Multimodal Event-based Task Load Estimation from Wearables

Siyuan Chen
School of Electrical Engineering and Telecommunications
The University of New South Wales
Sydney, Australia
siyuan.chen@unsw.edu.au

Julien Epps
School of Electrical Engineering and Telecommunications
The University of New South Wales
and *Data61, CSIRO*, Sydney, Australia
j.epps@unsw.edu.au

***Abstract*— Humans always engage multiple modalities when performing tasks, such as eye activity, speech and head movement, which contain rich information indicative of task load that can help understand and predict human psychological state and behavior. In recent research into multimodal signal processing, the ideas of sequence- and coordination-based event features have been proposed, which explicitly utilize the interaction information among different modalities. In this paper, we propose event intensity and event duration-based features, which capture the extent and duration of onset events that denote major changes in behavior signal. These features are combined with sequence- and coordination-based event features to achieve state-of-the-art performance in assessing task load levels and load types. In experimental work, we collected eye activity, speech and head movement data from 24 participants during cognitive, perceptual, physical and communication tasks. Results suggest that by fusing these four compact, interpretable event-based features, strong accuracy can be achieved: 84% for two load level classification, 89% for four load type classification and 76% for 8-class classification, outperforming conventional statistical features and deep neural network self-learned features by up to 9% and 25% respectively. These features do not need to be selected during training and can generalize well for different participants and different task types.**

Keywords— Task load, eye activity, speech, head movement

I. INTRODUCTION

Task load estimation plays an important role in understanding user mental state and behaviours. When we are performing daily tasks, e.g. thinking, seeking, speaking, moving, our cognitive system is continuously exposed to different load levels and types. Its capacity determines our task performance and behaviors. While this mental state may be obvious for the person who undertakes the tasks, it is difficult for others to estimate since it is very implicit. This motivates research into using physiological and behavioral signals to estimate task load, since understanding mental state can greatly facilitate cooperation between people or between users and interactive systems [1], and help design novel interfaces to enhance and mediate human computer interaction [2].

Efforts in physiological and behavioral computing have often focused on three research directions, aiming to improve psychophysiological state estimation and proliferate applications in different contexts. The first is the measurement mode: mobile or non-mobile. Experiments in non-mobile settings often include e.g. EEG, ECG, GSR, respiration [4-6], where body movements are restricted to a certain degree to avoid significant interference from movement. Other measures, such as pupillary response, speech, body movement [4,7-9], which can be acquired from mobile eye trackers, microphones, and Inertial Measurement Units (IMUs) can be easily used in mobile contexts. An example of a mobile device is shown in Fig. 1. They are suitable for longitudinal studies and daily usage. However, environmental noise may be significant, and mining useful information is challenging. Other research efforts were devoted to understanding and extracting effective features from different modalities, which not only contain psychological meaning but also can sufficiently discriminate the psychophysiological state of interest from others [5,8,9]. A well-known example is the mean pupil diameter during tasks for cognitive load estimation [10]. Employing different learning approaches is the third research direction. Features from a single modality or multiple modalities have been explored in different classification approaches or fused at different stages in the machine learning process [7,8,11,12] in order to improve mental state estimation.

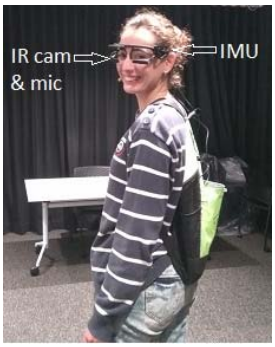


Fig. 1. A wearable eyewear system (Pupil Labs [3] is the main frame) integrating a microphone and an IMU is promising to collect eye activity, speech and head movement for mobile daily usage.

In this paper, we focus on processing multiple modalities which can be acquired in mobile settings, including eye activity, speech and head movement, for task load type and level estimation based on previous work. We extend recent research into features based on change events [13] to explore a further two event-based features, fully utilizing the interaction information among multiple modalities to improve task load estimation greatly. We also investigate the estimation performance of using event-based features and features learnt from deep learning networks. This addresses the open question of whether deep learning networks can successfully self-learn useful information for task load estimation from physiological and behavioral signals, which can help design deep learning network architectures specifically for physiological and behavioral computing.

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

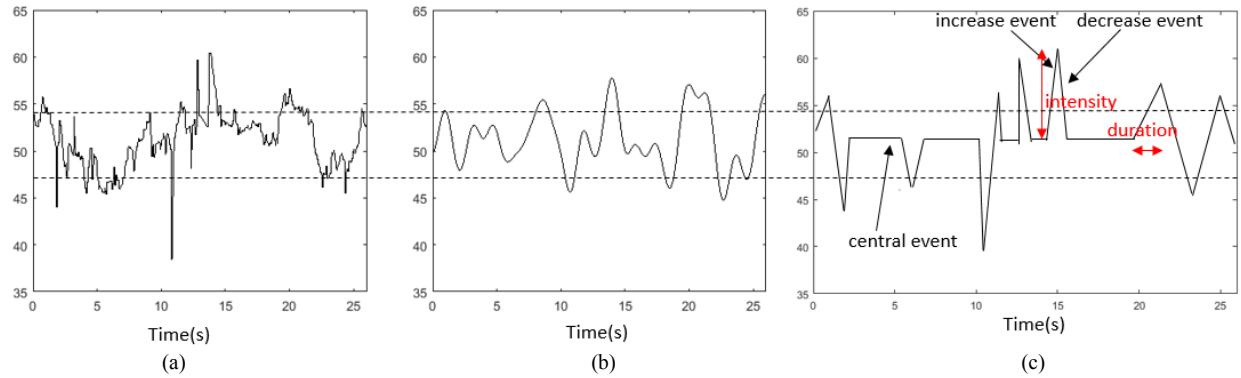


Fig. 2. An illustrative example showing (a) the raw pupil size signal and (b) the results of wavelet decomposition denoised [14,15] and (c) detection of increase, decrease and central events and their corresponding intensities using the atomic event detection algorithm [13].

II. Background and Related Work

A. Task Load Level and Type Recognition

Task load is generally believed to be produced due to the limited capacity of working memory in the cognitive system [16-18]. When temporal visual, auditory and other information are present, the central executive controls working memory and long-term memory to receive, store, organize and retrieve theses information. The bottleneck in working memory results in this processing being unable to occur simultaneously, causing a load on the cognitive system. This load in turn changes our physiology and behaviors. Performing multiple tasks at the same time [17] or performing a single task where information elements are highly interactive [18] can quickly deplete the limited resources, hence producing high task load levels.

Most studies using physiological and behavioral signals for task load estimation only considered load level recognition within a particular task. For example, Zhang et al. [9] employed eye gaze, EEG, and peripheral physiology (GSR, EMG etc.) to classify two load levels during virtual driving tasks, where load was induced by different speed, responsiveness, and weather conditions, and achieved the best accuracy around 84%. Fridman et al. [11] input eye images recorded in real world driving tasks to a deep leaning neural network, and achieved 86% accuracy discriminating 0-, 1- and 2-back working memory tasks during driving. In a recent study [19], it was found that even with the same features, the estimation accuracies of task load levels may significantly vary depending on task load type. Recognizing the load level of perceptual tasks and communication tasks (above 90% accuracy) was easier than that of cognitive tasks and physical tasks (around 70% accuracy). Few studies [13,19] took different task types into account and trained task load models using different task load types. Consequences of this may include low estimation performance or 'out-of-vocabulary' problems for continuous and longitudinal task load estimation because in the research laboratory task load models are usually built based on a small set of short-time prototypical tasks, while real-world tasks are continuously changing and may comprise previously unseen variants.

In recent work, Epps et al. [20] suggested a four-dimensional task load framework to describe load assessment associated with general load types. In this framework, tasks can be analyzed by their attributes in terms of four load types: cognitive, perceptual, physical, and communication load, based on the Berliner task taxonomy [21], and their intensity (load level). This kind of dimensional intensity-based representation has already been used extensively to represent emotions using arousal and valence [22], and by the NASA-TLX task load index [23] where multiple aspects from temporal demand, physical demand, mental demand are self-rated to fully assess task load. However, in terms of task load modeling for load level estimation, few studies [19] considered different task load types and fully assessed load levels across different task types.

B. Multiple Modality Processing

Physiological and behavioral signals acquired from different modalities are often represented by numerical feature values, which contain both useful information and noise. Those signal components which represent fast and non-continuous change at each task instant or during a task period are often filtered out before feature extraction [15]. Fig. 2(a)(b) shows an example of such a smoothing effect. This denoising process requires knowledge of the noise and/or physiological/behavioral signal characteristics for meaningful (rather than heuristic) smoothing filter parameter setting. Statistical features in the time or frequency domain [5,7-9], which take every task instant value into account, are often hand crafted as inputs to task load models. Even for some signals which are discrete in nature, e.g. blink, statistical features have been extracted during a time window, i.e., blink rate or blink duration per second [5, 9] to estimate task load.

As opposed to utilizing continuous signals to extract features, a recent study [13] converted continuous physiological and behavioral signals to discrete events, following the form of those behavior events to extract effective features. These discrete events were generated based on the 'atomic' event detection algorithm [13], which detects increase, decrease, and central movement events. The corresponding event intensity was measured by the dispersion from a central value, as illustrated in Fig 3(c). The physiological rationale behind the three atomic events is that they indicate a balance change in two antagonist systems: when the signal is 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.

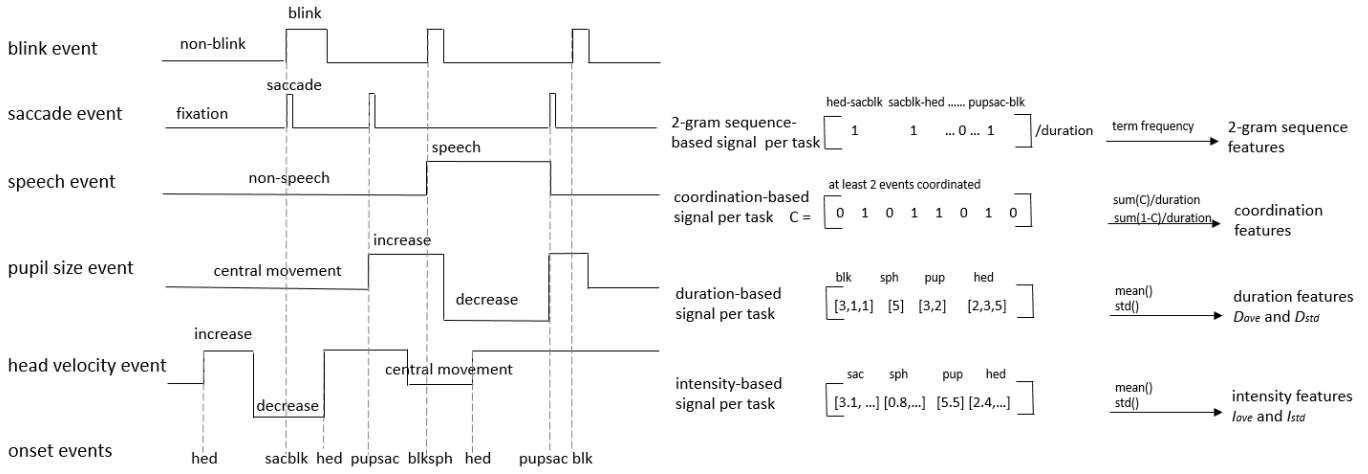


Fig. 3. An illustrative example showing (a) raw event signals (L_t) from five modalities, each extracted using the algorithms in [13] to compose onset events in sequence at the bottom and (b) sequence-based features, coordination-based features, duration-based features, and intensity-based features. To obtain the sequence-based features, 2-gram method is applied to the onset events during a task to form terms, each composed of 2 events. For all the tasks in training, each term is encoded as 1 or 0 to indicate whether the term exists in the task or not, then the count of each term is calculated as a 2-gram sequence feature and the feature dimension will depend on the training data. For coordination features, each event is firstly encoded as 1 or 0 to indicate whether this event is composed of at least 2 events coordinated or just one single event, then the count of the occurrence and non-occurrence of coordination is used as the feature. The proposed duration and intensity features are described by equations (4) and (6).

This approach is different from another common discrete encoding paradigm that uses *k*-means to cluster low-level descriptors into *k* clusters to represent them by a symbolic sequence of cluster indices [24], which cannot be easily interpreted. It is also different to event modeling based on task beginning and end timing [25]. The atomic event detection approach in principle allows partial reconstruction of a numerical representation from the events and their corresponding intensities, as shown in Fig 3(c): the events compactly preserve the general shape and remove all low-level variations between events, which is similar in principle to the aim of employing signal processing techniques such as wavelet decomposition, as shown in Fig. 2(b).

To explore these atomic events for task load estimation, one study [19] employed an *n*-gram bag-of-words behavioral model, which utilized the sequence events from different modalities to recognize task load levels. Specifically, the onset of blink, speech, saccade, pupil size increase and head angular velocity increase events were selected along time. Frequency of event sequences was used as a feature to input to an SVM classifier after feature selection in training. As opposed to an event sequence, another study [26] examined whether event coordination, i.e. the number of events occurring simultaneously, changes in different task load levels. Based on multimodal coordination dynamics studies (e.g. [27]), it was hypothesized that when task load was high, behavior events are more likely to co-occur. The statistical tests in [26] confirmed this hypothesis.

Regarding multiple modality processing in task load modelling, the aim has often been to find the best combinations of different features utilizing complementary information between multiple modalities [5,7-9]. By contrast, event-based features utilize the interaction among multiple modalities in terms of sequence and coordination and have been rarely studied for fusion in task load modeling. However, apart from employing event sequence features for task load estimation using machine learning in [13], task load estimation performance using coordinated event features is unknown yet.

C. Deep Learning for Mental State Estimation

Differently from hand-crafted features, which often rely on psychophysiological knowledge, deep learning models can automatically learn useful features, relying on their powerful learning ability and abundant information in big data. Current end-to-end deep learning has been applied to speech for emotion recognition [12]. The reported performance was better than using hand-crafted acoustic features. In their study [12], a raw 6-s long speech waveform sampled at 16kHz was directly input to a convolutional recurrent network with two convolutional layers and two LSTM layers. The two convolutional layers were expected to automatically learn effective features to replace conventionally engineered features. A visualization of gate activations demonstrated that the automatically learned features were highly correlated with the prosodic features that have previously been used to predict arousal.

Another end-to-end deep learning approach was applied to eye images for cognitive load recognition among *n*-back tasks during a driving scenario [11]. In this work, 90 temporally ordered raw eye images of size 64 × 64 pixels, cropped after face and eye recognition, were input to a 3D-CNN architecture and achieved 86% accuracy for three classes, which outperformed the accuracy using explicitly extracted pupil positions as input features to HMM models by around 8%.

With physiological and behavioral signals, effective hand-crafted features are usually always interpretable and generalize well based on psychophysiology studies. This generalization could be consistent with the deep learned features through end-to-end deep neural networks which require large amounts of data to learn. However, to guide the neural network to learn more advanced and effective features to improve task load estimation,

we need firstly to understand how neural networks perform with deep learned features compared with effective hand crafted features, then we may develop more effective deep learning network architectures for psychophysiological state estimation.

III. METHODS

The aim of our study is to fully utilize event-based features to improve task load estimation where different task load types are considered. We collected a dataset containing four types of task load (elaborated in Section IIIB) based on the four-dimensional task load framework [20]. Each load type contains two task load levels (low and high) for simplicity. Analysis was conducted under three conditions: (i) task load level recognition regardless of load type (2-class); (ii) task load type recognition (4-class) regardless of load level; (iii) task load level and type recognition (8-class) to fully understand the discriminability of the event-based features. The atomic events were extracted following the approach described in [13]. Sequence-based event features were the same as those in [19]. Coordination-based event features were also extracted according to [26]. New event-based features based on event duration and intensity were proposed and fused with the sequence and coordinated event features to fully evaluate the effectiveness of the event-based methods. End-to-end deep learning systems using raw events obtained from eye, speech and head movement data were used as baselines.

A. Proposed Event Duration and Intensity Features to Improve Task Load Estimation

Differently from previous work where only duration of blink and saccade were used (e.g. [14]), we also calculated the time elapsed from onset to offset for pupil increase and head velocity increase events. As shown in Fig. 2(c), a continuous signal during a task (S_t) can be segmented into an event signal L_t at time t, which includes increase, decrease and central events, corresponding to 1, -1 and 0 respectively in Fig. 3(a). According to the algorithm in [13], the central events were firstly obtained.

$$L'_t = \begin{cases} cen, & |v(t)| < v_{th} \\ non\_cen, & \text{otherwise} \end{cases} \quad (1)$$

where L'_t is the label of each t. Then the first derivative was calculated to decide the increasing (*incr*) and decreasing (*decr*) parts.

$$L_t = \begin{cases} cen, & \text{when } L'_t = cen \\ incr, \ \frac{dv(t)}{dt} > 0 & \text{when } L'_t = non\_cen \\ decr, \ \frac{dv(t)}{dt} \le 0 & \text{when } L'_t = non\_cen \end{cases} \quad (2)$$

Therefore, we can easily obtain the timestamps of the onset and offset of the nth increase events during a task, t_{1n} and t_{2n}. The signal of duration of increase event at time t is

$$Dur_t = \begin{cases} t_{2n} - t_{1n} & t = t_{1n} \\ 0 & otherwise \end{cases} \quad (3)$$

The event duration features per task, D_{ave} and D_{std}, are the average and standard deviation of the non-zero durations across a task. That is

$$D_{ave} = \frac{1}{N}\sum_{n=1}^{N} Dur_t \quad when\ t = t_{1n}$$

$$D_{std} = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(Dur_t - D_{ave})^2} \quad when\ \ t = t_{1n} \quad (4)$$

where N is the number of onset events t_{1n}. These features were extracted from each of the five modalities. However, most of the saccade durations were one frame, therefore only the four other modality durations were used, to form four dimensions for each duration feature.

The intensity of increase event at time t was

$$Int_t = \begin{cases} max(S_{t_{1n} \le t \le t_{2n}}) & t = t_{1n} \\ 0 & therwise \end{cases} \quad (5)$$

Similarly, the event intensity features per task were I_{ave} and I_{std} using

$$I_{ave} = \frac{1}{N}\sum_{n=1}^{N} Int_t \quad when\ t = t_{1n}$$

$$I_{std} = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(Int_t - I_{ave})^2} \quad when\ \ t = t_{1n} \quad (6)$$

These features were extracted from four modalities: saccade, pupil size, head movement, and speech signals while blink intensity (eyelid opening) was unavailable from current data. Therefore, four modality intensities were used to form four dimensions for each intensity feature.

B. Task Load Data Set

We collected one-hour eye activity, speech and head movement data from each of the twenty-four volunteer participants (14 males, 10 females, aged 18-25). They were asked to wear a wearable system and sit at a desk but were free to move any part of their body and to speak while completing the four types of tasks (approved by UNSW Human Research Ethics Advisory). The wearable system included a lightweight glasses frame on which a modified IR webcam was pointing towards the eye. The webcam recorded video (30 fps) and audio data (sampling rate: 44.1k) and was connected to a laptop with a USB cable. An IMU was attached to the participant's head by a head strap (in a product the IMU would be embedded in the glasses frame) and connected to the laptop with a USB cable. The IMU prototype consisted of an inertial measurement unit (MPU 9150) and output three-axis acceleration, angular velocity, and magnetic field strength at a rate of around 20 Hz. A 'scene view' camera was used to record all activities during the experiment for reference during annotation.

Four types of tasks were designed to induce four types of task load, namely cognitive load, perceptual load, physical load and communication load, representing some of our daily activities. 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 a very simple conversation or an object guessing game. The two load 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. Task durations varied in each individual task.

At the beginning of data recording, participants clapped their hands and nodded their head at the same time in order to synchronize all sensor signals in later processing. 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. After this, tasks in the four load types and two load levels were continuously presented in a counterbalanced order and completed by participants without breaks. At the end of a few selected tasks, subjective ratings (0-7) were solicited again and participants provided verbal answers. Following the completion of all tasks, participants were thanked and given a voucher.

In total, each participant completed 44 tasks. Among them, 22 tasks had low load level and 22 tasks had high load level. Irrespective of load levels, there were 14, 10, 10, and 10 tasks corresponding to cognitive, perceptual, physical and communication load types respectively. Considering both load type and load level, there were 7 tasks in the category of low cognitive load, 7 in high cognitive load, and 5 tasks in each of the remaining 6 categories. The induced load level was treated as the ground truth, and this assumption was verified using participants' subjective self-ratings of the task difficulties. The timestamps of each task were automatically recorded.

C. Signal Processing

The aim of the signal processing procedure was to obtain behavior events from five modalities during a task, including three natural events: blink, saccade, and speech onset, and two events converted from continuous physiological and behavioral signals: pupil size increase event and head velocity increase event. Then event-based features were extracted. The processing of these eye videos, speech waveforms and head movement data to obtain pupillary response, pupil center positions and blink, acoustic features and resampled 30 Hz head acceleration, angular velocity, and magnetic signals can be found in [28,29,13] respectively. Saccade and fixation were then separated from pupil center positions using dispersion-based algorithms [30] (1˚ of visual angle for at least 200 ms). Speech onset was obtained by thresholding voicing probability (one of the acoustic features) at 0.70. The increase, decrease and central movement events for pupil size and head velocity were obtained using the atomic head movement segmentation algorithm [13] as shown in equation (1-2), where the threshold for central movement was 3˚/s for head velocity and was mean pupil size during the first 0.5 sec of a task for pupil size. An illustrative example of these discrete event signals is shown in Fig. 3(a).

Four types of event-based features were extracted: sequence-event based features [19] where 2-gram was chosen, coordinated-event based features [26], event duration-based features, and the proposed event intensity-based features as described in equations (4) and (6). Fig. 3(b) illustrates the feature formulation and dimensions. It is worth mentioning that only the onset events were used as we assumed that they represent the occurrence of effort. Therefore, only the duration and intensity of the onset events were calculated. They can be interpreted as exerting and maintaining effort for each behavior, while the sequence- and coordinated-event features express the changes in multiple modality interaction in different task load contexts.

D. Multimodal Convolutional Recurrent Network Architecture

To recognize task load levels and load types, we input these hand-crafted event-based features into a deep neural network backend to learn task load models and evaluate the performance. Meanwhile, pseudo end-to-end deep learning was conducted where raw event signals from five modalities (shown in Fig. 3(a)), their intensity and duration of each event along time (equation (4) and (6)), and sequenced onset events (Fig. 3(a) bottom) were experimented with as inputs respectively. Effective features were expected to be automatically learned and passed to the same deep backend as the hand-crafted event-based features, and the accuracy was used as a baseline. The two multimodal deep neural network architectures are shown in Fig. 4(a) and (c), while for sequenced onset events, they were treated as a document and pre-processed as text data. They were encoded before being passed to the LSTM deep backend with a word embedding layer as shown in Fig. 4(b).

A long short-term memory (LSTM) network was chosen as the backend because physiological and behavioral signals are temporal data in nature, and the events extracted were also in sequence. The event-based features utilized the temporal information between multiple modalities to represent task load. We used the same LSTM deep backend in order to compare the effectiveness of the hand-crafted event-based features, conventional statistical features, and the automatically learned deep features for task load estimation. For the pseudo end-to-end deep learning baseline, we used the same convolutional recurrent network topology as the end-to-end speech emotion recognition in [12]. Comparing with the end-to-end cognitive load estimation in [11], we used one level of CNN rather than three levels and used CNN-LSTM rather than 3D-CNN.

E. Experiment

We firstly analyzed the subjective ratings of low and high load levels for each task type across all participants in order to show the validity of task load levels. Wilcoxon paired sign tests were conducted to confirm whether the two load levels were significantly different.

To estimate task load levels (2-class), task load types (4-class), and task load type and level (8-class), we used a leave-one-participant-out scheme. Therefore, 1012 tasks (23 (participants) × 44(tasks)) were used for training and 44 tasks for testing. To train the models, we set the number of epochs to be 30, batch size to be 32, and dropout rate to be 0.5 according to trial and error on one participant's data subset. The number of hidden neurons was 128 and 32 for LSTM1 and LSTM2 respectively. Each fully connected layer's number of hidden neurons was equal to the number of classes, i.e. 2, 4 or 8.

The four types of event-based features were extracted from each task. We firstly input them to the LSTM deep backend (Fig. 4(a)) individually to investigate their discriminability for task load level and type estimation. Then we fused them by concatenating all features to form a 43-dimensional feature vector (25 sequence features, 2 coordination features, 8 duration features, and 8 intensity features) and input them to the same LSTM deep backend for the three classification tasks.

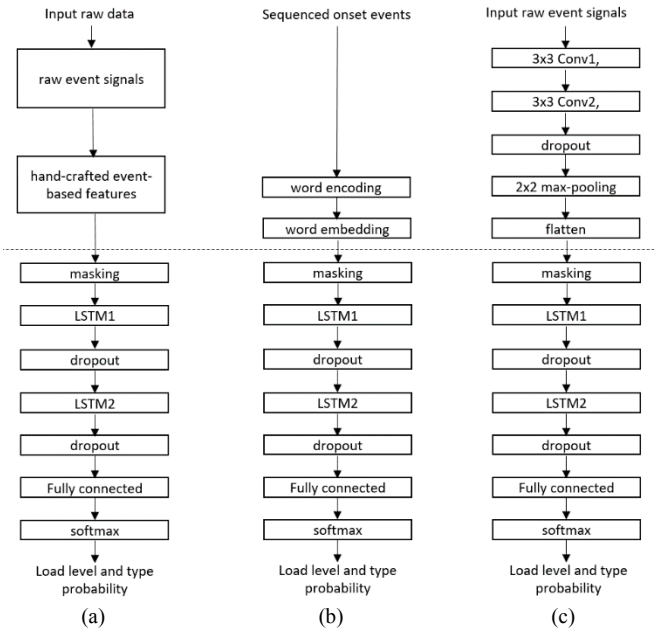


Fig. 4. The architectures of (a) multimodal LSTM for task load estimation, where the input can be hand-crafted event-based features (Fig. 3(b)), (b) multimodal LSTM where the input is the sequenced event signals (Fig. 3(a) bottom), and (c) CNN-LSTM where the input is the raw event signals (Fig. 3(a)) obtained from eye images, speech waveform and head movement.

Six baselines were employed. The first one adopted 157 conventional statistical features, which have often been seen for task load estimation, including mean and standard deviation of pupil size, blink, fixation and saccade features, head acceleration, angular velocity, magnetism values, in three axes, prosodic features, mel frequency cepstral coefficients (MFCC) features, and perceptual linear prediction (PLP) features to represent each task [19]. The best 25 features were selected during training using the neighborhood component analysis (NCA) feature selection method. These hand-crafted features were passed to the LSTM deep backend to compare with the proposed four types of event-based features.

The second baseline comprised the onset events along time in the form of words as shown the bottom line in Fig. 3(a). The performance was compared with that using the selected 25-dimensional 2-gram features since the sequence-based features were extracted from the same onset events.

The third baseline involved raw event signals from the five modalities shown in Fig. 3(a), containing all information about physiological and behavioral change, including the onset and offset of blink, saccade, and speech, pupil size and head velocity increase, decrease, as well as the duration of each event. In the fourth baseline, the intensity of each saccade, pupil size change and head velocity change event was employed, since they were not included in the raw event signals. The fifth baseline contained explicit event duration information, since this information is only implicit in the raw event signals from the five modalities. The last baseline concatenated all signals (43-D) above to provide complete information about eye activity, speech and head movement. Useful information is expected to be learned by the convolutional neural network to achieve better performance than the hand-crafted features. Sequence padding to make each task sequence the same length was used since the task duration varied.

All performances were compared using the average accuracies across 24 participants with a 95% confidence interval, since the amount of data in each class was nearly balanced.

IV. Results

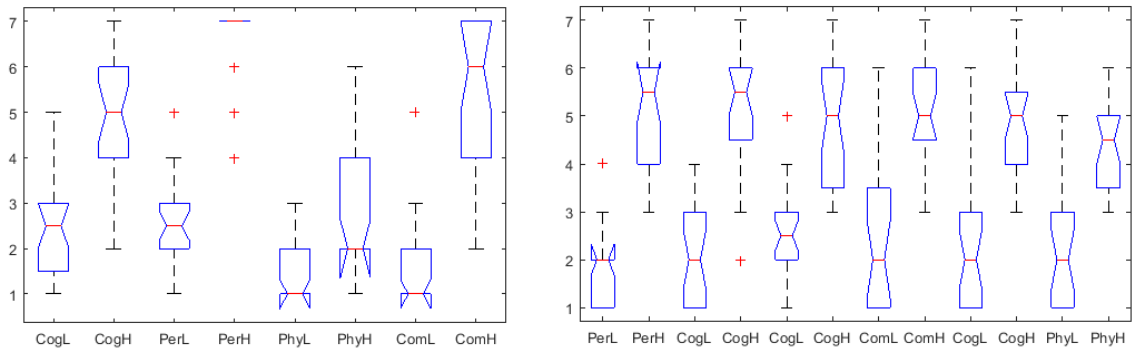
Fig. 5(a) shows the subjective rating of load level in each load type across 24 participants after they completed the first task of each category. Nonparametric Wilcoxon paired sign tests (two-sided) confirmed that in each task type, the perceived load level during the designed high load tasks was significantly higher than that in the designed low load tasks (Z = -4.3, -4.3, -3.5, -4.2 respectively, $p < 0.001$). Fig. 5(b) presents the subjective rating at the end of selected tasks. Wilcoxon paired sign tests (two-sided) also indicate that the two load levels were significantly different (Z = -4.3, -4.3, -4.1, -4.2, -4.3, -4.3, respectively, $p < 0.001$) at each time they were rated.

Table 1 presents the average accuracies with their 95% confidence intervals across 24 participants for task load level and load type estimations using the proposed event-based features and suggested baselines. We can see that the best task load level estimation performance was 84%, using the fused event-based features, which is far better than other methods. The best performance for load type estimation was 89% and for load level and type recognition was 76%, both of which were achieved using the fused event-based features. However, the performance of statistical features seems not significantly different from the two best accuracies.

V. Discussion

As expected, the subjective ratings across 24 participants after the first task of each category show that there are significant differences between the two designed load levels (Fig. 5(a)). Comparing with their ratings after the first task, there were some variations in the 24 participants' ratings at the end of each group of tasks (Fig. 5(b)). These variations might be due to learning effects or fatigue effects. However, the significant differences between low and high load levels did not change according to the statistical tests. Therefore, their psychological and behavioral signals are expected to be different under the different load levels. It seems reasonable to adopt the designed load levels as the ground truth in modelling and evaluation.

When examining the four event-based features individually (E1-E4 in the 2nd and 3rd block in TABLE 1), we found that three of them were better at discriminating load types than load levels. Among them, the sequence-based 2-gram features achieved the best accuracy, 82%, in classifying four task load types, followed by the proposed intensity-based event feature. This observation also applies to conventional statistical features (B1 in TABLE 1), which achieved 88%, the best accuracy except for fused features (the 4th block in TABLE 1), for task type recognition, better than load level recognition, 75%. This suggests that task load type also affects participants' physiological and behavioral signals and the performance is comparable with that in human activity recognition [24,31]. It could be beneficial to assess task load and type at the same time to avoid the confusion in interpreting them.



(a) (b)

Fig. 5. Boxplots of the subjective ratings over 24 participants for the four task load types for (a) subjective ratings conducted at the end of the first task of each category and (b) subjective ratings along time conducted at the end of selected tasks during experiment. 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.

TABLE I. AVERAGE ACCURACY (95% CONFIDENCE INTERVAL) SUMMARY FOR TASK LOAD LEVEL AND TYPE RECOGNITION USING A LEAVE-ONE-PARTICIPANT-OUT SCHEME AND USING EITHER LSTM OR CNN-LSTM NETWORK.

	Feature dimension	**Beck-end type**	**Load level regardless of type (2-class)**	**Load type regardless of level (4-class)**	**Load level and type recognition (8-class)**
B1: selected statistical feature per task based on training data	25	LSTM	0.75 (0.03)	0.88 (0.05)	0.73 (0.05)
B2: onset event words	1	CNN-LSTM	0.59 (0.05)	0.66 (0.08)	0.49 (0.07)
B3: raw event signals	5	CNN-LSTM	0.76 (0.04)	0.62 (0.10)	0.42 (0.07)
B4: event duration signals	4	CNN-LSTM	0.73 (0.04)	0.64 (0.08)	0.41 (0.06)
B5: event intensity signals	4	CNN-LSTM	0.76 (0.04)	0.70 (0.06)	0.56 (0.07)
B6: raw event signals + intensity + duration	13	CNN-LSTM	0.76 (0.04)	0.76 (0.08)	0.45 (0.05)
E1: selected sequence features per task 2-gram based on training data [19]	25	LSTM	0.72 (0.03)	0.82 (0.05)	0.57 (0.04)
E2: coordination features per task [26]	2	LSTM	0.79 (0.02)	0.60 (0.04)	0.54 (0.03)
F1: proposed sequence + coordination features per task	27	LSTM	0.80 (0.02)	0.86 (0.04)	0.70 (0.03)
E3: proposed duration features per task	8	LSTM	0.69 (0.03)	0.77 (0.04)	0.54 (0.04)
E4: proposed intensity features per task	8	LSTM	0.74 (0.02)	0.80 (0.03)	0.62 (0.03)
F2: proposed duration + intensity features per task	16	LSTM	0.77 (0.02)	0.85 (0.03)	0.66 (0.03)
F3: proposed coordination + duration + intensity features per task	18	LSTM	**0.84 (0.02)**	0.86 (0.03)	0.74 (0.04)
F4: proposed all event-based features per task	43	LSTM	**0.84 (0.02)**	**0.89 (0.03)**	**0.76 (0.03)**

On the contrary, only the coordination-based features (E2 in the 2nd block in TABLE 1) proposed in [26] (except baselines) were far better at classifying task load levels than task load types. The accuracy, 79%, was the best accuracy for load level recognition regardless of type, except the fused features. For the two proposed duration-based and intensity-based event features (E3 & E4 in the 3rd block in TABLE 1), they are capable of classifying load levels and load types, as their performances were comparable with the baseline performance for load level classification, and better than the baseline performance for load type classification, and for load level and load type classification (the 1st block in TABLE 1). When comparing the two proposed features, we found that the intensity-based feature was more effective as it ranked second among the four event-based features for load levels and load types classification respectively, and was best for both load levels and load types, 62% for 8-class classification. Since the information carried by these event-based features is different, this indicates that the proposed two features can contribute to task load assessment along with other event-based features.

To examine whether these four event-based features contain complementary information for both task load level and load type classification, we found that by concatenating them, the accuracy was significantly improved in classifying load levels (2-class), load types (4-class), and both load level and load type (8-class) as shown in the last row of the 2nd to 4th block in TABLE 1. The best accuracy was 84% for two load levels, 89% for four load types, 76% for both load level and type, outperforming all the baselines. Comparing with other studies, where only load levels were classified, our performance was comparable to theirs using a large number of modalities, including EEG [9] or using deep learning techniques with eye images [11]. It is worth noting that in our study, we trained the load level model with different tasks, unlike a single task in previous studies, therefore, our event-based features are more representative of different task types.

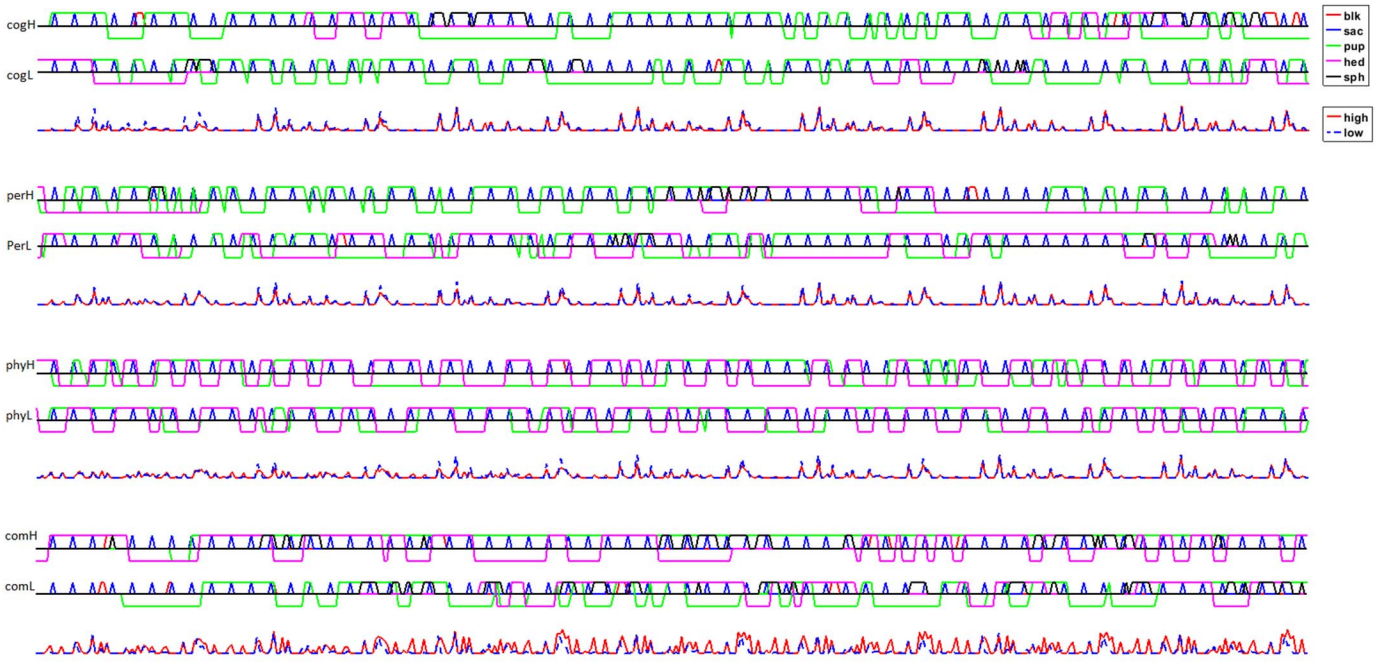


Fig. 6. A visualization of the gate activations (the third row in each group of three) in an intermediate layer (the masking layer, Fig. 4(c)) from a typical participant and the corresponding raw event signals (the first two rows in each group of three) obtained from blink, saccade, pupil size, head movement and speech data under the conditions of cognitive high (cogH) and low (cogL) load levels, perceptual high (perH) and low (perL) load levels, physical high (phyH) and low (phyL) load levels, and communication high (comH) and low (comL) load levels, respectively. The horizontal axis is time. Due to space limitations, only the first 20 seconds were plotted. This demonstrates how distinctive these features were in low and high load levels and across different task load types.

Among the baselines (the 1st block in TABLE 1), conventional statistical features also achieved good performance, especially for load type classification, and load level and type classification, where the performance was close to best. This is not surprising considering they are often employed in psychophysiological state recognition. However, it is important to mention that the 25-D statistical features (B1) were *highly data driven* since *different features* were selected in each of the leave-one-participant training process, while the 18-D event-based features (F3) were fixed for all participants. This is one advantage of the event-based features, which represent task load well across different tasks and different participants, indicated by small confidence intervals (0.02-0.04), as well as being compact and easily interpretable.

Surprisingly, the performance achieved by CNN-LSTM, i.e. self-learned features, was significantly worse than that using the fused event-based features. Specifically, when the input was the raw event signals, which the event-based features were also extracted from, its performance was 4% lower than that of the fusion of sequence-based and coordination-based event features in load level recognition, 24% lower in load type recognition, and 28% lower in both load level and type recognition. This indicates that the deep learning process did not learn better features than our hand-crafted features. We examined this closely by visualizing the gate activations of the intermediate layer before the LSTM1 layer from a participant whose performance was close to the average. As shown in Fig. 6, the features learned for low and high load levels were distinctive to some extent (i.e. the blue and red curves are not overlapping sometime), evidenced by the ordinary performance for load level classification (76%). However, the patterns learned for each of the four task load types are not very distinguishable. More specifically, the patterns for cognitive, perceptual, and physical task load look very similar. This explains why the performance of load type recognition and both load level and load type recognition were not very good, 62% and 42% respectively as shown in TABLE 1.

Other direct comparisons to examine whether the deep learned features were better than the hand-crafted features using the same signals included: the onset event words along time versus sequence-based event features during tasks (B2 vs E1); event duration values along time versus the duration-based features during tasks (B4 vs E3); and event intensity values along time versus intensity-based features during tasks (B5 vs E4). Except for the case of load level recognition, the event-based features (E1-E4) consistently achieved better accuracy than deep learned features (B2-B6) by 6-16%. For load level recognition, the deep learned features (B2-B6) achieved slightly better accuracy (2-4%) than duration and intensity-based event features (E3 & E4) but 13% worse than the sequence-based event feature (E1). The reason could be three-fold. It needs to learn a consistent pattern from four different task types, which is more difficult than learning the pattern from a single task type. Therefore, the pattern learned may not be effective enough if the data were not sufficiently large. Another reason could be that the patterns exhibited in detailed sequences were not very effective compared with the pattern extracted from the whole task. There could be a lot of impromptu behaviors during tasks. The last reason could be that the convolution process is not the best option for extracting useful information from these event signals, or psychological and behavioral signals in general.

Even with end-to-end deep leaning with the 9-dimensional raw head movement data or 10s raw speech data for load level

recognition in our initial experiment, we did not find better performance than using the raw event signals but consumed large computing resources. Knowing this, we can design new deep neural networks which are more suitable for psychological and behavioral signals to assess psychological state in the future.

One of the limitations is that we only used speech onset events from a large amount of acoustic data, so some information was ignored. For example, as the eye camera was close to the nose, sometimes deep breaths can be heard from the audio. As in the conventional statistical features, 39 MFCC features were used and three of them were found to be the most frequently selected from each leave-one-participant-out training. This could explain why the statistical features achieved good accuracy.

VI. CONCLUSION

In this work, we proposed two event-based features, namely, duration-based and intensity-based event features for task load assessment. Our results show that these two features have comparable discriminability to sequence-based and coordination-based event features, as well as the conventional statistical features in all or a few cases. When fusing the proposed two features with the other event-based features, we achieved state-of-the-art performance considering that different load types were used for task load assessment. This performance was better than that using the conventional statistical features and better than that employing a deep learning architecture, CNN-LSTM, with raw event signals as input in all cases. Considering the advantages of these event-based features, being more representative, compact, interpretable, and obtained from wearable sensors, it is promising to use them to assess task load in longitudinal and daily life to improve human and computer interaction. Future work could discover more behavior events and combine deep learning techniques to further improve the task load assessment performance.

ACKNOWLEDGMENT

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.

REFERENCES

[1] E. T. Solovey, D. Afergan, E. Peck, S. W. Hincks, R. J. K. Jacob, "Designing implicit interfaces for physiological computing: guidelines and lessons learned using fNIRS", ACM Trans. Comput.-Hum. Interact, 21, 6, Article 35, 27 pages, 2015.

[2] B. Dumas, D. Lalanne, S. Oviatt, "Multimodal interfaces: A survey of principles, models and frameworks", Human machine interaction, Springer, Berlin, Heidelberg, 2009, pp. 3-26.

[3] PupilLab hardware: https://pupil-labs.com/

[4] M. Soleymani, S. Asghari-Esfeden, Y. Fu, M. Pantic, "Analysis of EEG Signals and Facial Expressions for Continuous Emotion Detection", IEEE Transactions on Affective Computing, vol. 7, no. 1, 2016, pp. 17-28.

[5] E. Haapalainen, S. Kim, J.F. Forlizzi, A.K. Dey, "Psycho-physiological measures for assessing cognitive load", UbiComp, Copenhagen, 2010.

[6] G. K. Verma, U. S. Tiwary, "Multimodal fusion framework: A multiresolution approach for emotion classification and recognition from physiological signals", NeuroImage, 102, 2014, pp. 162-172.

[7] M. Wöllmer, M. Al-Hames, F. Eyben, B. Schuller, G. Rigoll, "A multidimensional dynamic time warping algorithm for efficient multimodal fusion of asynchronous data streams", Neurocomputing, 73(1-3), 2009, pp. 366-380.

[8] M. A. Hogervorst, A-M. Brouwer, J. B. F. van Erp, "Combining and comparing EEG, peripheral physiology and eye-related measures for the assessment of mental workload", Front Neuroscience, 8, 2014, p. 322.

[9] L. Zhang, J. Wade, D. Bian, J. Fan, A. Swanson, A. Weitlauf, Z. Warren, N. Sarkar, "Cognitive load measurement in a virtual reality-based driving system for autism intervention", IEEE Transactions on Affective Computing, 8, 2, 2017, pp. 176-189.

[10] J. Beatty, "Task-evoked pupillary responses, processing load, and the structure of processing resources", Psychological Bulletin, 91, 1982, pp. 276-292.

[11] L. Fridman, B. Reimer, B. Mehler, W. T. Freeman, "Cognitive Load Estimation in the Wild", CHI, Montreal, 2018.

[12] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, S. Zafeiriou, "Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network", ICASSP, Shanghai, 2016.

[13] S. Chen, and J. Epps, "Atomic Head Movement Analysis for Wearable Four-Dimensional Task Load Recognition", IEEE J. Biomed. Health Informatics, 23(6), 2019, pp. 2464-2474.

[14] P. Ren, A. Barreto, Y. Gao, M. Adjouadi, "Affective Assessment by Digital Processing of the Pupil Diameter", IEEE Transactions on Affective Computing, Vol. 4, No. 1, 2013, pp. 2-14.

[15] P. Celka, et al. "Wearable biosensing: signal processing and communication architectures issues", Journal of Telecommunications and Information Technology, 2005, pp. 90-104.

[16] A. Baddeley, "Working memory: looking back and looking forward", Nature Reviews Neuroscience, vol. 4, 2003, pp. 829-839.

[17] C. D. Wickens, "Multiple Resources and Mental Workload. Human Factors", The Journal of the Human Factors and Ergonomics Society, vol. 50, 2008, pp. 449-455.

[18] F. Paas, J. Van Merriënboer, "Instructional control of cognitive load in the training of complex cognitive tasks", Educational Psychology Review, vol. 6, 1994, pp. 351-371.

[19] S. Chen, J. Epps, "Task load estimation from multimodal head-worn sensors using event sequence features", IEEE Transactions on Affective Computing, 2019.

[20] J. Epps, S. Chen, "Automatic Task Analysis: Towards Wearable Behaviometrics", IEEE System, Man and Cybernetics Magazine, 4,4, 2018, pp. 15-20.

[21] K. J. Vicente, Cognitive work analysis: Toward safe, productive, and healthy computer-based work, CRC Press, p. 70, 1999.

[22] H. Gunes, and B. Schuller, "Categorical and dimensional affect analysis in continuous input: Current trends and future directions", Image and Vision Computing, 31(2), 2013, pp.120-136.

[23] S. G. Hart, L. E. Staveland, "Development of NASA-TLX (task load index): Results of empirical and theoretical research", In Hancock, P. S. and Meshkati, N. (Eds.), Human mental workload, Amsterdam: North-Holland, 1988, pp. 139-183.

[24] H. Steil and A. Bulling, "Discovery of everyday human activities from long-term visual behavior using topic models", UbiComp, 2015.

[25] W. Fitzgerald, R. J. Firby, M. Hannemann, "Multimodal event parsing for intelligent user interfaces", IUI, Miami, 2003.

[26] S. Chen, J. Epps, "Multimodal Coordination Measures to Understand Users and Tasks", TOCHI, 2020.

[27] J. Lagarde, J. A. S. Kelso, "Binding of movement, sound and touch: multimodal coordination dynamics". Experimental brain research, 173(4), 2006, pp. 673-688.

[28] S. Chen, J. Epps, "Efficient and Robust Pupil Size and Blink Estimation from Near-field Video Sequences for Human-Machine Interaction", IEEE Trans. Cybernetics, 44(12), 2014, pp. 2356-2367.

[29] openSMILE toolbox, http://opensmile.audeering.com/

[30] D. D. Salvucci, J. H. Goldberg, "Identifying Fixations and Saccades in Eye-tracking Protocols", ETRA, New York, 2000.

[31] L. Bao, S. S. Intille, "Activity Recognition from user-annotated acceleration data", Pervasive Computing, 2004, pp. 1-17.