MOCAS: A Multimodal Dataset for Objective Cognitive Workload Assessment on Simultaneous Tasks

Wonse Jo, Ruiqi Wang, Graduate Student Member, IEEE, Go-Eum Cha, Su Sun, Graduate Student Member, IEEE, Revanth Krishna Senthilkumaran, Daniel Foti, and Byung-Cheol Min, Senior Member, IEEE

Abstract—This paper presents MOCAS, a multimodal dataset dedicated for human cognitive workload (CWL) assessment. In contrast to existing datasets based on virtual game stimuli, the data in MOCAS was collected from realistic closed-circuit television (CCTV) monitoring tasks, increasing its applicability for real-world scenarios. To build MOCAS, two off-the-shelf wearable sensors and one webcam were utilized to collect physiological signals and behavioral features from 21 human subjects. After each task, participants reported their CWL by completing the NASA-Task Load Index (NASA-TLX) and Instantaneous Self-Assessment (ISA). Personal background (e.g., personality and prior experience) was surveyed using demographic and Big Five Factor personality questionnaires, and two domains of subjective emotion information (i.e., arousal and valence) were obtained from the Self-Assessment Manikin (SAM), which could serve as potential indicators for improving CWL recognition performance. Technical validation was conducted to demonstrate that target CWL levels were elicited during simultaneous CCTV monitoring tasks; its results support the high quality of the collected multimodal signals.

Index Terms—Affective computing, cognitive workload assessment, human-machine systems, human-robot teams, multimodal dataset.

I. INTRODUCTION

HUMANS serve as the core part of any human-machine interaction system; thus, the cognitive workload of human operators is a critical concern in the design and implementation of such systems [1]. Cognitive workload (CWL) can be defined as the quantitative amount of mental loads exceeding the operator’s ability to perform tasks [2]. It can also be considered as a mental strain to respond to tasks’ demands for performing specific tasks [3].

Previous research [4], [5], [6], [7], [8] has repeatedly shown that awareness and perception of human CWL can improve the performance of a human-machine interaction system as such awareness allows identifying and alleviating task errors that result from situations of task over-load or under-load. Monitoring the CWL of human operators in a human-robot team, for instance, enables the workload and autonomy levels of the robots to be adjusted as needed to help human operators efficiently and productively maintain their working state [9]. Similarly, recognizing the CWL of human drivers can improve the safety of human-vehicle systems [7].

Generally, CWL assessments can be categorized as subjective or objective. Subjective CWL estimation focuses on the self-assessment of one’s workload via subjective questionnaires such as the NASA-Task Load Index (NASA-TLX) [10]. Objective measurement of CWL relies on quantitative data concerning one’s physiological or behavioral responses when facing a certain level of workload. Accordingly, objective assessments can be further categorized as physiological or behavioral based on the data types utilized [11]. Physiological assessments perceive human biological metrics such as electroencephalography (EEG), functional near-infrared spectroscopy (fNIRS), skin temperature (SKT), and heart rate (HR), which change according to the reaction of the human nervous system to workload fluctuations. Behavioral assessments similarly analyze human behavioral responses such as facial landmarks, eye features, body gestures, and poses in relation to different levels of workload.

However, none of the aforementioned signals are optimal in terms of accuracy, robustness, and rapidity due to the drawbacks inherent in utilizing an unimodal data source [11]. For example, different sensors have different sensitivity to different task scenarios and to different human subjects, therefore it is non-trivial to identify one single metric that is both efficient and accurate for general scenarios [12]. Moreover, noise and the failure of a single-modality sensor can lead to serious errors and even invalidity of the recognition system. Multimodal...
fusion-based CWL assessments have been proposed as a solution to this problem; such assessments combine bio-signals from multiple modalities to provide latent and important information that is unobtainable from a single modal source, and have achieved better performance [11], [13], [14], [15], [16], [17], [18]. Additionally, personality traits have been shown capable of influencing observed responses when different human subjects face the same level of workload; that is, under a given level of CWL, different personality traits can generate physiological and behavioral signals with different features [19], [20], [21], [22]. In addition, human emotion has been proven to have a close relation with CWL [23], [24], [25]; for instance, one’s emotional state may directly affect CWL through expanding or shrinking cognitive resources [26]. Thus, investigating personal traits, human emotion, and their potential relation with CWL can broaden our awareness of individual differences and suggest additional indicators for use in CWL assessments, thereby improving the accuracy and robustness of those assessments.

Even as multimodal fusion methods are being increasingly utilized in the area of objective CWL assessment, the availability of open-source and sizable multimodal datasets remains limited, making it difficult for researchers to ensure their models and algorithms are fairly reproducible and verifiable. Furthermore, existing open-source datasets [20], [27], [28], [29], [30], [31], [32] were collected in the course of participants playing virtual games, such as dual N-back games [33], [34] and Multi-Attribute Task Battery (MATB) [35], [36], which weakens the applicability of those datasets to real-world CWL recognition in human-machine interactions; these virtual games often feature artificial scenarios or simplified tasks, which may not adequately represent the complexity and dynamics found in real-world situations. Also, they usually focus on specific cognitive processes, such as working memory or attention, and may overlook the interplay between multiple cognitive processes, emotions, and contextual factors that typically arise in real-world contexts [11], [37]. In addition, according to our best knowledge, no extant dataset covers behavioral data and emotional information (annotations), and only one [20] includes the personality traits of the human subjects, but this is limited to considering the cognitive loads and performance for game-based tasks that mean it is not suitable on realistic applications.

To fill the above-mentioned gaps, we constructed the Multimodal Dataset for Objective Cognitive Workload Assessment on Simultaneous Tasks (MOCAS). To better mimic real-world scenarios, we designed and utilized a simultaneous closed-circuit television (CCTV) monitoring task to elicit target cognitive load, where participants interact with a physical multi-robot system in real-time and may face potential realistic dynamics and latency caused by the robotic system. The data in MOCAS is based upon 21 human subjects and consists of physiological data collected from two wearable sensors, behavioral data obtained from a camera, subjective CWL annotations via the Instantaneous Self-Assessment (ISA) [38] and NASA-TLX, subjective emotion annotations via the Self-Assessment Manikin (SAM) [39], personal trait background surveyed from the Big Five Factor personality questionnaire [40], and raw screen video recordings. MOCAS represents a useful addition to current research in the field of multimodal fusion CWL assessment: to our best knowledge, MOCAS is the first open-access dataset to obtain both physiological and behavioral data along with both CWL and emotion annotations including the personal traits and background of subjects. All data was collected from off-the-shelf and user-friendly sensors; accordingly, an assessment model built from our dataset can be applied to everyday applications with ease and efficiency.

II. DATASET DESIGN

A. Dataset Design

MOCAS was designed with consideration of a task scenario typical in real-world human-machine systems (as illustrated in Fig. 1), in which one human subject undertakes a simultaneous CCTV monitoring task and multiple sensors track that person’s physiological and behavioral metrics. This dataset is intended to serve as an open-access, sizable, and multimodal data source for research in the field of objective assessment of cognitive workload, with the following aims:

- Support the study of cognitive workload recognition with a focus on real-world applications using human-machine systems.
- Encourage research on improving the accuracy of cognitive workload assessment using real-world applications of multimodal fusion approaches, especially combining physiological and behavioral data.
- Offer novel opportunities for investigating how awareness of a subject’s personality traits and emotion states can improve cognitive workload assessment performance.

Distinct from existing datasets related to cognitive workload, MOCAS evokes a target CWL with realistic CCTV monitoring tasks; it also includes more comprehensive multimodal data collected from both physiological and behavioral sensors, the personal backgrounds of participants (e.g., experience, preference, and personality traits), and emotion information (annotations). The major distinctions between existing CWL datasets and MOCAS are summarized in Table I. Compared to other related datasets listed in Table I, MOCAS stands out for its inclusion of multimodal data and features from physiological sensors and behavioral devices, including EEG, PPG, GSR, HR, ACC, and TEMP data, as well as facial video and monitor screen recordings. Additionally, MOCAS includes more diverse and realistic activities, which better reflect real-life scenarios. Furthermore, the dataset provides comprehensive annotations, such as labels for different activities, affective states, and cognitive load, that can be used to train and evaluate machine learning models for various applications, such as emotion recognition and mental workload estimation.

B. Ethics Statement

This study and the building of MOCAS was reviewed and approved by the Purdue University Institutional Review Board (IRB) with approval number IRB-2021-1813. The informed consent form for participants included the purpose and procedure of data collection, specific types of data collected, possible risks...
Fig. 1. Illustration of the design of MOCAS dataset. MOCAS was designed with consideration of a task scenario typical in real-world human-machine systems, in which one human subject undertakes a simultaneous CCTV monitoring task and multiple sensors track that person’s physiological and behavioral metrics.

| TABLE I | COMPARISON OF THE MOCAS DATASET WITH THE EXISTING COGNITIVE WORKLOAD ASSESSMENT DATASETS |
|---------------------------------|-------------------------------------------------------------------------------------|
| Name (year) | Number of Participants | Scenarios/Stimuli | Physiological Data | Behavioral Data | Personal Traits | Cognitive annotation | Emotional annotation |
|---------|------------------------|-------------------|-------------------|----------------|-----------------|---------------------|---------------------|
| CSAC [28] (2013) | 8 | Multi-attribute task battery (MATB) | EEG, EOG | X | X | O | X |
| STEW [27] (2018) | 48 | Single-session simultaneous capacity (SIMKAP) | EEG | X | X | O | X |
| MIND-COG [32] (2019) | 40 | Reading and math test | EEG, GSR, HR | X | X | O | X |
| WAUC [29] (2020) | 48 | Physical (biking and treadmill) and mental activities (MATB-II) | EEG, EOG, ACC, TEMP, GSR, BVP, RSP | X | X | O | X |
| Cog Load [20] (2020) | 23 | N-back game and video game | ACC, GSR, TEMP, RR | X | O | O | X |
| Snake [23] (2020) | 23 | Smartphone game: snake | ACC, GSR, TEMP, RR | X | O | O | X |
| CLAS [31] (2021) | 62 | Maze and logic problem | EEG, PPG, GSR, ACC | X | X | O | X |
| Tufts [42] (2022) | 68 | N-back game | fNIRS | X | X | O | X |
| MAUS [31] (2021) | 22 | N-back game | EEG, GSR, PPG | X | X | O | X |
| MOCAS (Ours, 2022) | 30 | Realistic CCTV monitoring task | EEG, PPG, GSR, HR, BVP, ACC, TEMP | Raw facial images, Facial features, Eye features, Mouse movement | O | O | O |

“X” means not included, and “O” means included.

or discomforts, compensation and benefits, protocols to protect privacy and confidentiality, and permission for potential usage of the collected data (publishing the dataset and conducting related research). Each participant was provided with the approved informed consent form upon their arrival at the experimental site for data collection, was asked to fully read and understand the content, and was asked to sign a written signature if s/he agreed to participate in the data collection procedure and approved the potential usage of the collected data. Participants also acknowledged that they had the right to terminate the data collection at any time they wanted.

All collected raw data, consent forms, and questionnaire forms (on which each participant was assigned and presented with a participant number) are kept separately and only accessible to the investigators of this study or authorized researchers who consent to the End User License Agreement (EULA) governing usage of dataset. Some aspects of participants’ electronic data were de-identified (e.g., faces blurred) as befit the permissions they gave.

C. Participant Recruitment

We reached out to potential participants by posting flyers all over Purdue campus, through social networking services, and by word of mouth. The investigators also used official contact lists (such as of faculty, staff, and students in the department) to recruit participants through emails. When contacted by potential participants, investigators gave further information about the experiment, confirmed participants’ availability, and sent back a confirmation email. When participants arrived at the site of
data collection, they were provided with the informed consent form, and investigators were responsible for answering all their questions and concerns.

Thirty participants were recruited and participated in our data collection study during March and April 2022; however, only 21 had their data included in our dataset based on their permissions. All recruited participants were students and faculty/staff at Purdue University, and their ages range from 18 to 37 years old (mean = 24.3 years, S.D. = 5.2 years). In order to encourage participants’ engagement in this experiment, we also compensated $15 for each participation and also provided additional compensations (e.g., amazon gift cards as prizes) to three participants who had the highest scores in the given tasks.

D. Stimuli Design

Video game scenarios such as N-back games and the Multi-Attribute Task Battery (MATB) are widely used for eliciting target cognitive workload levels but have limited relevance to real-world applications. In the interest of better mimicking real-world workload scenarios, we proposed a CCTV monitoring task scenario as the stimulus and collected physiological and behavioral signals under different workload levels. Such tasks, in which human operators monitor and control a display interface (or multiple interfaces simultaneously), are widely required in diverse human-machine system task scenarios such as security monitoring [43], [44], air traffic management [45], [46], [47] and performance checking [48], [49].

Here, participants were asked to monitor multiple video streams captured by multiple patrol robot platforms passing by multiple separated rooms located in multiple corridors (Fig. 2). As depicted in Fig. 3, participants were asked to monitor and detect two types of objects: (a) abnormal objects and (b) normal objects. Each room contained a random number of objects with a random assortment of types. The number and speed of robots could be changed to present different levels of stimuli (e.g., low, medium, and high workloads). The thresholds can be determined based on the number and speed of robots. For instance, a low cognitive workload level can be achieved with a small number of robots moving at a slow pace, while a medium cognitive workload level can be achieved with either two robots or faster-moving robots. On the other hand, a high cognitive workload level can be achieved by using many fast-moving robots. These thresholds were determined through multiple beta and pilot tests, which took into account factors such as the number and speed of robots.

During the task, the participant was asked to observe the webcam streams captured by the robots and to use a mouse to click on camera views that contain any abnormal objects; the associated graphical user interface (GUI) is shown in Fig. 3(a) (for more information, refer to [50]).

To incentivize accurate identification and acknowledge participants’ attentiveness and precision, we have designed a scoring system in which a participant is awarded 1 point for correctly identifying and clicking on an abnormal object. In contrast, failing to identify and click on an abnormal object results in a deduction of 3 points, which is a more severe penalty than the reward for correct identification. This scoring system aims
to emphasize the importance of avoiding errors in the task and to motivate participants to be vigilant and precise in identifying abnormal objects by balancing rewards for correct actions and penalties for incorrect actions.

The designed CCTV monitoring task was conducted in a room, approximately 5 m x 6 m x 3 m (width x depth x height), with a multi-robot system involving at least four mobile robot platforms [51] performing the patrol task (see Fig. 2(a)) to provide video streams: also in the room was a desk supporting a 23-inch monitor, a common wireless keyboard, and a mouse (see Fig. 2(b)) with which participants performed the monitoring task. During each collection period, only one participant was in the room. Participants were not allowed to directly observe the multi-robot system during the experiment, but could hear the sounds generated by its movements. However, we believe that the impact of these sounds on participants’ performance is minimal, as the sounds generated were consistent and unobtrusive, avoiding significant distractions. Moreover, the participants were informed about the presence of these sounds before the experiment, allowing them to acclimate to the environment. Furthermore, we intentionally included the presence of noise to simulate real-world conditions, thereby increasing the ecological validity of our study.

E. Measures

We assessed participants’ subjective cognitive workload using both ISA [38] and NASA-TLX [10] with weights, and also measured their subjective emotional state using SAM [39]. The ISA is a quick and simple assessment tool to measure mental workload with only five items (e.g., Underutilized, Relaxed, Comfortable, High, and Excessive). The NASA-TLX is a widely used assessment tool in physiology fields to measure task loads from five dimensions with 7-point scales on mental/physical/temporal demand, performance, effort, and frustration. The SAM is a non-verbal emotion assessment tool that directly measures the three domains related to emotional states. The details of aforementioned three measures can be found in Table II.

F. Apparatus of Data Collection

During the task, participants were required to wear two off-the-shelf wearable biosensors for physiological data collection, and their behavioral data was recorded through a webcam (i.e., Intel RealSense D435i) mounted on the monitor and a mouse as shown in Fig. 2(c):

- Emotiv Insight – captured 5-channel electroencephalogram (EEG), power spectrum (POW) (i.e., theta, alpha, beta, and gamma), and performance metrics generated from the EmotivPro SDK.
- Empatica E4 Wristband – captured photoplethysmography (PPG), 9-axis acceleration, skin temperature (SKT), electrodermal activity (GSR, or EDA), heart rate (HR) and the inter-beat interval (IBI) derived from PPG.
- Intel Realsense D435i – captured participant’s facial views and eye movements.
- Mouse – recorded participant’s mouse movements. Additionally, participant’s monitor screen, including CCTV video stream, mouse positions and status of pushed mouse buttons, was recorded for the experiment, enabling researchers to easily understand the status of the task performed by each participant. Table IV summaries collected data in our dataset.

As illustrated in Fig. 4, all sensors, devices and GUI programs used in the experiment were connected through Robot Operating System 2 (ROS2), where signals were collected as ROSbag2 files with synchronized time (e.g., ROS2 timestamp). The ROSbag2 format has more benefits than a traditional CSV format file in terms of collecting and analyzing the dataset [52], since it can ensure to synchronize the recording all topic data and to easily analyze the dataset by replaying both using a single ROSbag2 file.

G. Data Collection Procedures

Each participant’s data collection procedure was conducted in following three main stages: (1) Introduction stage, (2) Trial stage, and (3) Main stage. Fig. 5 illustrates the overall data collection procedures. Additionally, there is a supplementary video at https://youtu.be/BxVv7R9b70 that explains the details of this experiment design and procedures.

Introduction stage: At the beginning of the data collection, investigators checked again if the participant satisfied the qualifications for this study (such as not having medical history on mental, heart disorders, skin allergies, etc.). The satisfied participants then read and signed the informed consent, and filled out demographic and personality questionnaires. The demographic questionnaire collected participants’ age, gender, level of education, medical history, experiences in conducting video-based surveillance or monitoring tasks, and daily usages of interacting with a monitor that indicated their capability and ability. The personality questionnaire was based on Big Five Personality Test to categorize participants’ personal traits into five personality traits using IPIP Big-Five Factor Markers [53] as follows:

- Extraversion: a participant who has high scores tends to be outgoing/talkative/social, whereas one who has low scores tends to be reflective and reserved behavior.
- Agreeableness: a participant who has high scores tends to be friendly and optimistic, whereas one who has low scores tends to be critical and aggressive.
- Conscientiousness: a participant who has high scores tends to be careful and diligent, whereas one who has low scores tends to be impulsive and disorganized.
- Emotional stability (or neuroticism): a participant who has high scores tends to be sensitive and nervous, whereas one who has low scores tends to be resilient and confident.

| Categories | Name       | Scale Range | Description                                                                 |
|------------|------------|-------------|-----------------------------------------------------------------------------|
| Emotion    | SAM        | from -4 to 4| Two dimensions to assess emotional state: valence (from negative to positive) and arousal (from calm to excited). |
| Workload   | ISA        | from -2 to 2| Simple subjective workload assessment using five dimensions: Underutilized (2), Relaxed (1), Comfortable (0), High (1), and Excessive (2). |
| Workload   | NASA-TLX   | from 1 to 7 | Seven categories for measuring workloads: Underutilized (1), Relaxed (2), Comfortable (3), High (4), and Excessive (5). |

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
After that, investigators outlined the experimental process and gave instructions for each of the tasks that the participants need to complete, and then helped each participant wear and calibrate the physiological and behavioral sensors.

**Trial stage:** Following the introduction stage, participants were given time to get familiar with the hardware and software utilized and understand the tasks of this experiment. They conducted a trial experiment with one camera view and the minimum speed of the robots, which would not be included in the main experiment again. This stage was continued until participants fully understood and got familiar with the CCTV monitoring task in this study.

**Main stage:** The main stage has four repeated sub-phases as illustrated in the dashed box in Fig. 5: baseline phase (about 1 minute), main phase (about 2 minutes), evaluation phase (about 1 minute), and rest phase (optional). In each baseline phase, participants watched a white cross-line with a black background for 50 seconds, and then a countdown from 10 to 0, which helped participants minimize cognitive workload affected by their previous conditions [54], [55], [56], [57].

In each main phase, the participant operated one CCTV monitoring task selected from nine tasks at different workload levels decided by different combinations of three different numbers of camera views (one, two, or four cameras) and three different speeds of the multi-robot system (low, medium, and high speed). During each rest phase, participants were given a break if they requested one due to pain caused by the wearable sensors or cognitive load from the tasks. At that time, we provided enough time for participants to relieve stress and pain by removing the wearable sensors. Other participants who did not need rest time continued with the remaining tasks. Since this phase was optional, we did not collect physiological and behavioral signals...
Fig. 6. Graphical user interfaces for subjective questionnaires used in the user experiment: (a) SAM for measuring emotional states using Arousal and Valence, (b) ISA, and (c) NASA-TLX for measuring workloads.

TABLE III
SUMMARY OF THE MOCAS DATASET CONTENTS

| Number of participants | 21 (7 females and 14 males) |
|------------------------|------------------------------|
| Age                    | from 18 to 37 (mean=24.3 age, std.=5.2 age) |
| Average time per experimental task | mean= 3.52 mins, std.=0.28 mins |
| Personality traits     | Extraversion, Emotional stability, Agreeableness, Conscientiousness, and Intellect/Imagination |
| Physiological signals  | Empatica E4: BVP, CSR, HR, IBI, SKT Emotiv Insight: raw 5-channel EEGs, EEG band powers (i.e., \( \alpha, \beta, \gamma, \) and \( \theta \)), performance metrics |
| Behavioral features    | facial view (30 Hz), facial features & expressions (30 Hz), and Mouse positions & button clicking status (True or False) |
| Experiment status      | 0=loading phase, 1=baseline phase, 2=main phase, 4=Evaluation phase (SAM), 5=Evaluation phase (ISA), 6=Evaluation phase (NASA-TLX), and 7=Score phase |
| Experiment recording   | Screen record video (30 Hz) |
| Experiment scores      | Obtained scores, Success click, Failure click, and Success rate (=success clicks/all clicks) |
| Subjective annotations | SAM: two categories; valence (from negative to positive) and arousal domain (from calm to excited) with a range from -4 to +4 ISA: five categories; Underutilized (-2), Relaxed (-1), Comfortable (0), High (1), and Excessive (2) NASA-TLX: Seven categories for measuring workloads; Mental demand, Physical demand, Temporal demand, Performance, Effort, and Frustration with a range from 1 to 7 |

III. DATA RECORDS

A. Dataset Summary

Table III summarizes the MOCAS, which contains multi-modal data from 21 participants, including physiological signals, facial camera videos, mouse movement, screen record videos, and subjective questionnaires. Fig. 7 shows examples of physiological and behavioral data from the raw dataset. The total size of the dataset is about 722.4 GB which includes 754 ROSbag2 files.

Fig. 8 shows a folder tree of the online repository that displays directory paths and files for the MOCAS dataset. There are three major folders based on the types of the dataset:

- **1. subjective_information**: A single csv file containing subject’s demographic (e.g., age, gender, personality trait, and experience), answers of the subjective questionnaires used in this experience, and performance (e.g., scores, success click, failure click, and success rate).
- **2. raw_dataset_rosbag2**: Raw rosbag2 files saved into each sub-folder depending on the experimental factors (e.g., robot speed and the number of camera views) and data types (e.g., camera, signal, results, and screen). The last sub-folder of this major folder has two files: metadata.yaml and *.db.
- **3. pre-processed_dataset**: Downsamped CSV-format and Pickle-format [58], [59] files with 100 Hz sampling rate. The overall size of the CSV files and Pickle file is 51.9 GB and 32.8 GB, respectively. Each file has the annotations of the subjective questionnaires, the number of camera views, robot speeds, features of the physiological and behavioral data, and filtered physiological signals.

- **4. supplementary_code**: Contains the converter codes from raw ROSbag2-format into pickle and csv-format files, and supplementary codes used on this paper. The codes are written by Python under Ubuntu 20.04.
Fig. 7. Examples of the raw and cleaned collected physiological and behavioral data: raw and cleaned (a) BVP data and (b) GSR data, (c) features extracted from the GSR data (tonic and phasic components), (d) raw EEG signals, and behavioral data extracted from the facial videos with (e) closed eyes and (f) open eyes. The blue lines in (a) - (d) mean the collected data in the baseline phase and the orange lines mean the collected data in the main phase.

B. Pre-Processed and Downsampled Data

The raw ROSbag2 format files were converted to Pickel and CSV format with 100 Hz sampling rate, which were saved to two different folders based on the type of file format (i.e., CSV and Pickle [58]). Each file contains raw signals, each channel’s EEG signal divided from chunk EEG signals, cleaned BVP and GSR signals, and physiological features from the GSR signals (e.g., tonic and phasic domains).

For cleaning the raw BVP and GSR signals, we applied a bandpass filter to remove the noise from both the raw BVP and GSR signals through [60]. For the physiological features, we extracted tonic and phasic features from the raw GSR signals. They can provide valuable insights into the nervous system activity and emotional arousal of an individual [61]. The tonic feature is the baseline level of a physiological signal called Phasic Skin Conductance Response (SCR). To extract the tonic component from the raw GSR signal, we applied a moving averages algorithm for baseline correction to refine the tonic component. The phasic feature refers to the rapid and temporary changes in physiological signals that occur in response to specific stimuli or events, called phasic skin conductance responses (SCRs), which often associated with brief fluctuations in arousal or attention. To identify phasic events, we use peak detection algorithms that detect peaks in the GSR signal. The prominent peaks usually represent the SCRs by analyzing amplitude, rise time, and half-recovery time.

Furthermore, each file contains behavioral features from facial videos (e.g., Action units (AUs), probability and types of facial expression, and Eye Aspect Ratio (EAR)), as well as the results of the self-reporting questionnaires.

For extracting AUs and EAR, we first applied the facial detection algorithm to extract the face area from the facial videos and extracted each facial feature from the face area, called AUs (see Fig. 7(e) and (f)), through a face landmark (L) detection of Google MediaPipe [63]. Then, the EAR of each eye is calculated using six landmarks around the eye. The average EAR of eyes is measured using the following (1) [64]:

\[
EAR_{left} = \frac{||L_{160} - L_{144}|| + ||L_{158} - L_{153}||}{2 ||L_{33} - L_{133}||}
\]

\[
EAR_{right} = \frac{||L_{385} - L_{380}|| + ||L_{387} - L_{373}||}{2 ||L_{250} - L_{253}||}
\]

\[
EAR = \frac{(EAR_{left} + EAR_{right})}{2}
\]  

The details of all data in the pre-processed MOCAS dataset can be found in Table IV.
TABLE IV
DETAILS OF THE DATA IN THE PRE-PROCESSED MOCAS DATASET

| Device/Source           | Collected data                  | Channels | Sampling rate | Signal range [min, max] |
|-------------------------|---------------------------------|----------|---------------|------------------------|
| Empatica E4 Wristband   | BVP (PPG)                       | 1        | 64 Hz        | N/A                    |
|                         | CSK                             | 1        | 4 Hz         | [0.01μS, 100μS]         |
|                         | HR (from BVP)                   | 1        | 1 Hz         | N/A                    |
|                         | IBI (from BVP)                  | 1        | 1 Hz         | N/A                    |
| Emotive Insight         | EEG (with Contact Quality)      | 6        | 128 Hz       | N/A                    |
|                         | EEG band powers                 | 25       | 8 Hz         | [0.1]                  |
|                         | Performance metrics             | 7        | 1 Hz         | [0.1]                  |
| Intel RealSense D430i   | Action Unit (AU)                | 3 (x, y, average) | 1 Hz | N/A        |
|                         | Eye Aspect Ratio (EAR)          | 3 (left, right, average) | 1 Hz | N/A        |
| Subjective Annotations  | Cognitive Load (from ISA)       | 1        | N/A          | [-2.2]                 |
|                         | Cognitive Load (from NASA-TLX-Weighted) | 1 | N/A | [0.1] |
|                         | Emotion (from SAM)              | 2 (arousal, valence) | N/A | [-4.4] |
| GUI program             | Experiment states               | 1        | N/A          | [0.1]                  |

IV. ANALYSIS OF DATASET

A. Correlation Between Personality Traits and Physiological and Behavioral Signals

We calculated Pearson’s correlation coefficient ($\gamma$) to find the relationship between the participant’s personal traits and the mean of their physiological and behavioral data. $\gamma$ indicates the relationship between two variables with ranges from -1 to 1. The positive high $\gamma$ means a proportional relationship between both variables, whereas the negative high $\gamma$ means an inverse relationship [65]. Fig. 9 shows the results of the Pearson’s correlation coefficient, where the PE is an extraversion marker of the IPIP Big-Five Factor Markers, the PES is emotional stability, the PA is agreeableness, the PC is conscientiousness, and the PII is an intellect/imagination.

From the results of Pearson’s correlation, we observed a moderate relationship between personal traits and physiological and behavioral features. The moderate positive relationship ($\gamma > 0.3$) is found between PE and mean IBI, between PES and mean HR, and between the PC and mean HR, and between the PC and mean IBI and EAR. On the other hand, the moderate negative relationship ($\gamma < -0.3$) is found between PE and mean HR, between PA and mean EAR and IBI, between PC and mean HR, and between PII and mean GSR, SKR, and EAR.

B. Relationships Between Personality Traits, Gender, and Performance Measures

We also utilized Pearson’s correlation coefficient and independent sample t-test to examine the relationship between the participants’ personal characteristics and performance results, such as scores and success rate. As shown in Table V, participants with higher PE ratings demonstrated higher scores ($\gamma = .148$, $p<.05$) and a higher number of successful clicks ($\gamma = .236$, $p<.01$) during the experiment. Individuals with higher PA ratings had higher success rates in identifying abnormal objects. Furthermore, there was a positive correlation between participants’ PI ratings and the number of successful clicks ($\gamma = -.159$, $p<.05$).

The correlation coefficients between personality traits and performance, based on gender, are presented in Table V, highlighting distinct gender characteristics during the experiment. Among male participants, higher PE scores were positively correlated with higher scores ($\gamma = .204$, $p<.05$) and a higher frequency of successful clicks ($\gamma = .184$, $p<.05$). Also, higher PA scores among male participants were negatively correlated with the

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
number of successful clicks ($\gamma = -0.191, p < 0.05$). Among female participants, higher PA ratings showed a higher frequency of failure clicks ($\gamma = 0.314, p < 0.01$), while higher PI scores had a negative correlation with the number of successful clicks ($\gamma = -0.242, p < 0.05$).

We conducted an independent sample t-test between male and female participants to examine the consistency of the observed tendencies in Table V. The four performance measures demonstrated normality based on the Shapiro-Wilk test. As shown in Table VI, there were no significant differences between genders in the performance measures, as supported by minimal effect sizes (Cohen’s $d$); (score: $t_{185} = 326, p = 0.745$), success rate: $t_{185} = 0.072, p = 0.75$, and failure clicks: $t_{185} = 0.111, p = 0.912$). Therefore, we concluded that there were no significant differences in performance-related measures between genders.

### Table VI

| Measure                      | Male Mean | Male SD  | Female Mean | Female SD | t   | p   | Cohen’s $d$ |
|------------------------------|-----------|----------|-------------|-----------|-----|-----|-------------|
| Score                        | 72.04     | 40.05    | 65.98       | 40.72     | 326 | 0.745 | 0.047       |
| Success rate                 | 0.94      | 0.072    | 0.75        | 0.073     | -246| 0.001 | 0.337       |
| Successful clicks            | 9.07      | 31.108   | 8.84        | 35.104    | 492 | 0.024 | 0.074       |
| Failure clicks               | 6.68      | 8.63     | 6.33        | 8.26      | 111 | 0.912 | 0.17        |

It illustrates the results of the two-way rmANOVA for the ISA scores reported by participants. The scores were transformed into a range from -2 to 2, representing the string-type scores: Underutilized (-2), Relax (1), Comfortable (0), High (1), and Excessive (2). The Friedman test revealed significant differences among repeated observations, yielding a Chi-square ($\chi^2$) value of 88.336, with a significance level of $p < 0.001$. Further analysis of the ISA questionnaire was conducted separately for male and female participants, revealing noticeable differences. For male participants, the Friedman test showed a significant result ($\chi^2 = 57.313, p < 0.001$), while for female participants, the test yielded a significant result as well ($\chi^2 = 33.784, p < 0.001$). These findings suggest that the given task elicits distinct responses across different experimental conditions, regardless of the participants’ gender.

**NASA-TLX results for measuring cognitive workload:** Fig. 10 illustrates the results of the two-way rmANOVA results of the raw and weighted NASA-TLX scores. The weights used in the weighted NASA-TLX scores are $[0.5, 0, 0, 4, 3, 2, 1]$ in which the sequence is mental demand, physical demand, temporal demand, performance, effort, and frustration. From the two-way rmANOVA results using participants’ raw and weighted NASA-TLX scores, we found a main effect of the number of the camera views (raw: $F_{2,38} = 49.12, p < 0.001$, $\eta^2 = 0.72$ and weighted: $F_{2,38} = 59.29, p < 0.001$, $\eta^2 = 0.76$) and the robot speed on the cognitive workload (raw: $F_{2,38} = 21.06, p < 0.001$, $\eta^2 = 0.53$ and weighted: $F_{2,38} = 17.36, p < 0.001$, $\eta^2 = 0.48$), but there is no interaction between the two within-subject factors (raw: $F_{1,76} = 1.09, p = 0.37$, $\eta^2 = 0.05$ and weighted: $F_{1,76} = 0.86, p = 0.49$, $\eta^2 = 0.04$). We also conducted a validation of the results based on gender to determine if there was congruence between male and female participants’ responses. For male participants, both the raw and weighted NASA-TLX scores showed a significant main effect for two factors: the number of
camera views (raw: $F_{2.24} = 58.076$, $p < .001$, $\eta^2_p = .829$; weighted: $F_{2.24} = 54.922$, $p < .001$, $\eta^2_p = .821$) and the robot speed (raw: $F_{2.24} = 12.597$, $p < .001$, $\eta^2_p = .512$; weighted: $F_{2.24} = 8.648$, $p = .003$, $\eta^2_p = .419$). However, there was no interaction observed between these two factors (raw: $F_{4.48} = .474$, $p = .695$, $\eta^2_p = .038$; weighted: $F_{4.48} = .362$, $p = .770$). Likewise, the responses of female participants demonstrated congruence in the effects of the two factors. Significant main effects were observed for the number of cameras (raw: $F_{2.12} = 9.067$, $p = .010$ ($p < .05$), $\eta^2_p = .602$; weighted: $F_{2.12} = 12.728$, $p = .003$ ($p < .05$), $\eta^2_p = .680$) and the robot speed (raw: $F_{2.12} = 8.006$, $p = .018$, $\eta^2_p = .572$; weighted: $F_{2.12} = 9.256$, $p = .009$, $\eta^2_p = .607$) when analyzed separately. However, there was no significant interaction between these two factors (raw: $F_{4.24} = 1.048$, $p = .373$, $\eta^2_p = .149$; weighted: $F_{4.24} = .931$, $p = .402$, $\eta^2_p = .134$). These results indicate that individual factors significantly influence cognitive workload, but the interaction between factors does not significantly impact participants’ workload.

**SAM scores for measuring emotion state:** Table VIII summarizes the results of the two-way Friedman test conducted on the SAM scores reported by the participants, using a 7-point Likert scale. From the test results, it was found that the main experiment significantly differentiated participants’ responses in terms of arousal ($\chi^2 = 49.161$, $p < .001$) and valence ($\chi^2 = 63.282$, $p < .001$). These statistical findings hold true when the data are analyzed separately for each gender. Specifically, when considering arousal levels, both male participants ($\chi^2 = 35.113$, $p < .001$) and female participants ($\chi^2 = 18.719$, $p = .016$) exhibited statistically significant differences. Similarly, in terms of valence, both male participants ($\chi^2 = 45.116$, $p < .001$) and female participants ($\chi^2 = 21.553$, $p = .006$) showed statistically significant differences. These results indicate that the elicitation of emotions during the experiment is consistent across all genders.

### Table VII

**RESULTS OF THE TWO-WAY REPEATED-MEASURES ANOVA TEST BASED ON GENDER, RAW AND WEIGHTED NASA-TLX QUESTIONNAIRE SELF-REPORTING ANSWERS**

| Subjective questionnaires | Factor                                      | $D_{f_{e,ovr}}$ | $F$   | $P$    | $\eta^2_p$ | Note |
|---------------------------|---------------------------------------------|-----------------|-------|--------|------------|------|
| **All participants**      | Camera_number                               | (2, 38)         | 49.127| $p < .001$ | .721       |      |
|                           | Robot_speed                                 | (2, 38)         | 21.030| $p < .001$ | .525       |      |
|                           | Camera_number x Robot_speed                 | (4, 76)         | 1.094 | .360   | .054       |      |
| **NASA-TLX**              | Camera_number                               | (2, 38)         | 59.303| $p < .001$ | .757       |      |
|                           | Robot_speed                                 | (2, 38)         | 17.303| $p < .001$ | .477       |      |
|                           | Camera_number x Robot_speed                 | (4, 76)         | .875  | .468   | .044       |      |
| **Male**                  | Camera_number                               | (2, 24)         | 55.076| $p < .001$ | .829       |      |
|                           | Robot_speed                                 | (2, 24)         | 12.957| .001   | .512 **    |      |
|                           | Camera_number x Robot_speed                 | (4, 48)         | .474  | .695   | .038       |      |
| **NASA-TLX**              | Camera_number                               | (2, 24)         | 54.922| $p < .001$ | .821       |      |
|                           | Robot_speed                                 | (2, 24)         | 8.648 | .003   | .419 **    |      |
|                           | Camera_number x Robot_speed                 | (4, 48)         | .362  | .770   | .029       |      |
| **Female**                | Camera_number                               | (2, 12)         | 9.067 | .010   | .602 **    |      |
|                           | Robot_speed                                 | (2, 12)         | 8.006 | .018   | .572 **    |      |
|                           | Camera_number x Robot_speed                 | (4, 24)         | 1.048 | .373   | .149       |      |
| **NASA-TLX**              | Camera_number                               | (2, 12)         | 12.728| .003   | .680 **    |      |
|                           | Robot_speed                                 | (2, 12)         | 9.256 | .009   | .607 **    |      |
|                           | Camera_number x Robot_speed                 | (4, 24)         | .931  | .402   | .134       |      |

**Note:** $p < .05$

### Table VIII

**RESULTS OF THE TWO-WAY FRIEDMAN TEST ON GENDER, ISA, AND SAM**

| Subjective questionnaires | $\chi^2$ | $P$ |
|---------------------------|----------|-----|
| **All participants**      |          |     |
| ISA                       | 88.336   | $p < .001$ |
| SAM                       | 49.161   | $p < .001$ |
| ISA                       | 63.282   | $p < .001$ |
| Male                      |          |     |
| ISA                       | 57.313   | $p < .001$ |
| SAM                       | 35.113   | $p < .001$ |
| ISA                       | 45.116   | $p < .001$ |
| Female                    |          |     |
| ISA                       | 33.784   | $p < .001$ |
| SAM                       | 18.719   | .016 |
| ISA                       | 21.553   | .006 |

### D. Classification Evaluation

We conducted a comprehensive evaluation of the MOCAS dataset through a three-class cognitive workload classification. To this end, we utilized pre-processed features available in the pre-processed MOCAS dataset, which included EEG, EEG band powers (EEG_POW), EDA (or GSR), BVP, HR, EAR and AU (see Section III-B and Table IV for more information). In order to categorize the subjective cognitive workload annotations obtained from NASA-TLX questionnaires, we assigned them into three categories based on the literature [68], [69]. Specifically, scores ranging from 0-40 were labeled as “low” workload, scores from 40-60 as “medium” workload, and scores from 60-100 as “high” workload. Furthermore, we established and validated these score thresholds through multiple pilot tests, where the percentile statistics were empirically divided into three equal groups. The 33rd percentile was determined to be 37.78 and the 66th percentile was found to be 55.56. These values were then
The mean and standard deviation values of the three-class accuracy for both unimodal and multimodal fusion classifications under two evaluation schemes are detailed in Table IX. In the trial-independent evaluation, the highest achieved classification accuracy was 72.33%, reflecting the high quality of the data in the MOCAS dataset. Conversely, in the subject-independent evaluation, the highest accuracy decreased to 46.13%. This reduction is consistent with findings in previous studies [75], [76], [77] and can be attributed to the inherent challenges posed by the LOSO validation method. The LOSO approach, influenced by individual variances, complicates the development of a universally robust model. Moreover, our chosen classification network, the LF-LSTM, does not inherently adapt to these individual differences. Consequently, while the model demonstrates strong performance in a trial-independent context, its efficacy is diminished in the subject-independent scenario, reflecting the added complexity of generalizing across diverse individual datasets. This, in turn, provides an exciting venue for future research opportunities. The discrepancy in performance between trial-independent and subject-independent evaluations highlights the importance of developing more sophisticated models and techniques that can effectively account for individual differences. This area of research is particularly promising as it pushes the boundaries of current methodologies, encouraging innovation in personalized and adaptive modeling.

Subsequently, we conducted an independent two-sample t-test to compare the classification results across different modalities on the trial-independent evaluation. Generally, unimodal cognitive workload recognition using EEG_POW modality performs significantly better than using EAR modality ($p = .05$), EEG modality ($p = .01$), and other single modalities ($p < .001$). EEG-related modalities, including EEG and EEG_POW, perform well since EEG signals correspond more directly to different brain activities under different workload levels [78]. Compared with EEG modality, EEG_POW modality decomposes the raw EEG signals into component frequency bands, whose features are more intuitive and have fewer noises, leading to better classification results [79]. Moreover, EAR modality also achieves reasonable performance, we owe this to the fact that when facing different monitoring task levels where different numbers of camera views and robot speed, eye movements of participants would change correspondingly. Furthermore, classification using multimodal fusion significantly achieves better performance approximately rounded up to the previously mentioned ranges of 0–40, 40–60, and 60–100.

Taking inspiration from previous studies [70], [71] and recognizing the sequential nature of psychological and behavioral signals, we utilized a Long Short-Term Memory network (LSTM) [72] for the task of unimodal classification. This approach involved using a single modality as the input. We also employed a Late-Fusion Long Short-Term Memory network (LF-LSTM) [73], [74] for multimodal fusion classification, where multiple modalities were utilized as the input. As depicted in Fig. 11, our LF-LSTM model processed the features from each modality through an LSTM network to extract time-related unimodal features. These processed unimodal features of all modalities were then concatenated and passed through a one-dimensional convolutional neural network (CNN) to generate predictions. For further details and access to the code of our LF-LSTM implementation, refer to the Code Availability section.

To emulate real-world application scenarios, we constructed input samples for each modality as 2-D feature matrices derived from one-second segments. Our evaluation methodology and data partitioning focused on two primary types: trial-independent and subject-independent schemes. In the trial-independent scheme, all samples were randomly shuffled, followed by $k$-fold cross-validation (where $k = 5$). This process ensured that samples derived from the same monitoring task of a participant were distributed between the training and test sets. On the other hand, for the subject-independent scheme, we implemented the leave-one-subject-out (LOSO) cross-validation [75].
than using EEG_POW ($p = .01$) and any other single modalities ($p < .001$). This reflects the benefits of the multimodal fusion mentioned in Section Background and Summary.

Furthermore, there is considerable scope for enhancing classification performance on the MOCAS dataset. Key areas for improvement include the adoption of more sophisticated classification models, as suggested by recent studies [18], [80], [81], and refined data preprocessing techniques. Particularly for improving performance in subject-independent or LOSO evaluations, it is crucial to account for individual differences. Factors such as demographic information, personal traits, and task-specific experience should be considered. Advancements in transfer learning [77], user-specific attention mechanisms [82], [83], and few-shot learning [84] methodologies are poised to make significant contributions in this regard. By integrating these approaches, we can develop more robust models that better accommodate the unique characteristics of individual subjects, thereby improving classification accuracy in diverse and personalized settings.

V. DATASET ACCESS

In order to protect the sensitive data and privacy of human subjects (e.g., physiological signals and facial views), only authorized researchers who consent to the End User License Agreement (EULA) are allowed to download the MOCAS. The researchers who want to access the MOCAS should visit our website and download the ELUA document; https://polytechnic.purdue.edu/ahmrs/dataset. After reviewing and filling the document up, they should email it to and then request the access through Zenodo (https://zenodo.org/). Then, our research group will review and grant their access to our Zenodo repository having the downsampled MOCAS dataset, subjective information, and supplementary codes used in this paper. For sharing the raw dataset, we will sequentially invite their email address used in the Zenodo and the ELUA document to access raw dataset uploaded on an additional repository (Purdue BOX, https://purdue.box.com/v/mocas-dataset), due to huge size of the raw dataset.

A. Missing Data

Participant 2 (P2) discontinued experiments due to personal reasons, resulting in missing data for Task C and Task D from P2;

- $P_2$-mouse_cam_1-speed_60
- $P_2$-mouse_cam_2-speed_40

Out of 187 files, one file does not contain the Emotive Insight data due to sensor disconnection. However, it includes other physiological and behavioral signals;

- $P_3$-mouse_cam_4-speed_60_signal_0.db3,

where the term mouse in the file name refers to a type of input interface used in user experiments.

B. Code Availability

The code utilized for data format converting (ROSbag2 to CSV) and preprocessing with the code of LSTM and LF-LSTM used for classification validation is included in the supplementary code file of the dataset repository. Additionally, a command-line tool of the ROS2 middleware for replaying the ROSbag2 format files using the command of ros2 bag play [rosbag2_file_name] or rqt_bag is also available there.

VI. DISCUSSION & LIMITATIONS

We conducted experiments with real multi-robot systems to collect physiological and behavioral signals from human subjects while they watched real-time video streamed by the multirobot systems. However, we encountered several challenges during the data collection process, including ensuring stable and real-time data collection, keeping the participants engaged, and dealing with discomfort caused by the EEG headset. To address these issues, we used ROS2 for real-time computing, designed realistic missions and provided extra compensation to increase participant engagement, and offered breaks during the experiment to alleviate the discomfort caused by the EEG headset.

Then, we built the MOCAS dataset by collecting behavioral and physiological signals using practical wearable biosensors and devices in real-life scenarios. Compared to other existing datasets listed in Table I, our MOCAS dataset provides both raw physiological and behavioral data, as well as pre-processed features in real-time, from more realistic scenarios and stimuli. It also includes annotations of emotional states and cognitive load, which can be useful for training and evaluating machine learning models for different applications, such as emotion recognition and mental workload estimation. Moreover, the MOCAS dataset offers raw ROSbag files recorded during the entire experiment, enabling researchers to easily understand the situation and match it with the raw data. They can also test their prediction algorithms in real-time without requiring further experiments.

However, MOCAS dataset still has some limitations on the introduced dataset, so it should be considered before using this dataset as following limitations:

For the physiological signals, the commercial wearable biosensors used in this dataset tend to easily have unknown noises influenced by the participant’s movement. There are available open-source libraries to remove common noises for the biosensors, such as NeuroKit2 [60], ppysio [85] and BioSPPy [86]. The downsampled CSV files of our dataset were cleaned using the NeuroKit2 to remove the noises from the BVP and GSR signals.

For the behavioral data, the raw MOCAS dataset contains behavioral data obtained from the front facial videos of individuals who have agreed to share their data with the public. However, this dataset has a limitation in that it also includes behavioral data recorded from individuals wearing facial masks. This was due to the campus regulations during the data collection experiment, which mandated that everyone wear facial masks during the COVID-19 pandemic. Despite this limitation in the raw MOCAS, we extracted the behavioral features (such as EAR and AUs) using MediaPipe which can measure facial landmarks regardless of whether the individual is wearing a facial mask, then add the features on the prepossess MOCAS.
For the size of dataset and participants, this dataset comprises multimodal signals collected from 21 participants. We acknowledge that the number of subjects in our study is relatively small compared to certain existing datasets, with a total of 21 subjects included in the final analysis. As discussed in Section III, the sample size was constrained by subject availability and the inclusion criteria for the study. Additionally, explicit informed consent was obtained from all participants, resulting in the exclusion of certain potential subjects from the final analysis. In order to mitigate the limitations imposed by the dataset size, we employed $k$-fold cross-validation to construct a robust deep learning model for predicting human cognitive workloads. $k$-fold cross-validation is a widely recognized technique used to estimate the performance of a model on unseen data. It involves partitioning the dataset into $K$ equally sized folds and training the model $k$ times, with each fold serving as the test set once [87]. This method is extensively employed in machine learning research and practice to mitigate overfitting and maximize the utilization of available data. By leveraging this validation approach, our deep learning model achieved a predictive accuracy of 74.68% in categorizing three levels of cognitive workloads, namely low, medium, and high.

Other important limitation of this dataset is that it was collected from a restricted age range of participants. Although we aimed to select a diverse sample within this age range through official flyers and snowball sampling [88], our findings may not be generalizable to other age groups. This limitation is particularly relevant for phenomena known to vary across the lifespan, such as cognitive or physical abilities [89]. Therefore, future studies could consider expanding the age range to include a wider range of ages, which would provide a more comprehensive understanding of the phenomenon under investigation [90]. In addition, future research could explore potential age-related differences in our variables of interest to further enhance our understanding of the phenomenon [91]. Despite this limitation, our study offers valuable insights within the age range studied and sets the stage for future research to build upon our findings.

VII. CONCLUSION

We proposed a new multimodal dataset for human cognitive workload. The dataset includes physiological signals and behavioral features measured from 21 human subjects conducting the generalized CCTV monitoring task with a real multi-robot system. The physiological signals are acquired by the two wearable sensors, such as EEG, PPG, GSR, HR, IBI, SKT, and motion data. The behavioral features include eye ratios, facial expression, and facial action units extracted from a facial view of the webcam. The proposed dataset consists of raw and downsampled data, a summary of the participants’ information, and the results of the subjective questionnaires. The total size of the raw dataset is about 722.4 GB, including 754 rosbag2 files. For the downsampled dataset, `.csv` (Comma Separated Value) and `.pkl` (Pickle) file formats were converted from raw datasets with sampling rates of 100 Hz. The size was 51.9 GB and 32.8 GB respectively.

We also validated the quality of the dataset by analyzing the correlation between personality traits, physiological signals and behavioral features, evaluating the effects of within-subjects factors on the results of the questionnaires presented to subjects, and applying an LF-LSTM to classify the three-class cognitive workload classifications. As a result, we found that there are significant differences in factors within subjects (e.g., the number of camera views and robot speed) using statistical analysis, and also showed that the classification performance of the multimodal dataset outperforms that of the single-modal dataset through deep learning methods.

Additionally, we made the MOCAS dataset publicly available by uploading the dataset to the online repositories and codes used in this dataset. We hope that the proposed dataset can become a fundamental resource for other researchers to develop systems and algorithms in human cognitive workloads.

ACKNOWLEDGMENTS

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.

REFERENCES

[1] J. A. Adams, “Critical considerations for human-robot interface development,” in Proc. AAAI Fall Symp., 2002, pp. 1–8.
[2] C. D. Wickens, “Multiple resources and performance prediction,” Theor. Issues Ergonom. Sci., vol. 3, no. 2, pp. 159–177, 2002.
[3] B. Cain, “A review of the mental workload literature,” DTIC Doc., 2007.
[4] S. P. Marshall, “The index of cognitive activity: Measuring cognitive workload,” in Proc. IEEE 7th Conf. Hum. Factors Power Plants, 2002, pp. 7–7.
[5] C. J. Patten, A. Kircher, J. Östlund, L. Nilsson, and O. Svensson, “Driver experience and cognitive workload in different traffic environments,” Accident Anal. Prevention, vol. 38, no. 5, pp. 887–894, 2006.
[6] C. J. Patten, A. Kircher, J. Ostlund, and L. Nilsson, “Using mobile telephones: Cognitive workload and attention resource allocation,” Accident Anal. Prevention, vol. 36, no. 3, pp. 341–350, 2004.
[7] M. A. Just, P. A. Carpenter, and A. Miyake, “Neuroindices of cognitive workload: Neuroimaging, pupillometric and event-related potential studies of brain work,” Theor. Issues Ergonom. Sci., vol. 4, no. 1/2, pp. 56–88, 2003.
[8] M. Pomplun and S. Sunkara, “Pupil dilation as an indicator of cognitive workload in human–computer interaction,” in Proc. Human-Centered Comput., 2019, pp. 542–546.
[9] T. Mina, S. S. Kamam, W. Jo, and B.-C. Min, “Adaptive workload allocation for multi-human multi-robot teams for independent and homogeneous tasks,” IEEE Access, vol. 8, pp. 152697–152712, 2020.
[10] S. G. Hart and L. E. Staveland, “Development of nasa-tlx (task load index): Results of empirical and theoretical research,” in Advances in Psychology, vol. 52. Amsterdam, Netherlands: Elsevier, 1988, pp. 139–183.
[11] E. Debe et al., “Multimodal fusion for objective assessment of cognitive workload: A review,” IEEE Trans. Cybern., vol. 51, no. 3, pp. 1542–1555, Mar. 2021.
[12] F. Chen et al., “Multimodal behavior and interaction as indicators of cognitive load,” ACM Trans. Interactive Intell. Syst., vol. 2, no. 4, pp. 1–36, 2013.
[13] E. B. Coffey, A.-M. Brouwer, and J. B. van Erp, “Measuring workload using a combination of electroencephalography and near infrared spectroscopy,” in Proc. Hum. Factors Ergonom. Soc. Annu. Meeting, 2012, pp. 1822–1826.
[14] F. Putze et al., “Hybrid NIR-EEG based classification of auditory and visual perception processes,” Front. Neurosci., vol. 18, 2014, Art. no. 373.
[15] Y. Liu, H. Ayaz, and P. A. Shewokis, “Mental workload classification with concurrent electroencephalography and functional near-infrared spectroscopy,” Brain- Comput. Interfaces, vol. 4, no. 3, pp. 175–185, 2017.
Supervised Sequence Learning

IEEE Trans. Affect. Comput.

received the BS degree in computer science and engineering, the BE degree from Global Media School, Soongsil University, Seoul, South Korea, in 2016, and the MS degree in computer information and technology from Purdue University, USA. She is currently working towards the PhD degree in the Department of Computer and Information Technology with Purdue University, West Lafayette, IN, USA. Her research interests include social robotics, affective robotics, and multi-modal deep learning.

Wons Je received the BS degree in robotics engineering from the Beijing University of Chemical Technology, Beijing, China, in 2020. He is currently working towards the PhD degree with the Department of Computer and Information Technology, Purdue University, West Lafayette, IN, USA. His research interests include affective robotics/computing, human multi-robot interaction, environmental robotics, and field robotics.

Ruiqi Wang (Graduate Student Member, IEEE) received the BE degree in robotics engineering from the Beijing University of Chemical Technology, Beijing, China, in 2015. He is currently working toward the PhD degree with the Department of Computer and Information Technology, Purdue University, West Lafayette, IN, USA. His research interests include affective robotics/computing, human multi-robot interaction, environmental robotics, and field robotics.

Go-Eum Cha received the BS degree in computer science and engineering, the BE degree from Global Media School, Soongsil University, Seoul, South Korea, in 2016, and the MS degree in computer information and technology from Purdue University, USA. She is currently working towards the PhD degree in the Department of Computer and Information Technology with Purdue University, West Lafayette, IN, USA. Her research interests include deep learning & 3D geometry, SLAM, and 3D reconstruction.

Revanth Krishna Senthilkumaran is currently working toward the undergraduate degree in the Department of Electrical and Computer Engineering with Purdue University, West Lafayette, IN, USA. His research interests include human-robot interaction, embedded electronics, robot dynamics, and unmanned aerial systems.

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
Daniel Foti received the BA degree in biomedical engineering from Harvard University, in 2006. He completed his graduate studies from Stony Brook University, the MA degree in psychology, in 2008, and the PhD degree in clinical psychology, in 2013. He completed his predoctoral clinical internship with McLean Hospital in Belmont, MA. He joined the faculty in the Department of Psychological Sciences with Purdue University as an assistant professor, in 2013. He was promoted to associate professor with tenure, in 2018. His research interests include using psychophysiological measures to study cognition, emotion, and reward, as well as abnormalities in these processes that are associated with vulnerability to psychiatric illnesses. He was a recipient of the James C. Naylor Award for Teaching Excellence in Psychology, in 2019. He was named Director of the T32 Predoctoral Training Program through the Indiana Clinical and Translational Sciences Institute in 2021. He was elected to the board of directors for the Society for Psychophysiological Research, in 2021, and he was elected Treasurer, in 2022.

Byung-Cheol Min (Senior Member, IEEE) received the BS degree in electronics engineering and the MS degree in electronics and radio engineering from Kyung Hee University, Yongin, South Korea, in 2008 and 2010, respectively, and the PhD degree in technology with a specialization in robotics from Purdue University, West Lafayette, IN, USA, in 2014. He is currently an associate professor and University Faculty Scholar with the Department of Computer and Information Technology and the director of the SMART Laboratory, Purdue University. Prior to this position, he was a postdoctoral fellow with the Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, USA. His research interests include multirobot systems, human–robot interaction, robot learning, with applications in field robotics, and assistive technology and robotics. He was a recipient of the NSF CAREER Award, in 2019; the Purdue PPI Outstanding Faculty in Discovery Award, in 2019; the Purdue CIT Outstanding Graduate Mentor Award, in 2019; the Purdue Focus Award, in 2020; the Purdue PPI Interdisciplinary Research Collaboration Award, in 2021; the Purdue Corps of Engagement Award, in 2022. He was named a Purdue University Faculty Scholar, in 2021.