Task Load Estimation from Multimodal Head-Worn Sensors Using Event Sequence Features

Siyuan Chen, *Member, IEEE* and Julien Epps, *Member, IEEE*

Abstract—For longitudinal behavior analysis, task type is an inevitable and important variable. In this article, we propose an event-based behavior modeling approach and employ non-invasive wearable sensing modalities (eye activity, speech and head movement) to recognize task load level under four different task load types. The novelty lies in converting physiological and behavioral signals into meaningful events and utilizing their sequence across multiple modalities to distinguish load levels and types. We evaluated this approach on head-worn sensor data from 24 participants completing four different tasks for recognizing (i) low and high load level for a given task load type, (ii) low and high load level regardless of load type, and (iii) both load level and load type. Findings show that the recognition rate is reasonable in (i), close to chance level in (ii), and well above chance level in (iii) for 8 classes using participant-dependent and -independent schemes. Further, a fusion of the proposed event-based features and conventional continuous features achieved the best or similar performance in most cases. These results suggest that task type needs to be considered when using continuous features and that the proposed event-based modeling paradigm is promising for longitudinal behavior analysis.

Index Terms—Eye activity, speech, head movement, physiological sensing, task load, bag-of-words, topic models

1 Introduction

Affect is an essential experience of feeling, emotion, moods, attitude or other responses to external or internal stimulus events [1], [2], [32]. Perception of task difficulty is one such experience which pervades our daily life and forms an important component in affective response [3]. Modelling continuously variable affect opens the door to longitudinal behavior analysis, which requires computational models and systems to shift from a focus on instantaneous categorization to recognizing affect over time, e.g., in continuous valence and arousal spaces [1], [4].

Sensing of facial expression [4], speech [5] and EEG [4], for example, has been leveraged in research to date, since these physiological or behavioral signals can reflect altered affective states. Recent wearable system development [6], [29] has further facilitated continuous computing techniques, for example using wearable eye activity for cognitive load estimation [7], task analysis [3], [8], and human activity recognition [9], using a wearable Inertial Measurement Unit (IMU) to record head and body movement for human activity recognition [10] and task load estimation [11], [12], and using a microphone to record speech for emotion recognition [5], [13] and cognitive load estimation [14].

Although affective computing has made significant progress over the years, there are still challenges for continuous recognition and longitudinal analysis. First, most studies recognize affect states, whether in basic emotion categories or continuous dimensions, that are elicited in a single context from a task perspective. That is, affect is mainly induced and changed during the same task, e.g., watching images [15] and videos [4], speaking [5] or a single task [31]. However, for longitudinal and continuous analysis, expressions of affect can not only evolve and change over time but also over different task types. Concerning the experience of task load, few studies have recognized the load levels across different tasks systematically [3], [8] and interpreted the source of the load produced. However, for behavior analysis during human-human (e.g., remotely), human-computer or human-robot interaction, knowing the users' continuous workload changes and what kind of task is inducing their load, as shown in Fig. 1, can surely help improve interaction quality [34].

Second, since continuous computing requires continuous physiological or behavioral signal sensing, it may pose an inference problem in affect recognition. That is, it is uncertain whether it is affect rather than e.g., cognitive or functional requirements which significantly alters each instant of an affective signal [15]. One example is that body movements and even facial muscle activities were found to significantly contaminate EEG features in affect recognition [4]. Another related question is whether each instant of these physiological and behavioral signals represents affect rather than ordinary functional regulation. For example, as well as in response to arousing images or mental load, the pupil also dilates and contracts to light changes as a result of antagonistic mechanisms in the autonomic system [16]. Differently to instantaneous affect recognition, where only the strongest responses to a single stimulus are typically considered, continuous affect recognition employs

• *S. Chen and J. Epps are with the School of Electrical Engineering and Telecommunications, University of New South Wales, Sydney, NSW 2033, Australia. E-mail: {siyuan.chen, j.epps}@unsw.edu.au.*

Manuscript received 15 Feb. 2019; revised 17 Aug. 2019; accepted 9 Nov. 2019. Date of publication 26 Nov. 2019; date of current version 3 Sept. 2021. (Corresponding author: Siyuan Chen.)
Recommended for acceptance by T.-A. Percep.
Digital Object Identifier no. 10.1109/TAFFC.2019.2956135

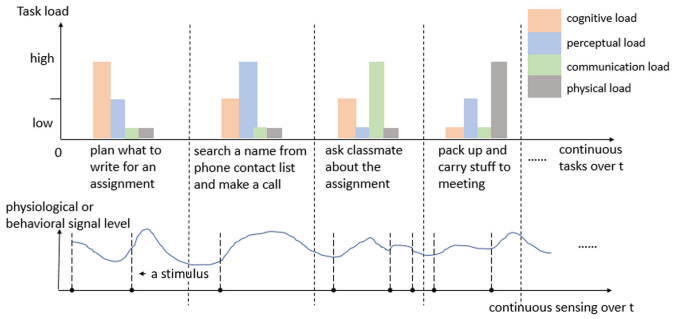

Fig. 1. An illustrative example showing different task types in continuous task load estimation, where each task is represented by four dimensions of task load using the framework in [6] and the corresponding physiological or behavioral signal changes in response to affect stimuli (dashed vertical lines with dark dots) of unknown timing and origin. In longitudinal tasks, there may be significant time periods between stimuli, and signal changes will occur that are unrelated to affect stimuli, so that it is questionable whether conventional signal-based analysis during non-stimulus time periods reflects meaningful affect states.

signals over time which include responses to discrete stimulus events as depicted as the dark dots in the bottom panel of Fig. 1. But during non-stimulus times, the reasons for signal changes may be difficult to identify.

The aim of this work is to investigate the recognition of task load level and type of task load in multiple task contexts, and the advantage of employing event-based modeling (only change events are utilized) over signal-based modeling (signal value at every instant is utilized for statistical features) in these task types. 'Continuous task load recognition' in this study refers to the estimation of task load levels during a series of different tasks performed on a continuous basis (task load type keeps changing) over a long time. In this study, low and high task load experiences were induced from cognitive, perceptive, light physical and communication activities. We used eye activity, speech, and head movement to continuously recognize the task load level and task load type using a novel approach based on multimodal topic modelling, which is focused on changes in behavior and proposed as more suitable for longitudinal affect recognition.

The contributions of this study include: (i) introducing an effective characterization of behavior using multimodal behavior event *n*-grams which is more intuitive, compact and meaningful than using the conventional signal-based paradigm, and on a per-task basis can achieve close performance to that using every moment of per-task data to make inferences in a continuous manner; (ii) employing eye, speech, and head activities which offer a non-invasive and relatively comprehensive wearable sensor combination, and are less constrained by environments; and (iii) automatically recognizing not only the perception of easy or difficult but also the type of task load.

2 Background

2.1 Affect Production and Task Type

Affect is a complex construct. Some psychologists have suggested that affective responses are pre-cognitive, that is, they occur before cognitive judgements without significant cognitive involvement [17], while others consider it to be post-cognitive, i.e., affect states such as liking or disliking, pleasure or displeasure are elicited only after a certain amount of cognition has been accomplished and are based on prior cognitive processes [18]. A more recent view combines both, arguing that initial affective reactions produce thoughts, then the thoughts produce affect [26].

This recent development in theoretical perspectives certainly affects the manner in which affect data should be collected. In the scenarios of task performance, the affect produced may be more related to mental state, which involves significant cognitive processing, e.g., the level of workload, engagement, fatigue, stress and pleasure as a result of tasks [20], [30]. Compared with the means of inducing emotion using images or videos [2], [4], [20], where experiences may be primary with little cognitive involvement, the experiences due to tasks are more covert and natural reactions as part of daily life. In such cases, annotation of instantaneous task load cannot be performed in the same way that emotions from facial expression or speech are often annotated, and task load must be self-rated.

Recognizing and interpreting workload is an established research area [6], often motivated by improving task performance, safety and wellness. For example, in [23], eye image sequences were used to recognize three levels of cognitive load induced by n-back tasks during a driving scenario using a CNN architecture. In [30], facial actions were analyzed to identify driving without and with cognitive load, which was induced by arithmetic tasks to avoid driver distractions. In [34], the effect of feedback of concurrent workload was investigated in human computer interaction in order to improve task performance. Meanwhile, research efforts in this area also focused on continuously recognizing task load and applying this to longitudinal contexts. For example, Chen et al. [3], [8] detected task transition breakpoints first to segment sensed signals, and then recognized the level of load within each segment, as opposed to the convention of employing fixed-length time windows, e.g., [4], which did not account for task transitions. Although it is possible to automatically segment tasks, there are few studies [3], [11] that have attempted to identify activity type during task windows or fixed time windows and determine the task type for task load estimation. It is of importance since task types keep changing in our daily life unlike the limited specific tasks typically used in research laboratories.

Recently, the authors [6] proposed a four-dimensional task load framework, where a task can be represented at a particular instant in terms of its perceptual, cognitive, physical, and conversational load levels, based on the Berliner task taxonomy, as illustrated in the top panel of Fig. 1. This framework allows task load to be recognized in a wider context of continuous tasks, regardless of task difference, and can represent any unseen task rather than being constrained to a fixed vocabulary of predefined tasks.

2.2 Topic Modeling for Behavior Analysis

Topic models have been widely used to discover human action categories in computer vision research, and recently they have been used on processing signals from wearable sensors to model human behavior. For example, in [21], signals acquired from heterogeneous smartphone sensors were processed to model atomic actions (like the tilt of a smartphone) as a "word" and a higher-level behavior as "phrases" in order to

verify user identity through their behavior patterns. In [9], a full range of eye movement (saccade and fixation) and blinks were encoded to discover ten human activities from long term visual behavior using latent Dirichlet allocation topic model (LDA). In [13], MFCCs and energy features extracted from speech were encoded using bag-of-audio-words for emotion recognition using support vector regression (SVR) and achieved better performance than deep learning approaches.

Regarding topic models in these studies, the common aspect in their processing is to encode low-level descriptors into a symbolic sequence, from which an *n*-gram bag-of-words representation is generated. This representation is then used to learn a model to infer the similarity to claimed users (identity recognition) [21], the activity categories (activity recognition) [9] or the arousal and valence values (emotion recognition) [13]. The most common encoding paradigm is to use *k*-means to cluster low-level descriptors into *k* clusters [9], [13], [21], so that a series of low-level descriptors can be represented instead by a symbolic sequence of cluster indices. There are also studies [9], [21] that have used a multi-level approach where the signal space was discretized with *k* granularity levels and then signals were encoded across these levels. There are some variants, for example, saccadic eye movements were encoded according to their direction and amplitude, fixation duration was encoded according to histogram bins while blink numbers were directly encoded into a string sequence to best characterize heterogeneous signals [9].

Although the bag-of-words representation may be a powerful computational linguistics method, one shortcoming in the encoding paradigm is that the parameter *k* is a data-driven parameter which may need to be adapted for different data. Even with the right choice of the parameter *k*, the 'word' produced is hardly interpretable. One recent study [11] proposed an approach to convert gyroscope signals into six meaningful atomic head movement events, and found they were more accurate and intuitive than statistical features aggregated across many time instants for sedentary activity analysis. As physiological and behavioral signals contain relevant information about behavior by nature, bag-of-words representations from multimodal signals should be meaningful and interpretable for representing subtle behavior changes as a result of affect change.

An interesting motivation for bag-of-words representations in behavior analysis is that some physiological and behavioral signals are event-like in nature, by contrast with a signal-based approach. For example, blink, saccade, fixation, and speech onset are all events whose sequence may be direct responses to internal or external stimuli, as opposed to continuous features extracted at every instant such as pupil diameter, speech energy, head trajectory, which can be meaningfully extracted whether there are affective/task stimuli or not. To date, the majority of affective signal analysis have used continuous features rather than event type signals.

2.3 Non-Invasive Wearable Sensing

Speech is a commonly studied modality in affect recognition e.g., [5], [13], which can be recorded with a wearable microphone. Some forms of eye activity, especially pupil diameter change, have been used for workload assessment studies e.g., [3], [7], [8], [23], [33], [31], which can be recorded with head-worn cameras. Only a couple of studies employed head movement using IMU in affective computing [12]. Other non-invasive werable sensing modalities such as EEG, EMG, EDA [2], [22], [33], require reliable contact with the skin and/or are more sensitive to body movements and are hence not included in our study. To recognize affect using these modalities, most studies tried to find effective statistical features through feature selection or learning procedure to achieve reasonable classification performance [2], [3], [4], [23]. However, to understand how modalities contribute to affect recognition, feature selection has often been applied to extracted signal-based features.

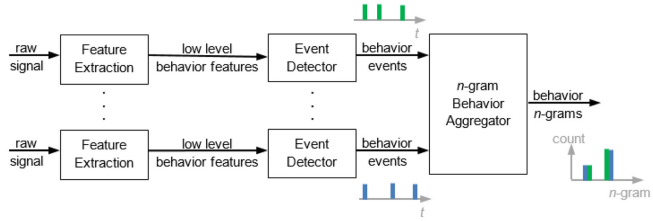

Fig. 2. Proposed event-based behavior modeling and feature extraction from multiple sensor signals using *n*-grams. In this example, *blue-green* and *green-blue* multimodal bigrams (words) are detected and then aggregated over a period of time to produce bigram counts, which can then be used as a (bag-of-words) feature vector for subsequent analysis/recognition. *n*-grams with high counts represent frequent sequences of behavior events. The multimodal *n*-gram based on event features shown here is relatively new in the affective computing literature, where continuous features are common; however, *n*-grams and associated methods are very well established in computational linguistics, for example.

Up to now, a few studies have attempted to convert low-level continuous signals into bag-of-words using a single modality, as mentioned in Section 2.2. However, there is no study which converted physiological or behavioral signals into behavioral events and integrated these with naturally occurring behavioral events like blink, fixation and saccade onset. More importantly, there is no study which utilizes the sequence of these behavioral events from different sensing modalities for continuous computing. Key questions are whether we can find the task load levels and types by analyzing the proposed words, and whether topic modelling is a better method for behavior analysis than using conventional continuous feature representations.

3 Proposed Event-Based Behavior Modeling

3.1 Overview

In the proposed event-based behavior modeling approach, each task is represented by a sequence of behavioral events, by analogy with a document represented by a bag-of-words using a word-document matrix, as depicted in Fig. 2. *N*-gram words are used to form terms which represent physiological and behavioral event sequences that occur during tasks. Each term is a dimension in the feature space used to distinguish load differences.

3.2 Feature Extraction and Event Detection

The purpose of feature extraction and event detection is to obtain behavior events within a single physiological or behavioral signal. There are usually two cases. One is that naturally occurring behavior events can be detected from raw signals

TABLE 1
Behavior Topics in Our Study

Level\topics	Cognitive load	Perceptual load	Physical load	Communication load
Low – easy to complete	Demand imposed on working memory to achieve task goals	Demand imposed on perceptual sense to achieve task goals	Demand imposed on the body to keep balance or achieve task goals	Demand imposed on phonological loop to achieve task goals
High – difficult to complete				

using customized algorithms. For example, blink can be detected from eye videos directly, fixation onset and saccade onset can be separated from pupil center trajectory, and speech onset can be located from speech waveforms – these all occur naturally as events.

The other case is that some physiological or behavioral signals are numerical and continuous nature, for example, pupil diameter and head location, and do not naturally generate event onsets and offsets. For these signals, discrete events can be generated by event detectors (see Fig. 2) triggered by significant changes of some kind. Considering head movement, differently to previous studies that used *k*-means to quantize signals, we used an atomic head movement analysis method [11] to convert continuous numerical signals to discrete events comprising increase, decrease, and central movement. Briefly, continuous head gyroscope signals were thresholded to obtain head 'central' movement (nonobvious movement or still) with a threshold of $3^\circ/\mathrm{s}$, based on experimental head stabilization studies. Then the first derivative was taken to obtain increase and decrease segments from their positive or negative values. Events from three axes were merged to obtain the final atomic events, and in [11] their frequency counts and intensity were used as features for classification. Although this method was proposed for head movement, in general, events can be similarly generated from other modalities, as long as the threshold for 'central' movement is suitably redefined to make it meaningful.

The physiological rationale behind the increase, decrease, and 'central' movement events is that they indicate a balance change in the antagonist sympathetic and parasympathetic systems [16], [22]. 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 the state of decrease indicate no sustained effort and/or the other autonomic system taking over to regulate the function. Therefore, increase, decrease, and 'central' events make more sense in terms of behavioral events than quantizing signals into an uncertain number of clusters using *k*-means that cannot easily be interpreted.

When looking into the difference between the proposed event features and conventional features, it can be observed that conventional signal-based features are extracted at every instant and focused on statistical meaning, while event-based features are focused on change and denote the 'edge' of some kind of behavior. The latter is of particular interest in longitudinal studies, where behavior may not change very much for long periods of time, but where it is critical to note any changes that do occur.

3.3 N-gram Bag-of-Words Aggregation

To use behavioral event sequences representing tasks instead of using every signal instant, we selected naturally occurring behavior onset events, and onset of increase events from continuous signals as the minimal subset words. This choice is because there is always an event offset corresponding to the onset of the same event, which is assumed to be less relevant to indicate mental effort and hence not investigated in this study, except when the offset of a decrease event shares the same instant of the onset of an increase event (decrease then increase). Further, increase events are an indicator of exerting efforts to cope with stimuli, which may indicate behavior changes. These words were collected according to the sequence of their occurrence from all signals. If two words occurred at the same time, they were concatenated to form a new word.

It is worth noting that the proposed bag-of-words model representation for behavior analysis is a little different to that used in information retrieval and natural language processing where some words like 'the' and 'and' are not important when representing text while in the proposed method, each word is meaningful and matters when representing behavior change, and we did not remove any word. As shown in Fig. 2, n-grams were then used to further chunk the words into terms. We hypothesize that the sequences of behavior events are different when task load and type are different. For example, it is possible that 'sac' is followed by 'head' more often than any other words for some task types. When the perceived task load is high, frequent sequences are less preserved, and some rare sequences appear, and the count of the frequent sequences will change.

By treating n-grams as words w, and tasks as documents d, we used a term document matrix of size $d \times w$ to represent a document topic. A document topic is regarded as the load levels and/or the load types of the task. In our study, as shown in Table 1, we set low and high task load levels as two topics, i.e., investigating the two-class problem of recognizing low vs. high task load, irrespective of the task type. When considering the task types in terms of cognitive, perceptual, physical and communication load as suggested in Section 2.1, we can extend the number of topics to be eight with their load levels (high and low) in each of the four load types. In terms of the feature dimensionality, the larger n is, the higher the feature dimension. To find effective features, we can select features based on their weights in terms of maximizing classification accuracy.

Above all, the notion behind using topic modelling for behavior analysis is that the sequence of behavior events can represent behavior changes during tasks.

4 Experiments

4.1 Task Load Stimuli

To investigate task load level and type recognition, we designed an experiment with the goal of collecting the

responses from eye activity, speech, and head movement during easy and difficult tasks and during four task types that are representative of diverse real-life tasks [6] (UNSW Human Research Ethics Advisory reference number 08/2014/23).

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 given targets from among pictures full of distractors, (iii) forearm lifting of two dumbbells with different weights, and (iv) holding conversations with the experimenter to complete simple conversation or an object guessing game.

In each task, two difficulty levels were created to induce low and high task load 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 duration of cognitive and perceptual tasks varied between participants, but each was no more than one minute. The physical and communication tasks were up to around one minute.

Each level of the four types of tasks was completed by participants first, followed by subjective rating (on a 7-point scale) of its difficulty at the end of each task to check the validity of the induced level. Then these tasks were continuously presented in a counterbalanced order and completed by participants without breaks. During the experiment, participants were seated at a desk, free to move any part of their body. The experiment lasted one hour.

4.2 Data Collection

Twenty-four participants (14 males, 10 females, aged 18-25) volunteered. Eye activity videos were filmed at 30 fps by a 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, as shown in Fig. 3a. Head movement was recorded using an IMU attached to the head by a head strap and connected to the laptop with a USB cable. The IMU prototype consisted of an inertial measurement unit (MPU 9150) and output 3-axis acceleration, angular velocity, and magnetic field strength at a rate of around 20 Hz.

In total, data from 44 tasks were collected from each of the 24 participants. Half were in low load and the other half were in high load. For the same load level, similar tasks were repeated five times except the cognitive tasks, which had seven repetitions. The timestamps of each task were automatically recorded to segment tasks. The mean (min, max) task duration in second for each load type and load level was 20.4 (14.2, 31.7) and 51.2 (23.0, 108.0) for low and high cognitive load; 29.1 (16.8, 47.6) and 125.6 (74.6, 147.4) for low and high perceptual load; 20.4 (18.8, 22.8) and 28.7 (25.0, 44.2) for low and high physical load; 31.7 (30.2, 34.0) and 61.4 (54.2, 65.2) for low and high communicative load respectively. The induced load level was treated as the ground truth, and this assumption was verified using participants' subjective self-ratings of the task difficulties.

In order to synchronize all signals for later processing, participants were asked to clap their hands and nod their head simultaneously at the beginning of the experiment. The sound in the eye video, large oscillations in the head IMU data, and clapping images in the scene video were utilized to identify the synchronization timestamps at the beginning of each recording. Using the automatically recorded timestamps of each task, signals within each task duration were automatically segmented. We also manually checked the segmented scene videos and eye videos to ensure correctness.

4.3 Event Detection and Bag-of-Words

We pre-processed near-field eye videos to obtain blink events, pupil center (relative to the head) and pupil size signals over time, using the self-tuning and dual ellipse fitting algorithms, with pupil size measurement accuracy of around 0.02 mm [24]. Saccade events were defined as eye movements between two fixations while fixations were extracted from the pupil center relative to the head employing dispersion-based algorithms using 1° of visual angle for at least 200 ms [19]. As visual perception and cognitive processing occur during fixation, saccade onset events indicate the end of cognition. Pupil size events over time were found by the atomic head movement segmentation algorithm in [11] where 'central' movement was set to be the pupil size baseline of the task (0.5 s) [7]. We used IMU readings of the head angular velocity of roll, pitch and yaw to obtain head movement events over time also using the atomic head movement segmentation algorithm in [11]. We processed speech recordings to obtain speech onset events over time determined by voicing probability > 0.7 using openSMILE [25] during each task. We selected the blink onset, saccade onset, speech onset, pupil size increase, and head angular velocity increase as the words. Fig. 3c shows the framework from which multimodal behavior event sequence bags-of-words were accumulated for task load analysis.

4.4 Data Processing and Analysis

4.4.1 Overview

Overall, from the three different sources of raw signals – IR eye images, head IMU data, and speech waveforms, we have extracted low level features – blink, pupil size, saccade from eye images, head velocity in three dimensions from IMU data, and voicing probability from speech. At this point, only blink and saccade can be directly taken as events. We have detected pupil size increase events from the continuous pupil size feature, head velocity increase events from the continuous 3-dimensional head velocity features, and speech events from the continuous voicing probability features as described in Section 4.3. After this event detection, we took the onset of each of these five events as a word and aggregated them according to the timeline as bag of words to represent a task. Neighborhood component analysis (NCA) was then employed to determine which terms best represent the task. An SVM classifier was finally employed to classify task load type and load level. More details are explained in Fig. 3d and the following description.

4.4.2 N-Gram Vector Space Model for Distinguishing Task Topics

Within the training data, all unique n-gram words were first located as unique terms. Since the task topics were overt in our analysis, rather than underlying topics to be discovered (as in some other research problems [9]), we used term

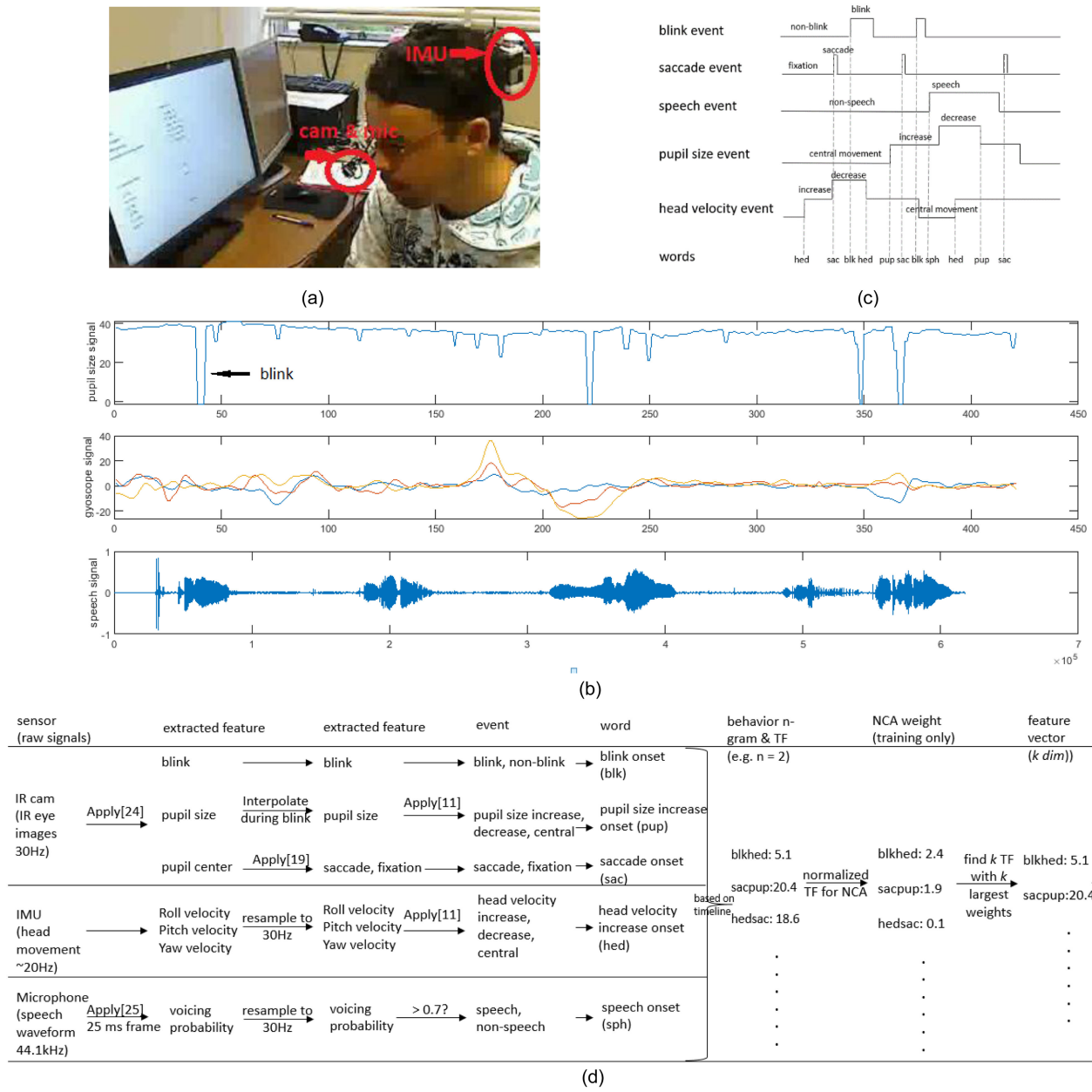

Fig. 3. (a) A participant wearing a webcam and IMU device, which recorded eye activity, speech, and head movement. Wearable device prototypes of this kind are now commercially available, as discussed in [6]. (b) Examples of extracted pupil size and blink signal (top) and recorded head gyroscope signals from three axes (middle) which were fed to the event detector in Fig. 2 and speech signals (bottom) from which voicing probability was extracted before being input to the event detector. (c) Proposed bag-of-words model for behavior modeling where the onset of blink, saccade, and speech events and the events for pupil size increase and head velocity increase were used as words, from which *n*-gram words were developed to represent behavior event sequences as terms. Note 'hed,' 'sac,' 'blk,' 'pup,' and 'sph' denote 'head,' 'saccade,' 'blink,' 'pupil,' and 'speech,' respectively. (d) shows the detailed data processing steps.

frequency (TF) vectors on a per-task basis as event-based features and their task annotations to train the task load type and level models. During testing, a new TF vector was decoded according to the unique *n*-gram words, then the TF vector was input to the trained classifier for the required task load level and type recognition.

Since the event-based feature dimension can be very high when n is large depending on the number of terms in training data, NCA was employed to select features (i.e., terms). NCA is a non-parametric embedding method where each feature's weight is learned while minimizing an objective function that measures the classification loss over the training data [27]. During training, we selected k features based on the top-scoring k feature weights learned from NCA, which can also build understanding of the contribution of specific n-gram words to task topic classification. Meanwhile, n is another variable in the bag of n-gram model to investigate for its impact on task topic recognition performance.

In this supervised learning, two classification schemes were employed. The participant-independent scheme was trained on 23 participants' data under each task topic and tested on the held-out participant's data, while the subject-dependent scheme was trained on the first half of tasks under each task topic for one participant only and tested on the remaining tasks within the same participant. The classifier used was SVM with RBF kernel, which is known for good

TABLE 2
157 Continuous Features Derived from Three Different Modalities

Modality	Activity	Features extracted during a task
eye	pupil	pupil_size_mean, pupil_size_std, pupil_orientation_mean, pupil_orientation_std, pupil_center_x_mean, pupil_center_x_std, pupil_center_y_mean, pupil_center_y_std
	blink	blink_rate, blink_duration_mean, blink_duration_std, blink_interval_mean,blink_interval_std
	eye movement	fixation_duration_mean, fixation_duration_std, saccade_rate, saccade_amplitude_mean, saccade_amplitude_std
head	acceleration	acceleration_x_mean, acceleration_x_std, acceleration_y_mean, acceleration_y_std, acceleration_z_mean, acceleration_z_std
	angular velocity	velocity_x_mean, velocity_x_std, velocity_y_mean, velocity_y_std, velocity_z_mean, velocity_z_std
	magnetism	magnetism_x_mean, magnetism_x_std, magnetism_y_mean, magnetism_y_std, magnetism_z_mean, magnetism_z_std
speech	prosody	speech rate, F0_mean, speech_intensity_mean, voice_probability
	MFCC (25 ms frames)	MFCC0_mean, MFCC1_mean, …, MFCC12_mean, MFCC0_std, MFCC1_std, …, MFCC12_std, MFCC0_delta_mean, …, MFCC12_delta_mean, MFCC0_delta_std, MFCC12_delta_std, MFCC0_ acceleration_mean, …, MFCC12_ acceleration_mean, MFCC0_ acceleration_std, …, MFCC12_ acceleration_std
	PLP (25 ms frames)	PLP0_mean, PLP1_mean, …, PLP5_mean, PLP0_std, PLP1_std, …, PLP5_std, PLP0_delta_mean, …, PLP5_delta_mean, PLP0_delta_std, PLP1_delta_std, …, PLP5_delta_std, PLP0_acceleration_mean, …, PLP5_ acceleration _mean, PLP0_acceleration_std, PLP1_ acceleration _std, …, PLP5_ acceleration _std

These were aggregated over time by calculating the mean and standard deviation or rate.

accuracy when working with high dimensional data. The two parameters, complexity and kernel scale, were tuned using Bayesian optimization during training.

Meanwhile, we employed a long short-term memory networks (LSTM) with an input layer, 1-layer LSTM (the hidden node size was tuned from among $\{3, 8, 13, 18, 23\}$, $\{5, 10, 15, 20\}$, and $\{9, 14, 19, 23\}$ for $2-, 4-$, and 8-class respectively), a fully connected layer, a softmax layer, and a classification output layer. The epoch number was 1000, batch size was the training sample size/3, optimizer was 'Adam', gradient clipping was 0.9, learning rate was 0.03. For a better performance for the 4- and 8-class cases in the participant-independent scheme, we applied a dropout rate of 0.5 to the LSTM layer and set the hidden node size to be 14 and 18 for the 4- and 8-calss respectively. In each case, we input a sequence of 25-d features (top 25 NCA weights) extracted from a series of 2-sec windows during tasks to train a sequence based deep model. The performance was also used as a baseline.

4.4.3 Continuous Features for Distinguishing Task Topics

In contrast with the event-based features, continuous features describe the detailed status of behavior signals (e.g., pupil dilation, head movement, speech intensity and rate) at every moment or in a short interval. Whether these details are meaningful to recognize task load is questionable, given that it is vulnerable to measurement and background noise since data from every instant is contributed.

To compare the impact of event-based features and continuous features on the performance of recognizing task load type and load level, we collected 157 statistical features including those that have often been reported as effective for different task classification from among eye activity [3], [6], [7], [8], [9], [15], [31], head movement [11], [12] and speech [13], [14]. These features are listed in Table 2 and were aggregated over time by calculating the mean and standard deviation or rate to form signal-based feature vectors. Except for the differences in feature vectors, other procedures including NCA and classification for continuous features were the same as those for event-based features.

4.4.4 Investigation of Task Load Level and Type Recognition

In this study, we were first interested in the performance of recognizing low and high task load levels in each of the four types of task load, i.e., four separate two-class classification problems, one per task type. This was to obtain a sense of the recognition performance in a single type, as most previous studies in affect recognition have done. We expected the accuracy to be above chance level since training and testing was conducted within the same task type.

For continuous task load recognition, task types are inevitably changing over time. If we are interested primarily in detecting high load, irrespective of the type of load, the classification accuracy could be poorer than that within an individual task type, since the different task types will introduce unwanted variability. By examining this, we can find out how important it is to identify task type.

Finally, we recognized low and high task load levels (2-class), the load types (4-class), and load level and load type simultaneously, i.e., four task load types in low and high load level respectively (8-class). We expected accuracy well above chance level performance since models also learned the type of load, which can help load level recognition. Apart from the continuous features and proposed event-based features, we combined them to see whether they encode complementary information in all cases.

5 Results and Discussion

5.1 Low and High Induced Load Level Verification

According to the comparisons of subjective ratings shown in Table 3, the high load level tasks all imposed higher loads on participants than the low load level tasks, as designed. Nonparametric Wilcoxon paired sign test confirmed that all pairs of load levels perceived by participants were significantly different for each load type, at a 0.01 significance level

TABLE 3
Mean ± 95% Confidence Intervals of Subjective Ratings From 24 Participants Using a 7-Point Rating Scale

Load type\level	Low	high
Cognitive load	2.54 ± 0.50	4.79 ± 0.60
Perceptual load	2.50 ± 0.44	6.70 ± 0.30
Physical load	1.50 ± 0.29	2.83 ± 0.62
Communication load	1.54 ± 0.41	5.54 ± 0.59

($Z = -4.3, -4.3, -3.5, -4.2$ respectively, $p < 0.001$). Different load levels were expected to change participants' physiological and behavioral signals.

5.2 Bag-of-Words for Task Load

Fig. 4 is an example of 1- and 2-gram TF counts across 24 participants for the load type and load level of each task topic. Due to the limited space for the x axis, only 40 terms are displayed and ordered according to NCA weight from highest to lowest. From this figure, we can see that for each type of task load, the pattern of event or event sequence is consistent in both low and high load levels, suggesting that similar behavior patterns occurred within the same type of tasks. However, some term frequencies vary between the two load levels, indicating that high task load affects our eye, head and speech activities. For example, during cognitive, perceptual and communication tasks, people tend to move their eyes (sac-sac term) more often when these tasks are difficult than when they are easy, while for physical tasks, when physical demands become greater, people tend to look at one location with less eye movement. This also indicates that due to different tasks, event sequences may not be consistent across all types of task load, so identifying task load type may help task load level recognition.

Among different types of task load, the patterns of event or event sequences are distinctive between each other, and it is evident that different types of task load require different eye, head and speech activity to coordinate. For example, the head moved (hed-sac term) more often in physical tasks than in the other three. Overall, the bigram TF in Fig. 4 demonstrates that event or event sequences can distinguish task load levels (two task topics) and identify both the load type and load levels (eight task topics).

Fig. 5 presents the word clouds from low and high task load levels regardless of load types using 2-gram words. As expected, the number of words present in the high task load is more than that in low task load, and this trend is consistent across 24 participants. This is probably because high load demands produce some efficiency or inefficiency in

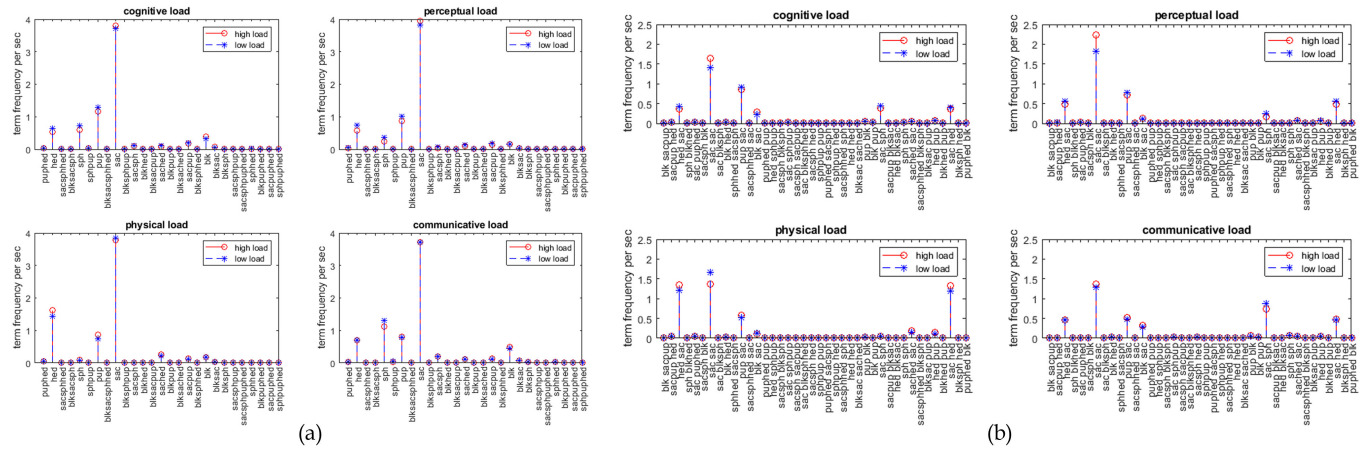

Fig. 4. Average (a) 1-gram and (b) bigram term-frequency normalized by task duration for different task topics across all tasks from 24 participants. The TF varies within and between different task load types; however, the difference between load types seems to be stronger than the differences within load levels. The terms in the x axis were ordered by the NCA weight, which was obtained by training two load levels, regardless of load type while leaving one subject out. Here 'sac' is a short for saccade, 'pup' for pupil increase, 'blk' for blink, 'sph' for speech onset, and 'hed' for head velocity increase.

Fig. 5. All bigram words pooled from 24 participants in (a) low task load (266 terms) and (b) in high task load (303 terms), showing that participants tend to have more unexpected behavior event sequences (more terms) and lower frequency count (smaller size) for some terms when the task load is high. The purpose is to show that the number and size of terms in low and high load levels is different as relevant information.

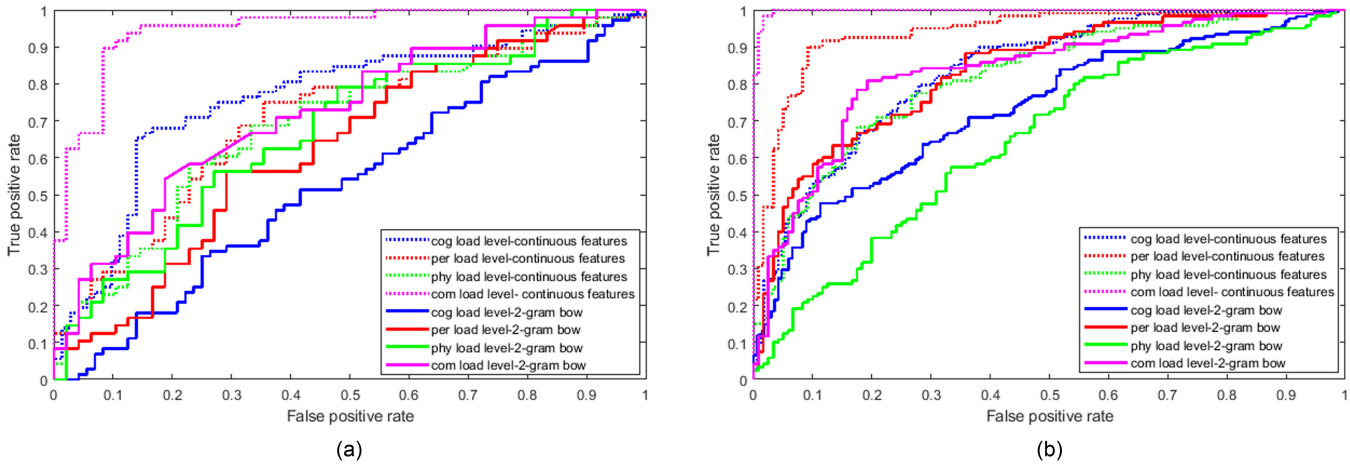

Fig. 6. ROC curves for low and high load level recognition in each type of task load, where bigram features whose NCA weights were in the top 25 were used in (a) participant-dependent and (b) participant-independent classification scheme.

performing mental and physical activity. As a result, a greater diversity of behavior patterns is introduced. The mean and 95 percent confidence intervals of the number of terms for low and high load levels is 115.8 ± 5.6 and 143.5 ± 8.3 respectively. A paired sample t-test $(t(23) = 11.48, p < 0.001)$ confirmed a significant difference between the two levels. This inevitably results in different TF for certain terms in low and high task load.

5.3 Task Load Recognition Within a Single Type

In most previous affect studies, the affective state was recognized in a single task type. Fig. 6 shows the Receiver Operating Characteristic (ROC) curve for recognizing low and high load levels within each type of task load when the bigram feature dimension was fixed at 25, that is, 2-gram features whose NCA weights were in the top 25 were used. To show the efficacy of the proposed event features, we compared the classification results with those using continuous features in both participant-dependent and -independent schemes where features whose NCA weight also ranked in the top 25 were used.

From this figure, we can see that in general classification results vary in different task types, with high performance using both types of features in communication load in both classification schemes. From Fig. 6a participant-dependent and Fig. 6b participant-dependent classification, employing continuous features achieved better performance than that using event features in low and high load level recognition in all load types. The average classification accuracies (95 percent confidence intervals) in all cases are detailed in Table 4. Wilcoxon signed rank tests confirm that the average accuracies using continuous features are significantly better $(p < 0.05)$ than those using event-based features in both participant-dependent and -independent schemes. Meanwhile they are both similar or slightly better than using LSTM sequence model, indicating the information of sequence works for task load level recognition.

The tentative reason for the proposed event features achieving worse performance than continuous features within a single task type is that the event features contain limited temporal information (only onsets) and the continuous features include subtle, low level measurements. For example, the only speech-based event feature is speech onset, while the continuous speech features include multiple prosody, MFCC and PLP measurements. These additional fine-grained measurements may contribute to the efficacy of the continuous speech features, especially for recognizing high vs. low task load of communication tasks. Thus, adding additional event features based on changes in these fine-grained measurements may lead to improved performance of the model using event features.

5.4 Task Load Recognition Regardless of Context

In order to recognize affect state continuously, we can ignore task type and let the models learn the feature patterns caused purely by low and high levels of task load. Fig. 7 demonstrates the classification accuracy for recognizing low and high task load levels regardless of the task load type, for different n-grams, at different dimensions. We also included the results from continuous features for comparison in both participant-dependent and -independent schemes. From the figure we can find that with a certain number of top-weighted features (whose weights were assigned during NCA in training), including more event features generally does not help improve classification performance much when the dimension is high enough. At dimension 25, as shown in Table 4, continuous features performed better than the proposed event-based feature in the participant-dependent and -independent schemes, confirmed by Wilcoxon signed rank test $(p < 0.05)$. A possible reason is that continuous features are good at distinguishing low and high load levels in communication tasks. When feature dimension is at 25, the communication load was correctly classified 77 and 86 percent of communication load tasks in average in participant-dependent and -independent respectively, which contributes a high accuracy in general. For the LSTM sequence model, it achieved a slightly lower accuracy than continuous features and event-based features in the participant-dependent scheme and significantly poorer accuracy in the participant-independent scheme, indicating the sequence information was hardly shared by all participants.

TABLE 4
Average Accuracy $\pm 95\%$ Confidence Interval Summary for Task Load Recognition for all Combinations of Context Setting Using Continuous Features and Event-Based Features (Dimension $k = 25$), Respectively

	Participant-dependent				Participant-independent			
	load level for specific type (2-class)	load level (2-class)	load type (4-class)	Load level and type (8-class)	load level for specific type (2-class)	load level (2-class)	load type (4-class)	load level and type (8-class)
continuous features (25d)	0.72(0.08), 0.61(0.12), 0.68(0.10), 0.52(0.13), 0.68(0.08), 0.57(0.15), 0.91(0.06), 0.81(0.08). 0.74(0.04)*. 0.63(0.08)	0.68(0.05)* 0.59 (0.06)	0.93 (0.03) 0.46 (0.04)	0.72(0.04) 0.27 (0.03)	0.73(0.07), 0.74(0.07), 0.90(0.04), 0.76(0.05), 0.74(0.06), 0.78(0.06), 0.98(0.01), 0.93(0.04). **0.84(0.02)***. 0.80(0.03)	0.77(0.03)* 0.52 (0.01)	0.87 (0.05) 0.76 (0.06)	**0.73 (0.06)** 0.41 (0.03)
1-gram (25d)	0.62(0.08), 0.59(0.07), 0.59(0.07), 0.66(0.06), 0.69(0.07), 0.70(0.08), 0.69(0.10), 0.63(0.07). 0.65(0.05)*. 0.64(0.03)	0.56(0.04)* 0.55 (0.03)	0.80 (0.04) 0.43 (0.04)	0.49 (0.05) 0.22 (0.03)	0.67(0.05), 0.48(0.02), 0.82(0.04), 0.63(0.07), 0.67(0.10), 0.77(0.05), 0.80(0.06), 0.77(0.07), 0.74(0.03), 0.66(0.03)	0.74 (0.02) 0.51 (0.01)	0.85 (0.04) 0.72 (0.04)	0.63 (0.04) 0.34 (0.03)
2-gram (25d)	0.53(0.08), 0.51 (0.05) 0.61(0.09), 0.57 (0.06) 0.66(0.12), 0.71 (0.09) 0.67(0.08), 0.60 (0.09) 0.62(0.04). 0.60 (0.04)	0.57 (0.05) 0.53 (0.04)	0.82 (0.05) 0.44 (0.04)	0.43 (0.05) 0.23 (0.04)	0.66(0.05), 0.49 (0.04) 0.73(0.04), 0.65 (0.06) 0.62(0.08), 0.67 (0.06) 0.80(0.05), 0.73 (0.05) 0.70(0.03). 0.63 (0.03)	0.72 (0.03) 0.51 (0.01)	0.84 (0.04) 0.69 (0.04)	0.60 (0.03) 0.34 (0.03)
3-gram (25d)	0.53(0.09), 0.54 (0.07) 0.57(0.07), 0.60 (0.06) 0.61(0.09), 0.57 (0.11) 0.55(0.06), 0.54 (0.08) 0.57(0.04). 0.56 (0.04)	0.58 (0.04) 0.53 (0.03)	0.85 (0.04) 0.44 (0.02)	0.43 (0.05) 0.21 (0.02)	0.72(0.03), 0.50 (0.04) 0.79(0.06), 0.56 (0.07) 0.62(0.07), 0.76 (0.07) 0.75(0.05), 0.76 (0.05) 0.72(0.03). 0.65 (0.02)	0.72 (0.02) 0.51 (0.01)	0.84 (0.03) 0.64 (0.04)	0.60 (0.04) 0.32 (0.03)
1-,2-,3-gram fusion (7d,9d,9d)	0.59(0.08), 0.61(0.07), 0.66(0.10), 0.73(0.08) 0.65(0.03)	0.60 (0.05)	0.82 (0.03)	0.49 (0.05)	0.67(0.04), 0.77(0.05), 0.65(0.07), 0.79(0.05) 0.72(0.02)	0.71 (0.03)	0.85 (0.04)	0.59 (0.04)
2-,3-gram fusion (12d,13d)	0.54(0.08), 0.62(0.08), 0.63(0.10), 0.58(0.10) 0.59(0.04)	0.57 (0.05)	0.80 (0.05)	0.46 (0.04)	0.70(0.04), 0.77(0.05), 0.65(0.07), 0.81(0.06) 0.73(0.03)	0.73 (0.03)	0.84 (0.03)	0.58 (0.04)
1-,2-,3-gram, continuous features fusion (4d,7d,7d,7d)	0.76(0.07), 0.63(0.09), 0.68(0.07), 0.97(0.03) **0.76(0.04)***	0.68 (0.06)*	0.96 (0.02)	0.72 (0.05)	0.70(0.06), 0.85(0.05), 0.75(0.06), 0.96(0.04) 0.81(0.03)	**0.78 (0.02)**	**0.89 (0.04)**	0.72 (0.04)
2-,3-gram, continuous features fusion (8d,8d,9d)	0.75(0.07), 0.62(0.09), 0.69(0.09), 0.97(0.03) **0.76(0.04)**	**0.73 (0.04)**	**0.96 (0.02)**	**0.74 (0.04)**	0.73(0.06), 0.86(0.04), 0.72(0.07), 0.95(0.03) 0.81(0.02)	**0.78 (0.02)**	**0.89 (0.04)**	0.72 (0.04)

*If the dimension $k < 25$ for event-based features, then the performance of the highest dimension was used. The accuracies in black and blue were obtained using the SVM and LSTM sequence model respectively. The value under a bar is the mean of the above four accuracies of cognitive, perceptual, physical and communicative load in order. * indicates that the accuracies of the 2-class classification are significantly different at the 0.05 level.*

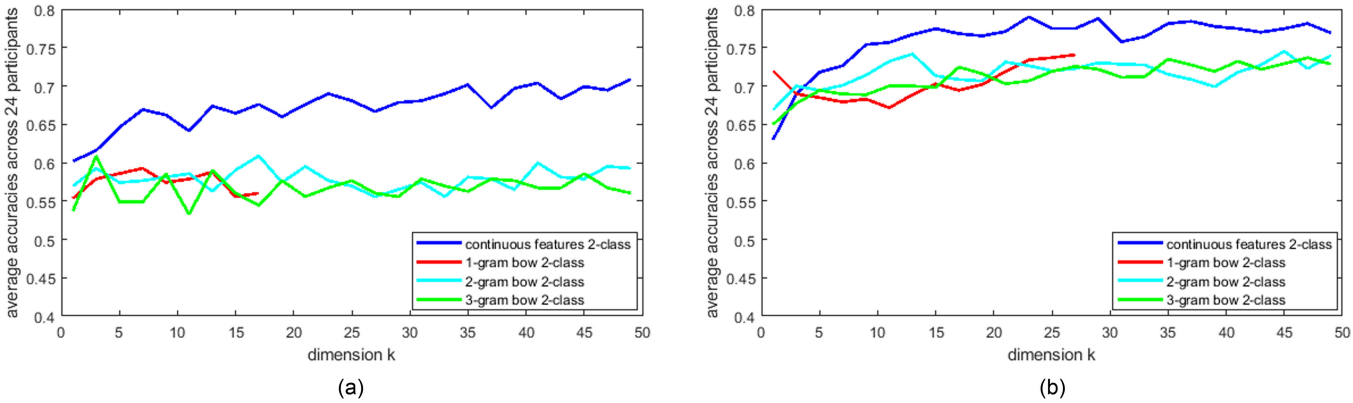

Fig. 7. Average accuracy versus feature dimension *k* (i.e., *k* features whose NCA weights ranked in the top *k*) for low and high load level behavioral feature *n*-gram recognition for various values of *n*, with all task load types pooled in (a) participant-dependent and (b) participant-independent classification scheme.

To answer the question of whether considering the type of task load can result in higher classification performance in task load level recognition, we found that at dimension 25, continuous features performed significantly better in load level recognition considering specific load types than that when load type was not considered, as shown in Table 4

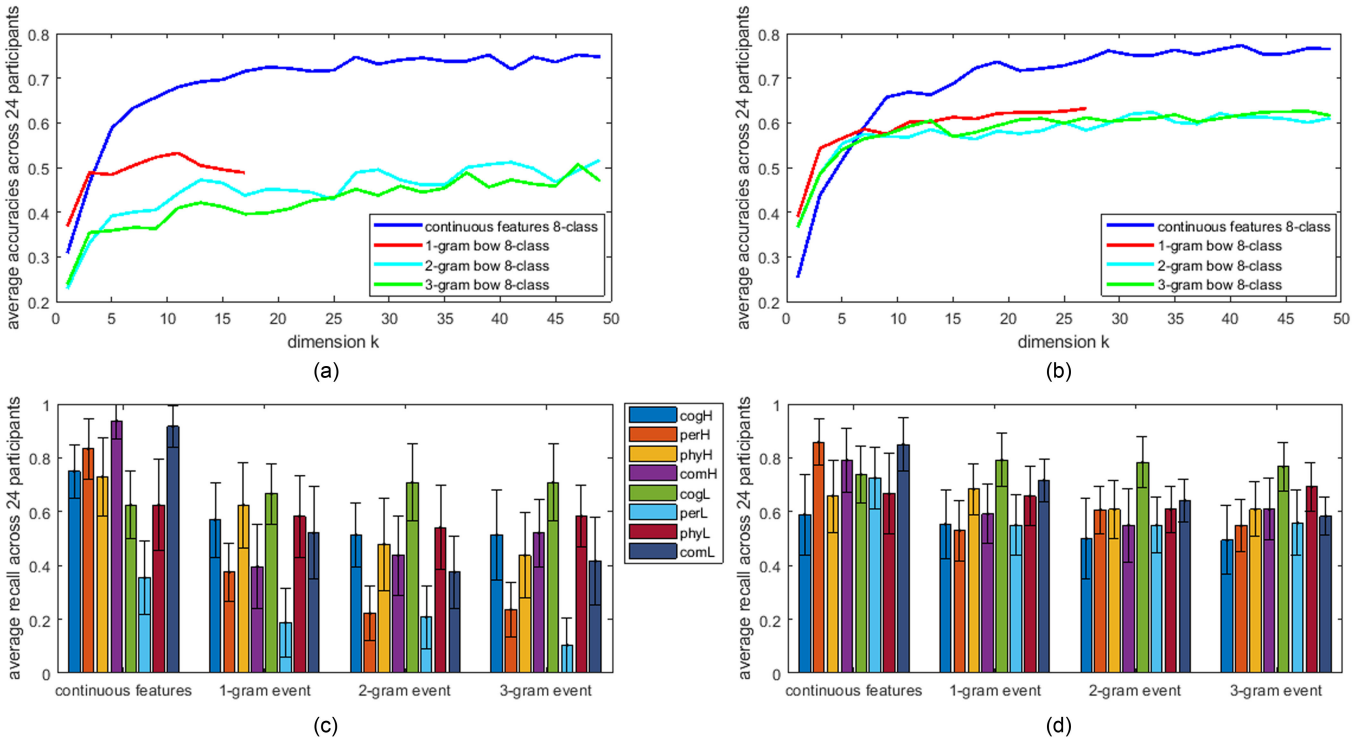

Fig. 8. Average accuracy versus feature dimension *k* (i.e., *k* features whose NCA weights ranked in the top *k*) for low and high load level and four task load type recognition in (a) participant-dependent and (b) participant-independent classification schemes. The average recall performance at a feature dimension of 25 is shown for (c) participant-dependent and (d) participant-independent classification schemes. If the dimension $k < 25$ in 1-gram event-based feature, then the performance of the highest dimension was used.

and indicated by a Wilcoxon signed rank test ($p < 0.05$). For event-based features, only 1-gram (non-sequence) in the participant-dependent scheme can reach this conclusion. This shows that by considering the type of task load, recognition of low and high task load levels can be easier, especially for continuous features.

5.5 Task Load Type and Load Level Recognition

The purpose of recognizing task load types and task load levels at the same time is to provide additional contextual information about what *causes* the load experienced. Figs. 8a, 8b shows the 8-class classification performance using event-based features in participant-dependent and -independent schemes, including a comparison with continuous features. It is evident that continuous features are superior to event-based features in participant-dependent and -independent schemes after dimension 5. In all cases, they are well above the chance-level accuracy of around 12.5 percent. Figs. 8c, 8d and Table 4 presents specific details for $1-, 2-$, and 3-gram event and continuous features at a feature dimension of 25. Interestingly, continuous features recognize communication load very well in both participant-dependent and -independent schemes, while recognizing low perceptual load is difficult to all features in participant-dependent scheme and high cognitive load in participant-independent scheme. However, overall, cognitive load recognition is the most difficult according to its ranking among the four load types. For the LSTM sequence model, from Table 4, we found that it achieved very poor performance, indicating that the sequence information does not help task load type recognition.

In terms of the choice of n for event-based features, 1-gram performed better than 2- and 3-gram in load level recognition regardless of load type when the dimension was low, probably due to a few events that are sensitive to task load difference. However, caution should be given to 1-gram as events might be more likely to result from non-affect stimuli, such as illuminance change, or due to task requirements, while the sequences in 2- and 3-grams may be less likely affected.

In terms of feature fusion, the fusion of 2-, 3-gram and continuous features achieved the best accuracies at dimension 25 in participant-dependent case as shown in Table 4. Wilcoxon signed rank tests ($p < 0.05$) suggest that using this fusion is significantly better than using continuous features alone in 2-class load level recognition and significantly better than using event-based feature alone in 8-class load level and type recognition. This indicates complementary information between these two feature types. That is, the continuous features can provide information on fine-grained measurements and subtle changes, while the event features are more robust against noise and capture changes in behavior pattern at a longer time frame.

Table 5 shows the features listed among the top 20 as ranked by NCA for all participants during leave-one-participant-out training. We can find that events from each modality have contributions to classification and their sequences matter, while for continuous features, behavior signals from head and speech were more important than physiological signals such as pupil and blink. This probably explain why continuous features achieved better performance as head and speech can be often affected by environment noise and stimulus presentations. Further studies are needed to find out more.

TABLE 5
Features Which Were Among the Top 20 Ranked By NCA From 24 Participants and Shared by all of Them in the 8-Class Participant-Independent Scheme

3-gram word	2-gram word
sac-sac-sac	sac-hed
sac-sph-sac	sph-sac
sac-hed-sac	sac-sac
sac-pup-sac	sac-pup
blk-sac-sac	hed-sac
sac-sac-hed	sac-sph
hed-sac-sac	
sac-sac-sph	**1-gram word**
sac-sac-blk	hed
Continuous feature	puphed
head_ velocity_x_std	sph
head_ velocity_y_std	pup
head_magnetism_y_std	blk
head_magnetism_z_std	sac
speech rate	sached
MFCC6_ mean	blksph
MFCC1_ delta _std	blksac
MFCC13_ acceleration _std	sacsph

Overall, these results demonstrate that the proposed event features capture most information about load level and load type from continuous features and can also provide certain complementary information in participant-dependent scheme. Meanwhile, sequence features have the attribute of being less sensitive to load type for load level recognition in participant-dependent and -independent schemes, suggesting that behavior event sequences may be more similar across load type than continuous feature, a very appealing hypothesis for future investigation.

Furthermore, in general, recognizing cognitive load is more difficult than other types of load probably because most mental activity is covert, resulting in fewer behavioral cues embodied in physical events. Further research is needed to improve classification performance to make 8-class load level and load type recognition at least as good as the 4-calss performance shown in Table 4, in order to be commercially viable in applications. For longitudinal modelling, by recognizing task load types, we hope to extract behavior (event sequences) in a broad range of tasks where the patterns of behavior may be consistent in each load type. Such behaviors may be found implicitly using a deep learning approach with a large dataset; however the current small dataset is not suitable to use deep learning.

5.6 Limitations and Future Work

One limitation of this study is that in communication tasks, speech features were extracted from all audio data during conversations, which inevitably involved the low voice of the experimenter. Although this occurs elsewhere in affective computing, e.g., [28], it might introduce some bias to the classification here, e.g., different pause time in conversations, which may reflect individual differences more than communication load. Second, although we designed the task activity to involve one type of load as specifically as possible, it is difficult to find tasks that contain only a single pure type of load. Therefore, the load types mentioned in this study are the *main* load type, rather than the *only* load type imposed on participants, e.g., the communication load task herein includes some cognitive load. Furthermore, as we rely on task design and subjective ratings for the load level ground truth, there may be a fatigue or learning effects changing the physiological or behavior data over time, although we did not find evidence that the load level ground truth changed over time by examining limited subjective ratings evolving with one hour. Finally, the tasks were completed continuously within one hour, which is shorter than the time aimed for everyday and longitudinal settings. As the proposed event features require less storage space to collect and less computational power to analyze than continuous features, while preserving most information related to affective states, this potentially leads to more affordable data collection and processing in longitudinal studies, or in-the-wild studies with limited resources.

It is worth mentioning that the breadth of different tasks may be uncountable in real life contexts, since single tasks can be overlapping and interleaving, but the number of load types produced from these tasks might be limited. The four types of load in this paper were proposed in [6] in order to represent different tasks [11], as shown in Fig. 1. In other words, for each task, we should recognize the load levels of every dimension of the four-dimensional task load at the same time. However, as the focus of this paper was the proposed event approach in the context of continuous task load and type change, we only recognized the load levels of the main task load type rather than all four types and left it to future work. Meanwhile, the characteristics of event sequences and their dimension will be further studied, and the duration of event will be added into event-based features to improve performance.

6 Conclusions

In this work, we proposed an event-based behavior modeling approach for continuous task load estimation. It converts physiological and behavioral signals from eye activity, speech and head movement into meaningful behavior events, such as blink, saccade, speech onset, pupil size increase, and head velocity increase, utilizes their sequence to form words, and employs their term frequency as event-based features to recognize both low and high load level and four task load types.

Classification experiments demonstrated that the proposed event-based feature can achieve almost comparable accuracies to the continuous features in 2-class load level recognition regardless of load type (71-74 percent), 4-class load type recognition (84-85 percent) with dimension 25 in a participant-independent scheme, while they were substantially lower in accuracy for 8-class load level and type recognition (58-63 percent) (in these schemes, the event features may suffer more from the reduced training data). The conventional continuous features achieved significantly better performance than the proposed event-based features in participant-dependent and -independent schemes, while the most effective features were from head and speech, which may reflect task and scene differences. However, by concatenating 2-, 3- gram features with continuous features, we achieved the best or similar accuracies in all cases except load level in each load type recognition in the participant-independent scheme, suggesting that event-based features contain beneficial complementary information to

continuous features. For the LSTM sequence model, it achieved some similar performance in task load level recognition, but very poor performance in load level and type recognition.

When we ignored task type changes in load level estimation and trained all the four task type data, classification performance dropped by 6 and 7 percent for continuous features in participant-dependent and -independent basis respectively. The general implication of these results is that for recognizing task load levels, we need to take the task type into account for continuous features while this may not be necessary for 2- and 3-gram event-based features.

This paper suggests a different behavior representation using discrete behavior events rather than using continuous features, in which data from every instant contributes. Our study is the first work showing the promising performance achieved using multimodal behavior event-based features and the merit of this approach in task load level recognition under different task load types. This may be an important research step towards longitudinal behavior analysis in the future.

Acknowledgments

This work was supported in part by US Army ITC-PAC, through contract FA5209-17-P-0154. Opinions expressed are the authors' and may not reflect those of the US Army. The authors thank I. Wang for collecting the data for this study. The authors also thank the reviewers for their constructive suggestions.

References

[1] M. Karg, A. Samadani, R. Gorbet, K. Kuhnlenz, J. Hoey, and D. Kulic, "Body movements for affective expression: A survey of automatic recognition and generation," *IEEE Trans. Affect. Comput.*, vol. 4, no. 4, pp. 341–359, Oct.-Dec. 2013.

[2] R. A. Calvo and S. D'Mello, "Affect detection: An interdisciplinary review of models, methods, and their applications," *IEEE Trans. Affect. Comput.*, vol. 1, no. 1, pp. 18–37, Jan. 2010.

[3] S. Chen, J. Epps, and F. Chen, "Automatic and continuous user task analysis using eye activity," in *Proc. Int. Conf. Intell. User Interfaces*, 2013, pp. 57–66.

[4] M. Soleymani, S. Asghari-Esfeden, Y. Fu, and M. Pantic, "Analysis of EEG signals and facial expressions for continuous emotion detection," *IEEE Trans. Affect. Comput.*, vol. 7, no. 1, pp. 17–28, Jan.-Mar. 2016.

[5] Z. Huang, J. Epps, and E. Ambikairajah, "An investigation of emotion change detection from speech," in *Proc. INTERSPEECH*, 2015, pp. 1329–1333.

[6] J. Epps and S. Chen, "Automatic task analysis: Towards wearable behavio-metrics," *IEEE Syst. Man Cybern. Mag.*, vol. 4, no. 14, pp. 15–20, Oct. 2018.

[7] S. Chen, J. Epps, and F. Chen, "An investigation of pupil-based cognitive load measurement with low cost infrared webcam under light reflex interference," in *Proc. Int. Conf. IEEE Eng. Med. Biol. Soc.*, 2013, pp, 3202–3205.

[8] S. Chen and J. Epps, "Blinking: Towards wearable computing that understands your current task," *Pervasive Comput.*, 12, no. 3, pp. 56–65, 2013.

[9] H. Steil and A. Bulling, "Discovery of everyday human activities from long-term visual behavior using topic models," in *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput.*, 2015, pp. 75–85.

[10] S. C. Mukhopadhyay, "Wearable sensors for human activity monitoring: A review," *IEEE Sens. J.*, vol. 15, no. 3, pp. 1321–1330, Mar. 2015.

[11] S. Chen and J. Epps, "Atomic head movement analysis for wearable four-dimensional task load recognition," *IEEE J. Biomed. Health Inform.*, vol. 23, no. 6, pp. 2464–2474, Nov. 2019.

[12] R. M. Makepeace and J. Epps, "Automatic task analysis based on head movement," in *Proc. EMBC*, 2015, pp. 5167–5170.

[13] M. Schmitt, F. Ringeval, and B. Schuller, "At the border of acoustics and linguistics: Bag-of-audio-words for the recognition of emotions in speech," in *Proc. Interspeech*, 2016, pp. 495–499.

[14] T. F. Yap, J. Epps, E. Ambikairajah, and E. H. C. Choi, "Voice source under cognitive load: Effects and classification," *Speech Commun.*, vol. 72, pp. 74–95, 2015.

[15] S. Chen and J. Epps, "Automatic classification of eye activity for cognitive load measurement with emotion interference," *Comput. Methods Programs Biomedicine*, vol. 110, no. 2, pp. 111–124, 2013.

[16] R. F. Stanners, M. Coulter, A. W. Sweet, and P. Murphy, The pupillary response as an indicator of arousal and cognition, *Motivation Emotion*, vol. 3, pp. 319–340, 1979.

[17] R. B. Zajonc, "Feelings and thinking: Preferences need no inferences," *Amer. Psychologist*, vol. 35, no. 2, pp. 151–175, 1980.

[18] R. S. Lazarus, "Thoughts on the relations between emotions and cognition," *Amer. Physiologist*, 37, no. 10, pp. 1019–1024, 1982.

[19] D. D Salvucci and J. H. Goldberg, "Identifying fixations and saccades in eye-tracking protocols," in *Proc. Symp. Eye Tracking Res. Appl.*, 2000, pp. 71–78.

[20] S. Zhalehpour, O. Onder, Z. Akhtar, and C. E. Erdem, "BAUM-1: A spontaneous audio-visual face database of affective and mental states," *IEEE Trans. Affect. Comput.*, vol. 8, no. 3, pp. 300–313, Jul.-Sep. 2017.

[21] J. Zhu, "Mobile behaviometrics: Behavior modeling from heterogeneous sensor time-series," PhD dissertation, Carnegie Mellon Univ., Pittsburgh, PA, USA, 2014.

[22] K. Wac and C. Tsiourti, "Ambulatory assessment of affect: Survey of sensor systems for monitoring of autonomic nervous systems activation in emotion," *IEEE Trans. Affect. Comput.*, vol. 5, no. 3, pp. 251–272, Jul.-Sep. 2014.

[23] L. Fridman, B. Reimer, B. Mehler, and W. T. Freeman, "Cognitive load estimation in the wild," *Proc. CHI Conf. Human Factors Comput. Syst.*, 2018, Art. no. 652.

[24] S. Chen and J. Epps, "Efficient and robust pupil size and blink estimation from near-field video sequences for human-machine interaction," *IEEE Trans. Cybern.*, vol. 44, no. 12, pp. 2356–2367, Dec. 2014.

[25] F. Eyben, F. Weninger, M. Wollmer, and B. Schuller, "Open-Source Media Interpretation by Large feature-space Extraction," 2016. [Online]. Available: http://opensmile.audeering.com/

[26] J. S. Lerner and D. Keltner, "Beyond valence: Toward a model of emotion-specific influences on judgement and choice," *Cognition Emotion*, vol. 14, no. 4, pp. 473–493, 2000.

[27] W. Yang, K. Wang, and W. Zuo, "Neighborhood component feature selection for high-dimensional data," *J. Comput.*, vol. 7, pp. 161–168, 2012.

[28] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions," in *Proc. 10th IEEE Int. Conf. Workshops Autom. Face Gesture Recognit.*, 2013, pp. 1–8.

[29] S. Patel, H. Park, P. Bonato, L. Chan, and M. Rodgers, "A review of wearable sensors and systems with application in rehabilitation," *J. Neuroengineering Rehabil.*, vol. 9, no. 21, 2012.

[30] A. Yuce, H. Gao, G. L. Cuendetm, and J. Thiran, "Action units and their cross-correlations for prediction of cognitive load during driving," *IEEE Trans. Affect. Comput.*, vol. 8, no. 2, pp. 161–175, Apr.-Jun. 2017.

[31] P. Ren, A. Barreto, Y. Gao, and M. Adjouadi, "Affective assessment by digital processing of the pupil diameter," *IEEE Trans. Affect. Comput.*, vol. 4, no. 1, pp. 2–14, Jan.-Mar. 2013.

[32] M. Munezero, C. S. Montero, E. Sutinen, and J. Pajunen, "Are they different? Affect, feeling, emotion, sentiment, and opinion detection in text," *IEEE Trans. Affect. Comput.*, vol. 5, no. 2, pp. 101–111, Apr.-Jun. 2014.

[33] L. Zhang, *et al.*, "Cognitive load measurement in a virtual reality-based driving system for autism intervention," *IEEE Trans. Affect. Comput.*, vol. 8, no. 2, pp. 176–189, Apr.-Jun. 2017.

[34] H. A. Maior, M. L. Wilson, and S. Sharples, "Workload alerts - Using physiological measures of mental workload to provide feedback during tasks," *ACM Trans. Comput.-Hum. Interact.*, vol. 25, no. 2, 2018, Article 9, 30 pages.

Siyuan Chen (M'11) received the ME degree from the Roy. Melbourne Inst. Technol. (RMIT) University, Melbourne, Australia, in 2008 and the PhD degree from the University of New South Wales, Sydney, Australia, in 2014. From 2014 to 2015, she was a postdoctoral fellow with the School of Computing and Information Systems, the University of Melbourne. She was a recipient of Australia Endeavour Fellowship in 2015, which funded her to conduct a 6-month research visit at Inria, Sophia Antipolis, France, during 2015-2016. Currently, she is a research fellow in the School of Electrical Engineering and Telecommunications, the University of New South Wales, Sydney, Australia. Her research interests include eye activity computing and analysis, cognitive load modeling, machine learning for human-centered computing, and intelligent computing systems.

Julien Epps (M'97) received the BE and PhD degrees from the University of New South Wales, Sydney, Australia, in 1997 and 2001, respectively. He was a postdoctoral fellow at the University of New South Wales. From 2002 to 2004, he was a senior research engineer with Motorola Labs, where he was engaged in speech recognition. From 2004 to 2006, he was a senior researcher with National ICT Australia, Sydney. He then joined the UNSW School of Electrical Engineering and Telecommunications, New South Wales, Australia, in 2007, as a senior lecturer, and is currently a professor. He is also a contributed researcher at Data61, CSIRO, Australia. He has authored or coauthored more than 200 publications and serves as an associate editor for the *IEEE Transactions on Affective Computing*. His current research interests include characterization, modelling, and classification of mental state from behavioral signals, such as speech, eye activity, and head movement.