frac{p(S|X_S)}{p(S)} \quad (2)$$

RESULTS AND DISCUSSION

Due to the small dataset (12 participants), we used the holdout method (Kim, 2009) for cross-validation for testing the performance of our proposed method. In each run of the holdout, we randomly selected data of 3 participants as the testing dataset and data of the remaining 9 participants as the training dataset. To find the best number of hidden states, we varied the number of hidden states from 2 to 10 for HMM and ran 100 holdouts for each number of hidden states. The results indicate that 4 was the best number of hidden states. We ran 100 holdouts for HMM and SVM in order to compute the prior knowledge of their performance ($p(H|W_L), p(S|W_L), p(S)$).

We then ran another 100 holdouts to combine our proposed Bayesian inference model and the baseline HMM and SVM models. Precision, recall and F_1 score were used as performance metrics. Precision is the number of true positives divided by the number of true positives + false positives. Recall is the number of true positives divided by the number of true positives + false negatives. For our multi-classification problem, the precision is the mean precision of all classes and the recall is the mean recall of all classes. $F_1 = 2 \cdot \text{precision} \cdot \text{recall}/(\text{precision} + \text{recall})$. Table 1 shows the mean and standard error of each performance metric for different methods. The results show that the proposed model achieved a 0.69 F_1 score, 0.70 precision and 0.69 recall. Table 2 shows the pairwise t-test comparing the three models. The results indicate that our proposed Bayesian inference model had better prediction than either the HMM or the SVM model alone.

Table 1: Performance of HMM, SVM and Bayesian Inference

	HMM	SVM	BI
F_1	0.655 ± 0.008	0.581 ± 0.005	0.693 ± 0.009
Precision	0.660 ± 0.009	0.583 ± 0.005	0.699 ± 0.009
Recall	0.650 ± 0.008	0.578 ± 0.005	0.687 ± 0.008

Table 2: Statistical analysis results of model comparison

	BI vs HMM	BI vs SVM
F_1	$t(1,99) = 6.89, p < .001$	$t(1,99) = 10.54, p < .001$
Precision	$t(1,99) = 6.081, p < .001$	$t(1,99) = 10.75, p < .001$
Recall	$t(1,99) = 6.826, p < .001$	$t(1,99) = 10.19, p < .001$

CONCLUSION

Workload management is critical for teleoperation tasks, as high workload would lead to sub-optimal task performance and even task failures (Lu et al., 2019). Various computational models have been developed to estimate human's workload by analyzing different types of physiological signals (Hogervorst et al., 2014).

In this paper, we proposed a Bayesian inference model leveraging different computational models for different types of physiological signals, i.e. HMM for gaze trajectory and SVM for changes of pupil size. Experimental results showed that our proposed method estimated human's workload using only 4-second physiological data and achieved 0.69 F_1 score, 0.70 precision and 0.69 recall.

Our results should be viewed in light of the several limitations. First, pupil dilation is sensitive to ambient light levels. As

our experiment is conducted in a controlled environment with constant ambient light, it is unclear if our algorithm will be generalized to environments with varying ambient light. Second, we did not compare our algorithm with other baseline methods such as Fridman et al. (2018).

In our future work, we aim to reduce the time window for workload estimation and combine other computational models for other physiological signals such as heart rate variability (HRV) in order to achieve more accurate workload estimation performance in real-time. We will also incorporate our workload estimation in the design of an adaptive shared-control autonomy.

Acknowledgement

We acknowledge the technical and financial support of the Automotive Research Center (ARC) in accordance with Cooperative Agreement W56HZV-14-2-0001 U.S. Army Tank Automotive Research, Development and Engineering Center (TARDEC) Warren, MI.

REFERENCES

Ayaz, H., Shewokis, P. A., Bunce, S., Izzetoglu, K., Willems, B., and Onaral, B. (2012). Optical brain monitoring for operator training and mental workload assessment. *Neuroimage*, 59(1):36–47.

Backs, R. W., Lenneman, J. K., Wetzel, J. M., and Green, P. (2003). Cardiac measures of driver workload during simulated driving with and without visual occlusion. *Human Factors*, 45(4):525–538.

Burke, J. L., Murphy, R. R., Coovert, M. D., and Riddle, D. L. (2004). A field study of human–robot interaction in the context of an urban search and rescue disaster response training exercise.

Calinon, S. (2016). A tutorial on task-parameterized movement learning and retrieval. *Intelligent Service Robotics*, 9(1):1–29.

Calinon, S. and Billard, A. (2005). Recognition and reproduction of gestures using a probabilistic framework combining pca, ica and hmm. In *Proceedings of the 22nd international conference on Machine learning*, pages 105–112. ACM.

Chang, C.-C. and Lin, C.-J. (2011). Libsvm: A library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2(3):27.

Chen, S. and Epps, J. (2013). Automatic classification of eye activity for cognitive load measurement with emotion interference. *Computer methods and programs in biomedicine*, 110(2):111–124.

Chen, W., Jaques, N., Taylor, S., Sano, A., Fedor, S., and Picard, R. W. (2015). Wavelet-based motion artifact removal for electrodermal activity. In *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 6223–6226. IEEE.

Coral, M. P. (2016). Analyzing cognitive workload through eye-related measurements: A meta-analysis.

Di Nocera, F., Camilli, M., and Terenzi, M. (2007). A random glance at the flight deck: Pilots' scanning strategies and the real-time assessment of mental workload. *Journal of Cognitive Engineering and Decision Making*, 1(3):271–285.

Febbo, H., Liu, J., Jayakumar, P., Stein, J. L., and Ersal, T. (2017). Moving obstacle avoidance for large, high-speed autonomous ground vehicles. In *2017 American Control Conference (ACC)*, pages 5568–5573. IEEE.

Fridman, L., Reimer, B., Mehler, B., and Freeman, W. T. (2018). Cognitive load estimation in the wild. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, page 652. ACM.

Girard, A. R., Howell, A. S., and Hedrick, J. K. (2004). Border patrol and surveillance missions using multiple unmanned air vehicles. In *2004 43rd IEEE Conference on Decision and Control (CDC)(IEEE Cat. No. 04CH37601)*, volume 1, pages 620–625. IEEE.

Hogervorst, M. A., Brouwer, A.-M., and Van Erp, J. B. (2014). Combining and comparing eeg, peripheral physiology and eye-related measures for the assessment of mental workload. *Frontiers in neuroscience*, 8:322.

Kim, J.-H. (2009). Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap. *Computational statistics & data analysis*, 53(11):3735–3745.

Liu, J., Jayakumar, P., Stein, J. L., and Ersal, T. (2017a). Combined speed and steering control in high speed autonomous ground vehicles for obstacle avoidance using model predictive control. *IEEE Transactions on Vehicular Technology*, 66(10):8746–8763.

Liu, Y., Ayaz, H., and Shewokis, P. A. (2017b). Multisubject "learning" for mental workload classification using concurrent eeg, fnirs, and physiological measures. *Frontiers in human neuroscience*, 11:389.

Lu, S., Zhang, M. Y., Ersal, T., and Yang, X. J. (2019). Workload management in teleoperation of unmanned ground vehicles: Effects of a delay compensation aid on human operators' workload and teleoperation performance. *International Journal of Human–Computer Interaction*, pages 1–11.

Luo, R., Hayne, R., and Berenson, D. (2018). Unsupervised early prediction of human reaching for human–robot collaboration in shared workspaces. *Autonomous Robots*, 42(3):631–648.

Recarte, M. A. and Nunes, L. M. (2003). Mental workload while driving: effects on visual search, discrimination, and decision making. *Journal of experimental psychology: Applied*, 9(2):119.

Reimer, B. (2009). Impact of cognitive task complexity on drivers' visual tunneling. *Transportation Research Record*, 2138(1):13–19.

Rozo, L., Silverio, J., Calinon, S., and Caldwell, D. G. (2016). Learning controllers for reactive and proactive behaviors in human–robot collaboration. *Frontiers in Robotics and AI*, 3:30.

Schwarz, G. et al. (1978). Estimating the dimension of a model. *The annals of statistics*.

Solovey, E. T., Zec, M., Garcia Perez, E. A., Reimer, B., and Mehler, B. (2014). Classifying driver workload using physiological and driving performance data: two field studies. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 4057–4066. ACM.

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

Wu, T.-F., Lin, C.-J., and Weng, R. C. (2004). Probability estimates for multi-class classification by pairwise coupling. *Journal of Machine Learning Research*, 5(Aug):975–1005.

Xu, X. (2014). *Analysis on mental stress/workload using heart rate variability and galvanic skin response during design process*. PhD thesis, Concordia University.

Zhang, Y., Owechko, Y., and Zhang, J. (2004). Driver cognitive workload estimation: A data-driven perspective. In *Proceedings. The 7th International IEEE Conference on Intelligent Transportation Systems (IEEE Cat. No. 04TH8749)*, pages 642–647. IEEE.