COGNITIVE LOAD ESTIMATION WITH BEHAVIORAL CUES IN HUMAN-MACHINE INTERACTION

by

Go-Eum Cha

A Thesis

Submitted to the Faculty of Purdue University

In Partial Fulfillment of the Requirements for the Degree of

Master of Science

Department of Computer and Information Technology

West Lafayette, Indiana

December 2020

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF COMMITTEE APPROVAL

Dr. Byung-Cheol Min, Chair

Department of Computer and Information Technology

Dr. Baijian Yang

Department of Computer and Information Technology

Dr. Dawn Laux

Department of Computer and Information Technology

Approved by:

Dr. John A. Springer

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to all who supported me during my masters' program. During my research, I tried a lot of things I haven't done. This work and my Masers' program would not have been possible without the help and support of all. It was an excellent opportunity to meet good people in the end.

I owe my deepest gratitude to my academic advisor Dr. Byung-Cheol Min for his support and patience he has provided throughout the research. Discussions with you have always been a great source of inspiration. I would also like to thank Dr. Baijian Yang and Dr. Dawn Laux for being a member of my thesis committees and giving me valuable advice on the research. This research could not be completed without helpful advice and support. Also, I wish to thank Wonse Jo with whom I have had the pleasure to work during this study.

And last but not least, I would like to show appreciation for my family for their encouragement and support.

This material is based upon work supported by the National Science Foundation under Grant No. IIS-1846221. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

EAR	Eye Aspect Ratio
EEG	Electroencephalography
AWS	Amazon Web Services
rmANOVA	repeated measure Analysis of variance
ROS	Robot Operating System
HMI	Human-Machine Interaction
HCI	Human-Computer Interaction
HRI	Human-Robot Interaction
SVM	Support vector machine
HRC	Human-Robot Collaboration
IL	Intrinsic Load
EL	Extraneous Load
GL	Germane Load
EEG	Electroencephalography
GSR	Galvanic Skin Response
ECG	Electrocardiogram
EMG	Electromyography
HOG	Histogram of Oriented Gradient
ANOVA	Analysis Of Variance
DT	Decision Tree
KNN	K-Nearest Neighbor
BT	Boosted Tree
LDA	Linear Discriminant Analysis
3D-CNN	Three-Dimensional Convolutional Neural Network
HMM	Hidden Markov Model
ConvLSTM	ConvolutionalLSTM
TP	True-Positive
TN	True-Negative

FP	False-Positive
FN	False-Negative
TICC	Toeplitz Inverse Covariance-Based Clustering
GUI	Graphical User Interface
3DResNet	3D Residual Networks

ABSTRACT

Detecting human cognitive load is an increasingly important issue in the interaction between humans and machines, computers, and robots. In the past decade, several studies have sought to distinguish the cognitive load, or workload, state of humans based on multiple observations, such as behavioral, physiological, or multi-modal data. In the Human-Machine Interaction (HMI) cases, estimating human workload is essential because manipulators' performance could be adversely affected when they have many tasks that may be demanding. If the workload level can be detected, it will be beneficial to reallocate tasks on manipulators to improve the productivity of HMI tasks. However, it is still on question marks what kinds of cues can be utilized to know the degree of workload. In this research, eye blinking and mouse tracking are chosen as behavioral cues, exploring the possibility of a non-intrusive and automated workload estimator. During tests, behavior cues are statistically analyzed to find the difference among levels, using a dataset focused on three levels of the dual *n*-back memory game. The statistically analyzed signal is trained in a deep neural network model to classify the workload level. In this study, eye blinking related data and mouse tracking data have been statistically analyzed. The one-way repeated measure analysis of variance test result showed eye blinking duration on the dual 1-back and 3-back are significantly different. The mouse tracking data could not pass the statistical test. A three-dimension convolutional deep neural network is used to train visual data of human behavior. Classifying the dual 1-back and 3-back data accuracy is 51% with 0.66 F1-score on 1-back and 0.14 on 3-back data. In conclusion, blinking and mouse tracking are unlikely helpful cues when estimating different levels of workload.

CHAPTER 1. INTRODUCTION

This research examines different mental workload levels and observes the actual degree of workload by analyzing spontaneous behavioral responses from eye activities and mouse usages in a human-machine interaction setting.

1.1 Background

Estimating human cognitive load, or simply workload, has been studying extensively over the last decades. O'Donnel and Eggemeir (1986) defined that human workload can be identified as limited mental capacity of individuals' when they perform tasks, which could be measured as expanded capacity. In other words, the human ability to complete tasks is vulnerable since the level of human cognition cannot be exponential. For example, when we were continuously assigned to a duty that requires high-level demand, working on the task would negatively affect our ability to accomplish as time goes by. Finally, it makes us exhausted and causes lower performance.

Operative performance has been described in several approaches in terms of contextual factors—the reason why connecting performance and mental workload is because of workload derivation. Borghini, Astolfi, Vecchiato, Mattia, and Babiloni (2014) summarized the relationship between mental workload, situation awareness, or comprehending the situation, and operative performance as a combination of a negative and positive proportional relationship. When the mental workload increases, the increased workload will decrease the ability to aware of the circumstances; therefore, performance will be reduced. Debie et al. (2019) outlined the correlation of mental workload and performance as negative impacts from depletion factors, such as stress and fatigue, which can affect humans from an external source. Even though there is no dominant causation of mental workload and performance, it is acceptable that estimating cognitive workload is crucial to improvise a human operator's performance level.

What would happen when individuals are interacting with machines? Not only are machines limited to traditional vehicles, but also the range of machines covers any computers or robots. The major field on estimating cognitive load has been done in car drivers or aircraft pilots

(Borghini et al. (2014); Fridman et al. (2018); Peruzzini, Tonietti, and Iani (2019), and Benedetto et al. (2011)). Driving cars or aircraft is reliant on human workload, and their cognitive status involves fatigue or drowsiness, which causes a harmful safety issue. Measuring human demand when working on activities is crucial. Correspondingly, when it comes to operators of human-robot scenario manipulating multiple machines, measuring workload becomes more essential not only for safety issues but also for enhancing the performance.

Considering the nature of the human workload, which is an ambiguous concept that is unable to measure concisely numerically, researchers have examined a considerable amount of cues from humans and have built metrics to measure the vague subject. When assessing cognitive demand, various measurements have been conducted, such as behavioral, physiological, or combination of dissimilar information, as multi-modal inputs (Debie et al. (2019)). The cognitive effort has been analyzed from physiological or neurophysiological sensors such as electroencephalography (EEG; Borghini et al. (2014)), galvanic skin response (GSR; Nourbakhsh, Wang, Chen, and Calvo (2012)), and Functional Near-Infrared Spectroscopy (fNIRS; Herff et al. (2014)). Although several studies investigating workload have been carried out on physiological sensors, researchers have no consensus on which cues will be the most beneficial to measure workload. While Cech and Soukupova (2016); Peruzzini et al. (2019) have mentioned that a non-intrusive system will be needed. However, what kinds of human cues will be worthy in terms of workload is still a big open question because load estimation can vary from tasks and personal behavior patterns.

In the situation where several cues are still used in measuring the amount of workload, recent developments in non-verbal behavior detection have led to flourishing the quality of interactions. As a part of observable human behaviors, Longo (2011) employed mouse movement tracking as one of the important signals to monitor cognitive states' fluctuation. Also, eye blinks have been stated as a promising sign to infer mental workload. Tsubota, Kwong, Lee, Nakamura, and Cheg (1999) articulated a significant relationship between brain activation and eye blinking by capturing cortical area functional magnetic resonance imaging, which means eye blinking variations will result from when human dealing with given information. Moreover, Wascher, Heppner, Möckel, Kobald, and Getzmann (2015) noted that the execution of blinks is unlikely related to stochastic occurrences at an everyday moment, emphasizing blinks as a useful measurement of cognitive processing. Even though eye-related movements are simple, mental load indicators vary, including blink duration, blink frequency, saccade rate, and pupil size.

Taking into account that the human workload cannot be easily quantified as numeric values, diverse researchers have stimulated different levels of workload by designing resembled or the same environments of demanding cognitive ability. (Benedetto et al. (2011); Fridman et al. (2018); Sampei, Ogawa, Torres, Sato, and Miki (2016)) When they measure behavioral cues, a multidimensional scale, NASA Task Load Index (NASA-TLX; Hart and Staveland (1988)) has been utilized to have self-assessment from respondents. Eventually, behavioral observation and rating workload of participants in the experiment can be done concurrently.

The repeated simple eye blinks, which have temporal variations and spatial inequalities, may be challenging to understand at the machine level. Thanks to the drastic development of deep learning technology, neural network structure enables to learn continuous observation data. For example, when detecting eye-related behavior, Fridman et al. (2018) showed that microscopic eye movement could also be trained through two different features, image-based one and numeric value-based one. Taking advantage of neural networks, this study proposes to show eye blinking duration is a notable feature to predict human workload variations.

To sum up, the aim of this study is to evaluate and validate the relationship between eye blinking activity and human cognitive load. Stimulating workload at different levels, statistical data analysis between the result of levels, and learning data to estimate workload in the realistic environment will be done in this study.

1.2 Problem Statement

When individuals perform a task with machines, the quality of interaction could be adversely affected when a collaborative system puts much burden on humans. People may tell how much they struggle with a given task; however, such behavior is less frequent. In collaborating with a machine, it is vital to know the amount of work because if the amount of human mental load can be known on the system, the appropriate amount of work will be distributed.

1.3 Research Question

- Will eye blinking duration be different when we conduct various levels of tasks?
- Will mouse movement be effective in differentiating the different levels of cognitive load?
- Using a simple behavioral cue, would cognitive demand be classified into specific levels?

1.4 Significance

When human collaborates with machines, operators workload should be analyzed through their observable cues. This study will show whether eye blinking and mouse tracking could be meaningful cues measuring cognitive demands.

1.5 Assumptions

The assumptions for this study include:

- NASA-TLX questionnaire reflects the degree of human workload.
- The results of facial landmark prediction are robust.

1.6 Limitations

There are two limitations to this study. First, even though a metric of rating workload is widely used, it is hard to measure cognitive demand in numerical values. Second, people may react differently according to their tendencies when they are given the same workload demand.

1.7 Delimitations

The workload level will be divided into three classes: low, medium, and high, rather than calculated into specific numerical values. Regarding the possibility of reaction variance, each level's result will be statistically analyzed as average values.

1.8 Summary

In human-machine or human-computer, and human-robot interaction, analyzing workload by behavioral cues, specifically eye blinking, is critical to measure how much individuals feel demanding in the given tasks. Given that behavioral signals reflect a person's cognitive activity, the signals will be analyzed and trained to estimate the degree of cognition in three levels.

CHAPTER 2. REVIEW OF LITERATURE

This chapter illustrates a review of relevant literature in the following order: human-machine interactive environment, workload estimation, eye blinking, and cognitive task classification with eye blinking activity.

2.1 Human-Machine Interaction

Human-Machine Interaction (HMI) indicates automated systems that interact with people. As machines such as computers and robots increase, the demand for better interaction between devices and people is increasing accordingly. Researchers have focused on human beings, the main subjects of interaction, for enhancing the sensitivity of communication. In this circumstance, much attention has been drawn to finding out the trivial determinants like eye movement activity of individuals in various environments, including Human-Computer Interaction (HCI) or Human-Robot Interaction (HRI).

Ohn-Bar and Trivedi (2014) mainly investigated automotive interfaces with hand gestures. They proposed a contactless driver assistance system watching the seat of drivers. The system collected RGB-Depth camera streams of the cues of drivers to classify hand movements and recognizing users. After the system segmentized hand motions by their histogram changes, a support vector machine (SVM) was employed in classification. The interface aimed to develop a vision-based interface anticipating user customization and a contact-free interface. Interest in HMI is not limited to hands. Fahim et al. (2020) applied eye movements and head gestures to help people who have hand disabilities. Since people who are not able to use a mouse cursor have limited accessibility, the researchers translated head gestures and blinking to the location and the clicking event, respectively. When recognizing head gestures and blinking, an accelerometer and gyro-sensor were adopted with a camera. When blinking detected, a convolution neural network was selected, as Figure 2.1 presents. The accuracy of their proposed model was 95.36% in identifying human blinking. The performance of their method has not surpassed the use of conventional mouses within the user study with ten participants.

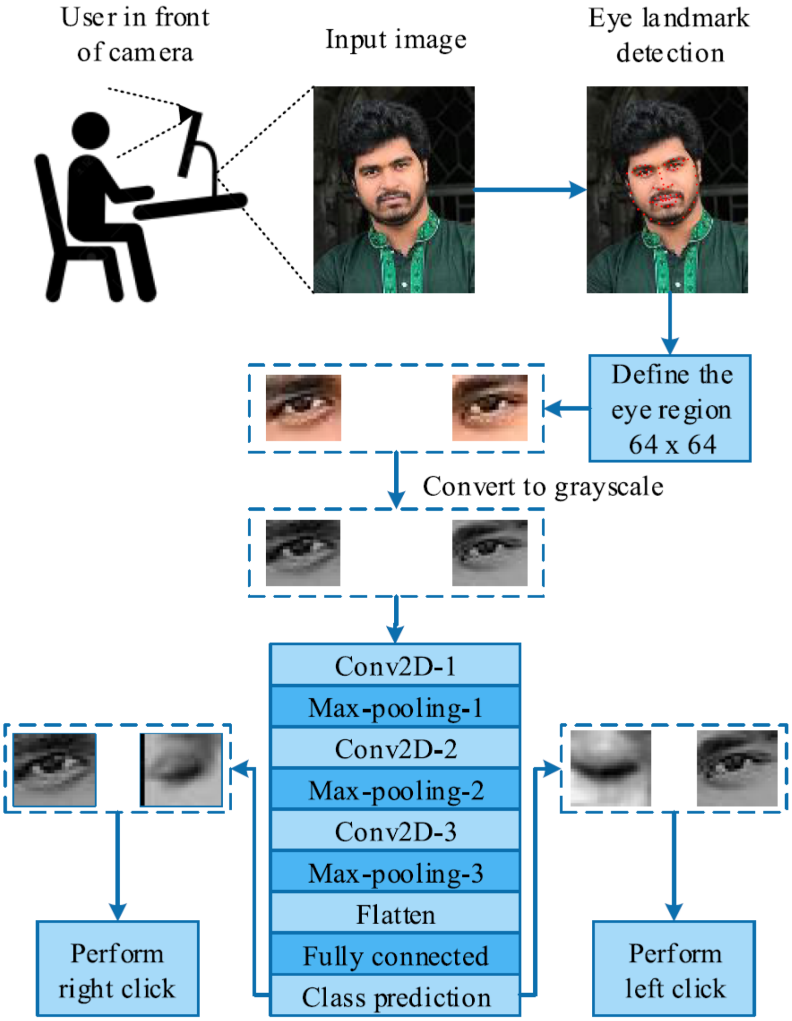


Figure 2.1. An assistive application based on eye movements and head gestures by Fahim et al. (2020) (©IEEE2020)

Human-Robot Interaction (HRI) is one of the most popular areas to enhance the interaction between humans and robots. McColl, Jiang, and Nejat (2016) conducted a study for a social robot to recognize affective states of individuals measuring accessibility, the level of openness, and rapport toward the robot. A Kinect camera was selected to read two- and three-dimensional human body language to estimate static body pose. Robot behaviors were differently presented based on the level of accessibility and speech of individuals. A questionnaire from 24 participants showed robots' expression as neutral or positive ratings, not showing a negative attitude. The level of familiarity toward robots of participants has not been illustrated in the research. However, the striking point is the degree of openness with robots can be measured by just learning human postures and speech.

Especially, Human-Robot Collaboration (HRC) is the most popular area associated with the manufacturing industry. The types of collaboration aim to share the skills of humans and robots and to improve reliability, safety, and performance. Collaborative robots, or Cobots, are used not only for industrial purposes like subtask allocation (F. Chen et al. (2013); Sadrfaridpour and Wang (2017)) or motion planning (Lasota and Shah (2015)) but also for collaboration such as assistive robots (Mukai et al. (2010)). In manufacturing contexts, humans adaptability and flexibility are necessitated due to the low flexibility of robots (F. Chen et al. (2013); Mukai et al. (2010)). F. Chen et al. (2013) conducted a study for optimal task allocation when humans and robots work as co-workers. Effective scheduling tasks become critical because humans can lose concentration and make mistakes derived from fatigue. They revealed which algorithms effectively distribute tasks when a person and a cobot face each other, called a hybrid assembly system. They insisted that the collaboration between robots and people will become important in manufacturing based on the results. Sadrfaridpour and Wang (2017) also studied with an assembly task setting when individuals and cobots are in the vicinity. They experimented by measuring the degree of trust between two types of workers, which is a social factor, as well as the physical interaction, giving feedback by expressing robot emotions on the screen according to the process. Based on these studies, interest in social HRI elements is increasing, while attention to HRC is gradually increasing in the industrial field.

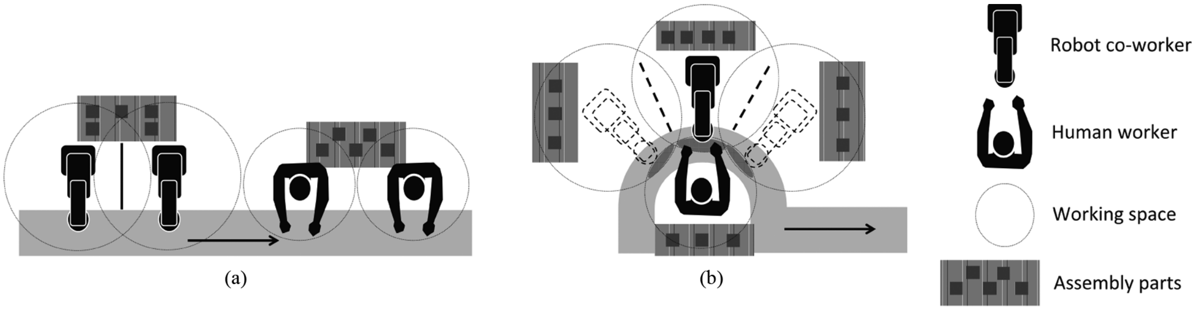


Figure 2.2. An example of human and robot collaboration scenario of assembly task. (F. Chen et al. (2013); ©IEEE2013)

2.2 Cognitive Load Estimation

Human cognitive load, in other words, workload, has been studied for decades. Sweller (1988) defined cognitive load as a capability of problem-solving, mainly focused on learning effectiveness. The problem-solving procedure has several aspects: First, the number of tasks is relevant to cognitive load, considering the human capacity to deal with several tasks in a row is limited. Then, suppose a person grasps the problem-solving procedure. In that case, the degree of the schemata acquisition, knowing how to solve problems, can differ depending on the experience. The author defined the acquisition gap as a gauge to determine novices and experts, which means that the degree of the experience is related to gaining schemata. Debue and Van De Leemput (2014) simplified three different cognitive as intrinsic load (IL), extraneous load (EL), germane load (GL). IL indicates the number of tasks that can initiate a load on a person. At the same time, EL refers to the previous stage of schemata acquisition, and GL can be explained as the cognitive load of gaining schemata of problem-solving. However, the expertise of each individual cannot be easily measured in experiments. Moreover, assumptions might need to encode each load to which indicator of quantifiable ratings. Therefore, in this study, the workload will be divided into different levels.

2.2.1 Behavioral Measures

Behavioral measures indicate actions or expressions of humans, which are extrinsic and observable. For example, human pose and development, behavior patterns when manipulating input devices, and facial activities can be significant hints to presume cognitive progressions. Some researchers conducted studies on human movement, eye blinking, but not only did they focus on blinks, but they measured as many as possible from the eye region, including blink rate, blink duration, and average pupil size (Benedetto et al. (2011); S. Chen, Epps, Ruiz, and Chen (2011); Li et al. (2020)). Although a single behavioral measure has its own limitations, which are subjective to the measurement noise contrary to physiological or neuro-physiological measurement and has low efficiency, it will be worth to analyze when it is joined to other cues.

2.2.1.1 Eye Blinking

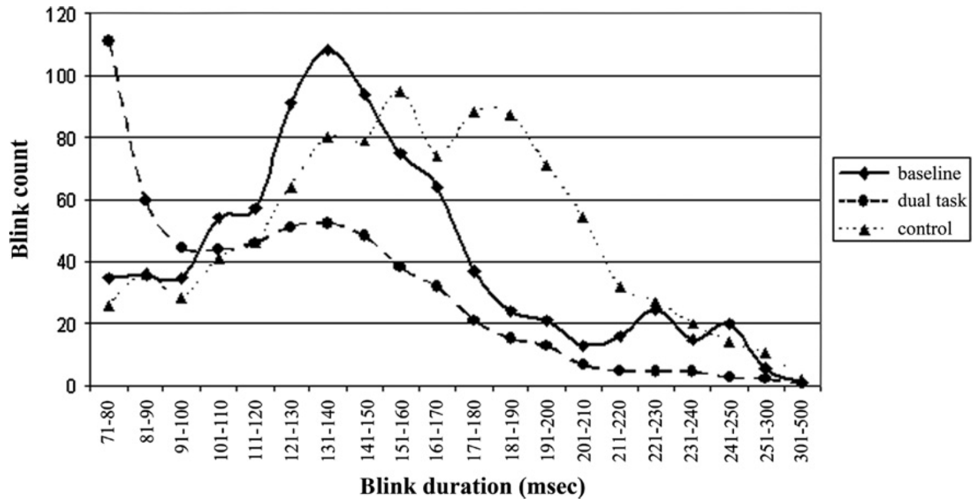


Figure 2.3. Eye blinking analysis by Benedetto et al. (2011) (©2010 Elsevier Ltd.)

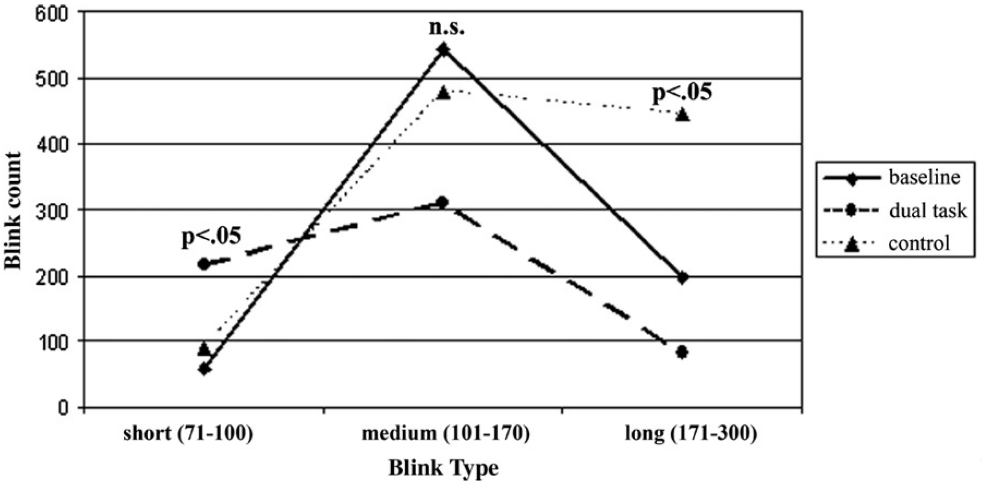


Figure 2.4. A heatmap of mouse tracking Benedetto et al. (2011) (©2010 Elsevier Ltd.)

A considerable amount of research has been published on eye blinking measurement and workload over three decades (Benedetto et al. (2011); Boehm-Davis, Gray, and Schoelles (2000); Fridman et al. (2018); Stern and Skelly (1984)). Benedetto et al. (2011) showed analysis results of eye blinks that the duration of eye blinks are different in ordinary states and when tasks are given to participants, as Figure 2.3 depicts. However, eye blinking duration has no significance during the designed experiment setup, Benedetto et al. (2011) classified three groups of eye blinks, as Figure 2.4 shows. Benedetto et al. (2011); Boehm-Davis et al. (2000); Wascher et al. (2015) have in common sense that eye blink rate is increased when the cognitive process is finished in discrete experimental setups. Even though Benedetto et al. (2011) stated that the studies which have been done performed differently and even used analysis methods are not the same, we can conclude that eye blinking can be a significant and meaningful cue in terms of cognitive workload measurement.

2.2.1.2 Mouse Movement

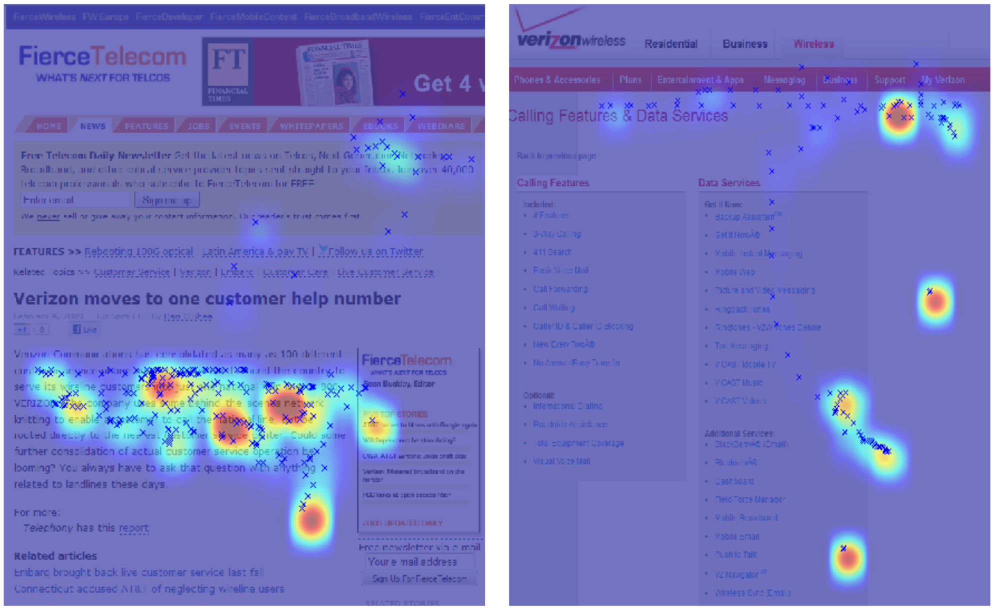


Figure 2.5. A heatmap of mouse tracking (Guo and Agichtein (2012); ©ACM2012)

Guo and Agichtein (2012) focused on mouse usage in web sites to know individuals' interest. As Figure 2.5 shows, creating a heatmap can represent how much time a user is interested in a specific field on a screen. The relevancy of data on the screen can be known with the mouse-tracking data, analyzing dwelling time. Rheem, Verma, and Becker (2018) conducted a study to find relevance between cognitive load and mouse usage. In the experiment, participants were asked to perform two kinds of tasks at the same time. One task is a mouse motor task to catch three different sizes that appeared on the screen; another one is arithmetic calculations in two levels. The study showed that slower responses and relatively inactive trajectory resulted from a higher cognitive load, showing the possibility of mouse tracking for estimating cognitive load.

2.2.2 Multimodal measures

The reason for measurements with different cues is that a single signal has limitations of sensor failures, has various sources of noise and the degree of demands on a person, and unrepresentative of a large number of subjects (Debie et al. (2019)). In this section, multimodal measures only related to eye-related behavior regarding workload estimation were reviewed. Rozado (2015) fused EEG and pupilometry to detect cognitive workload. They conducted an experiment asking participants to repeatedly perform arithmetical calculations, retrieving pupil diameter and EEG signal variations. It turns out that combining EEG signal and pupilometry outperformed in terms of accuracy other than single cues. In this context, combining more than two signals will be beneficial to estimate workload levels more accurately than using a single movement.

2.3 Facial Feature Extraction

Several studies have been made to detect human facial components or facial landmarks. With respect to the component detection, Haar Cascade Classifier of OpenCV is one of the well-known approaches. The haar-feature based classifiers catch the region of similar features that are already trained. However, the exposed region cannot answer facial components' exact positions, giving blunt areas of targeted objects. Facial landmark detection prospers automatic face-related analysis tasks, such as face action recognition (Pfister, Li, Zhao, and Pietikäinen (2011); Zhu, Lei, Yan, Yi, and Li (2015)).

In this context, Dlib (King (2009)), which is based on Histogram of Oriented Gradient (HOG) and SVM, is one of widely utilized in facial landmark detection task. A number of studies have used Dlib; however, gradient vectors in HOG are unlikely to cover all different angles of face detection. Contrary to the landmark detection result of a frontal face, even though the same frame is given to two detectors, Dlib cannot return any value of tilted head direction, while

OpenPose can. Compared to the HOG and SVM based approach, OpenPose (Cao, Simon, Wei, and Sheikh (2017)) supports facial keypoints detection using multiple views to project the precise position of human landmarks (Simon, Joo, Matthews, and Sheikh (2017)). Each key point can be extracted by the given number in Figure 2.6. Given a trained model from techniques, we would expect the result of facial feature extraction will be comparably robust.

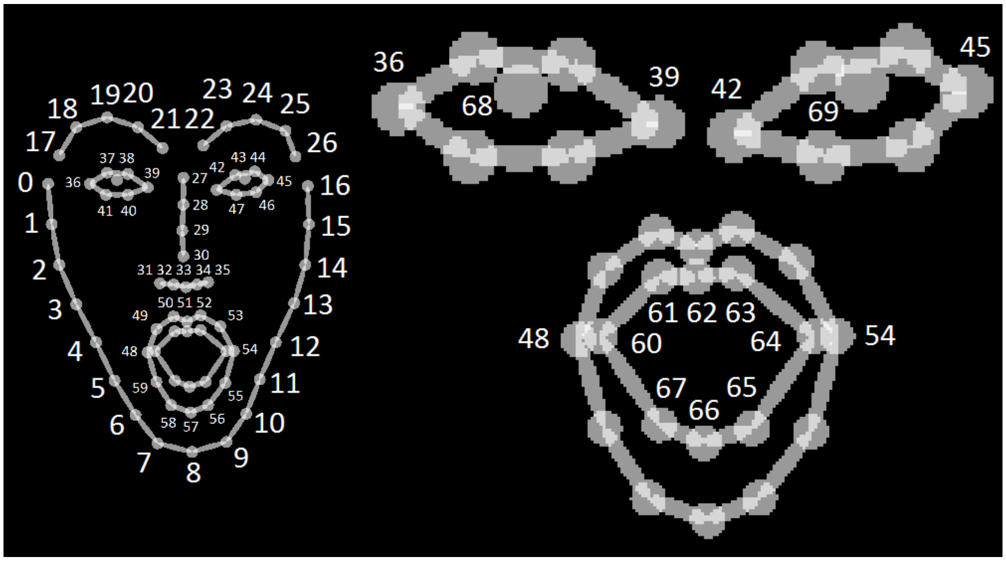


Figure 2.6. An example of facial key points in *CMU-Perceptual-Computing-Lab/openpose* (n.d.) (©CMU-Perceptual-Computing-Lab)

2.4 Eye Blinking Detection

2.4.1 Landmark-based Eye Blinking Detection

Cech and Soukupova (2016) proposed an eye detection algorithm using facial landmarks. The eye-related features are mainly six points, both horizontal ends $p1, p4$ and four curve points $p2, p3, p5$, and $p6$ as in the Figure 2.7. Taking advantage of that essential points can be detected with the open-sourced libraries, Cech and Soukupova (2016) calculated eye blinking with a simple Euclidean distance equation. The calculated Eye Aspect Ratio (EAR) value decreases when eyes are closing two vertical distances between $p2$ *and* $p6$ and $p3$ *and* $p5$ gradually decreases, as Figure 2.7. After both EAR values are calculated, the average of the values is compared with a specific threshold. Finally, the authors set the thresholding value to 0.2.

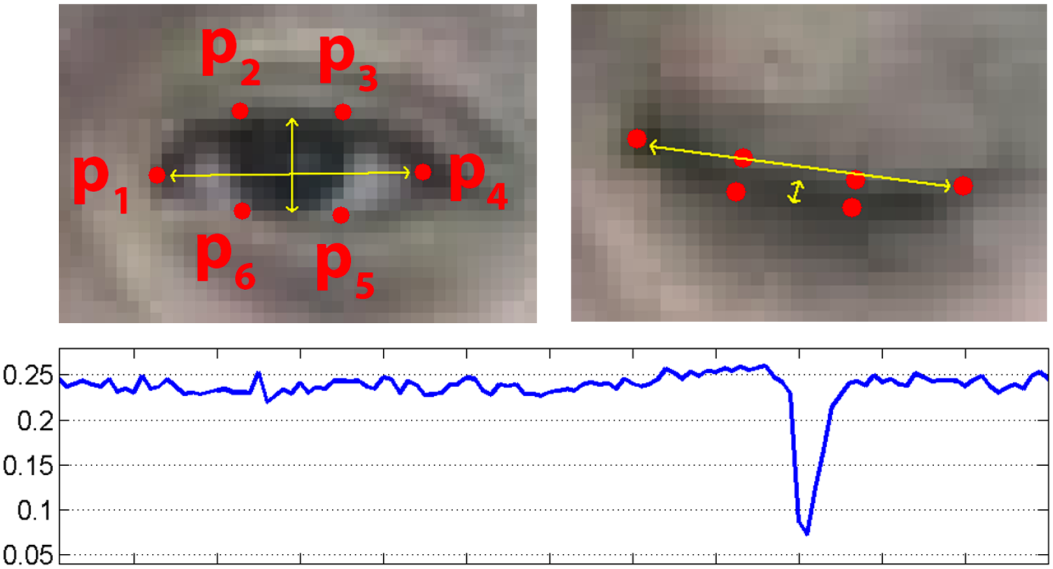


Figure 2.7. Eye landmark prediction on left eye (Cech and Soukupova (2016); ©Computer Vision Winter Workshop)

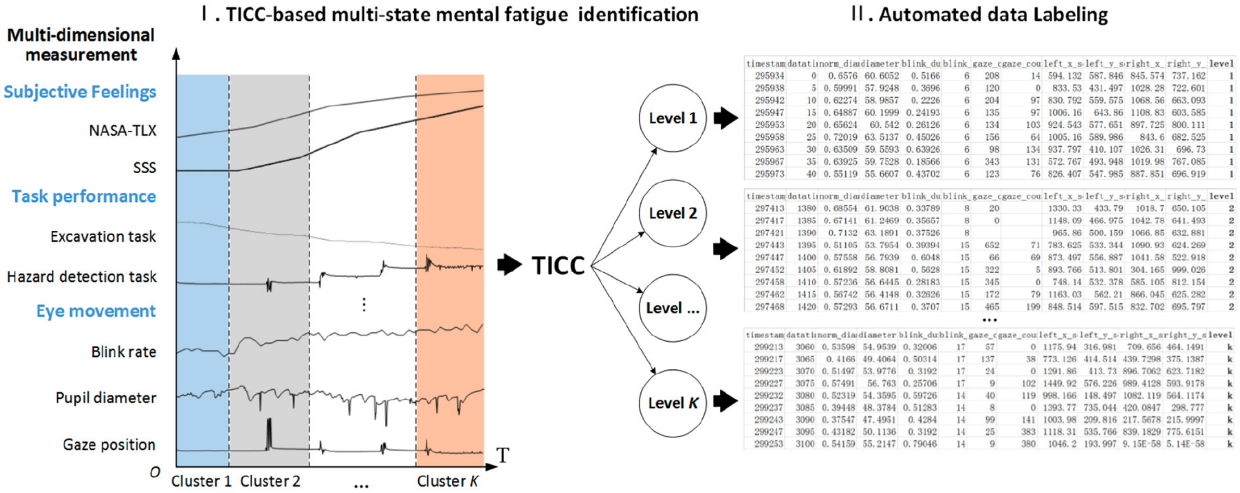
2.4.2 Importance of Statistical Analysis

Coral (2016) brought up a lack of statistical analysis metric about eye-related behavior, mentioning that only reviewing p-test value could be not sufficient. Debie et al. (2019) indicated that statistical analysis on sensor measurement results would be needed when the authors reviewed studies of fusion-based cognitive workload assessments. For example, Benedetto et al. (2011); Boehm-Davis et al. (2000); Wascher et al. (2015) performed ANOVA tests to verify significance between baselines and controlled experiments. If participants' behavior in experiments was measured over twice, the measures are considered dependant variables considered treatments to observe any discrepancy between more than two observations (Singh, Rana, and Singhal (2013)). Mainly, Benedetto et al. (2011); Boehm-Davis et al. (2000); Rozado (2015) performed repeated measure ANOVA tests, considering collected data are dependent and measured repeatedly. An assumption for rmANOVA, sphericity assumption has to be satisfied with the rmANOVA test. The assumption is to show the necessities of population variances are equal to the test environment since the experiment trials are randomly selected from the general people (Singh et al. (2013)).

2.4.3 Learning methods of eye-related features

Several methodologies have been applied to train eye-related features such as eye blinking or pupil movements. Considering eye movements are spontaneous development of temporal and spatial representation, features were selected in diverse ways. In this section, the literature review has been done with respect to the relationship between eye-related activities and workload estimation.

2.4.3.1 Machine Learning and Deep Learning Li et al. (2020) utilized a supervised learning approach when predicting mental fatigue. In the scenario of construction equipment manipulating, eye-behavior-related features, such as blink rate, blink duration, pupil diameter, and gaze position, were chosen altogether from an eye tracker to be used as time-series data. The sequential data are divided into several levels in terms of data labeling. When the derived measurements are classified into different levels, Toeplitz Inverse Covariance-Based Clustering (TICC) was used. Each level consists of time-based sequential data, as Figure 2.8 shows. The extracted data and label went into classification algorithms, SVM, Decision Tree (DT), K-Nearest Neighbor (KNN), Boosted Tree (BT), and Linear Discriminant Analysis (LDA). The authors found that SVM outperforms other algorithms with an accuracy of at least 80%. However, the authors stated that short 1 hour experiments would not be evident compared to operators who work more than 8 hours. On top of that, additional sensors that resemble eyeglasses will need to extract pupil movement as well.



timestam	datati	norm_dia	diameter	blink_du	blink_	gaze_c	gaze_cou	left_x_s	left_y_s	right_x_	right_y_	level
295934	0	0.6576	60.6052	0.5166	6	208	14	594.132	587.846	845.574	737.162	1
295938	5	0.59991	57.9248	0.3696	6	120	0	833.53	431.497	1028.28	722.601	1
295942	10	0.62274	58.9857	0.2226	6	204	97	830.792	559.575	1068.56	663.093	1
295947	15	0.64887	60.1999	0.24193	6	135	97	1006.16	643.86	1108.83	603.585	1
295953	20	0.65624	60.542	0.26126	6	134	103	924.543	577.651	897.725	800.111	1
295958	25	0.72019	63.5137	0.45026	6	156	64	1005.16	589.986	843.6	682.525	1
295963	30	0.63509	59.5593	0.63926	6	98	134	937.797	410.107	1026.31	696.73	1
295967	35	0.63925	59.7528	0.18566	6	343	131	572.767	493.948	1019.98	767.085	1
295973	40	0.55119	55.6607	0.43702	6	123	76	826.407	547.985	887.851	696.919	1

timestam	datati	norm_dia	diameter	blink_du	blink_	gaze_c	gaze_cou	left_x_s	left_y_s	right_x_	right_y_	level
297413	1380	0.68554	61.9038	0.33789	8	20		1330.33	433.79	1018.7	650.105	2
297417	1385	0.67141	61.2469	0.35657	8	0		1148.09	466.975	1042.78	641.493	2
297421	1390	0.7132	63.1891	0.37526	8			965.86	500.159	1066.85	632.881	2
297443	1395	0.51105	53.7954	0.39394	15	652	71	783.625	533.344	1090.93	624.269	2
297447	1400	0.57558	56.7939	0.6048	15	66	69	873.497	556.887	1041.58	522.918	2
297452	1405	0.61892	58.8081	0.5628	15	322	5	893.766	513.801	304.165	999.026	2
297458	1410	0.57236	56.6445	0.28183	15	345	0	748.14	532.378	585.105	812.154	2
297462	1415	0.56742	56.4148	0.32626	15	172	79	1163.03	562.21	866.045	625.282	2
297468	1420	0.57293	56.6711	0.3707	15	465	199	848.514	597.515	832.702	695.797	2

...

timestam	datati	norm_dia	diameter	blink_du	blink_	gaze_c	gaze_cou	left_x_s	left_y_s	right_x_s	right_y_s	level
299213	3060	0.53598	54.9539	0.32006	17	57	0	1175.94	316.981	709.656	464.1491	k
299217	3065	0.4166	49.4064	0.50314	17	137	38	773.126	414.514	439.7298	375.1387	k
299223	3070	0.51497	53.9776	0.3192	17	24	0	1291.86	413.73	896.7062	623.7182	k
299227	3075	0.57491	56.763	0.25706	17	9	102	1449.92	576.226	989.4128	593.9178	k
299232	3080	0.52319	54.3595	0.59726	14	40	119	998.166	148.497	1082.119	564.1174	k
299237	3085	0.39448	48.3784	0.51283	14	8	0	1393.77	735.044	420.0847	298.777	k
299243	3090	0.37547	47.4951	0.4284	14	99	141	1003.98	209.816	217.5678	215.9997	k
299247	3095	0.43182	50.1136	0.3192	14	25	383	1118.31	535.766	839.1829	775.6151	k
299253	3100	0.54159	55.2147	0.79046	14	9	380	1046.2	193.997	9.15E-58	5.14E-58	k

Figure 2.8. TICC used mental fatigue identification (Li et al. (2020); ©2019 Elsevier B.V.)

Fridman et al. (2018) adopted two methods, the three-dimensional convolutional neural network (3D-CNN) model, as Figure 2.9 shows, and the Hidden Markov Model (HMM). The features fed in each model are eye image sequence and extracted pupil position, respectively, divided by the level of verbal cognitive tasks. Contrary to numeric features mentioned in Section 2.2.1, image sequence can be another type of feature which does not require feature extraction stages. The authors defined 3D-CNN as early temporal fusion because image concatenation is given to the neural net structure. The frame per second (FPS) is downsampled from 30fps to 15fps and temporarily concatenated into 90 frames to hold temporal and spatial features of 6 seconds, showing examples as Figure 2.10. HMM is described as a late temporal fusion as in the model features are mixed in Markov chains with random variables that states are not identifiable. One critical downside of HMM is that the inference result could be affected by arbitrarily given parameters. As a result, 3D-CNN outperformed HMM with 86.1% accuracy. However, even though Fridman et al. (2018) stated that the two models would be available as open-sources, it was impossible to access the implemented models while this study was in progress. The 3D-CNN structure can be one possible neural net structure in terms of workload estimation.

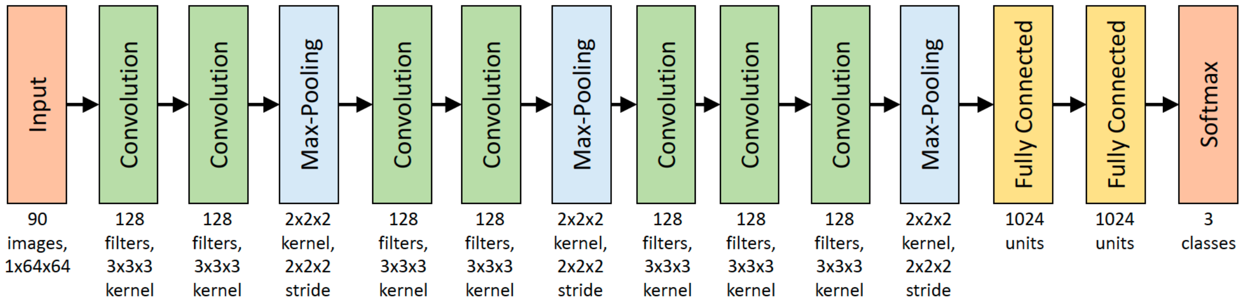

Input
Convolution
Convolution
Max-Pooling
Convolution
Convolution
Max-Pooling
Convolution
Convolution
Convolution
Max-Pooling
Fully Connected
Fully Connected
Softmax
90 images, 1x64x64
128 filters, 3x3x3 kernel
128 filters, 3x3x3 kernel
2x2x2 kernel, 2x2x2 stride
128 filters, 3x3x3 kernel
128 filters, 3x3x3 kernel
2x2x2 kernel, 2x2x2 stride
128 filters, 3x3x3 kernel
128 filters, 3x3x3 kernel
128 filters, 3x3x3 kernel
2x2x2 kernel, 2x2x2 stride
1024 units
1024 units
3 classes

Figure 2.9. The three-dimensional convolutional nerual network (3D-CNN) model structure (Fridman et al. (2018); ©ACM2018)

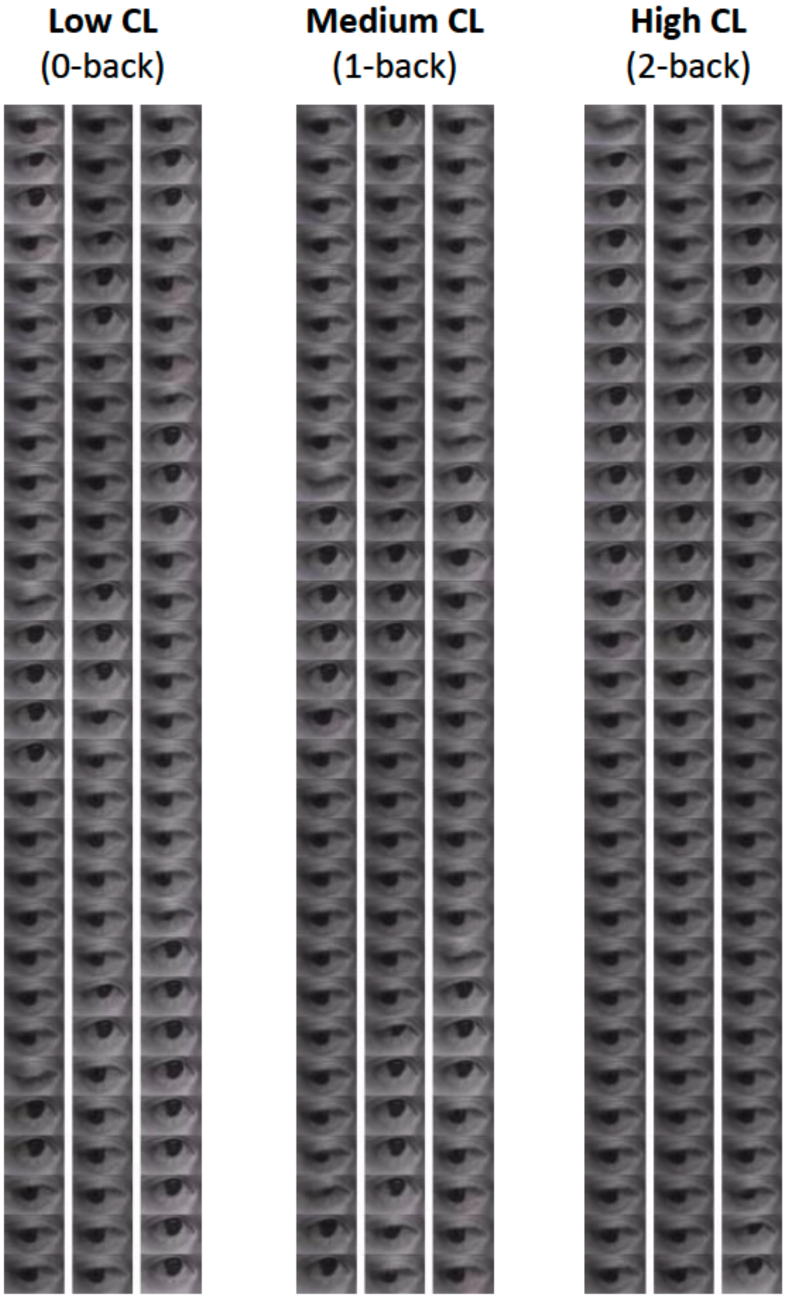


Figure 2.10. Eye blinks of three different cognitive loads (Fridman et al. (2018); ©ACM2018)

2.5 Summary

This chapter elaborated a review of the literature regarding the concept of human-robot collaboration, workload estimation related to behavioral signals, and several learning methods from eye-related features.

CHAPTER 3. RESEARCH METHODOLOGY

The goal of this research is to establish a workload estimation model with the images of significant behavioral cues with visual images. In this chapter, an overview of the proposed structure and the phases of research are introduced. The research overview consists of the following order: an explanation of the available dataset from a preliminary study, preprocessing, data evaluation, and a deep learning model. The final part demonstrates how metrics will evaluate the result of workload estimation.

3.1 Dataset

A preliminary study was conducted with 30 human subjects for data collection and evaluation under Purdue IRB #1812021453. For this preliminary study, two cameras were employed and positioned at the participant's front and side. Participants were asked to wear sensors, including electrocardiogram (ECG), electromyography (EMG), and GSR.

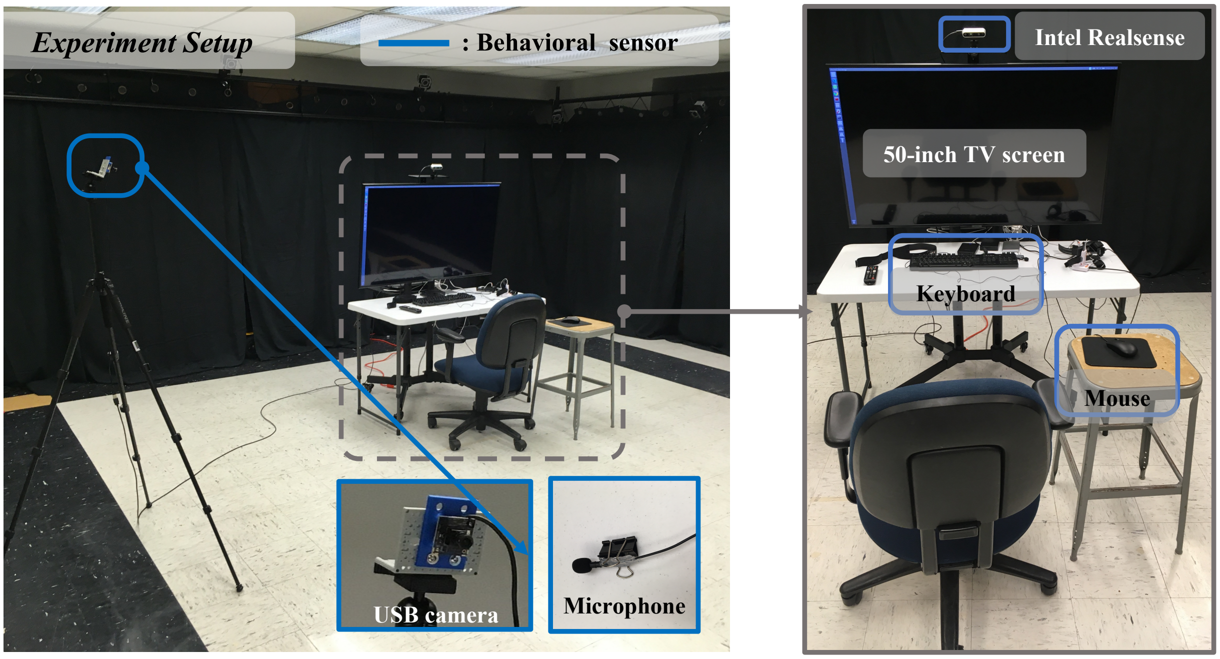
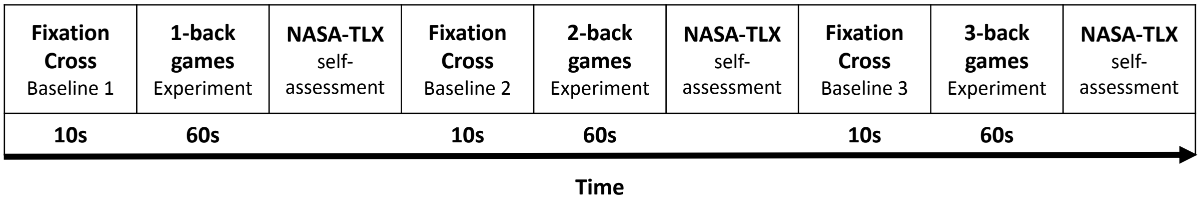


Figure 3.1. User study setup (Jo et al. (2020))

3.1.1 Experimental Setup



Fixation Cross Baseline 1	**1-back games** Experiment	**NASA-TLX** self-assessment	**Fixation Cross** Baseline 2	**2-back games** Experiment	**NASA-TLX** self-assessment	**Fixation Cross** Baseline 3	**3-back games** Experiment	**NASA-TLX** self-assessment
10s	**60s**		**10s**	**60s**		**10s**	**60s**	

Figure 3.2. Workload test procedure

The user study setup is presented in Figure 3.1. After counting down ten to one, the dual *n*-back games (Hampson, Driesen, Skudlarski, Gore, and Constable (2006)) were done consecutively. The dual *n*-back games derive human cognitive load, which requires participants to remember the previous signs presented in a front screen or played through a speaker, as Figure 3.3 shows. Two randomly selected test methods, position and audio cues, are presented. During the test, participants were asked to click the left mouse button if a visible box in the screen is the same as the one shown *n*-steps before, and the right button if given question matched in an auditory way. When clicking one of the mouse buttons, participants do not need to move the mouse or press any keyboard buttons. When completed the memory test games, participants were asked to complete the NASA-TLX self-assessment rating their perceived cognitive state during the games, Figure 3.4 shows. NASA-TLX rating consists of six components, mental demand, physical demand, temporal demand, performance, effort, and frustration. After completing each level of the *n*-back game, participants can answer how much it was demanding they felt during the game as a self-assessment. Each component is answered with scrolling bars to indicate demands from very low to very high. For example, if the degree of mental demand is much less than medium, a participant can scroll the bar of the mental demand section near very low. This procedure is done on three different levels, 1-back, 2-back, and 3-back, as Figure 3.2 depicts. During the user study, the two types of camera data record participants' behavioral responses simultaneously. Sensors record the physiological reactions at the same time.

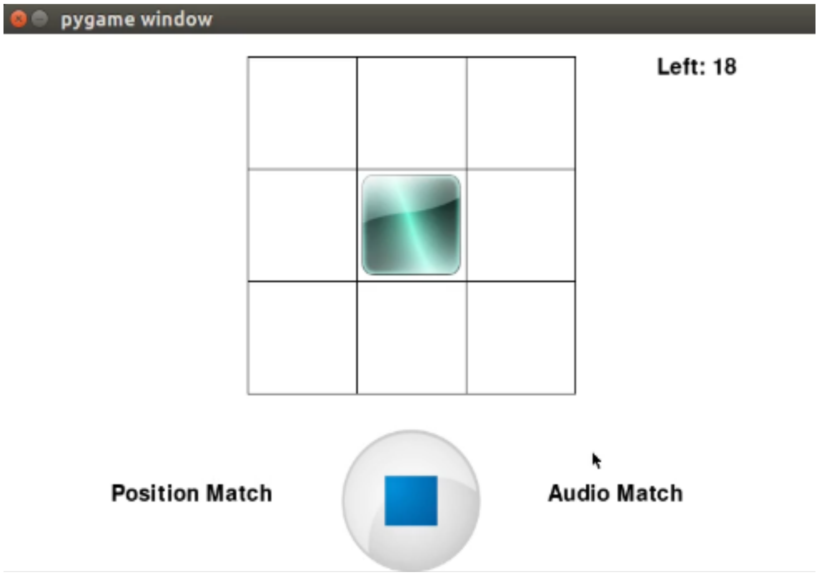
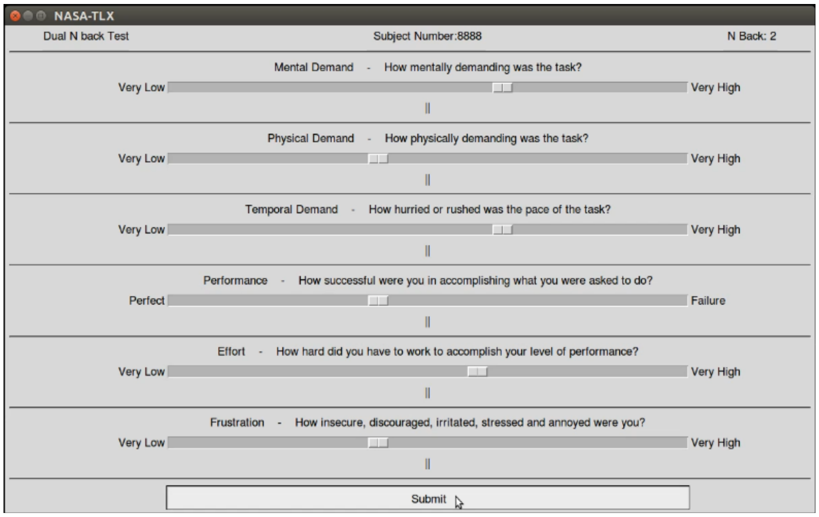


Figure 3.3. Dual *n*-back Graphical User Interface (GUI) in the experiment (Jo et al. (2020))



Dual N back Test Subject Number:8888 N Back: 2

Mental Demand - How mentally demanding was the task?
Very Low Very High

Physical Demand - How physically demanding was the task?
Very Low Very High

Temporal Demand - How hurried or rushed was the pace of the task?
Very Low Very High

Performance - How successful were you in accomplishing what you were asked to do?
Perfect Failure

Effort - How hard did you have to work to accomplish your level of performance?
Very Low Very High

Frustration - How insecure, discouraged, irritated, stressed and annoyed were you?
Very Low Very High

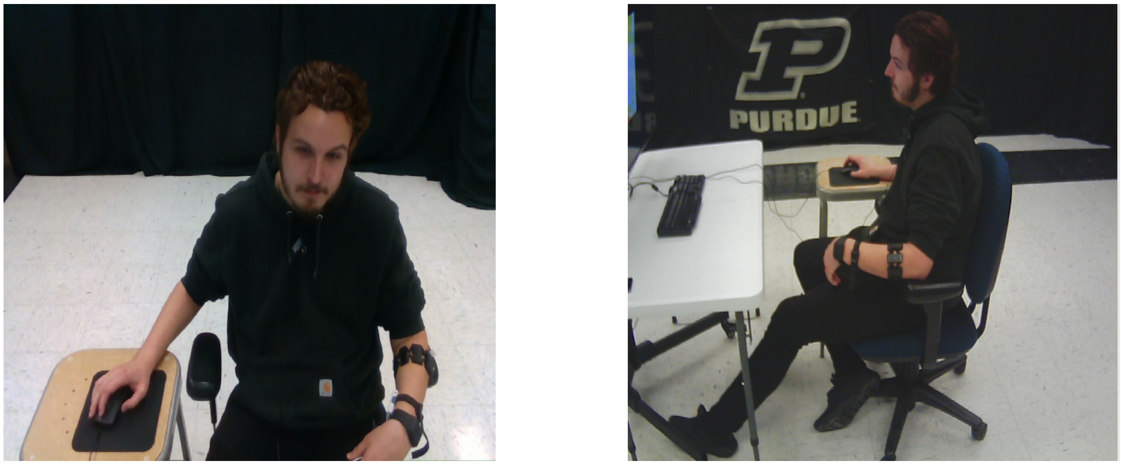
Figure 3.4. NASA-TLX GUI in the experiment (Jo et al. (2020))

As a result of the workload test, following datatypes were collected:

- Video recordings from the front: In order to get psychological and physiological response, the participants' facial expressions and gestures of upper body were recorded, as shown in Figure 3.5(a).

- Video recordings from the side: As displayed in Figure 3.5(b), a side view of the participant was also recorded to monitor the behavior of participants. This recording displayed any observable changes when they sat down.

- Input/Output device measurements: Voice recordings and mouse tracking of the participants were collected during the workload tests.



(a) Front view of a participant and features

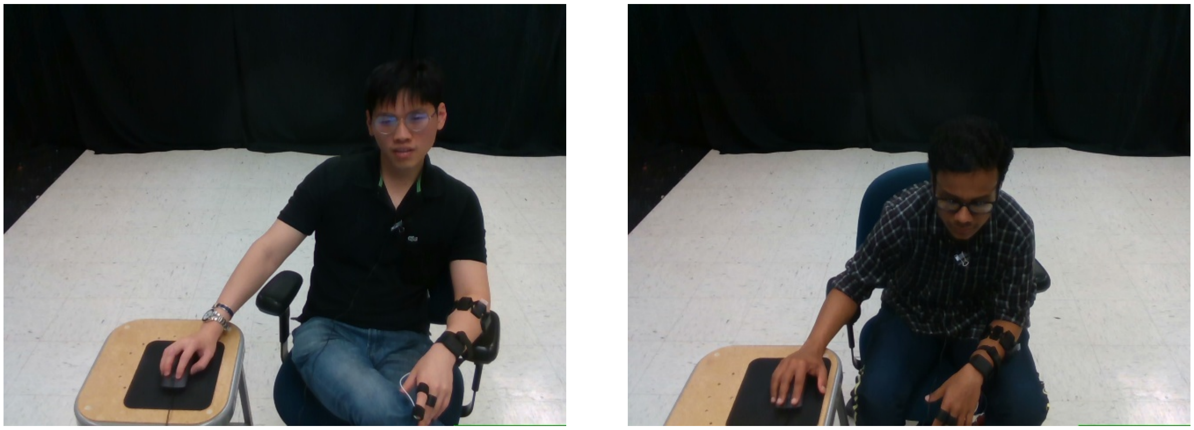
(b) Side view of a participant and features

Figure 3.5. An example of camera images from the front and the side

3.1.2 Data Selection

The part of 16 sets of data among 30 participants is selected to extract a more precise inference of the eye region. The excluded data has two main issues. As Figure 3.6(a) shows, object occlusions with reflected lights around eye region (e.g. wearing eyeglasses). Or, as Figure 3.6(b) presents, the irregular position of the front camera disturbs to get eye-related information The two main issues cause the low accuracy of the open-source eye landmark detection. Therefore, analysis of eye detection was preceded, and data of 16 participants were selected according to the result. When each participant conducted *n*-back games with three distinct levels, the front camera stream and mouse tracking information were chosen.

Mouse tracking has been analyzed from 23 sets of participants. The main reason for excluded data is data error at each experiment. Three types of data are expected among three different levels, but one experiment of data cannot be examined in 6 participant sets. Mouse tracking has been analyzed based on saved position data and ROS time stamps.



(a) Example of eye occlusion and light reflection

(b) Example of irregular position

Figure 3.6. Examples of excluded data

3.1.3 Data Extraction for Analysis

The ROS environment is selected to proceed with the data extraction phase following the rosbag file format of the dataset. The procedure of data extraction is depicted in Figure 3.7. When extracting eye-related data, an open-source library, OpenPose (Cao et al. (2017), *CMU-Perceptual-Computing-Lab/openpose* (n.d.)) is chosen. Each frame is supplied as an input of OpenPose to get facial landmarks. An example of extracting facial landmarks is shown in Figure 3.13. Besides, the library was not supported in the ROS environment. One necessity is that the library is needed to be converted to ROS compatible version to use. Therefore, another open-sourced project of OpenPose wrapper by Zhang and Travers-rhodes (n.d.) has been applied into the feature extraction module and optimized for the characteristics of the dataset. The whole structure of ROS-based data extractor is presented in Figure 3.8.

The library supports the prediction results of given images with high probability, as Figure 3.11 indicates the possible eye landmarks, enabling the use of EAR mentioned in Section 2.2.1.1.

The Equation 3.1 shows how six feature points were calculated. When participants blink their eyes as Figure 3.12 presents, each EAR from the eyes and the mean value of two EAR were stored numerically. Based on the extracted EAR value in a frame, eye blinking can be detected, which enables to measure eye blinking duration in milliseconds. On the other hand, the front camera stream is converted into a group of single images. The size of the extracted images is 640x480, and the number of images is 88,776.

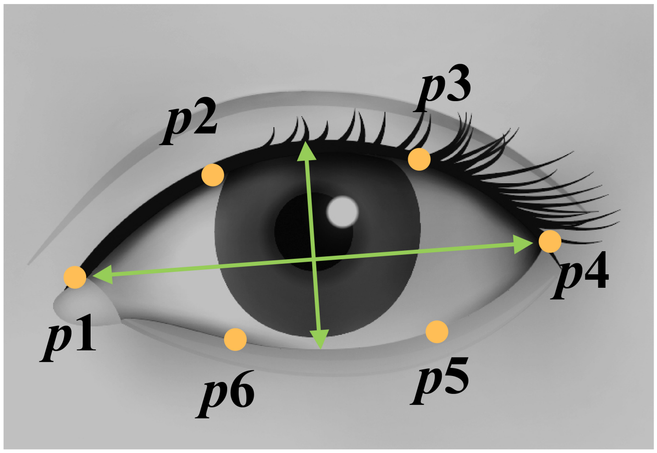


Figure 3.11. Eye landmark prediction on left eye

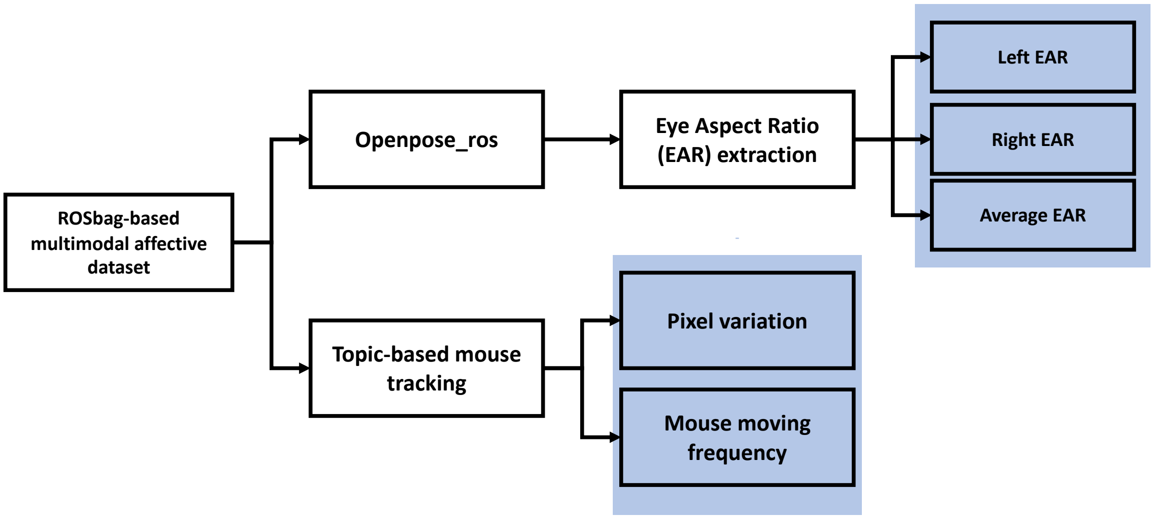

ROSbag-based multimodal affective dataset
Openpose_ros
Eye Aspect Ratio (EAR) extraction
Left EAR
Right EAR
Average EAR
Topic-based mouse tracking
Pixel variation
Mouse moving frequency

Figure 3.7. Data extraction procedure on eye-related and mouse-related behavior

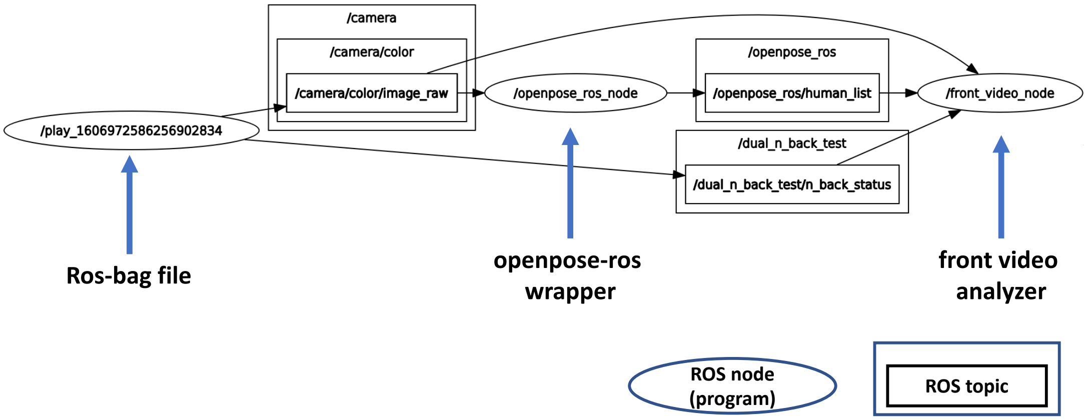


Figure 3.8. ROS-based data extractor

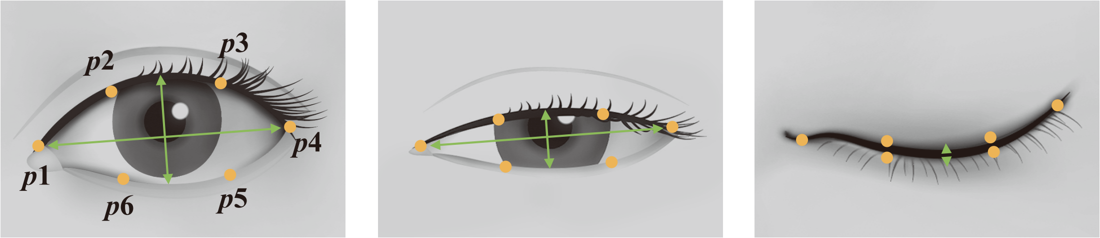
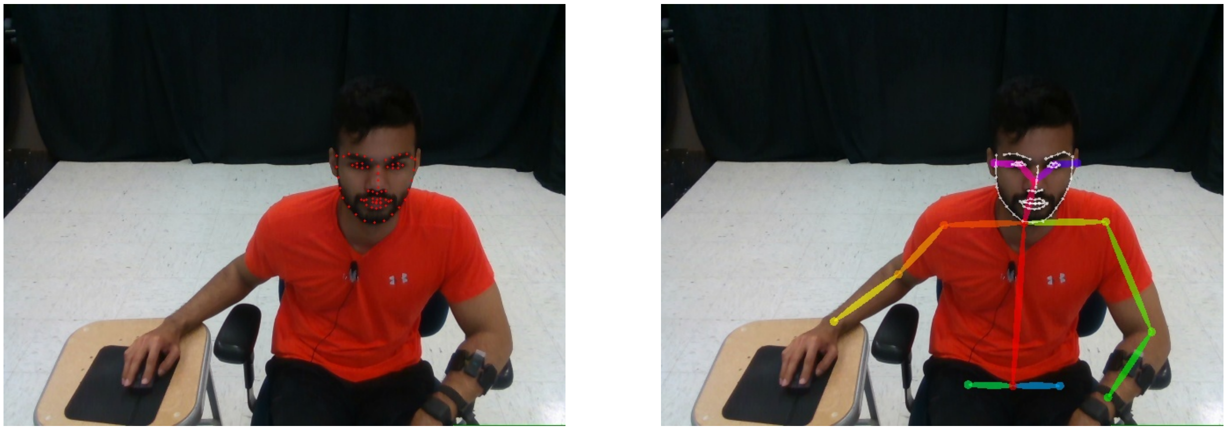


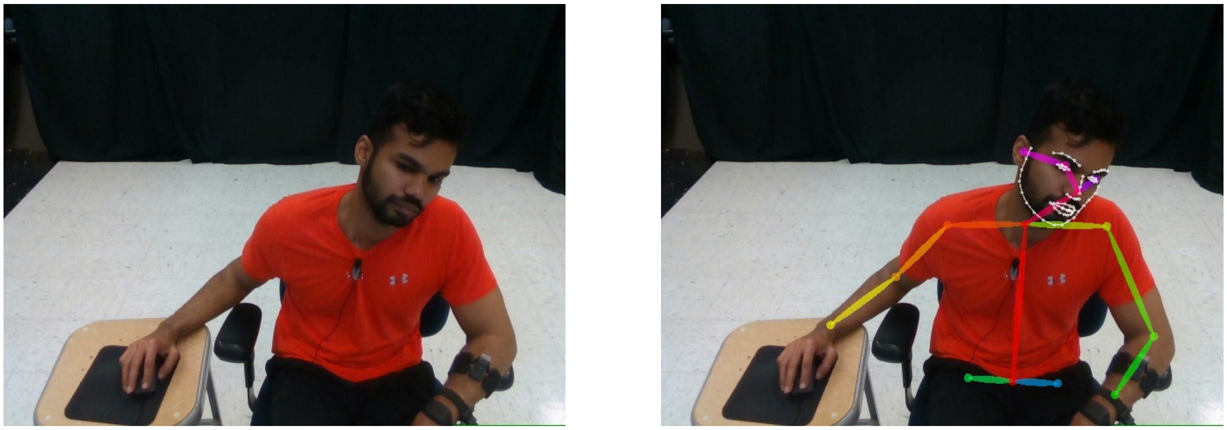
Figure 3.12. Eye Aspect Ratio (EAR) calculation based upon eye closing



(a) Dlib result of a frontal face

(b) OpenPose result of a frontal face

Figure 3.9. A result of facial landmark detection of a frontal face



(a) Dlib result of a tilted face

(b) OpenPose result of a tilted face

Figure 3.10. A result of facial landmark detection of a tilted face

$$Eye\ Apsect\ Ratio(EAR) = \frac{\|p2 - p6\| + \|p3 - p5\|}{2\,\|p1 - p4\|} \tag{3.1}$$

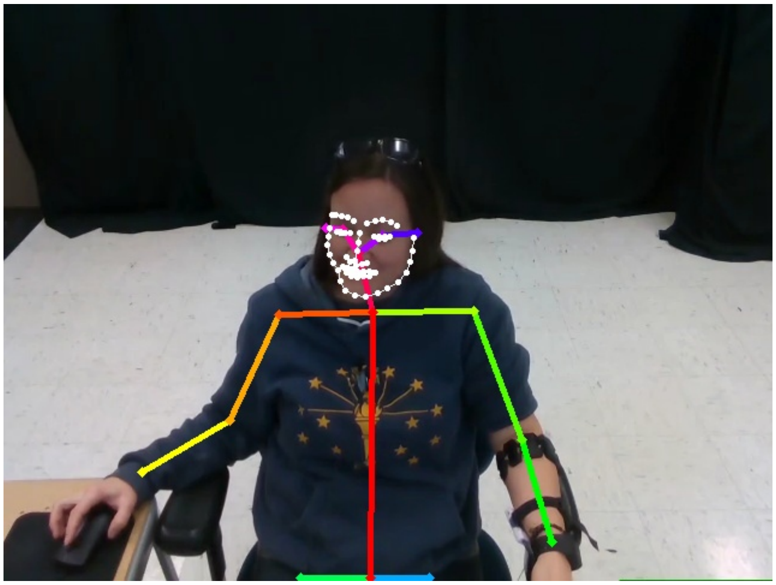


Figure 3.13. Eye landmarks extraction

Mouse tracking data has been extracted with respect to moving frequency and position changes. The data on a mouse moving frequency and position changes are based on the ROS-bag dataset topics designed to send signals indicating a participant moved the mouse to the specific position during the test, as Figure 3.14 illustrates. The position data has the coordinate of the mouse at specific ROS time stamps. Participants' mouse operation can be measured by combining the position of the mouse and the timestamps. An example of mouse-tracking data can be shown as Figure 3.15.

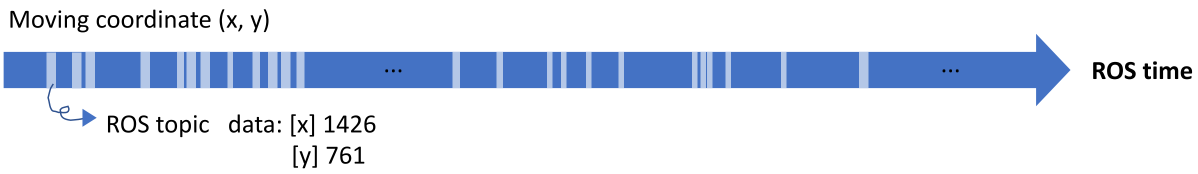


Figure 3.14. ROS topic structure of mouse tracking

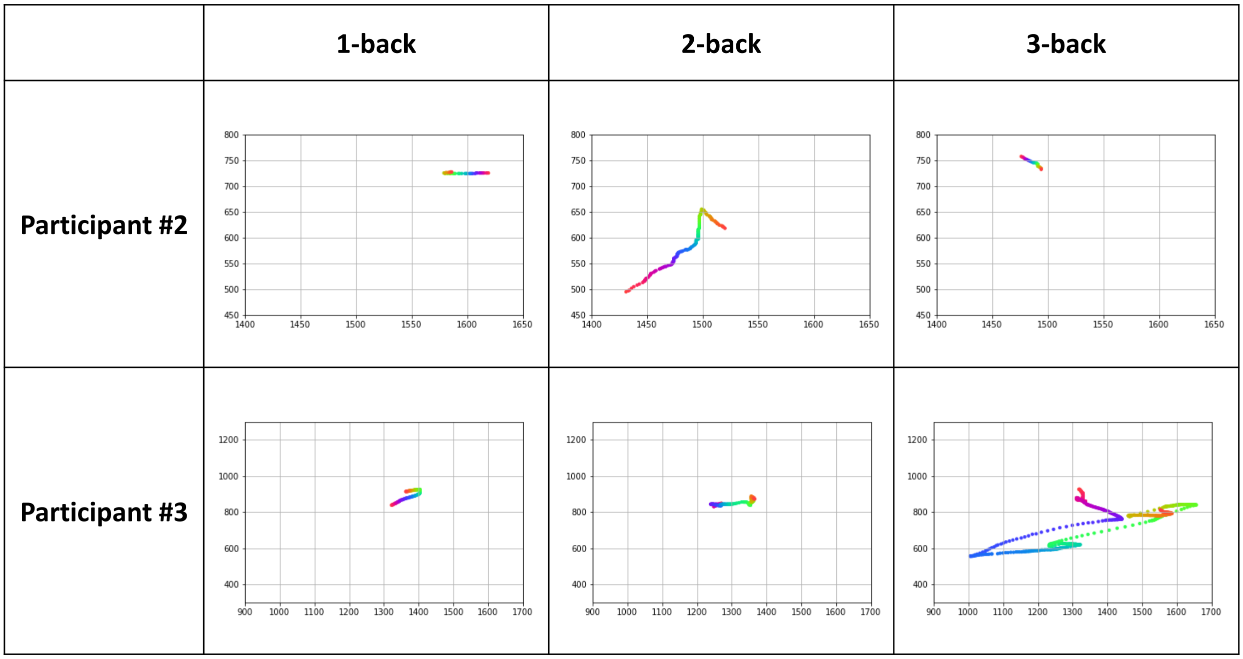


Figure 3.15. Examples of mouse tracks

3.1.4 Statistical Analysis

After extracting behavioral features, statistical tests are done to verify that retrieved information is valid. The blinking duration of each stage and each 10 seconds time frame will be statistically evaluated by one-way repeated measure analysis of variance (rmANOVA; Benedetto et al. (2011); Singh et al. (2013)) since the dataset in Section 3.1 is based on the individuals' dependant response of different *n*-back level. The purpose of statistical evaluation is to determine whether there is a significant difference between baseline and three different levels of *n*-back games. On top of that, every ten seconds of data from baseline to n-back games will be analyzed to discover any workload development during the experiment. In the rmANOVA test, measurement results have been assumed that sphericity is already satisfied.

3.1.4.1 Hypotheses

3.1.4.1.1. Hypothesis 1

The first hypotheses for rmANOVA are the following:

> Null hypothesis H_0: The response of participants in terms of each behavioral cue will have the same measurement during three different levels of the n-back test.
>
> Alternative hypothesis H_α: The behavioral cue of participants will be different during three different n-back tests.

3.1.4.1.2. Hypothesis 2

The second hypotheses for rmANOVA are the following:

> Null hypothesis H_1: The participants' response of each ten seconds from the baseline and *n*-back games will be the same.
>
> Alternative hypothesis H_β: Each measurement of participants will be different during three different *n*-back tests from baseline measurements.

If any behavior measurements are assessed to be of statistical significance in different levels' tests, the given data can be trained with the selected deep neural network to estimate workload. In that case, the visual data will be an input of one of the learning models described in the following Section 3.2.

3.1.5 Data Extraction for Deep Learning

If a behavioral signal is analyzed to be statistically meaningful, the deep learning process is processed according to the statistical result. For deep learning training to estimate the different level workload, image-based input is chosen. The extracted images are transformed to have only eye area information through a preprocessing to train them in the deep learning structure. Since the angle of the face is highly likely different with time, the face region should be aligned to compare frames. The algorithm of facial alignment is followed from Rosebrock (2017). As in Figure 3.16, the slanted head direction in an individual frame is aligned based on the eyes' center points, obtaining a two-dimensional rotation matrix to correct images by the affine transformation. When taking two eye center points, facial landmarks are extracted by OpenPose (*CMU-Perceptual-Computing-Lab/openpose* (n.d.)). Accordingly, the newly updated two eye center points are calculated by a matrix dot operation with the rotation matrix and two original eye center points. Based on the updated eye centers, two eye regions are cropped as 64x64 images from both sides.

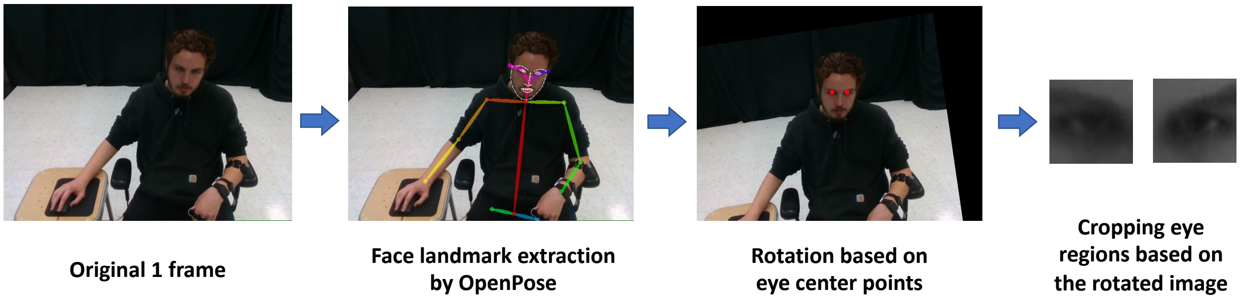

Original 1 frame
Face landmark extraction by OpenPose
Rotation based on eye center points
Cropping eye regions based on the rotated image

Figure 3.16. Eye regional image extraction with angle correction

3.2 Deep Learning Approach

The purpose of 3D-CNN is to classify the different level of human cognitive load from the experiments. This section explains how input for learning has been created, a selected model, and what metrics are used.

3.2.1 3D-CNN Training

This study adopts the 3D-CNN model from Fridman et al. (2018) to estimate workload in this study. As Figure 3.17 shows, the extracted sequential gray-scaled images from the ROS-bag dataset are integrated in 90 frames. Since each image has a 3-dimensional format with 1-channel, the 90 integrated images are in 4-dimension, indicating eyelid movement activity of 3 seconds. The 4-dimensional cubes can be derived from each set of participants' data. Each 3-second sequence is made into five-dimensional chunks that could be fed in the 3D-CNN model. The five-dimensional input can be represented as following:

$$(Sample\ number, Frame\ count, Image\ width, Image\ height, Channel)$$

The sample number illustrates the number of extracted sets of spatial eye features for 3 seconds. The frame count indicates the number of multiplying 30fps and duration. The remaining part, image width, image height, and channel can be derived from individual images. The structure of 3D-CNN model is presented in Table 3.1. The network uses the combination of three-dimensional convolutional layers, max-pooling, and fully connected layers. In the last part, a softmax layer is to have a classification result of 3 different workload levels. Each convolutional layer uses 128 features with a 3x3x3 kernel. Besides, max-pooling layers utilize a 2x2x2 kernel and a 2x2x2 stride at each stage. Up to the author's experience, a problem was found that 3D-CNN occupied a lot of memory while using NVidia RTX 2080 Super GPU, which has an 8-gigabyte capacity. However, this memory issue can be solved using cloud computing resources, for example, using Amazon Web Services (AWS).

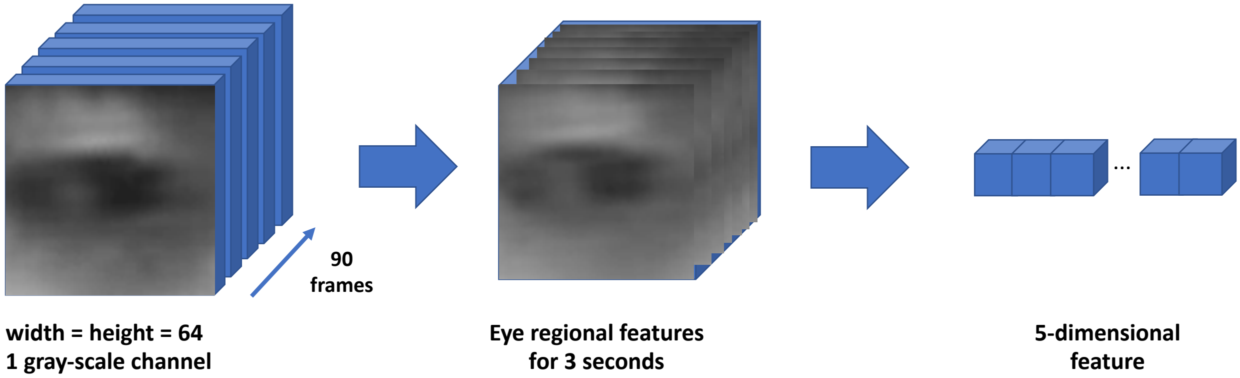


Figure 3.17. The steps of creating 5-dimensional eye regional input

Layer (type)	Options	Shape
Input Layer	None	(Sample number, 90, 64, 64, 1)

Convolution3D	features=128
	kernel=3
		(Sample number, 90, 64, 64, 128)

Convolution3D	features=128
	kernel=3
		(Sample number, 90, 64, 64, 128)

Maxpooling3D	kernel=2
	strides=2
		(Sample number, 45, 32, 32, 128)

Convolution3D	features=128
	kernel=3
		(Sample number, 45, 32, 32, 128)

Convolution3D	features=128
	kernel=3
		(Sample number, 45, 32, 32, 128)

Maxpooling3D	kernel=2
	strides=2
		(Sample number, 22, 16, 16, 128)

Convolution3D	features=128
	kernel=3
		(Sample number, 22, 16, 16, 128)

Convolution3D	features=128
	kernel=3
		(Sample number, 22, 16, 16, 128)

Convolution3D	features=128
	kernel=3
		(Sample number, 22, 16, 16, 128)

GlobalMaxpooling3D	None	(Sample number, 128)
Fully Connected	1024	(Sample number, 1024)
Fully Connected	512	(Sample number, 512)
Softmax	3	(Sample number, 3)

Table 3.1. *The structure of 3D-CNN (Fridman et al. (2018))*

3.2.2 Evaluation metrics

F1-score is widely used to evaluate the accuracy of machine learning or deep learning models. In this proposed study, as seven universal emotions, each emotion-labeled class has four cases. We define them as true-positive (TP), true-negative (TN), false-positive (FP), and false-negative (FN), respectively. For example, true-positive means for when the given data indicates the eye blinking activity of 1-back, and the result of the predictor indicates perceived information is about 1-back test. In calculating F1-scores, accuracy is essential, but it is more important not to make incorrect detection results. There are two terminologies to indicate the detection rate and accuracy, as recall (Eq. 3.2) and precision (Eq. 3.3).

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \tag{3.2}$$

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \tag{3.3}$$

$$F1 - Score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{3.4}$$

3.3 Summary

This chapter discussed the methodology of the proposed study. The dataset from preliminary research on workload derivation test, statistical analysis and hypotheses, a deep learning approach for detecting workload based on temporal and spatial data from extracted images, and a method for evaluating strategy are presented.

CHAPTER 4. RESULTS AND EVALUATION

In this chapter, steps of verifying extracted data, selecting behavioral measures, preprocessing results, and evaluating trained data are presented. The analysis of behavioral measurements based on the dataset in Section 3.1 is explained in Section 4.1. The evaluation of workload result is depicted in Section 4.2 will be presented and explained based on proposed methodology in Section 3.2.1.

4.1 Statistical Analysis

As mentioned in Section 2.4.2, according to the need for data analysis, this section describes the statistical analysis results. First, data validation between workload level and NASA-TLX questionnaire result is presented. Accordingly, statistical analysis of behavioral measurements from data is presented. The purpose of the analysis to find valid behavioral cues to estimate different level of workload. The statistical analysis is done with the IBM SPSS Statistics software.

4.1.1 Data validation between workload and NASA-TLX questionnaire result

After collecting all the data, we analyzed the dual *n*-back game score sand the questionnaire results created by the participants after experiments of each level are completed. Specifically, as Figure 4.1 shows, *n*-back game score, depicted in the blue color, gradually

Rating	1-back	2-back	3-back
Game score	100	55.56	40
Mental demand	40	60	70
Physical demand	40	65	75
Temporal demand	35	55	65
Performance	40	60	70
Effort	40	65	75
Frustration	50	60	70

Table 4.1. *Median values of the dual n-back game score and NASA-TLX questionnaire*

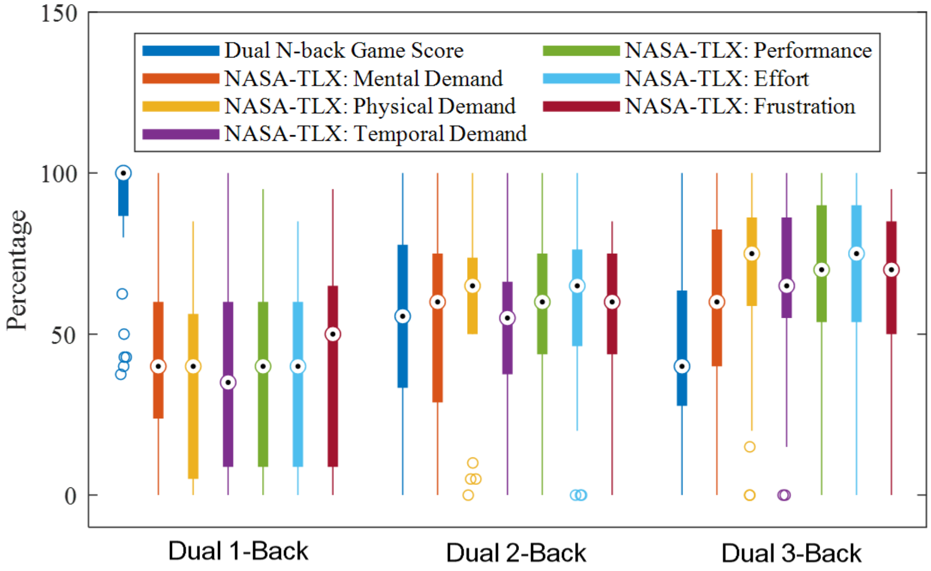


Figure 4.1. The results of the Dual *n*-Back game score and NASA-TLX questionnaire according to the level of the workload (Jo et al. (2020))

dropped as the game level increased (1-back: 100; 2-back: 55.56; 3-back: 40). As Table 4.1, shows, NASA-TLX ratings show that the dual 1-back has the lowest ratings, while the dual 3-back has slightly the highest ratings. The result means that participants' workload is relatively high at the dual 3-back test than other levels. The median values are increased in each NASA-TLX rating items. Considering the game score and questionnaire result, the dual *n*-back game could successfully derive the different cognitive load levels. Based on the analysis, the dual 1-back can be considered as low-level workload, 2-back as medium one, and 3-back as high-level workload. Each component of NASA-TLX is measured with one-way rmANOVA test. As Table 4.2 represents, All tests showed significance, which means the results of self-assessment in three levels are statistically different.

	F-test	p-value	Significance
Game score	F(2, 56) = 36.622	p = .000 (p <.05)	Significant difference
Mental demand	F(2, 56) = 36.416	p = .000 (p <.05)	Significant difference
Physical demand	F(2, 56) = 4.137	p = .021 (p <.05)	Significant difference
Temporal demand	F(2, 56) = 13.684	p = .000 (p <.05)	Significant difference
Performance	F(2, 56) = 29.178	p = .000 (p <.05)	Significant difference
Effort	F(2, 56) = 17.911	p = .000 (p <.05)	Significant difference
Frustration	F(2, 56) = 33.616	p = .000 (p <.05)	Significant difference

Table 4.2. *rmANOVA results of NASA-TLX*

4.1.2 Data analysis on measurements

As extracted measures from the dataset in Section 3.1, each measure, blink frequency, blink duration, mouse moving frequency, and mouse position changes, is examined to find statistical difference between the workload level. Table 4.3 shows mean and standard deviation values of measures. The values in parenthesis indicate standard deviation values. While eye blink frequency has minor fluctuation in mean values (1-back: 18 counts, 2-back: 18.25 counts, 3-back: 18.44 counts), blink duration, mouse moving frequency, and mouse position changes have variations among levels. Blink duration shows gradual changes among levels, while mouse moving frequency has slightly decreased. The dual 2-back mouse position changes was the lowest than other levels.

Measure	Task		
	1-back	2-back	3-back
Blink frequency [count]	18 (8.17)	18.25 (11.16)	18.44 (9.80)
Blink duration [milliseconds]	409.30 (103.48)	438.44 (95.29)	473.14 (112.8)
Mouse moving frequency [count]	3.11 (3.05)	3.23 (3.56)	5.61 (8.25)
Mouse moving duration [nanoseconds]	0.87 (1.69)	0.35 (0.23)	0.39 (0.37)
Mouse position changes [pixel]	4.19 (4.56)	3.40 (1.70)	4.10 (2.70)

Table 4.3. *Mean and standard deviation values of levels*

Mean value of each measure from each experiment is evaluated by one-way rmANOVA to find whehter the result is significantly different among levels. The result of significance is shown in Table 4.4. If p-value is less than 0.05, null hypothesis H_0 in Section 3.1.4.1.1 is rejected, in other words, the alternative hypothesis, there are difference across the experiment levels, is valid. Blink frequency has no significant difference among levels ($F(2,30) = 0.038$, $p >.05$). Mouse moving frequency has also no significance across the experiment levels ($F(2,44) = 2.036$, $p >.05$). The time duration of moving a mouse did not show any significance ($F(2,44) = 1.889$, $p >.05$). Lastly, mouse position changes have no significant gaps between levels ($F(2,44) = 0.419$, $p >.05$). The only significant measure is blink duration across the levels, showing p-value less than 0.05. According to the rmANOVA test result, blink duration can be meaningful cues predicting people's workload.

Based on the result of the table 4.3, the each *n*-back result is compared to the result of other level games. The Table 4.5 presents the comparison of each level. One finding from the rmANOVA test is that there is a meaningful difference between 1-back and 3-back with a significance of 98% confidence. There is no significance between the pair of 1-back and 2-back, and 2-back and 3-back.

Measure	F-test	p-value	Significance
Blink frequency	$F(2,30) = 0.0386$	$p = .9621$ ($p >.05$)	No significance
Blink duration	$F(2,30) = 3.50$	$p = .0430$ ($p <.05$)	Significant difference across levels
Mouse moving frequency	$F(2,44) = 2.036$	$p = .143$ ($p >.05$)	No significance
Mouse moving duration	$F(2,44) = 1.889$	$p = .163$ ($p >.05$)	No significance
Mouse position changes	$F(2,44) = 0.419$	$p = .660$ ($p >.05$)	No significance

Table 4.4. *F-test result of measurements*

Level	**Source**	**Sum of Squares**	**df**	**Mean Square**	**F-value**	**Significance**
1-back & 2-back	level	6795.713	1	6795.713	1.559	.231
	Error(level)	65399.775	15	4359.985		
2-back & 3-back	level	9629.219	1	9629.219	2.015	.176
	Error(level)	71667.708	15	4777.847		
1-back & 3-back	level	32603.610	1	32603.610	6.699	.021
	Error(level)	73007.891	15	4867.193		

Table 4.5. *rmANOVA results of each level comparison*

4.1.3 Analysis on blink duration

Based on the statistical analysis in Section 4.1.2, eye blinking duration has been analyzed slicing experiment results in ten seconds as same as the duration of baseline. Here, each baseline refers to the ten seconds when participants were asked to watch the fixation cross before the dual *n*-back tests were initiated.

Following null hypothesis H_1 in Section 3.1.4.1.2, measured eye blinking duration of baseline and each dual *n*-back in the same time period, ten seconds, is compared. Generally, the mean blink duration values are higher at each baseline, showing the longest duration at 3-back games (1-back: 725.81ms, 2-back: 735.35ms, 3-back: 912.00ms). The small difference found according to the Table 4.6 is blinking duration shows slight higher values among time frames in *n*-back tests.

The Table 4.7 shows a comparison result between same time periods of baseline and experiment. All different levels show a significant difference between each duration, referring to the data from baseline and experiment are different (1-back: $F(6,90) = 4.988$, $p <.05$; 2-back: $F(6,90) = 3.846$, $p <.05$; 3-back: $F(6,90) = 2.197$, $p <.05$). Therefore, the alternative hypothesis H_β is selected, meaning eye blinking reaction of participants are different from baseline and dual *n*-back test.

Measurement	Baseline	*n*-back test					
		Duration 1	Duration 2	Duration 3	Duration 4	Duration 5	Duration 6
1-back	725.91 (538.58)	345.06 (196.04)	339.79 (163.40)	427.62 (227.35)	402.97 (166.11)	453.25 (290.72)	313.09 (190.40)
2-back	735.35 (593.40)	348.75 (240.87)	424.49 (174.46)	493.13 (384.96)	361.99 (188.78)	318.39 (205.60)	331.66 (248 34)
3-back	912.00 (1195.75)	409.56 (188.80)	477.36 (359.03)	488.10 (159.17)	480.94 (267.68)	445.69 (219.02)	363.23 (245.81)

Table 4.6. *Average and variance values of every ten second frame*

Level	**Source**	**Sum of Squares**	**df**	**Mean Square**	**F-value**	**Significance**
1-back	Measure	1885788.67	6	314298.111	4.988	.000
	Error(measure)	5670664.57	90	63007.384		
2-back	Measure	2089690.73	6	348281.788	3.846	.002
	Error(measure)	8149446.94	90	90549.410		
3-back	Measure	3196104.11	6	532684.019	2.197	.05
	Error(measure)	21822583.7	90	242473.152		

Table 4.7. *rmANOVA results of levels in ten seconds frame*

4.1.4 Summary of Statistic Analysis

First of all, the eye blinking duration is only a significant behavioral cue when estimating different human cognitive load levels. Table 4.5 verifies that participants' response in the 1-back and 3-back game are distinct, while two consecutive level experiments do not show significance. As analyzed in Table 4.7, there is a clear difference between when a person is not given any task and when a task is performed. All in all, eye blinking activity could be a promising behavioral cue to have automated estimation of human cognitive workload.

4.2 Result

This chapter shares three items of the result of this study: the result of image preprocessing, training, and a confusion matrix of a trained model based on 3D-CNN neural network structure.

4.2.1 Image Preprocessing

Based on the preprocessing stage in Section 3.1.5, the number of extracted eye images of both side are about 177,034. The original images are about 88,776, expecting the aligned the number of images were 177,552. The loss of images was 60,518.

4.2.2 Training Steps and Result

The training model is implemented in TensorFlow with four batch-size and 20 epoch due to extensive memory consumption. The training phase has been done in AWS Deep Learning AMI (Ubuntu 18.04), utilizing p2.xlarge instance with Tesla K80 GPU, taking approximately 3.5 hours. The five-dimensional, temporal, and spatial features are made to 910 samples. The ratio of training and testing data has been divided into 80-20, 728 samples of training, and 182 samples of testing. During the training, the best result of training based on value accuracy is saved. The result of the 19th epoch among 20 trials is finally chosen as the best result with 39.041% value accuracy. When evaluating the model with previously split test data, the accuracy is about 38% with F1-score 0.45, 0.14, and 0.40 in the order of 1-back to 3-back. Following the statistical result that there is a statistical difference between dual 1-back and 3-back between them, binary classification training is done. The accuracy of the trained model is 51% with F1-score 0.66 and 0.14 in the order of the dual 1-back and 3-back.

	Level	Precision	Recall	F1-score	Accuracy
3-level classification	1-back	0.32	0.78	0.45	0.38
	2-back	1.00	0.08	0.14	
	3-back	0.52	0.32	0.40	
2-level classification	1-back	0.50	0.97	0.66	0.51
	3-back	0.71	0.07	0.14	

Table 4.8. *Precison, recall, F1-score of trained data*

4.3 Discussion

With interest in behavioral signals other than sensors when observing human responses, behavioral cues during three different levels of *n*-back games have been examined to estimate human cognitive workload. Among available behavioral signals from the multimodal affective dataset, eye blinks and mouse tracking have been chosen as promising ones.

In Table 4.1, there is a big difference between dual 1-back and 2-back games when looking at the game scores and the NASA-TLX self-assessment score. It is worth noting that during the experiment, some participants experienced difficulty in playing 2-back games after they completed the previous ones. The difficulty caused drastic falls between game scores. In Table 4.2, the numerical gaps between 2-back and 3-back are lesser than the difference between 1-back and 2-back, which means participants showed their mental burden at the first two games more. One guess about the smaller difference between 2-back and 3-back is that some participants even gave up their 3-back games if they experienced pressure or frustration on 2-back games. It seems that memorizing visual or phonetic signs two steps before might cause workload in the HMI setting.

Their difficulties become remarkable when they use the mouse in the experiment. In Table 4.3, the number of mouse usage times gradually increased as the workload level increases. The distance of mouse position has the lowest mean and standard values at the 2-back games. The mouse relevant data on 2-back indicates that their usage comparably decreased from 1-back with lower moving duration and position changes. The lower time of moves and changes show participants made fewer moves at the mouse usage. Even though they were required to click buttons on the mouse, not relocating it, their activities in 2-back decreased because they might focus on the games more to have accurate answers, which refrains them from using redundant behavior responses. At the 3-back, all mouse data were increased. The increased value may be caused by individuals who gave up the test at the level. The value changes of blink frequency and blink duration in Table 4.3 are unexpected because we expected that value would decrease when participants play more complicated games. The frequency and duration values are slightly increased, which means when their workload increases, they blink longer.

However, an interesting point found in Table 4.5. Although some participants expressed their mental burden transition of 1-back to 2-back, there is no statistical significance on their blink duration. The statistical difference between 1-back and 3-back exists. As Li et al. (2020) stated in their discussion, given tasks could not arouse all workload levels in short experiments. Each game conducted about 60 seconds, which might not be enough to trigger different cognitive states. Taking into consideration that consecutive games did not show their significance, the experiment duration could be longer, or the level could be increased.

When looking at Table 4.6, blink duration is longer in the baseline. During the experiment divided into ten seconds as same as the baseline, participants' blinking duration increased as time went by and decreased at the last part of the experiments. Considering the variance of each measure, it is yet difficult to say all participants showed similar patterns. However, it is an intriguing point that some participants showed the longest eye blinking in the middle of experiments at the third duration, while others showed the shortest eye blinking at the 2-back games. With the result of Table 4.7, blink duration in baseline and experiment are significantly different. It would be able to distinguish human blinks in natural and task environments.

Regarding the results of deep learning, the accuracy is about 38%. Paying attention to the F1-scores calculated for each level, we can see that the score of 2-back games is significantly lower than other games. The result aligns with the statistical output of Table 4.5 that the 2-back game result does not have a statistical difference from the outcome of 1-back and 3-back games. Another trial is training 1-back and 3-back game results except for 2-back ones. The trial with two different levels is because the dual 1-back and 3-back show a significant difference from the one-way rmANOVA test. However, the accuracy of the newly trained model is 51%, which is a similar result of choosing one between two.

There are two possible steps to improve the model's accuracy. It would be a challenging point to measure generalized estimation of the human workload because of individual characteristics. One trial is that estimating human behavior patterns compared to the baseline. In the experiment, the time focusing a fixation cross was utilized as the baseline. It would be possible to classify human behavior when they naturally present in the HMI setting. Based on the result during baseline, we could expect that each participant falls in a specific category. For example, some individuals tend to blink longer, while others do not. If we can get the relevant point of eye blink duration between baseline and experiment, it might be able to optimize the workload classification problem. Another trial is increasing the amount of data. In the training phase, 16 sets of blinks were utilized. When utilizing three sets of data, the accuracy was about 76%. The higher accuracy result aligns with the result of Fridman et al. (2018) that they used 90 samples with six seconds of eye behavior in training in the same neural net structure. It could be concluded as more prolonged eye behavior should be examined when estimating human workload. Still, it is not easy to generalize that the trained model with reduced data can classify all participants' different workload levels with possibilities of overfitting. If the data increases, it might be helpful to find a pattern of blinking.

Taking into account that physiological or neuro-physiological measurements can be other cues to estimate workload as well, different modalities can be joined together. Not only confined the estimation into behavioral measures, but other methods also can be utilized. A multimodal approach will be an alternative one to improve the assessment of workload. On top of that, one of the original objectives not stated in the previous sections was running a real-time workload estimator in the local environment. When loading the trained model, it requires greater memory consumption due to larger dimensional inputs than other neural networks. It would discourage using estimation of physiological or neuro-physiological sensors. Therefore, other neural networks considering spatiotemporal features such as Convolutional LSTM (ConvLSTM; Shi et al. (2015)) or 3D Residual Networks (3DResNet; Hara, Kataoka, and Satoh (2017)) might be alternatives to find whether different networks would work in terms of reducing memory consumption.

CHAPTER 5. CONCLUSION

This chapter illustrates the conclusion of the study, challenges during this study, and expected future study.

5.1 Conclusion

This study demonstrates the human cognitive load estimation with behavioral cues. As behavioral measures, two simple data have been chosen, mouse tracking data and eye blink. Based on the statistical analysis, we found a significant difference in eye blinking duration between the lowest and the highest levels with 98% occurrence. On the contrary, the mouse-tracking data analysis did not significantly differ among the dual *n*-back levels. The data of eye blinking is given to a 3d convolutional deep neural networks to be learned. Overall, the trained model can assess with 39.04% accuracy in three workload classification, while another model trained with the dual 1-back and 3-back data in 51% accuracy. We tried to find the relationship between the workload and human behavioral signals, but in conclusion, it is unlikely to estimate the workload with mouse tracking or eye blinking.

5.2 Challenges

In this study, the front videos and mouse tracking data were utilized as an interesting data stream to analyze. Regarding mouse-tracking data, there was a relatively rare disturbance. However, the eye region was solely based on recorded pixels, which can be distorted from the field of view of cameras, occluded from some objects, or partially missing due to face rotation. In the dataset, the distance and angle from the front camera to participants were not constant, as mentioned in 3.1.2, which might require different image correction techniques and different EAR threshold values per each participant. When aligning facial regions, it seems that the alignment should not have been done in the 2D environment, but in the 3D environment as mentioned by Fridman et al. (2018).

On top of that, one issue regarding facial landmark detection unexpectedly computed key points differently. To be more specific, when a single video was evaluated by a ROS-based calculation, the extracted facial points, and accordingly, the EAR value was not expected to have the same value at discrete trials. Considering the nature of the open-source library trained from a neural network structure, it produces non-identical values at each trial. When it comes to the problem requiring a precise value-based decision, it could be challenging to have accurate estimation results.

Maior, das Chagas Moura, Santana, and Lins (2020) stated that EAR threshold is not sufficient to be applied to people in a different condition, especially, natural eye openness is different. Even the Rosbag-based Multimodal Affective Dataset has different angles at each participant, which will cause considerable changes when dealing with various participants' behavioral data. Taking into consideration that facial landmark detection computes different values at each trial, even the EAR value from one after another trial cannot be the same.

5.3 Future Study

This study explored two types of behavioral signals, eye blinking and mouse tracking. However, vision-based data have flourished data, such as pose and its development, facial expressions, and gaze. As the expected progress, extracting data, analyzing them statistically if needed, training models, and testing them will be the same. Moreover, not limited to behavioral cues but expanded to physiological sensors, multimodal analysis can be expected as a future study.

REFERENCES

Benedetto, S., Pedrotti, M., Minin, L., Baccino, T., Re, A., & Montanari, R. (2011). Driver workload and eye blink duration. *Transportation research part F: traffic psychology and behaviour*, *14*(3), 199–208.

Boehm-Davis, D. A., Gray, W. D., & Schoelles, M. J. (2000). The eye blink as a physiological indicator of cognitive workload. In *Proceedings of the human factors and ergonomics society annual meeting* (Vol. 44, pp. 6–116).

Borghini, G., Astolfi, L., Vecchiato, G., Mattia, D., & Babiloni, F. (2014). Measuring neurophysiological signals in aircraft pilots and car drivers for the assessment of mental workload, fatigue and drowsiness. *Neuroscience & Biobehavioral Reviews*, *44*, 58–75.

Cao, Z., Simon, T., Wei, S.-E., & Sheikh, Y. (2017). Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 7291–7299).

Cech, J., & Soukupova, T. (2016). Real-time eye blink detection using facial landmarks. *Cent. Mach. Perception, Dep. Cybern. Fac. Electr. Eng. Czech Tech. Univ. Prague*, 1–8.

Chen, F., Sekiyama, K., Cannella, F., & Fukuda, T. (2013). Optimal subtask allocation for human and robot collaboration within hybrid assembly system. *IEEE Transactions on Automation Science and Engineering*, *11*(4), 1065–1075.

Chen, S., Epps, J., Ruiz, N., & Chen, F. (2011). Eye activity as a measure of human mental effort in hci. In *Proceedings of the 16th international conference on intelligent user interfaces* (pp. 315–318).

Cmu-perceptual-computing-lab/openpose. (n.d.). CMU-Perceptual-Computing-Lab. Retrieved from `https://github.com/CMU-Perceptual-Computing-Lab/openpose`

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

Debie, E., Rojas, R. F., Fidock, J., Barlow, M., Kasmarik, K., Anavatti, S., … Abbass, H. A. (2019). Multimodal fusion for objective assessment of cognitive workload: a review. *IEEE Transactions on Cybernetics*.

Debue, N., & Van De Leemput, C. (2014). What does germane load mean? an empirical contribution to the cognitive load theory. *Frontiers in psychology*, *5*, 1099.

Fahim, S. R., Datta, D., Sheikh, M. R. I., Dey, S., Sarker, Y., Sarker, S. K., … Das, S. K. (2020). A visual analytic in deep learning approach to eye movement for human-machine interaction based on inertia measurement. *IEEE Access*, *8*, 45924-45937.

Fridman, L., Reimer, B., Mehler, B., & Freeman, W. T. (2018). Cognitive load estimation in the wild. In *Proceedings of the 2018 chi conference on human factors in computing systems* (pp. 1–9).

Guo, Q., & Agichtein, E. (2012). Beyond dwell time: estimating document relevance from cursor movements and other post-click searcher behavior. In *Proceedings of the 21st international conference on world wide web* (pp. 569–578).

Hampson, M., Driesen, N. R., Skudlarski, P., Gore, J. C., & Constable, R. T. (2006). Brain connectivity related to working memory performance. *Journal of Neuroscience*, *26*(51), 13338–13343.

Hara, K., Kataoka, H., & Satoh, Y. (2017). Learning spatio-temporal features with 3d residual networks for action recognition. In *Proceedings of the ieee international conference on computer vision workshops* (pp. 3154–3160).

Hart, S. G., & Staveland, L. E. (1988). Development of nasa-tlx (task load index): Results of empirical and theoretical research. In *Advances in psychology* (Vol. 52, pp. 139–183). Elsevier.

Herff, C., Heger, D., Fortmann, O., Hennrich, J., Putze, F., & Schultz, T. (2014). Mental workload during n-back task—quantified in the prefrontal cortex using fnirs. *Frontiers in human neuroscience*, *7*, 935.

Jo, W., Kannan, S. S., Cha, G.-E., Lee, A., & Min, B.-C. (2020). Rosbag-based multimodal affective dataset for emotional and cognitive states. *arXiv preprint arXiv:2006.05102*.

King, D. E. (2009). Dlib-ml: A machine learning toolkit. *The Journal of Machine Learning Research*, *10*, 1755–1758.

Lasota, P. A., & Shah, J. A. (2015). Analyzing the effects of human-aware motion planning on close-proximity human–robot collaboration. *Human factors*, *57*(1), 21–33.

Li, J., Li, H., Umer, W., Wang, H., Xing, X., Zhao, S., & Hou, J. (2020). Identification and classification of construction equipment operators' mental fatigue using wearable eye-tracking technology. *Automation in Construction*, *109*, 103000.

Longo, L. (2011). Human-computer interaction and human mental workload: Assessing cognitive engagement in the world wide web. In *Ifip conference on human-computer interaction* (pp. 402–405).

Maior, C. B. S., das Chagas Moura, M. J., Santana, J. M. M., & Lins, I. D. (2020). Real-time classification for autonomous drowsiness detection using eye aspect ratio. *Expert Systems with Applications*, 113505.

McColl, D., Jiang, C., & Nejat, G. (2016). Classifying a person's degree of accessibility from natural body language during social human–robot interactions. *IEEE transactions on cybernetics*, *47*(2), 524–538.

Mukai, T., Hirano, S., Nakashima, H., Kato, Y., Sakaida, Y., Guo, S., & Hosoe, S. (2010). Development of a nursing-care assistant robot riba that can lift a human in its arms. In *2010 ieee/rsj international conference on intelligent robots and systems* (pp. 5996–6001).

Nourbakhsh, N., Wang, Y., Chen, F., & Calvo, R. A. (2012). Using galvanic skin response for cognitive load measurement in arithmetic and reading tasks. In *Proceedings of the 24th australian computer-human interaction conference* (pp. 420–423).

Ohn-Bar, E., & Trivedi, M. M. (2014). Hand gesture recognition in real time for automotive interfaces: A multimodal vision-based approach and evaluations. *IEEE Transactions on Intelligent Transportation Systems*, *15*(6), 2368-2377.

O'Donnel, C., & Eggemeir, F. (1986). Workload assessment methodology: Chapter 42. *Handbook of Perception and Human Performance. II*, 1–49.

Peruzzini, M., Tonietti, M., & Iani, C. (2019). Transdisciplinary design approach based on driver's workload monitoring. *Journal of Industrial Information Integration*, *15*, 91–102.

Pfister, T., Li, X., Zhao, G., & Pietikäinen, M. (2011). Recognising spontaneous facial micro-expressions. In *2011 international conference on computer vision* (pp. 1449–1456).

Rheem, H., Verma, V., & Becker, D. V. (2018). Use of mouse-tracking method to measure cognitive load. In *Proceedings of the human factors and ergonomics society annual meeting* (Vol. 62, pp. 1982–1986).

Rosebrock, A. (2017, March). *Face alignment with opencv and python.* Retrieved from `https://www.pyimagesearch.com/2017/05/22/face-alignment-with-opencv-and-python/`

Rozado, D. (2015). Combining eeg with pupillometry to improve cognitive workload detection.

Sadrfaridpour, B., & Wang, Y. (2017). Collaborative assembly in hybrid manufacturing cells: An integrated framework for human–robot interaction. *IEEE Transactions on Automation Science and Engineering*, *15*(3), 1178–1192.

Sampei, K., Ogawa, M., Torres, C. C. C., Sato, M., & Miki, N. (2016). Mental fatigue monitoring using a wearable transparent eye detection system. *Micromachines*, *7*(2), 20.

Shi, X., Chen, Z., Wang, H., Yeung, D.-Y., Wong, W.-K., & Woo, W.-c. (2015). Convolutional lstm network: A machine learning approach for precipitation nowcasting. *Advances in neural information processing systems*, *28*, 802–810.

Simon, T., Joo, H., Matthews, I., & Sheikh, Y. (2017). Hand keypoint detection in single images using multiview bootstrapping. In *Cvpr.*

Singh, V., Rana, R. K., & Singhal, R. (2013). Analysis of repeated measurement data in the clinical trials. *Journal of Ayurveda and integrative medicine*, *4*(2), 77.

Stern, J. A., & Skelly, J. J. (1984). The eye blink and workload considerations. In *Proceedings of the human factors society annual meeting* (Vol. 28, pp. 942–944).

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive science*, *12*(2), 257–285.

Tsubota, K., Kwong, K. K., Lee, T.-Y., Nakamura, J., & Cheg, H.-M. (1999). Functional mri of brain activation by eye blinking. *Experimental eye research*, *69*(1), 1–7.

Wascher, E., Heppner, H., Möckel, T., Kobald, S. O., & Getzmann, S. (2015). Eye-blinks in choice response tasks uncover hidden aspects of information processing. *EXCLI journal*, *14*, 1207.

Zhang, K., & Travers-rhodes. (n.d.). *firephinx/openpose_ros.* Retrieved from `https://github.com/firephinx/openpose\_ros`

Zhu, X., Lei, Z., Yan, J., Yi, D., & Li, S. Z. (2015). High-fidelity pose and expression normalization for face recognition in the wild. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 787–796).