Multimodal Coordination Measures to Understand Users and Tasks

SIYUAN CHEN, University of New South Wales
JULIEN EPPS, University of New South Wales & Data61, CSIRO

Physiological and behavioral measures allow computing devices to augment user interaction experience by understanding their mental load. Current techniques often utilize complementary information between different modalities to index load level typically within a specific task. In this study, we propose a new approach utilizing the timing between physiology/behavior change events to index low and high load level of four task types. Findings from a user study where eye, speech, and head movement data were collected from 24 participants demonstrate that the proposed measures are significantly different between low and high load levels with high effect size. It was also found that voluntary actions are more likely to be coordinated during tasks. Implications for the design of multimodal-multisensor interfaces include (i) utilizing event change and interaction in multiple modalities is feasible to distinguish task load levels and load types and (ii) voluntary actions should be allowed for effective task completion.

CCS Concepts: • **Human-centered computing → HCI theory, concepts and models**;

Additional Key Words and Phrases: Multimodality, task load, eye, speech, head

ACM Reference format:
Siyuan Chen and Julien Epps. 2020. Multimodal Coordination Measures to Understand Users and Tasks. *ACM Trans. Comput.-Hum. Interact.* 27, 6, Article 42 (November 2020), 26 pages.
https://doi.org/10.1145/3412365

1 INTRODUCTION

Estimating user mental state in our daily life is becoming more important. The increasing role of visual and verbal information in many contemporary tasks has resulted in a dominance of mental components over their physical counterparts [Sharples & Megaw 2015]. When interacting with devices, humans increasingly play the role of supervisor and decision-maker, which potentially leads to increased demands on human mental activity [Maior Wilson & Sharples 2018].

Understanding users mental state is an important aspect of Human–Computer Interaction (HCI) and for human-centric design [Afergan et al. 2014; Yuksel et al. 2016, Oviatt et al. 2004]. Contrary to users' physical state which often involves body movements, where user behaviors can be modelled through events generated by manipulating a mouse or keyboard [Myers et al. 2000], understanding

This work was supported in part by the U.S. Army ITC-PAC, through contract FA5209-17-P-0154.
Authors' addresses: S. Chen, School of Electrical Engineering and Telecommunications, University of New South Wales, High Street, Kensington, NSW 2052, Australia; email: siyuan.chen@unsw.edu.au; J. Epps, School of Electrical Engineering and Telecommunications, University of New South Wales & Data61, CSIRO, High Street, Kensington, NSW 2052, Australia; email: j.epps@unsw.edu.au.

1073-0516/2020/11-ART42 $15.00
https://doi.org/10.1145/3412365

users' mental state is challenging since mental activity is often covert. However, physiological and behavioral signals are persistent data sources which often can reflect mental and physical state at any time. They become an ideal part of computing interfaces for modelling user behaviors and providing context awareness.

In the past decades, physiological and behavioral signals have been studied as input modalities in, for example, voice-, eye-gaze-, gesture-based interactions or multimodal interaction [Oviatt 1999]. New classes of applications which employ more physiological and behavioral signals have emerged to mediate HCI and augment user experience [Fairclough 2009], for example, affective interfaces [Petta at al. 2011], adaptive learning interfaces [Yuksel et al. 2016], or multimodal-multisensor interfaces [Dumas et al. 2009]. They were designed to utilize different physiological and behavioral signals to reflect user's current mental states as computer inputs to determine what to display next. Among these interfaces, the development of brain computer interfaces (BCI), where various brain sensors [Yuksel et al. 2016] were used to obtain neural activities, has been notable. Non-BCI interfaces have employed a variety of physiological and behavioral signals, such as electrocardiogram (ECG), electromyogram (EMG), galvanic skin response (GSR), pupillary response, and paralinguistic speech, to explore user mental state [Chen et al. 2012].

Mental load reflects one aspect of mental state during tasks. As one long-term goal for HCI, estimation of mental load will soon be needed outside the confines of controlled research labs, which poses challenges on continuous assessment. Current studies often assess mental load during highly controlled, specific task stimuli which might not easily generalize to continuously changing tasks with unknown stimuli in real life. It is also evident that some physiological and behavioral signals such as eye activity, speech, and head movement exhibit different patterns in different task types due to task requirements [Chen and Epps 2014A; Huttunen et al. 2011], which might lead to incorrect interpretations for unseen or unknown tasks. Therefore, the question of how to utilize physiological and behavioral signals to assess mental load for interaction mediation remains open.

The article proposes a new multimodal measure to assess mental load under different task types which are promising for the multimodal-multisensor interfaces proposed by Oviatt et al. [2017]. We employed three modalities with few restrictions on user activity—eye activity, speech, and head movement. A wearable system continuously records these multimodal signals in different task types (mental and light physical activities). We investigate the disparity in coordination between these modalities and propose a multimodal measure to answer the research question of how task load levels and load types change the coordination between these modalities and whether the proposed multimodal measure is robust and effective to assess task load.

2 RELATED WORK

In this section, we survey prior research on mental load assessment. We discuss current mental load measures focusing on those from eye activity, speech, and head movement and a few multimodal measures. Finally, we discuss the knowledge and theory from human motor coordination research to develop the rationale for our proposed measure.

2.1 User Mental Load and Task Load

Assessing user mental load is an important aspect of understanding users during tasks. This line of research builds significantly on the developed models and theories on how information is processed—received, organized, stored, and retrieved—in the human cognitive system, which help us understand how mental load is generated. For example, Kahneman's Capacity Model for attention [Kahneman 1973] illustrates how attention (mental effort) is allocated and affected by the allocation policy, and the evaluation of demands on limited capacity; Baddeley's Multi-component Model for working memory [Baddeley 2003] describes how the central executive controls

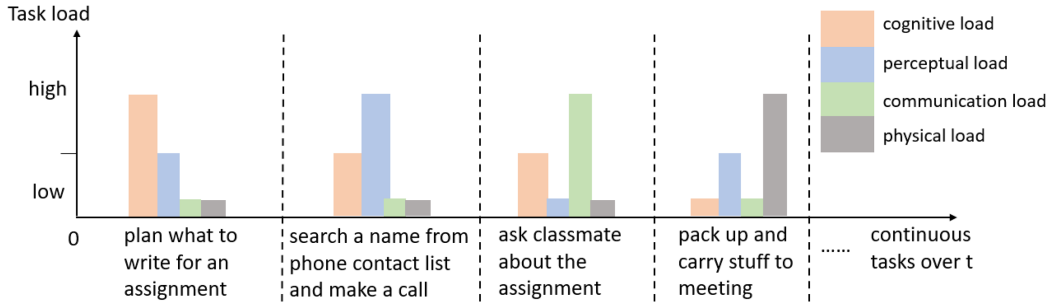

Fig. 1. An illustrative example showing the four-dimensional task load framework, where each task is assessed by four load types and the intensity (level) of each type exposed on users.

working memory (stores temporal visual and auditory or verbal information) and long-term memory (crystallized knowledge); Multiple Resource Theory [Wickens 2008] describes the tasks involving different dimensions and levels of different resources can cause high demands; Cognitive Load Theory [Paas & Van Merrienboer 1994] illustrates the relationship between causal factors, cognitive capacity and assessment factors; and the Load Theory of Attention [Lavie et al. 2004] demonstrates the occurrence of cognitive control and bottleneck of information processing under different perceptual load conditions. Although they have different foci on mental load, all theories agree that humans have limited attention and working memory, and mental load is significantly related to task demands and the corresponding mental effort. Meanwhile, they highlight that different types of tasks, such as visual (or perceptual load) and verbal (communication load), may affect working memory or resources in a different manner.

Although these models and theories do not circumscribe the type of tasks, previous studies have practically employed specific task types, e.g., n-back task, arithmetic task, or search task, sometimes with a dual task to assess mental load. Differently to these designed tasks, task types in real life continuously change with little notice. For instance, in a typical office work context, we may sit at a desk to plan, search information, attend a meeting and talk to colleagues. The task demands shift from cognitive task, perceptual task, light physical task to communication task. Therefore, previous research findings [e.g., Appel et al. 2018; Fridman et al. 2018] which used a specific task to estimate mental load might only be applicable to some moments under the same task type, assuming that the type of this specific task is already known, and the algorithms are suitable for all tasks like the mental load assessment systems based purely on recall and subjective input such as NASA-TLX [Hart and Staveland 1988] or SWAT [Reid and Nygren 1988].

Recognizing the human activities of daily life [e.g., Steil and Bulling 2015] is one way to know a user's current task type and hence assess mental load according to the signal patterns under this task type; however, human activity and its variants can be unlimited, not falling neatly into pre-determined task categories. Recently, a four-dimensional task load framework [Epps & Chen 2018] was proposed to describe all possible activity (tasks) attributes as shown in Figure 1, where task loads are evaluated in terms of the following four attributes: cognitive, perceptual, physical, and communication, based on the Berliner task taxonomy [Vicente, 1999]. That is, the attributes of a specific activity can be represented in terms of the task load intensity for each load type. For example, an arithmetic task can be represented by high cognitive load, medium perceptual load, low physical load, and low communicative load. The cognitive load is high because we calculate numbers using working memory. The perceptual load is medium is because we need to see the numbers. This representation opens the opportunity for longitudinal and continuous computing, as we only need to evaluate four task load types, rather than actually identify every specific task,

and provides a compact and generic interface for human task status context awareness in HCI systems.

In summary, previous studies often focused on a specific task while generalizing their insights to unlimited tasks in our daily life is very challenging. Few studies have examined multiple task types from cognitive, perceptual, physical, and communicative tasks at the same time to investigate mental load levels and assess the efficacy of physiological and behavioral measures regard or regardless of the four task load types.

2.2 Physiological and Behavioral Measures

Mental state assessment conventionally relies on a manual and subjective process involving tedious diary-style recordings and expensive expert consultation [Czerwinski et al. 2004]. For mental load assessment, it often includes subjective rating, performance scores, or response times which are obtained after task completion and have the shortcomings of being either subjective, manual, or non-continuous [Epps and Chen 2018 for a review]. Automatic assessment is expected to change these conventional approaches by providing quantitative measures, prompted by recent progress on low-cost sensing technologies to obtain users' physiological and behavioral signals. This line of research has the advantage of being real time and objective, but often suffers from interference from noise sources (either external environments or internal nervous system). Currently, it is often researched under specific tasks and segmented moments under controlled laboratory conditions.

As for wearable, non-invasive sensing modalities, eye activity, speech and head movement were each found to be correlated with mental load and/or physical load. Among them, eye activity, including pupillary response (involuntary response), blink (involuntary response), and eye movement (fixation and saccade, voluntary activity) [Chen et al. 2011; Di Nocera et al. 2007], is often used to infer user cognitive load during a specific task. Common measures include simple but effective statistical features such as mean, median, and variance [e.g., Appel et al. 2018; Pfleging et al. 2016; Chen Epps and Chen 2013A]. Other measures, e.g., energy or power [Chen and Epps 2013B], Nearest Neighbor Index [Di Nocera et al. 2007], requiring signal processing techniques were also explored. It is worth noting that eye activity patterns may be different in different types of tasks. For example, in comparisons between a cognitive task and a perceptual task of different load levels, Chen and Epps [2014A] found that pupil diameter and blink rate increased as cognitive load level increased only when perceptual load was low. When perceptual load increased, blink rate decreased and pupil diameter was saturated. The inconsistency has also been found in eye movement in different task contexts [Foulsham 2015]. This may result in, for example, that fewer saccades may indicate lower load in search tasks but it also may indicate higher load during reading tasks [Liversedge and Findlay 2000]. All these suggest that task type may affect eye activity patterns and we need to know the task type in advance in order to correctly interpret these patterns as discussed in Section 2.1.

Speaking is a common activity in our life (voluntary activity). With progress in artificial intelligence technology, people increasingly communicate via intelligent systems or applications, e.g., smart phone or voice assistant. To assess mental load using speech, paralinguistic information is often used, including prosody dynamics (e.g., pitch, formant frequencies, intensity, and speech rate), spectral representations (e.g., mel-frequency cepstral coefficients (MFCC)), or glottal excitation flow patterns [Yin and Chen 2007; Quatieri et al. 2015]. It is worth mentioning that cognitive load was often estimated during a specific task where required events occurred or sentences were read out. Differently, Huttunen, et al. [2011] separated three types of cognitive load in a simulated combat flight—situational awareness load (perceiving environment elements), information load (receiving information via radio communication), and decision load (making single/multiple choices), and investigated pitch and intensity change. Their results demonstrated

that pitch increased when cognitive load increased but the increasing slope was different in the three load tasks. Meanwhile, the pattern of intensity of situation awareness load was different to that of the other two, which was not increasing as load level increased. This indicates that the patterns of some measures may change with different task type. These measures can be used with some confidence for studying tasks with the same attributes, but it is unknown whether they can be applied to different task types.

Compared with eye activity and speech, there are a limited number of studies which have employed head movement (voluntary activity) recorded by Inertial Measurement Unit (IMU) for mental load assessment. Makepeace and Epps [2015] used statistical features extracted from acceleration and gyroscope signals for cognitive load measurement. Recently, Chen and Epps [2019] used an event analysis method to convert continuous gyroscope signals to discrete atomic events comprising increase, decrease, and central movement of head movement, and took the frequency counts and intensities of these events as measures to assess the four-dimensional task load levels. These measures were found to be more accurate and explainable than statistical features aggregated across many time instants in a sedentary context. The increase, decrease, and central events may be interpreted as a result of two antagonist systems competing in the nervous system (e.g., the interaction of the two skeletal eyelid muscles produces blinks [Irwin and Thomas, 2010], and the interaction of the dilator and sphincter muscles controls pupil dilation [Steinhauer et al. 2004]), reflecting the cognitive system state change by nature. However, this research direction has not been fully explored.

Regarding multiple modalities for mental load assessment, studies often focused on achieving high load level classification accuracy by combining different measures from different modalities, where complementary information was utilized. Modalities that have often been compared and combined are EEG, eye-related measures, and EMG, ECG, and GSR [e.g., Haapalainen et al. 2010; Hogervorst et al. 2014]. These studies aimed to explore the best modalities and the best modality combinations for mental load assessment in different specific tasks. However, they often achieved inconsistent conclusions [e.g., Haapalainen et al. 2010; Hogervorst et al. 2014]. The often-used method to combine these multimodal data is feature-level fusion or decision-level fusion, which is a process of selecting and aggregating effective statistical features. The interplay information between different modalities, such as the coordination between them, which has often been studied in motor control, has not been explored.

To summarize, as presented in Table 1, single modalities such as eye activity, speech, and head movement were found to be related to mental load in specific tasks. Statistical measures are often used to assess load levels; however, due to their numerical nature, these measures often vary between studies, and sometimes, the patterns may change in different task types. Multimodal methods currently utilize the complementary information from different modalities for load level assessment, while the interplay information such as the coordination between modalities has not been explored.

2.3 Human Motor Coordination

Motor coordination is a phenomenon where effectors and movements between different body parts are accurately coordinated in time to control task-related states of the body and environment to achieve goals [Diedrichsen et al. 2010]. There are a plethora of studies in the discipline of neural science and psychology aiming to understand the mechanism of the neural system and motor control-related disorders. The consensus among them is that motor control involves cognitive processes such as anticipating and updating task aspects in order to plan, inhibit, monitor, and correct movements [Nowak et al. 2017; Diedrichsen et al. 2010; Shadmehr et al. 2010; Rigoli et al. 2012]. Empirical evidence from children with Developmental Coordination Disorder [Rigoli et al.

Table 1. A Summary of Related Measures to Assess Task Load

Method	Common measures or examples	Advantage and disadvantage
Subjective rating, performance score, and reaction time	Point categorical scale Error rate Time to complete [see Epps and Chen 2018 for a review]	Easy to implement, directly relate to tasks, but being subjective, manual or non-continuous, are undertaken after tasks so as to not interfere tasks.
Pupillary response (involuntary response)	Statistical measures such as mean [e.g., Appel et al. 2018; Pfleging et al. 2016; Chen, Epps and Chen, 2013A] energy, or power [Chen and Epps 2013B]	Can be real time and objective, but often suffer from interference from noise sources. Patterns may be inconsistent, e.g., pupil diameter increases with cognitive load [Steinhauer et al. 2004] only when perceptual load is low [Chen and Epps, 2014A].
Blink (involuntary response)	Blink rate, duration, and inter blink duration [Chen and Epps 2014A; Hogervorst et al. 2014]	Can be real time and objective, but often suffers from interference from noise sources. Patterns may be inconsistent, e.g., blink rate increases with cognitive load [Chen and Epps, 2014A] and decreases with perceptual load [Irwin and Thomas, 2010; Chen and Epps, 2014A].
Fixation and saccade (voluntary response)	Number and duration of fixations and saccades [Chen et al. 2011, Chen Epps and Chen 2013A], and Nearest Neighbor Index [Di Nocera et al. 2007]	Can be real time and objective, but patterns may be inconsistent in different task contexts [Foulsham 2015], e.g., fewer saccades may indicate lower load in search tasks but may also indicate higher load during reading tasks [Liversedge and Findlay 2000].
Speech (voluntary response)	Prosody dynamics (e.g., pitch, formant frequencies, intensity, and speech rate), spectral representations (e.g., MFCC), or glottal excitation flow patterns [Yin and Chen 2007; Quatieri et al., 2015]	Can be real time and objective, but patterns may be inconsistent in different task context, e.g., intensity change trend is different in three types of cognitive load in a simulated combat flight [Huttunen et al. 2011].
Head movement (voluntary response)	Statistical measures of velocity and acceleration [Makepeace and Epps 2015] Frequency counts and intensities of discrete atomic events [Chen and Epps 2019]	Can be real time and objective but has been less studied for task load.
Multiple modalities (involuntary/voluntary response)	Statistical measures of pupil, GSR, heat flux, ECG, EEG, and heart rate [Haapalainen et al. 2010] Statistical measures of EEG, GSR, respiration, ECG, pupl, and blink [Hogervorst et al. 2014]	Can be real time and objective, but findings may be inconsistent in different task contexts, e.g., in one study, heat flux and ECG were the best two features and fusing them achieved 81% accuracy [Haapalainen et al. 2010] while another study shows EEG and eye features were the best two and fusing them achieved 91% accuracy [Hogervorst et al. 2014].

2012] suggests that motor coordination is related to executive function. More precisely, it is more closely related to visuospatial working memory than verbal working memory. Nowak et al. [2017] believe that like motor coordination, coordinated activity among elements is the essence of effective performance in all operational levels of human activity, from neural, behavioral, mental function to social dynamics. In other words, to perform an effective function, specific elements (neurons, movements, thoughts, and feelings, and individuals) are repeatedly assembled and disassembled as needed in response to task demands and environmental constraints. These studies and models clearly indicate that concurrent activation of certain networks in the brain could be measured by synchronization of low-level elements, such as movements or physiological changes in our body. Therefore, synchronization of low-level elements may reflect the load imposed on the cognitive network system.

Eye movement is an enduring interest in sensorimotor studies, and the coordination between eye and head has been well studied to understand how our visual and motor skills unite in tasks. Einhauser et al. [2007] provided quantitative evidence of the extent of eye and head movement coordinated in terms of velocity and direction in natural viewing conditions of walking, staying at home and at the train station. They found that around 40% of the time eye and head movements co-occurred. Meanwhile, at least 20% of eye movements were in an opposite direction to the head movement to compensate for gaze stabilization, but the dominant function of eye-head coordination was to support head movement in directing gaze synergistically. However, few studies explored speech movements together with other modalities. Dromey and Benson [2003] studied how speech movements were influenced by specific distractors. They found that motor task demand (hand movement) significantly reduced lip displacement and velocity. This influence was different to that imposed by linguistic and cognitive distractors, which only caused low consistency of speech movements. Nevertheless, although motor coordination occurs in our everyday tasks, there is little study into the coordination between different modalities to assess task load. Multimodal coordination in this study refers to the coordination of motor responses from different sensory when interacting with the world, e.g., eye movement, eyelid movement, head movement, pupillary response, and verbal response. In this article, we realize the concept practically by automatically detecting combinations of events across different modalities that are proximal in time.

The rationale for using multimodal coordination measures for task load assessment is based on limited capacity theory and the human motor control theory of optimal control. Limited capacity theory suggests that demands of concurrent activities must be met by a finite pool of processing capacity (or mental resources) [Kahneman 1973; Pashler 1994; Wickens 2008]. Some operations might be interfered with, e.g., speaking while doing complex hand movement, or might be synergetic, e.g., the eye and head as aforementioned. Since virtually every natural task involves motor responses like eye, head movement, and speech, higher task demand can mean less capacity for these planning and acting with these modalities, hence task load may change the extent of modality coordination.

The computational framework of optimal control theory is currently a general theory of motor coordination to explain why and how movement coordinates. The essence of the theory is that motor coordination can be understood as the solution of optimizing behavior with respect to biologically relevant task goals [Diedrichsen et al. 2010]. From this theory, if certain movements are expensive for the motor system to produce, they are planned and performed with the least effort in the optimization process. This may increase coordination between different modalities in tasks. When task load is high, the temporal coherence may be changed in order to achieve new optimization under this condition. Although this subsection has focused on coordination of explicitly measured motor movements, there is significant evidence of increased synchrony among EEG channels under higher load levels [Zarjam et al. 2013; Palva and Palva 2016]. Bearing in mind that motor

movements originate from the motor cortex, these results also provide strong support for the relationship between motor coordination and task load. It is worth mentioning that our study focuses on the coordination between physiological and behavioral signals from one individual for task load analysis. This is different to studies examining the interpersonal and interactional synchrony with physiological and behavioral signals for rapport engaged interactions [Reidsma et al. 2010].

In summary, motor coordination occurs in everyday task and may bring about persistent patterns in physiological and behavioral signals such as eye, head and speech. The limited capacity theory and optimal control theory suggest that the coordination between multiple modalities might change according to task load type and load level. The motivated us to investigate multimodal coordination measure for task load assessment.

3 EXPERIMENT DESIGN

To assess whether the proposed multimodal coordination measure is reliable across different task types, we design four types of tasks in two different load levels. We focus on event-based measures and examine their coordination across multiple modalities, and analyze their feasibility as effective measures to index load change in the four task types.

3.1 Hypotheses

H1: Fewer coordinated actions occur when load is high than when load is low per unit time, and more coordinated actions per task occur when load is high than when load is low.

Since actions, such as starting speech, saccade, or head movement, are planned prior to motor execution by an active cognitive mechanism, when load is high, according to limited capacity theory, within a time unit, there is less room for these actions to be planned and executed which leads to less coordinated actions. However, from the task duration perspective, the same actions are required to be conducted in the same task type regardless of its load level, although a high load task usually requires more time than a low load task of the same type. Based on optimal control theory, to be efficient, there should be more coordinated actions to save time. As mentioned earlier, findings from EEG studies [Zarjam et al 2013; Palva and Palva 2016] suggest that when task load is high, there are more areas activated, indicating that high load means more synchrony, therefore, more actions should be coordinated.

H2: Coordination between involuntary–involuntary actions and involuntary–voluntary actions are more likely to occur when load is high than when load is low.

For involuntary actions, such as pupil size increase and blink, which are a result of agonist–antagonist system as responses to environments and tasks, there are few studies to show that they possess the same resource as voluntary actions, such as speech, saccade, and head movement, to our best knowledge. We assume that they require less capacity resource due to their automated nature. Based on optimal control theory, we develop this hypothesis.

H3: Regardless of the load level, these four task types possess a significantly different coordinated event index.

Regarding different task types, since the main required sensorimotor responses are different in these four distinct task load types, for efficient task completion or minimizing effort, different coordinated actions can be expected to appear, which can help assess task load type.

H4: If H1 is true, the density of coordinated actions per unit time in the middle of the task is lower than that at the beginning and end of tasks.

Human information processing theory [Kahneman 1973; Baddeley 2003] says that information is received from sensors before being processed by the cognitive system, so we believe that this mostly occurs at the beginning of a task. The middle of the task is the most effortful time due to the demands of problem solving. At the end of task, the problem has been solved and load is

released. Considering this dynamic change in information processing during a task [Bailey and Iqbal 2008], we believe load level also changes accordingly. That is, the load level is low at the beginning and end of a task while it is high during the middle of the task. Therefore, to further understand whether the temporal coordination of actions agrees with load level change during a task, we develop this hypothesis.

3.2 Wearable System Using Eye, Speech, and Head Activity

The wearable system we used was equipped with a non-invasive infrared (IR) camera, a microphone, and an IMU, embedded in a single glasses form-factor device to acquire eye, speech, and head activity, respectively. The camera was an off-the-shelf webcam with a built-in microphone. We modified it to be IR sensitive and able to record near-field eye videos with a resolution of 640×480 pixels at a frame rate of 30 fps and with audio sampled at 44.1 kHz. The IMU prototype consisted of an IMU (MPU 9150) and output three-axis acceleration, angular velocity, and magnetic field strength at a rate of around 20 Hz. This combination of modalities has the advantage of being non-intrusive (no direct contact with the skin) and less sensitive to unwanted variability from body movements compared with ECG, EMG, EEG, GSR and the like, and being mounted on a single integrated eyewear device, but has been less researched than other modality combinations.

To acquire eye activity, the pupil center, pupil diameter, and blink were firstly extracted from near-field IR eye images. They were then resampled to 20 Hz. Then from pupil center, eye movement (i.e., fixation and saccade) were obtained. Pupil size were interpolated during blink. The algorithm for continuous estimation of pupil diameter from near-field IR eye images was the self-tuning and dual ellipse fitting algorithm, with pupil diameter measurement accuracy of around 0.02 mm and blink detection accuracy of around 99% (Chen and Epps, 2014B]. Saccade events were extracted from the pupil center relative to the head employing dispersion-based algorithms using one degree of visual angle for at least 200 ms [Salvucci and Goldberg 2000]. From the recorded speech, we obtained speech segments (10 ms) over time using Voice Activity Detection techniques with the open source software openSMILE [Eyben et al. 2016]. After resampling it to 20 Hz, we set voicing probability to >0.7 to obtain voice activity. From the IMU recordings, only the head angular velocity of roll, pitch, and yaw was used for subsequent processing. As the data collected were not uniformly sampled, they were resampled to 20 Hz for multimodal processing.

Among all the eye, speech, and head activities, some are naturally event-based, like blink, saccade, and speech, while pupil size and head angular velocity are numerical and continuous in nature. To convert these continuous signals into meaningful events, we used the atomic head movement segmentation algorithm [Chen and Epps 2019] to segment them into increase, decrease, and "central" movement events. The threshold used for "central" movement was 3°/s for head angular velocity based on experimental head stabilization studies. The threshold for "central" movement events from pupil size was the pupil size baseline of the task, which was the average pupil size during the first 0.5 seconds of task beginning. The physiological rationale behind these three events is that they indicate a balance change in the antagonist sympathetic and parasympathetic systems (Section 2.2). When events are in a state of increase, it means efforts from one system have been made to overcome the resistance from the other, while events in a state of decrease indicate no sustained effort and/or the opposing autonomic system taking over to regulate the function. From this perspective, increase, decrease, and "central" events are an intuitive way to interpret behavioral events.

3.3 Proposed Multimodal Coordination Measures

We selected the blink onset, saccade onset, speech onset, pupil size increase, and head angular velocity increase as the events for coordination investigation. The reason for selecting onset only is

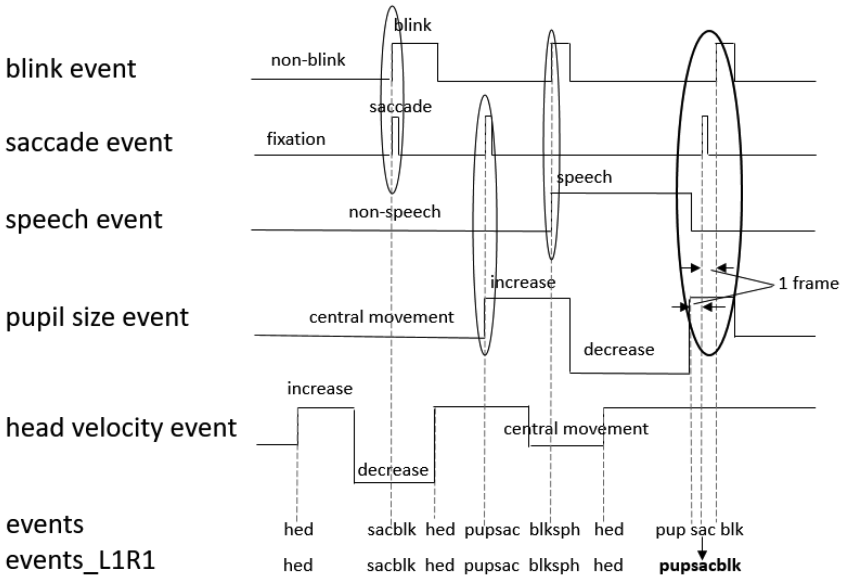

Fig. 2. The onset of blink, saccade, and speech events and the events for pupil size increase and head velocity increase were combined into a single event sequence. Events that occurred at the same time formed a new event representing a coordinated action. Events_L1R1 denoted events formed using a tolerance window of 2 frames (to account for synchronization and resolution mismatches). Note that "hed," "sac," "blk," "pup," and "sph" denote "head," "saccade," "blink," "pupil," and "speech" events, respectively.

that increase events are an indicator of exerting efforts to cope with stimuli, which may indicate the start of coordination. These events were collected according to the sequence of their occurrence from all signals. If more than one event occurred at exactly the same time, they were concatenated to form a new coordinated event (order-insensitive), which is the coordinated action we focus on in this study. Figure 2 shows the framework of accumulating multimodal events and forming coordinated events for task load analysis. The number of all distinct coordinated events (i.e., coordinated event index) is the multimodal coordination measure for a given task. For example, in Figure 2, the measure has value 3 if using the "events" scheme and 4 if using the "events_L1R1" scheme. The proposed measures indicate some degree of interaction between the multiple modalities, and are novel in mental load assessment.

3.4 Task Contexts

Four types of tasks were designed to induce four types of task load, namely, cognitive load, perceptual load, physical load, and communication load. These tasks were (i) solving a set of addition problems presented visually and giving the answers verbally, (ii) searching for a given target from pictures full of distractors, (iii) forearm lifting of two dumbbells with different weights, and (iv) holding conversations with the experimenter to complete a simple conversation or an object guessing game. These tasks or task instructions were displayed on a PC monitor in a sequence by running a MATLAB script, with which the task timestamps were also automatically recorded.

In each task, two difficulty levels were created to induce low and high task loads in participants. The two levels were manipulated by changing the difficulty of the addition problems, the size and number of the distractors, the weight of the dumbbells, and requirements for only yes/no answers (low load) or asking questions (high load), respectively. The task completion duration of cognitive

and perceptual tasks varied between participants. The physical and communication tasks were fixed to a maximum of around 1 minute.

3.5 Participants and Data

A total of 24 participants (14 males, 10 females, aged 18–25) volunteered, who met the criteria of being over 18 years old and not wearing glasses. They were required to wear the wearable system and sit at a desk, with full freedom to move any part of their body and speak while completing the four types of tasks (UNSW Human Research Ethics Advisory reference number 08/2014/23). At the beginning of the protocol, they clapped their hands and nodded their head at the same time in order to synchronize all signals for later processing from the audio, head movement, and eye images. Next, they completed each level of the four types of tasks, with full explanations provided by the experimenter next to them, followed by subjective ratings (0–7) of task difficulty at the end of each trial to check the validity of the induced level. Then these trials were continuously presented in a counterbalanced order and completed by participants without breaks. Following completion of all tasks, they were thanked and given a voucher.

In total, the experiment lasted 1 hour including introduction and debriefing, where each level of the tasks was repeated five times except the cognitive tasks, which had seven repetitions. The timestamps of each task were automatically recorded. The induced load level was treated as the ground truth, and this assumption was verified using participants' subjective self-ratings of the task difficulties.

Data were obtained from eye videos, speech, and IMU recordings during tasks. Eye activity videos were filmed at 30 fps by a small modified IR webcam mounted on a pair of lightweight glasses frames, pointing toward the eye. Speech files were acquired from the eye videos. There was also a "scene view" camera used to record all activities during the experiment for reference during annotation, as shown in Figure 3. Head movement was recorded using an IMU attached to the head by a head strap and connected to the laptop with a USB cable.

3.6 Analysis Methods

The analysis was conducted on a per-task basis. The first thing we did was to manually locate the timestamp in the scene video when participants started clapping, the timestamp in the sound waveform when the clapping began, and the timestamp where the vibration in the y-axis acceleration began suggesting nodding. As IMU data were already synchronized because they were recorded by the same device and eye videos and audio files were also synchronized because they were recorded by the camera, we used the correspondence of the timestamps to synchronize all the signals in processing. Due to different sampling rates in different devices, we resampled the raw pupil data from camera, raw IMU data, and segmented speech data to 20 Hz. With the automatically recorded timestamps of each task and the manually identified synchronization timestamps from the beginning of the scene video, audio file and IMU readings, signals within each task duration were segmented and manually checked to ensure a rough correctness. These data from each task were then processed by the system described in Sections 3.2 and 3.3, and the coordinated events during each task were obtained.

Coordinated event index was used as a dependent variable to answer the hypotheses in Section 3.1. It is worth noting that this is different to the count of coordinated events, which means the frequency of an event in a task (i.e., the y-axis in Figure 5): we calculated the coordinated event index (i.e., the number of distinct coordinated events shown in the x-axis in Figure 5) (normalized by individual task duration and no normalization) and averaged them across all trials in each task load level of each load type for each participant. With task load type and load level as independent variables, we performed repeated two-way measures analysis of variance (ANOVA) for the

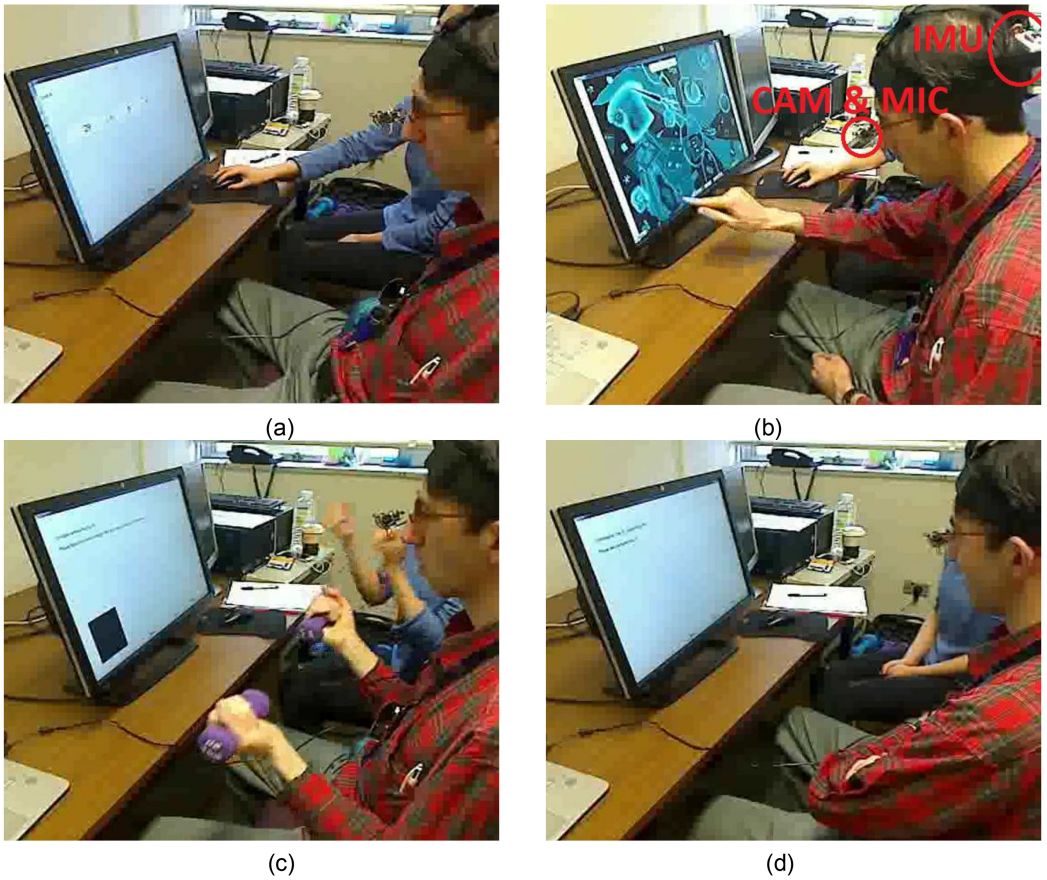

Fig. 3. A participant wearing camera, microphone, and IMU during (a) cognitive task, (b) perceptual task, (c) physical task, and (d) communication task.

difference between the load level and load type conditions for hypotheses H1 and H3 to find out whether they are good task load measures.

To examine the distribution of coordinated actions during a task, we segmented each task into the beginning, middle, and end epoch by setting thresholds of (a) the first 10% and last 10% of task duration accordingly and (b) the first second and last second of a task, in order to find out whether the difference depends on task duration. The choice of 1 second was due to the possible response time to a stimulus, and the 10% of task duration in our dataset ranged from around 1 to 6 seconds. We then counted the coordinated event index in these segments regardless of task load type and load level, and conducted a one-way repeated ANOVA to determine whether there were significant differences between the task beginning, middle and end segments, to test hypothesis H4, that is, whether they are sensitive to load change during tasks.

Then we investigated the composite of coordinated events by categorizing the coordinated events into the different combinations of involuntary actions (pupil and blink) and voluntary actions (speech, saccade, and head) based on whether it is controllable. The conditional probabilities of involuntary coordinated with involuntary actions, involuntary coordinated with voluntary actions, and voluntary coordinated with voluntary actions under different task conditions were calculated. The equation used is $P(A \cap B|C)$ where A and B are the voluntary and involuntary

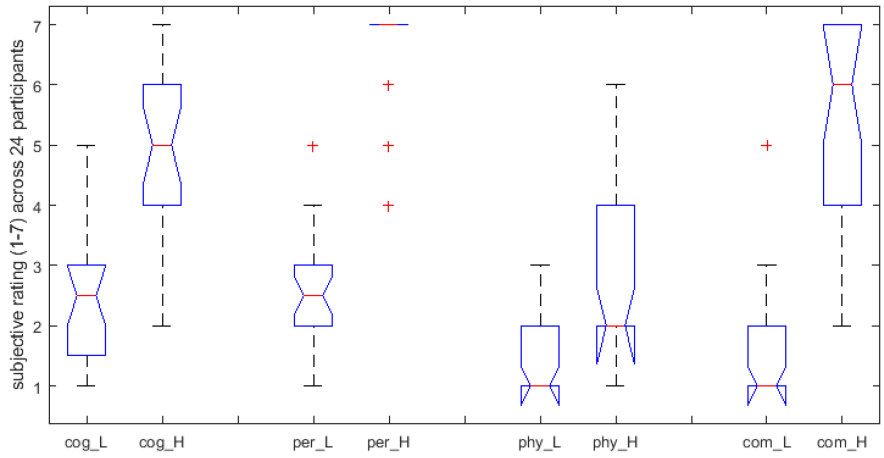

Fig. 4. Boxplot of the subjective ratings over 24 participants for the four task load types. Note "cog," "per," "phy," and "com" denote "cognitive load," "perceptual load," "physical load," "communicative load," respectively, and "L" and "H" denote "low" and "high," respectively.

action, respectively, and C is the condition of each load level at each load type. Chi-square goodness-of-fit tests were then conducted to confirm whether the occurred probability when load is high is from the same distribution when load is low, and whether they follow the chance-level expectation of the occurrence of the composite elements to prove hypothesis H2, that is, whether task load affects the coordination event type. Lastly, we found the common coordinated events which occurred in 75% of all the tasks from all participants to further understand the generality of the coordination measures.

However, as time is critical for coordinated event generation, it is difficult to ensure that all signals were perfectly aligned. To account for the possible error due to synchronization and resolution, we allowed neighboring events within a tolerance window of n frames (one frame is 0.05 s) (n = 1,2) on the left and right of the event when forming coordinated events. We named them events_Ln, events_Rn, and events_LnRn, where L and R denotes the tolerance window on the left or right respectively and n denotes the length of the window. Figure 2 shows an example of events_L1R1. All sets of coordinated events went through the same analysis to determine whether this temporal tolerance affected the conclusions.

4 RESULTS

4.1 Subjective Ratings of Load Levels in Each Load Type

To verify that the designed task load levels induced the loads as expected, we examined the ratings of perceived load from each participant. Figure 4 shows the subjective ratings of task load across 24 participants for different task load types. Nonparametric Wilcoxon paired sign tests (two-sided) confirm that in each task type, the perceived load level during the designed high load tasks was significantly higher than that in the designed low load tasks ($Z = -4.3, -4.3, -3.5$, and -4.2 for cognitive load, perceptual load, physical load, and communication load, respectively, $p < 0.001$ for all).

4.2 Coordinated Events for Different Load Levels in Different Load Type

To obtain an overview of the coordinated events in the four types of task load and how often each coordinated action occurs per second, we present Figure 5 where the coordinated events were aggregated across all tasks from all participants in each task load type. We can see that across the

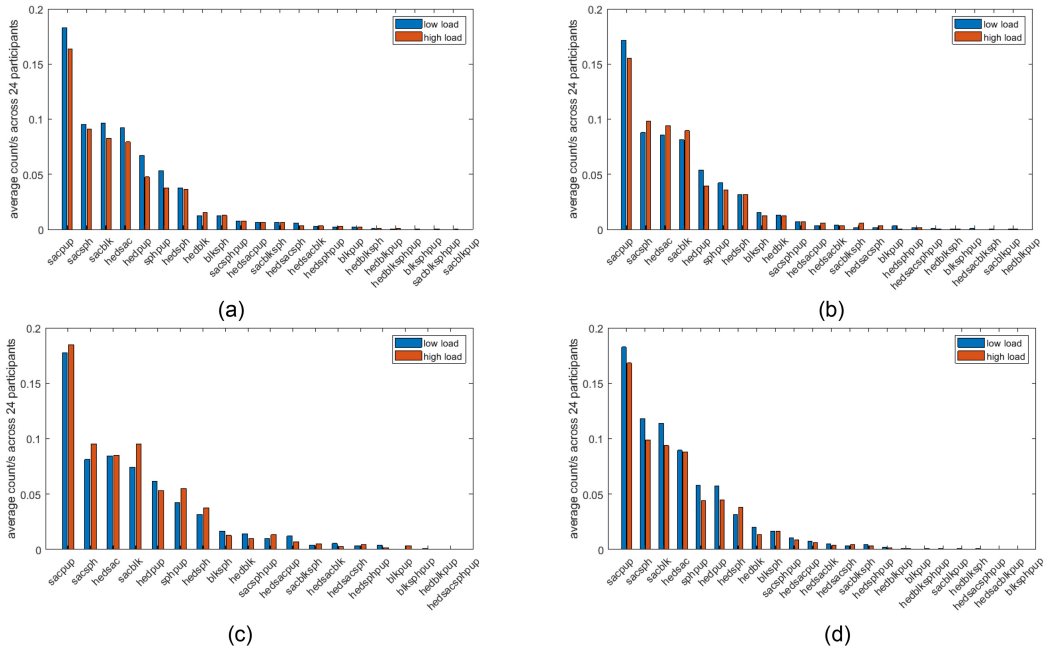

Fig. 5. The count of each of the coordinated events aggregated from all tasks from all participants, normalized by task duration and averaged across all participants for (a) cognitive load tasks, (b) perceptual load tasks, (c) physical load tasks, and (d) communicative load tasks. This demonstrates that the occurrence of particular coordinated events is different for different task types and load levels, a total number of 22, 22, 19, and 23 coordinated events occurring across all participants and across both load levels in the cognitive, perceptual, physical, and communicative task, respectively. In general, more distinct coordinated events occurred in high load level than in low load level in each of the task load types, the statistical significance of which can be found in Figure. 6.

different load types, the count of different modality change events is different. The coordination of saccade and pupil size increase and the coordination of saccade and head velocity increase are among the highest four counts in the four task load types, but they have different counts per second values and different count per second gaps in low and high load level of different load types. While the coordination of saccade and speech onset significantly depends on the task type and level, having the highest count in low level of communication tasks. Meanwhile, when simply aggregating all distinct coordinated events occurred in all tasks from all participants, we have 22, 22, 19, and 23 distinct coordinated events occurred regardless of low and high load level for cognitive, perceptual, physical, and communicative load task, respectively.

To investigate the hypotheses regarding the difference in modality coordination in different task load levels (H1) and different task load types (H3), we first examined the total number of distinct coordinated events, which indicates the total coordination effort. Figure 6 shows the boxplot of coordinated event index across 24 participants due to task load type and load level in the cases of without and with normalization by task duration. We can see that the trends of the proposed multimodal coordinate measures for low and high load level are consistent in different task load types. A repeated two-way ANOVA test found that significantly more coordinated events occurred during tasks when the task load was high than when load was low ($F(1,23) = 260.16$, $p < 0.0001$, $\eta^2 = 0.91$), while significantly fewer coordinated events occurred per second when task load was

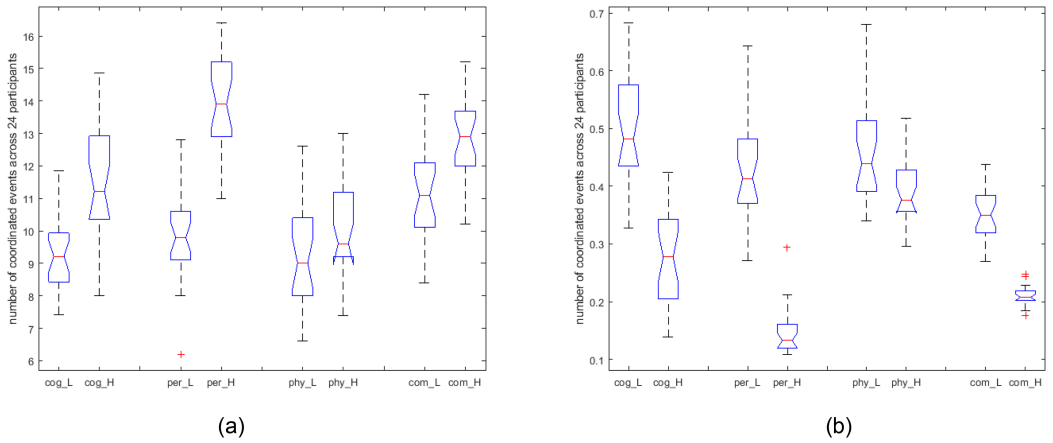

Fig. 6. Coordinated event index across 24 participants (a) without and (b) with normalization by each task duration for the low and high load level of each load type tasks. Participants usually took longer to complete high load tasks, but coordinated event index was higher than for low load tasks.

high than when load was low ($F(1,23) = 378.86$, $p < 0.0001$, $\eta 2 = 0.94$). There were significant differences between task load types in both cases of without and with normalization ($F(3,69) = 57.18$ and 50.46 respectively, $p < 0.0001$, $\eta 2 = 0.71$ and 0.69, respectively). Meanwhile, the interaction between task load type and load level was also significant regardless of normalization ($F(3,23) = 25.09$ and 56.26 respectively, $p < 0.001$), indicating a consistent trend.

When coordinated events were generated with different tolerance windows, this did not change the findings, although a few more coordinated events were generated in all conditions in general.

4.3 Coordinated Events Distribution During Task Segments

To examine the hypothesis about the difference in modality coordination in different task segments (H3), we utilized the proposed multimodal coordination measure with task duration normalization since it possesses higher $\eta 2$ than the measure without normalization. Figure 7 shows coordinated event index per second distributed in the segments of task beginning, middle, and end for the cases of using a percentage threshold (10%) and a fixed time duration threshold (1 second). We can see that the trend is also similar in the two cases with higher coordinated events per second (lower effort) in task beginning than that in task middle and end, but the difference is more evident using a fixed time duration threshold than using a percentage threshold. A repeated one-way ANOVA confirmed that there exists a significant difference between the three segments in both cases ($F(2,46) = 9.19$ and 46.2 for Figure 7(a) and (b), respectively, $p < 0.001$ for both). Post-hoc tests (with Bonferroni adjustment) confirmed that coordinated event index per second in the beginning of a task was larger than that in the middle of a task ($p < 0.001$) and the end of a task ($p < 0.01$), while there was no significant difference between that during the middle and end of a task ($p = 0.38$ and 0.02 for Figure 7(a) and (b), respectively) in both cases. However, some sets of coordinated events generated with tolerance windows affected this finding, where the coordinated event index per second during the end of a task was significantly lower than that in the middle of a task and that in the beginning of a task ($p < 0.01$), indicating that most effort occurred at the end of tasks.

4.4 Composite Coordinated Events

To test the hypothesis regarding the coordination difference in modality types in different task load levels (H2), we categorized pupil size increase and blink as involuntary types while saccade,

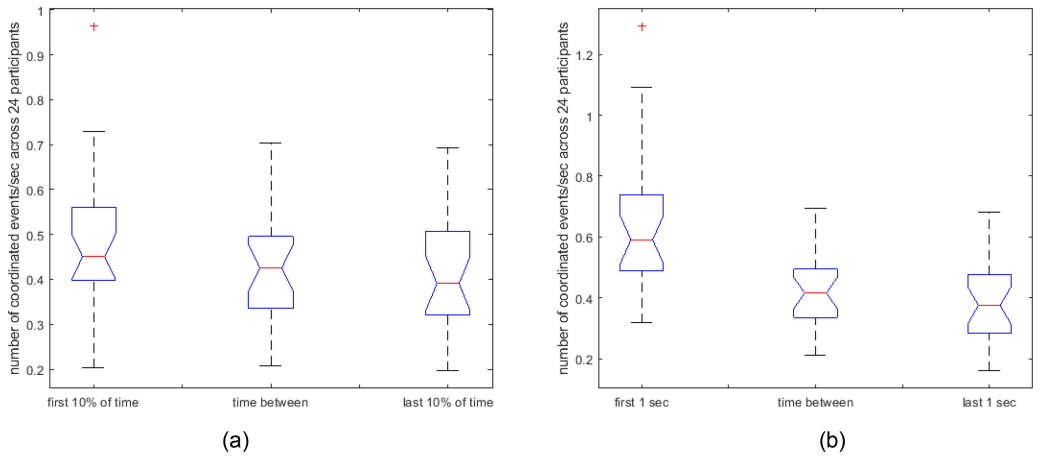

Fig. 7. The distribution of coordinated events occurred per second during the beginning, middle, and end of tasks. In (a), the beginning and end of a task were considered as the first 10% and last 10% of task duration respectively, while in (b) they were considered as the first one second and last one second respectively.

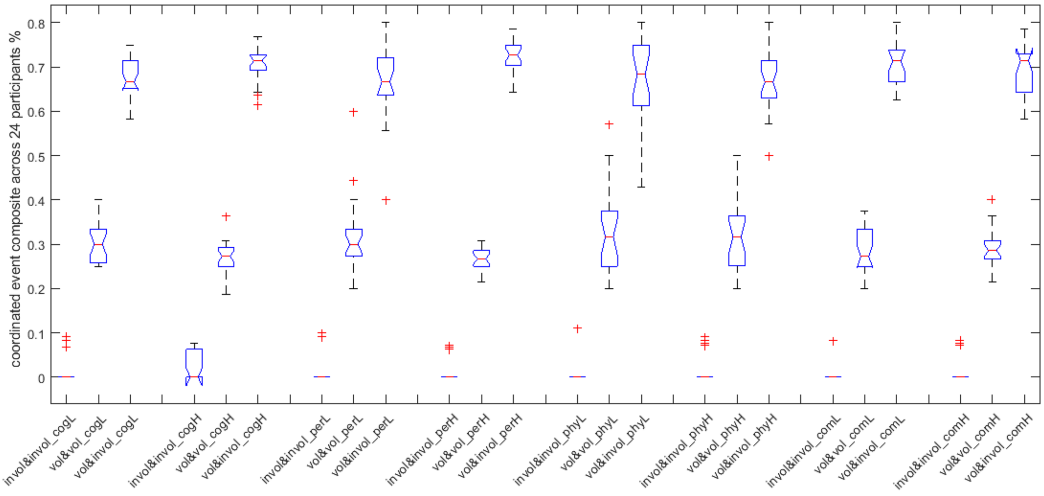

Fig. 8. The percentages of voluntary events only, involuntary events only and voluntary with involuntary, which made up of the coordinated events, under the conditions of different load type and load levels to show which kind of actions are easily coordinated.

speech onset, and head velocity increase were treated as voluntary, and investigated their percentage composition. Figure 8 demonstrates the composite coordinated events in percentages by the combinations of voluntary events and involuntary events. From this figure, we can find that the trend is greatly consistent regardless of task load levels and load types. That is, voluntary events co-occurring with involuntary events were always the most frequent combination, in a range of 0.4–0.8 (average 0.69), while involuntary events co-occurring with involuntary events is always the least frequent combination, in a range of 0–0.1 (average 0.0125). The chance-level expectation of occurrence of involuntary events only, voluntary events only, and involuntary and voluntary combination was 3.6% (1 possible involuntary event combinations out of 26 possible combinations of all events), 16% (4 possible voluntary event combinations out of 26 possible combinations of all

Table 2. Common Events Occurring in More than 75% of the Tasks Completed by all Participants

Load level	Common events occurring in > 75% of the tasks
low cognitive load	sacpup
high cognitive load	sacpup, hedsac
low perceptual load	sacpup
high perceptual load	sacpup, hedsac, hedpup, sacsph, sacblk, sphpup
low physical load	sacpup
high physical load	sacpup
low communicative load	sacpup, hedsac, sacsph
high communicative load	sacpup, hedsac, sacsph, sacblk, sphpup, hedsph, hedpup

events) and 80.8%, respectively. In this study, when task load is high, the probability calculated from all high load tasks from all 24 participants for the involuntary only events, voluntary only events, and voluntary with involuntary events was 1.0%, 30.1%, and 68.9%, respectively. When task load was low, these were 1.7%, 28.0%, and 70.3% respectively. The chi-squared goodness-of-fit test for composite percentages in two load levels indicated that there was no significant difference between the two distributions ($\chi^2(2) = 3.26$, $p = 0.20$) varied in load level. However, the chi-squared goodness-of-fit test for composite percentages from all tasks and from chance-level expectation of occurrence confirmed that there was a significant difference between these two distributions ($\chi^2(2) = 331.3$, $p < 0.001$). Meanwhile, using coordinated events generated with tolerance windows did not change these findings.

4.5 Common Coordinated Events

To further understand whether all participants share the same coordinated events and the same multimodal coordination measures in all tasks, we explored the common events among all tasks from all participants. Table 2 lists the most common coordinated events, found in 75% or more of tasks from all 24 participants. It is evident that the synchronization of saccade and the event of pupil size increase occurred in at least 75% of all tasks while the synchronization of saccade and the event of head velocity increase occurred in at least 75% of all high load level tasks. Meanwhile, high load tasks resulted in more coordination across multiple modalities in cognitive, perceptual and communicative load tasks but equal coordination in physical load tasks when compared with low load tasks. The generated coordinated events during high load level were different in different task load types but were more related to head event. Coordinated events generated with different tolerance windows only added a few more common events but did not change these conclusions.

5 DISCUSSION

5.1 Impact of Task Load Level and Load Type on Action Coordination (H1 & H3)

The results of coordinated event index varying with different load levels and load types shown in Figures 5 and 6 enable us to build an empirical understanding of the proposed multimodal coordination measure. First, different tasks require different motor responses to deal with a wide diversity of stimuli. These motor responses not only differ in modality usage across task load level but also across task load types. The exact coordination events occurred may heavily depend on the task requirements. For example, speech was required in cognitive and communicative load tasks while it was not in perceptual and physical load tasks. Therefore, speech-related coordination occurred slightly more often in cognitive load and communicative load tasks than in the other two

counterparts. Speech-related coordination, e.g., sacsph, has different occurrence rate in cognitive load and communicative load tasks, and the trend in low and high load level is also different to the other two counterparts. Meanwhile, the count of hedsac has a different trend in low and high load level in cognitive and perceptual load tasks as shown in Figure 5(a) and (b) and little difference was observable in the perceptual and communicative load task as shown in Figure 5(c) and (d). These make it hard to index load levels when the specific task is not known. These also indicate that the count of exact coordinated events is very task-specific and therefore it is not suggested in our proposed method.

Instead, the total number of distinct coordinated events is proposed as one multimodal coordination measure. It seems less task-specific, at least the trend of the measure in low and high load level is consistent in all the four task load types. As shown in Figure 5, most distinct coordinated actions, especially two coordinated actions, were largely shared in low and high load level within the same task type where the load level was verified by participants' subjective rating results shown in Figure 4. However, some of them, especially three or four coordinated actions, did not appear in both load levels. Regardless of which actions are coordinated, the number of distinct coordinated events is an index to distinguish task load levels. This is more evident in Figure 6(a). The statistical test result confirmed the significant difference in the measure for low and high load level. Meanwhile, the effect size ($\eta 2$) is as high as 0.91, suggesting that the large variations were due to task load levels. However, the measure is also different between the four task load types, reflecting different modality usage in load types. From Figure 6(a), we can see that without normalization, coordinated event index is significantly different between physical load and communication load, while the difference between cognitive load and perceptual load is smaller. Nevertheless, ANOVA tests indicated that the proposed measures can distinguish at least two different load types. The effect size ($\eta 2$) is as high as 0.71, suggesting the large variations were due to task load type.

Compared to the multimodal coordination measure without normalization, coordinated event index per second is an even less task-specific measure with consistent trends in low and high load level. As shown in Figure 6(b), regardless of task load type, coordinated event index per second in high load level was always smaller than that in low load levels, confirmed by the ANOVA test. The effect size ($\eta 2$) was 0.94, higher than that for non-normalized coordination measure, indicating that most variance is attributable to the factor of load level. There is also a statistically significant difference in this normalized measure between the four task load types, confirmed by the ANOVA test. It is worth emphasizing that after normalization, the effect size dropped from 0.71 to 0.69, suggesting less task-specific property. Visually, we can see that it is due to smaller difference between physical load and communication load while slightly larger difference between cognitive load and perceptual load than these in non-normalized measure. It is also worth mentioning that the advantage of using the normalized measure lies in one or two thresholds to separate low and high load level regardless of task load types. As shown in Figure 6(b), a threshold of 0.35 events per second can easily separate load levels regardless of task load types except physical tasks. This could be more related to the ability of the brain to process information in a unit time, and removes task specificity in load level assessment to some extent. Although our experiments were limited to only four specific tasks, this measure shows good generalizability in at least these three specific tasks. Nevertheless, for both multimodal coordination measures, hypothesis H3 is supported; regardless of the load level, different task types possess a significantly different number of distinct coordinated actions.

It is also interesting to explore the trend exhibited in low and high load level using the normalized and non-normalized coordination measures. If we do not take task duration into account, coordinated event index is always larger during high load level than during low load level. That is, to cope with high load, more actions are coordinated in effort. Although the increased effort

can be possibly transferred to an increase of the count of single actions during longer time without coordination, it does not concur with optimal control theory. When a time unit is considered, there is a lower coordinated action rate in high load level, meaning that limited resources might be the bottleneck, which agrees with multiple resource theory in that when mental load is high, the pool of multiple resources can be used becomes scarce, which slows down modality coordination. Therefore, hypothesis H1 is supported by these results; fewer coordinated actions occur when load is high than when load is low per unit time, and more coordinated actions per task occur when load is high than when load is low.

Regarding the ability to distinguish load type and load level at the same time, we found that the effect size was larger for load level using the normalized number of coordinated actions than using non-normalized, but it was smaller for load type. It seems that it is difficult to have high discriminability and high diagnosticity at the same time.

5.2 Impact of Task Segments on Action Coordination (H4)

To understand user task load change during tasks, a good measure should be sensitive to slight load level change during a task, where mental effort is unlikely to be constant during task beginning, middle and end. The results of Figure 7 demonstrate that the proposed measure—coordinated event index per second—can reflect the dynamic load level change during a task regardless of the factors of type and load level. However, no matter whether segmenting a task into the beginning, middle, and end using the criteria of 10%/90% of task duration or using the first/last one second of each task, a statistically significant difference was essentially only present between the task beginning and the remaining two segments. If a larger number of coordinated events per second indicates a lower load level, then the results can be interpreted as participants always experiencing a lower load at task onset, followed by the load escalating to a higher level, maintained until the task ends. This is similar to the expectation that load level would rise at the beginning of the task, but different to the expectation for the task offset, i.e., that load level would drop [Bailey and Iqbal 2008]. Therefore, in this study, hypothesis H4 is partly supported. A tentative explanation after reviewing some "scene view" videos could be that the annotated beginning and end of task did not strictly agree with the exact beginning or end of the task from the participants' point of view as the experimenter manipulated the mouse to end or begin a task in order to reduce participants' actions unrelated to tasks. It could also because that some participants had not found the answers or completed the tasks when time was up and it might be hard to release the mental load immediately.

When comparing the two methods for segmenting tasks into beginning/middle/end, we found that the impact of using 1 second was more evident than that using 10% by observing the median of each segment shown in Figure 7, because the measurement gap between task beginning and non-task-beginning in Figure 7(b) is larger than that in Figure 7(a). Therefore, it seems more appropriate to consider the first one second of a task as the "task beginning," when the mental effort in processing information is least.

Since in these experiments, like many others in the literature, we only have a load level ground truth for the whole task rather than moment-by-moment because of the challenges of task design and subjective rating. In future, it would therefore be good to have a physiological measure that might reflect dynamic changes without requiring some time to rise and fall (e.g., GSR) to measure dynamic load change effectively.

5.3 Impact of Action Type on Action Coordination (H2)

It is of interest to understand how modalities are coordinated, which could not only help understand user behavioral preference but also potentially help improve interaction quality. The results of Figure 8 exhibit a high similarity in terms of the coordinated event composites across

different task load types and load levels. Comparing the composite percentage in low load level, the percentage of voluntary event coordination in high load level dropped by around 2% and the percentage of voluntary with involuntary events coordination increased by over 1%. However, the chi-squared goodness-of-fit test confirmed that the difference was insignificant, indicating that load level changes did not influence the multimodal coordination type, and has no ability to distinguish task load level. Since this coordination composite is not affected by load level, hypothesis H2 is unsupported. This result is surprising. From the supported hypothesis H1, when task load is high, more coordination between different modalities appears. This means that they coordinate in the same way as that in low load level, that is, by a similar amount of voluntary events coordination, involuntary event coordination, and voluntary and involuntary event coordination, regardless of the high likelihood of voluntary and involuntary events coordination. The tentative reason could be that since tasks in low and high load level have the same style of stimulus presentation and task requirements that only differ in mental effort exertion, the physiological and cognitive systems tend to complete tasks in the same manner. Another reason could be that the newly generated multimodal coordination events due to high load occur at a trivial rate compared with the large amount of multimodal coordination events occurring in low load tasks in order to complete any task in an efficient way.

It is also interesting to find that the composite percentage of multimodal coordination diverges from the chance-level expectation of their occurrence. According to Figure 8, around 68% of the coordinated actions contained at least one voluntary action (speech, saccade, or head velocity increase) and at least one involuntary action (pupil size increase or blink) at the same time, while around 30% comprised all voluntary actions, and only 2% at most comprised all voluntary actions. These percentages are around 11% lower, 14% higher, and 1% lower than the chance-level expectation of respective coordination type. This suggests that there are some coordination types that occur all the time, e.g., saccade and head related coordination, while some occur seldom, e.g., pupil size increase and blink (pupil size was linearly interpolated during blink and the increase event depends on the pupil size before and after blink) in this study.

Since voluntary actions are often associated with the need to complete tasks and require planning in advance while involuntary actions may not, the composite of coordinated actions reflects the optimal minimum efforts in the cognitive system if all the actions cannot be avoided. However, the impact of task load level on action coordination seems to only make more different actions coordinated, but coordinated in the same composite way.

5.4 Most Frequent Action Coordination

Table 2 highlights the most frequent coordinated actions. Among them, head movement and saccade coordination nearly occurred in every high load task, which agrees with the fact that they have been observed in quite a few studies [Einhauser et al. 2007; Pelz at al. 2001]. Meanwhile, saccade and pupil size increase also occurred together more than 75% of the tasks in this study. The former coordination could be interpreted as head velocity increase facilitating eye movement to complete saccades for the optimal efficiency point of view since both are voluntary actions. While the latter coordination could have multiple interpretations. For example, a shift of fixation can result in local pupil light reflex due to light intensity change or near reflex due to distance change. It also could be because of mental effort in motor planning for where to fixate the eye, which then results in pupil size increase and saccade occurring at the same time. It is worth noting that saccade is the most frequently occurring event in our selected modality events. Therefore, it has the largest chance to coordinate with other events. Nevertheless, the coordination of saccade and blink is only seen in high load level of perceptual and communicative tasks. The tentative reason is that since visual information will be missed during blink and saccade, they intend to

occur together to reduce the chance of visual information loss when vast visual information is important, but saccade and blink cannot be avoided in a short time. Overall, from Table 2, we can see that saccade and head plays a great role in multimodal coordination during tasks.

Another important observation from Table 2 is that as load level increases, some new coordinated actions appear (e.g., hedpup, sacsph, and hedsph), and these new coordinated actions occur in above 75% of tasks within a certain type. The possible reason for no new common events in physical high load tasks could be due to time alignment because the coordinated events generated with different tolerance windows consistently show more coordinated two actions in high load level than in low load level. Meanwhile, among the new coordinated actions due to high load, the coordination of actions is varied, and half of them are related to the head. They occur in above 75% of tasks, which could be due to demands from tasks or experimental setup, e.g., moving the head forward to see better or keep the body balanced or speak faster. To us, what actions are coordinated is not important, but the occurrence in high load level outnumbers that than in low load level. Another observation is that there was no coordination of more than two events occurring in above 75% tasks although we can see that more than two events coordinated in some tasks from Figure 5. Overall, more commonly appearing coordinated events involve in high load level of tasks, which can indicate load level change in a certain task load type to some degree.

5.5 Limitations

There are some limitations in our experiment. Although we intended to design each of the tasks to involve one type of load as specifically as possible, it is difficult to find tasks for experimental use that contain only a single, pure type of load. Throughout our study, we refer to the four task types in terms of the dominant load type induced by the tasks. Meanwhile, in each load type, participants were required to do the designed tasks with designed low and high load levels rather than complete free-style tasks. However, this could be overcome by future studies since this study is just the first step of understanding the proposed measure's characteristics under different task load types. Although we employed a typical office context during the laboratory data collection, participants were always sitting in a chair, and physical activities such as standing up, walking, and sitting down, common movements were not included. Lastly, saccade as a voluntary action has the highest rate of occurrence followed by blink (involuntary action), which are naturally more likely to occur simultaneously with other actions, hence it may induce some bias in action coordination.

6 IMPLICATIONS FOR HCI

Our study bears implications for the design of multimodal-multisensor interfaces [Dumas et al. 2009; Oviatt et al. 2017]. Multimodal-multisensor interfaces aim to process and interpret information from various sensory and communication channels for natural human/machine interactions, just like humans inherently use multiple modalities to interact with the world. One important feature of the multimodal user interface is that it accommodates various modalities as input streams including speech, gestures, facial expression, and continuously interprets continuous multimodal information. Therefore, multimodal systems have the potential to be robust because of combining various modality information, to be personalized due to the ability of understanding user and context, and to be used in mobile computing [Dumas et al. 2009].

To design an efficient multimodal interface, a few theoretical principles need to be considered in the framework of input multimodal fusion, output multimodal fission, and user model and context [Dumas et al. 2009]. Our work of assessing task load level and load type can contribute to input multimodal fusion and user models and context in the design of multimodal-multisensor interfaces. Firstly, the proposed multimodal coordination measures utilized the temporal coordination information between different modalities in a novel feature-level fusion process and demonstrated

the potential to robustly interpret user task load level in different task load types using the modalities from eye, speech, and head movement. Specifically, for the measure of coordinated event index per time unit, a specific threshold can identify low and high load level irrespective of specific load types to some extent, as shown in Figure 6(b), indicating good generalizability in these three specific tasks. This effectiveness of this measure can improve system robustness for task load assessment when users freely perform different load type tasks during interaction with machines, which meets one of the guidelines for multimodal user interface design in terms of error prevention and handling [Reeves et al. 2004]. Unlike the proposed measures in our study, current task load measures based on physiological manifestations were often assessed in a specific task and may present conflicting trends for load levels in different tasks. Examples have been found in blink rate in different load levels in cognitive load tasks and in perceptual load tasks [Chen and Epps 2014A]. Some measures may also have different change rates in similar load levels for different load types [Huttunen et al. 2011]. In this study, we also arrive at a similar conclusion if we use the count of coordinated action as a measure, as the y axis shown in Figure 5. These measures presented different patterns in different load types appear to be only valid for indexing load levels within a designated task. If they are used in multimodal interface design, users may be limited to certain tasks or may lose interest to the interface.

The idea of assessing user load level in each of the four-task load type—cognitive, perceptual, physical and communicative load—is to maximize human performance, increase efficiency, flexibility, and user satisfaction based on understanding their cognitive and physical abilities, limitations, and their current contexts in terms of load types. This agrees with the insights for multimodal interaction in terms of the advantage of multimodal systems [Oviatt 1999] where multimodal systems should not set limitations on specific users or their preferences, tasks, and environments. It can also help examine whether multimodal interfaces meet the guidelines for multimodal user interface design regarding the ability of interfaces to adapt to different contexts, user abilities and to be used easily [Reeves et al. 2004]. In our study, the proposed multimodal coordination measure without normalization demonstrated the ability to distinguish task load types, where the effect size was large (0.71) (Figure 6(a)). If the proposed measures can be integrated into multimodal input recognizers and reliably identify load level and load types, continuously assessing user task load within an identified task load type can greatly facilitate user profile and context awareness in human–computer systems.

The modalities we used in this study were eye, speech, and head movement, which are closely linked with communication and can be used in mobile contexts. The system setup is relatively easy as off-the-shelf webcams and IMUs can be used and attached on the head. The computation of the measure can be demanding, requiring an accurate synchronization, but the proposed measure has good interpretability to different task contexts or evaluations of interaction techniques. Their coordination style can also help design natural multimodal interfaces when using them both as communication input and assessing mental state input. As found in one of the ten myths of multimodal interaction [Oviatt 1999], users can interact multimodally with different communication inputs but do not always do so, depending on the action type, and different communication inputs do not often co-occur temporally. This may be true for communication input. However, as assessing mental input in our study, we found that voluntary actions are more likely to be coordinated temporally than their chance-level expectation in order to complete tasks, which was not influenced by task load level. With the proposed multimodal coordination method, the implication is that to assess user mental state, we need to collect both voluntary and involuntary signals to make the measures usable. Meanwhile, a diverse set of allowable (e.g., not restrict body movement) or encouraged (e.g., provide something to motivate uses actively explore) voluntary actions may facilitate task completion. Overall, the findings reported in this study could potentially affect the

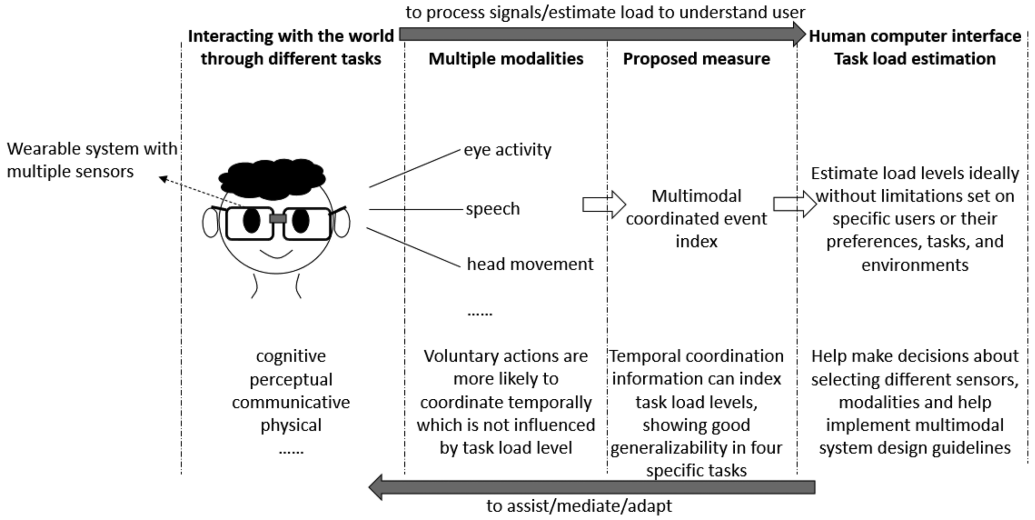

Fig. 9. A summary of how the proposed coordination event index can be interpreted and integrated into multisensory and multimodal system interface design.

decisions of selecting physiological and behavioral signals from a multimodal interface and for multimodal system design, as summarized in Figure 9.

Finally, we provide some guidelines to use the proposed coordinated event index to estimate task load. First, being aware of the type of tasks is a first key step to generate understanding about task load for users in different scenarios (Principle 1). This can allow identification of task load types, such as cognitive, perceptual, physical, and communicative load, and levels of each load type that may help in explaining why some physiological and behavioral signals work more effectively than others, or why some moments are more effective than other moments when task load type changes unnoticedly. This holistic view of task load type and level also can help to generalize the use of multimodal multisensory interfaces and the use of appropriate analytics when task load type change cannot be avoided (Principle 2). The multiple modalities recommended include the motor responses that commonly occur in tasks, such as eye activity, speech, and head movement, which consist of both voluntary and involuntary actions. In more technical terms, these multimodal data that can be automatically captured by wearable sensors can be processed to find change events, particularly blink, saccadic movement, pupil size increase, and head movement increase events. In all cases, focus on how many temporally coordinated events occur rather than what actions are coordinated as the latter may vary between different use scenarios (Principle 3). Without considering task duration, the coordination of event index during a task is high if the task load level is high indicating more coordination for efficiency. However, normalizing coordination of event index by time is recommended. The index value will be low if task load level is high, indicating scarce resources per unit time to respond. No matter how many modalities are captured from multiple sensors, an important consideration is how fast the events of interest occur, which determines the sampling frequency for data collection. Choice of an appropriately high sampling frequency for each sensor is vital to avoid missing events of interest, but very high sampling frequency will decrease the processing efficiency of the analytics (Principles 4). Accurate time alignment from different sensors is also critical for locating temporally coordinated events. However, a duration of tolerance window of up to 0.2 s is also acceptable, which did not change the conclusions found in this study.

7 CONCLUSIONS

In this study, we proposed multimodal coordination measures to understand user current task load type and load level using physiological and behavioral signals recorded with wearable devices out of the tightly controlled research lab environment. Differently to previous studies, where physiological and behavioral measures were investigated under a single task context and investigated the most informative modality or the best combination of multiple modalities working together, our study is the first to utilize the interaction information between modalities and analyze different type of task loads to assess the ability to index load levels across different load types.

Our results show that high load tasks can increase coordinated event index compared with that in low load tasks. However, if this measure is normalized by task duration, in a time unit, coordinated event index in high load tasks is lower than that in low load tasks. These observations agree with multiple resource theory, optimal motor control theory, and observations from EEG studies where widely separated areas of cortex were activated at the same time during tasks. Another important message is that coordinated event index per second can index load level regardless of task load type, more effectively than indicating load type. Good sensitivity of the proposed measure was also shown in reflecting dynamic load level changes from task beginning to middle segments regardless of load type in our data. Finally, we found that voluntary actions are more likely to be coordinated during tasks since their composite percentage was significantly different to the chance-level expectation; however, there was no evidence showing that it was affected by load levels.

All our findings suggest that the proposed multimodal coordination measures are promising for continuous user mental state assessment in terms of load level and load type. The measures may provide important automatic task load context for future multimodal-multisensor interaction systems.

ACKNOWLEDGMENTS

Opinions expressed are of the authors and may not reflect those of the U.S. Army.

REFERENCES

S. Sharples and T. Megaw. 2015. Definition and measurement of human workload. In *Evaluation of Human Work*. John R. Wilson and Sarah Sharples (Eds.). CRC Press.

H. A. Maior, M. L. Wilson, and S. Sharples. 2018. Workload alerts-using physiological measures of mental workload to provide feedback during tasks. *ACM Transactions on Computer-Human Interaction* 25, 2 (2018), 9–30.

D. Afergan, E. M. Peck, E. T. Solovey, A. Jenkins, S. W. Hincks, E. T. Brown, R. Chang, R. J. K. Jacob. 2014. Dynamic difficulty using brain metrics of workload. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'14)*. 3797–3806.

B. F. Yuksel, K. B. Oleson, L. Harrison, E.M. Peck, D. Afergan, R. Chang, and R. J. K. Jacob. 2016. Learn piano with BACh: An adaptive learning interface that adjusts task difficulty based on brain state. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'16)*. 5372–5384.

S. Oviatt, R. Coulston, R. Lunsford. 2004. When do we interact multimodally?: Cognitive load and multimodal communication patterns. In *Proceedings of the 6th ACM International Conference on Multimodal Interfaces*. 129–136.

B. Myers, S. E. Hudson, and R. Pausch. 2000. Past, present, and future of user interface software tools. *ACM Transactions on Computer-Human Interaction* 7, 1 (2000), 3–28.

S. Oviatt. 1999. Ten myths of multimodal interaction. *Communications of the ACM* 42, 11 (1999), 74–81.

B. Dumas, D. Lalanne, and S. Oviatt. 2009. Multimodal interfaces: A survey of principles, models and frameworks. In *Human Machine Interaction*. Springer, Berlin, 3–26.

S. Fairclough. 2009. Fundamentals of physiological computing. *Interacting with Computers* 21, 1–2 (2009), 133–145.

P. Petta, C. Pelachaud, and R. Cowie. (Eds.). (2011). Emotion-oriented systems: The humaine handbook. *Cognitive Technologies Series*. Springer.

F. Chen, N. Ruiz, E. Choi, J. Epps, M.A. Khawaja, R. Taib, and Y. Wang. 2012. Multimodal behavior and interaction as indicators of cognitive load. *ACM Transactions on Interactive Intelligent Systems* 2, 4 (2012), 22.

S. Chen and J. Epps. 2014a. Using task-induced pupil diameter and blink rate to infer cognitive load. *Human-Computer Interaction* 29, 4 (2014), 390–413.
K. Huttunen, H. Keranen, E. Vayrynen, R. Paakkonen, and T. Leino. 2011. Effect of cognitive load on speech prosody in aviation: Evidence from military simulator flights. *Applied Ergonomics* 42, 2 (2011), 348–357.
S. Oviatt, B. Schuller, P. Cohen, D. Sonntag, and G. Potamianos. 2017. The handbook of multimodal-multisensor interfaces, 1 *Foundations, User Modeling, and Common Modality Combinations*. Morgan & Claypool, 1–15.
D. Kahneman. 1973. *Attention and Effort*. Prentice-Hall, Englewood Cliffs, N.J.
A. Baddeley. 2003. Working memory: Looking back and looking forward. *Nature Reviews Neuroscience* 4, 10 (2003), 829–839.
C. D. Wickens. 2008. Multiple resources and mental workload. *Human Factors, The Journal of the Human Factors and Ergonomics Society* 50, 3 (2008), 449–455.
F. Paas and J. Van Merriënboer. 1994. Instructional control of cognitive load in the training of complex cognitive tasks. *Educational Psychology Review* 6, 4 (1994), 351–371.
N. Lavie, A. Hirst, J. W. de Fockert, and E. Viding. 2004. Load theory of selective attention and cognitive control. *Journal of Experimental Psychology: General* 133, 3 (2004), 339–354.
T. Appel, C. Scharinger, P. Gerjets, and E. Kasneci. 2018. Cross-subject workload classification using pupil-related measures. In *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications (ETRA'18)*. 4.
L. Fridman, B. Reimer, B. Mehler, and W. T. Freeman. 2018. Cognitive load estimation in the wild. In *Proceedings of the 2018 SIGCHI Conference on Human Factors in Computing Systems (CHI'18)*.
S. G. Hart and L. E. Staveland. 1988. Development of NASA-TLX (task load index): Results of empirical and theoretical research. In *Human Mental Workload*, Hancock, P.S. and Meshkati, N. (Eds.). North-Holland, Amsterdam, 139–183.
G. G. Reid and T. E. Nygren. 1988. The subjective workload assessment technique: A scaling procedure for measuring mental workload. *Advances in Psychology* 52 (1988), 185–218.
H. Steil and A. Bulling. 2015. Discovery of everyday human activities from long-term visual behavior using topic models. In *Proceedings of the 17th ACM International Conference on Ubiquitous Computing (UbiComp'15)*.
J. Epps and S. Chen. 2018. Automatic task analysis: Towards wearable behaviometrics. *IEEE System, Man and Cybernetics Magazine* 4, 4 (2018), 15–20.
K. J. Vicente. 1999. Cognitive work analysis: Toward safe. In *Productive, and Healthy Computer-based Work*. CRC Press, 70.
M. Czerwinski, E. Horvitz, and S. Wilhite. 2004. A diary study of task switching and interruptions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'04)*. 175–182.
S. Chen, J. Epps, N. Ruiz, and F. Chen. 2011. Eye activity as a measure of human mental effort in HCI. In *Proceedings of the 16th International Conference on Intelligent User (IUI'11)*. 315–318.
B. Pfleging, D. K. Fekety, A. Schmidt, and A. L. Kun. 2016. A model relating pupil diameter to mental workload and lighting conditions. In *Proceedings of the 2016 Conference on Human Factors in Computing Systems (CHI'16)*. 5776–5788.
S. Chen, J. Epps, and F. Chen. 2013a. Automatic and continuous user task analysis using eye activity. In *Proceedings of the International Conference on Intelligent User Interfaces (IUI'13)*. 57–66.
S. Chen and J. Epps. 2013b. Automatic classification of eye activity for cognitive load measurement with emotion interference. *Computer Methods and Programs in Biomedicine* 110, 2 (2013), 111–24.
S. Chen and J. Epps. 2014b. Efficient and robust pupil diameter estimation and blink detection from near-field video sequences for human machine interaction. *IEEE Transactions on Cybernetics* 44, 12 (2014b), 2356–67.
T. Foulsham. 2015. Eye movements and their functions in everyday tasks. *Eye* 29, 2 (2015), 196–199.
S. P. Liversedge and J. M. Findlay. 2000. Saccadic eye movement and cognition. *Trends in Cognitive Sciences* 4, 1 (2000) 6–14.
B. Yin and F. Chen. 2007. Towards automatic cognitive load measurement from speech analysis. In *Proceedings of the International Conference on Human-Computer Interaction*. Springer, Berlin, 1011–1020.
T. F. Quatieri, J. R. Williamson, C. J. Smalt, T. Patel, J. Perricone, D. D. Mehta, and J. Palmer. 2015. Vocal biomarkers to discriminate cognitive load in a working memory task. In *Proceedings of the 2015 INTERSPEECH*.
R. M. Makepeace and J. Epps. 2015. Automatic task analysis based on head movement. In *Proceedings of the IEEE Engineering in Medicine and Biology Conference*. 5167–5170.
S. Chen and J. Epps. 2019. Wearable four-dimensional task load recognition based on atomic head movement. *IEEE Journal of Biomedical and Health Informatics* 23, 6 (2019), 2464–2474.
D. E. Irwin and L. E. Thomas. 2010. Eyeblinks and cognition. In *Macquarie Monographs in Cognitive Science. Tutorials in Visual Cognition*, V. Coltheart (Ed.). Psychology Press, 121–141.
S. R. Steinhauer, G. J. Siegle, R. Condray, and M. Pless. 2004. Sympathetic and parasympathetic innervation of pupillary dilation during sustained processing. *International Journal of Psychophysiology* 52, 1 (2004), 77–86.
E. Haapalainen, S. Kim, J. F. Forlizzi, and A. K. Dey. 2010. Psycho-physiological measures for assessing cognitive load. In *Proceedings of the 12th ACM International Conference on Ubiquitous Computing (UbiComp'10)*. 301–310.
M. A. Hogervorst, A-M. Brouwer, and J. B. F. van Erp. 2014. Combining and comparing EEG, peripheral physiology and eye-related measures for the assessment of mental workload. *Frontiers in Neuroscience* 8 (2014), 322.

J. Diedrichsen, R. Shadmehr, and R. B. Ivry. 2010. The coordination of movement: Optimal feedback control and beyond. *Trends in Cognitive Sciences* 14, 1 (2010), 31–39.

A. Nowak, R. R. Vallacher, M. Zochowski, and A. Rychwalska. 2017. Functional synchronization: The emergence of coordinated activity in human systems. *Frontiers in Psychology* 8 (2017), 945.

R. Shadmehr, M. A. Smith, and J. W. Krakauer. 2010. Error correction, sensory prediction, and adaptation in motor control. *Annual Review of Neuroscience* 33, 1 (2010), 89–108.

D. Rigoli, J. Piek, R. Kane, and J. OOsterlaan. 2012. An examination of the relationship between motor coordination and executive functions in adolescents. *Developmental Medicine & Child Neurology* 54, 11 (2012), 1025–1031.

W. Einhauser, F. Schumann, S. Bardins, K. Bartl, G. Boning, E. Schneider, and P. Konig. 2007. Human Eye-head Co-ordination in Natural Exploration. *Network: Computation in Neural Systems* 18, 3 (2007), 267–297.

C. Dromey and A. Benson. 2003. Effects of concurrent motor, linguistic, or cognitive tasks on speech motor performance. *Journal of Speech, Language, and Hearing Research* 46, 5 (2003), 1234–1246.

H. Pashler. 1994. Dual-task interference in simple tasks: Data and theory. *Psychological Bulletin* 116, 2 (1994), 220.

S. Palva and J. M. Palva. 2016. The role of local and large-scale neuronal synchronization in human cognition. In *Multimodal Oscillation-based Connectivity Theory*. Springer, 51–67.

P. Zarjam, J. Epps, F. Chen, and N. H. Lovell. 2013. Estimating cognitive workload using wavelet entropy-based features during an arithmetic task. *Computers in Biology and Medicine* 43, 12 (2013), 2186–2195.

B. P. Bailey and S. T. Iqbal. 2008. Understanding changes in mental workload during execution of goal-directed tasks and its application for interruption management. *ACM Transactions on Computer-Human Interaction* 14, 4 (2008), 1–28.

D. D. Salvucci and J. H. Goldberg. 2000. Identifying fixations and saccades in eye-tracking protocols. In *Proceedings of the 2000 Symposium on Eye Tracking Research & Applications*. 71–78.

F. Eyben, F. Weninger, M. Wollmer, B. Schuller. 2016. Open-source media interpretation by large feature-space extraction. Retrieved from http://opensmile.audeering.com/.

J. Pelz, M. Hayhoe, R. Loeber. 2001. The coordination of eye, head, and hand movements in a natural task. *Experimental Brain Research* 139, 3 (2001), 266–277.

D. Reidsma, A. Nijholt, W. Tschacher, and F. Ramseyer. 2010. Measuring multimodal synchrony for human-computer interaction. In *Proceedings of the IEEE 2010 International Conference on Cyberworlds*. 67–71.

L. M. Reeves, J. Lai, J. A. Larson, S. Oviatt, T. S. Balaji, S. Buisine, P. Collings, P. Cohen, B. Kraal, J.-C. Martin, M. McTear, T. V. Raman, K. M. Stanney, H. Su, and Q. Y. Wang. 2004. Guidelines for multimodal user interface design. *Communications of the ACM* 47, 1 (2004), 57–59.

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

Received December 2019; revised July 2020; accepted July 2020