Real-Time Assessment of the Cross-Task Mental Workload Using Physiological Measures During Anomaly Detection

Guozhen Zhao, Yong-Jin Liu, *Senior Member, IEEE*, and Yuanchun Shi, *Senior Member, IEEE*

***Abstract*—The ability to detect anomalies in perceived stimuli is critical to a broad range of practical and applied activities involving human operators. In this paper, we propose a real-time physiological-based system to assess the cross-task mental workload during anomaly detection. Forty participants were recruited to detect anomalous images from a set of different distracting images (Task I) and abnormal activities from surveillance videos (Task II). In Task I, the task difficulty levels were manipulated by changing the number of anomalies/distracting stimuli (15, 21, 28, or 36) with and without time constraints (i.e., $4 \times 2 = 8$ task difficulty levels). Physiological and behavioral data from four task difficulty levels were divided into four categories according to subjective ratings of the mental workload. The support vector machine (SVM) classifiers were trained on these data to predict the mental workload categories of: 1) the same four task difficulty levels (within level); and 2) the other four task difficulty levels in Task I (cross level). Within-level classifications (with an average of 95.29%) were more accurate than cross-level classifications (average of 72.2%), which were much more accurate than random level classifications (25%). In Task II, the same participants monitored one, two, or four video clips simultaneously in accordance with three task difficulty levels. The same physiological signals were processed for real-time recognition of a participant's mental workload after he or she completed each activity detection task. The three-class SVM classifiers were trained on physiological data from Task I to predict the mental workload categories of the Task II (cross task), achieving an overall classification accuracy of 53.83%, compared to a 33.33% accuracy at random. These results are discussed in terms of their implications for developing situation-aware recognition systems of the mental workload and adaptive human–computer interaction platforms.**

***Index Terms*—Anomaly detection, cross task, human–computer interaction, mental workload, physiological measures, workload classification.**

Manuscript received June 29, 2016; revised April 30, 2017, September 1, 2017, and December 28, 2017; accepted January 22, 2018. Date of publication February 20, 2018; date of current version March 13, 2018. This work was supported in part by the National Key Research and Development Plan (2016YFB1001200) and in part by the National Natural Science Foundation of China (31771226, U1736220, 61725204, 61521002). This paper was recommended by associate editor Yili Liu. *(Corresponding author: Yong-Jin Liu.)*

G. Zhao is with the CAS Key Laboratory of Behavioral Science, Institute of Psychology, Beijing 100101, China, and also with the University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: zhaogz@psych.ac.cn).

Y. J. Liu and Y. Shi are with the Tsinghua National Lab for Information Science and Technology, Department of Computer Science and Technology, Tsinghua University, Beijing 100044, China (e-mail: liuyongjin@tsinghua.edu.cn; shiyc@tsinghua.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/THMS.2018.2803025

I. Introduction

ANOMALY detection in images/videos is applicable in a variety of domains and has been widely studied in computer vision, pattern recognition, and engineering psychology. A number of automatic anomaly detection algorithms have been proposed [1]. However, there is little work that studies the human factors in the process of anomaly detection. In many anomaly detection situations, such as human–computer monitoring, the role of a human operator is critical. We list following three examples.

1) *X-ray security screening:* The security checkpoints for X-raying passenger bags are the key element in transportation security. Despite great improvements in technological equipment (e.g., high-resolution X-ray machines), the decision as to whether a piece of luggage can enter the gate or not is still made by a human screener under a critical time limit [2].
2) *Surveillance video analysis:* To train machine-learning algorithms to analyze an extremely large amount of surveillance video data and detect abnormal activities, previously observed activities need to be analyzed, annotated, and profiled by human experts, thus requiring substantial human operator effort to obtain an accurately labeled training dataset in a limited amount of time [1].
3) *Medical diagnosis:* Tasks such as radiological diagnosis of chest X-rays or CT images aim to detect a large number of anomalies that signal different diseases. Although computer-aided diagnosis techniques have been developed, these images frequently need to be judged by doctors themseland also with the ves [3]. In many medical contexts, there is time pressure on radiology doctors to act quickly because of the sheer volume of work.

The ability to detect anomalies in perceived stimuli by a person is critical in these applications. Anomalous features are likely to weigh heavily in the cognitive tasks of visual search, signal detection, pattern classification, and discrimination. Detecting subtle anomalies in naturalistic stimuli is challenging even after extensive experience. For example, in [4], medical students and expert diagnosticians were asked to describe photographs showing prototypical symptoms of pancreatitis. A surprising finding was the widespread failure of participants to notice supposedly obvious features.

During anomaly detection, a human operator knows what should be considered normal items and discriminates anomalies

from a field of normal stimuli. This requires many resources and close cooperation between the stimuli perception, attention, working memory, and decision-making ability involved [5]. Because anomalies are typically characterized by ambiguity as to what should be considered a feature, a person may focus on features that are intrinsically salient but irrelevant to the task [3], which increases the task difficulty and complexity.

Behavioral measures, such as detection accuracy and reaction time (RT), might not be reliably sensitive to changes in task demand. Implicit is the belief that as task difficulty increases, anomaly detection performance usually decreases: reaction time and error increase, and fewer anomalies are detected per unit time. However, there exists strong evidence to show that this is not always the case for monitoring, vigilance, or troubleshooting applications [6]. As a result, mental workload (MW) that is more sensitive to quantify the mental cost of performing anomaly detection tasks is introduced in this paper to predict human operator and system performance.

Recently, physiological signals have attracted attention in the quest to understand a human operator's cognitive and mental processes in performing a task (e.g., [7]–[9]) due to their accurate and perceiver-dependent objective data. Physiological signals are measured via instruments that read bodily events, such as electroencephalography (EEG), heart rate change, electrodermal activity (EDA), and cardiac output. EEG-based indices are sensitive to subtle changes in the mental workload. However, traditional EEG systems are somewhat difficult to use, requiring preparation of the skin, application of conductive gel, and cleaning of the cap afterwards. Although alternative electrode systems (e.g., dry electrode) have been developed, it still takes time to provide evidence of good signal quality, comparable to that of standard gel-based electrodes [10]. As a comparison, peripheral physiological signals provide alternative approaches for quick and practical application. Recent developments in the field of wearable sensors and systems have resulted in devices that require less preparation time and are more comfortable for the wearer [11].

A key challenge in the development of a physiological-based system to recognize the mental workload is the establishment of a generalized model for different difficulty levels and tasks that the human operator has not yet experienced. The cross-task mental workload classification is the process in which classifiers are trained on physiological features under one task and applied in other tasks to recognize mental workload states under different taskloads. In this paper, we present a real-time physiological-based adaptive system to assess the cross-task mental workload during anomaly detection, and we achieve better classification accuracies than chance level. The proposed real-time assessment method and system can be used to develop situation-aware recognition systems of the mental workload and adaptive human–computer interaction platforms. For example, if a human screener in an X-ray security checkpoint is recognized as having a high level of mental workload, a warning message can be automatically sent to the security officer to suggest a work shift or adjust the task demands (e.g., reduce the amount of luggage X-rayed) via an adaptive human–machine interface.

II. Related Work

A. Mental Workload

Mental workload reflects the interaction of mental demands imposed on operators by tasks they attend to [12] or the mental cost of accomplishing the task demands [13]. Mental workload depends upon the human operator and the interaction between the operator and task. The same task demands do not result in an equal level of the workload for all individuals. Individuals can adapt their behaviors and cope with increasing demands to keep the performance at the same level with an increase in effort.

Workload results from the aggregation of many different demands, and there is no single measure that can evaluate all of its components [14]. Although mental workload is difficult to directly observe, the previous research has suggested that it can be inferred from the measurement of physiological processes [15]. Compared to subjective measures, physiological indices have better performances in the aspects of sensitivity, diagnostic ability, and nonintrusiveness, and they provide online methods for measuring mental workload in the practical and applied activities involving human operators [16].

Both time-domain and frequency-domain measures of electrocardiograms (ECGs) have been utilized to obtain information about the workload. In the time domain, the average heart rate (i.e., the number of beats per minute) during task performance compared to a rest-baseline measurement is an accurate measure of metabolic activity [17]. Compared to time-domain analysis, the frequency analysis of heart rate variability (HRV) provides additional information regarding the biological control mechanisms [18]. A decrease in power in the midfrequency band (also called the "0.1-Hz component" after the main frequency component) has been shown to be related to mental effort and task demands [19]. It is sensitive to not only task–rest differences but also relatively low or moderate changes in the mental workload. Second, endogenous eye blinks measured by an electrooculogram (EOG) are meaningful indicators of the mental workload. Previous studies have concluded that increased eye blink rate is the most useful in the assessment of visual demands [20]. Other eye blink data, such as blink duration and latency, have been analyzed and used as workload measures in a series of studies [21]. Third, measures of respiration (RSP) provide an index of energy expenditure. Evidence has been found that the respiration rate increases as a result of increased mental workload or temporal demands [22]. Fourth, Kramer [23] concludes in his review that EDA appears to be sensitive to general information processing. Both mean skin conductance response (SCR) amplitude and frequency have been extracted and used as indices of the mental workload in the literature [24]. Finally, the photoplethysmogram (PPG)-based vascular response index (e.g., a relative amplitude ratio between two contours in a PPG waveform) and changes in blood pressure variability have been employed to assess cognitive load and mental stress [25].

B. Mental Workload Classification

The aforementioned various physiological measures make it possible to recognize different levels of mental workload [7],

[26]–[28]. For example, Zhang *et al.* [7] proposed an adaptive support vector machine (SVM)-based method to classify operator mental workload into a few discrete levels. Physiological signals were recorded continuously while subjects performed a process control operation. Combining adaptive exponential smoothing algorithms and bounded SVM methods, their model was able to recognize mental workload every 5 seconds with higher temporal resolution and cross subject and cross-trial generalizability. Wilson and Russell [26] used an artificial neural network (ANN) to classify the mental workload of highly trained operators in a simulated air traffic control task. The authors manipulated different levels of task difficulty by varying either the volume of traffic or the complexity of traffic at each of three levels. The ANN model was trained for approximately 6 h for each operator and achieved an average classification accuracy of 80%. Classification accuracies improved to an average of 85.8% for ANNs trained on within-difficulty manipulation (i.e., ANNs were trained on the three levels of traffic volume and tested on the same three levels).

Moreover, physiological-based systems have been proposed to monitor an operator's mental workload in real time [29]– [31]. For example, Wilson and Russell [29] used ANNs to classify operator states on a multitask combination of manual tracking, visual and auditory monitoring, and the dynamic resource allocation task. Two difficulty levels were manipulated by varying the number of events that occurred within a 5-min trial. The trained ANNs were used to determine the difficulty level of the task while it was being performed in two 5-min blocks per level. They achieved real-time classification accuracies ranging from 82% under the low-workload condition to 86% under the high-workload condition. These findings and systems facilitate the development of real-time assessment and classification of mental workload based on physiological signals that detect performance degradation or breakdown in safety-critical human–machine systems.

Recently, the state of the art research tried to establish a generalized model for different difficulty levels and tasks that the operator has not yet experienced. Baldwin and Penaranda [31] first attempted to train classifiers with two difficulty levels in one working memory task, and then, predict two workload levels in a different working memory task when the operator has little experience in performing either task. Classification accuracies were higher when the classifiers were trained on examples from the same task ($M = 87.1\%$) than a set containing the to-be-classified task ($M = 85.3\%$). Cross-task classification accuracies (44.8% on average) were much lower than the within-task accuracies, indicating consistent misclassification for certain tasks in some individuals (50% accuracy at random).

To improve cross-task classification accuracies, in this study, we manipulated a number of task difficulty levels to increase the differences in mental workload. Two anomaly detection tasks were designed to contain a combination of visual monitoring, pattern discrimination, manual tracking, and resource allocation among different visual stimuli. Moreover, we set up time constraints under certain experimental conditions and provided additional bonuses for participants with better detection performance. These approaches aimed to evoke a higher level of arousal, mental effort, and energy demands, which are associated with mental workload [32]. We propose a physiological-based system to assess: 1) within- and cross-level mental workload during anomalous image detection; and 2) real-time cross-task mental workload during anomaly activity detection. Our system achieves high within-level classification accuracies of 95.29%, on average, and acceptable cross-level accuracies of 72.2%, on average, compared to 25% for random chance. Additionally, our cross-task classification accuracies (53.83% on average) were much higher than random chance (33.33%).

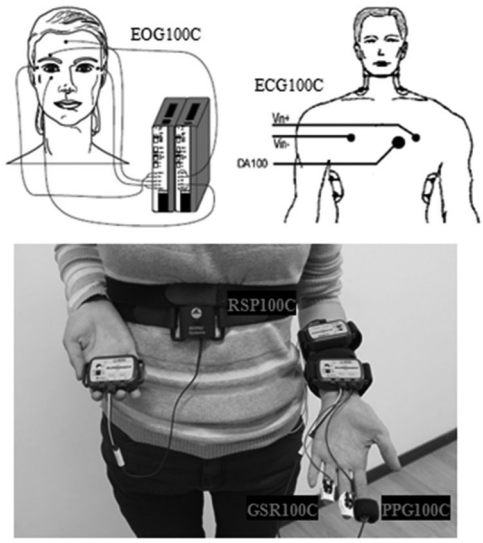

Fig. 1. Electrode locations and connections for EOG (upper left), ECG (upper right), RSP (bottom, real amplifier module), GSR (bottom, real amplifier module), and PPG (bottom, real amplifier module).

III. Real-Time Physiological-Based System of the Cross-Task Mental Workload

The objective of this study is to establish a real-time, generalized, and multimodal physiological system to recognize different levels of mental workload during anomaly detection. To achieve this objective, we employ various physiological measures and behavioral responses to train an SVM classifier when the operator detects anomalous images. Trained classifiers are tested on data from different difficulty levels of the same task and from an abnormal activity detection task to verify the effectiveness and generalizability of our system.

A. Acquisition of Physiological Signals

The proposed system employs five amplifier modules to acquire physiological signals, including an electrocardiogram (ECG100C), electrooculogram (EOG100C), respiration pneumogram (RSP100C), electrodermal activity (GSR100C), and photoplethysmogram (PPG100C) (see Fig. 1). The ECG100C records a standard LEAD I with an additional R-wave detector to calculate the real-time R-R interval, R-wave amplitude, and heart rate. The EOG100C is recorded via Ag–AgCl electrodes placed above and below the participant's right eye to track his/her vertical eye movement. The RSP100C is placed around

the body at the level of maximum respiratory expansion (generally approximately 5 cm below the armpits) to measure abdominal or thoracic expansion and contraction. The GSR100C provides a measure of skin resistance by positioning two Ag–AgCl electrodes on the distal phalanges of the middle and index fingers. The PPG 100C indirectly indicates the point of maximal blood density in the respective location and blood pressure by comparing the point of R-wave onset in the ECG to the point of maximum blood density.

B. Signal Processing and Feature Extraction

All the physiological signals are recorded at a 1000-Hz sampling rate and later downsampled to 256 Hz to reduce the real-time data processing time. Basic filter methods with recommended settings are applied to remove linear trends and artifacts. We compute channel statistics after removing each piece of segmented data to determine whether to remove this segment or the entire channel. In the case that abnormal waveforms may still exist (e.g., a distorted ECG waveform), we select normal ECG waveform patterns that the user recorded in the resting period and set them as a template. Then, we compare each piece of segmented data with the template to remove abnormal data based on the degree of correlation.

An external 5-V digital input is recorded simultaneously to mark the beginning and end of each anomaly detection task. Specifically, the presence of each anomaly detection task and the time a participant clicks the button to continue the next task generates a transient voltage change. Based on these digital inputs, continuous physiological signals are segmented to successive epochs, and each epoch (i.e., a few seconds, depending on task completion time) contains only one anomaly detection task. In each epoch, the first and last 10% of the physiological signals from the beginning and end of each task are excluded from the analyses to eliminate transience.

The BIOPAC MP150 system provides various presets to calculate and extract features from raw physiological signals in real time. For example, after setting up an ECG channel, the system allows us to select the R-R interval, heart rate, and R wave amplitude presets. The preset and customized calculation functions are able to extract all time-domain features and a majority of frequency-domain features (e.g., HRV). The other frequency-domain features are computed using the BIOPAC script function. Data are smoothed using the Hanning window and are transformed into power spectra with fast Fourier transfer analysis. The power in the target frequency range is integrated to obtain frequency-domain features. Because of large individual differences in physiological signals, extracted features are normalized between 0 (representing no mental workload) and 1 (representing maximum mental workload) for further analysis.

C. Classification of the Mental Workload

All features are extracted from peripheral physiological responses based on the extensive literature review of the mental workload. As shown in Table I, there are 42 physiological indices in total (i.e., 9 ECG features, 5 EOG features, 11 RSP features, 12 GSR features, and 5 PPG features). For a detailed description of these 42 extracted features, see Section II—related work and the literature [6], [21], [23], [33], [34]. In addition, there are three behavioral indicators (i.e., detection reaction time, miss rate, and false alarm rate).

TABLE I
FEATURES EXTRACTED FROM PHYSIOLOGICAL AND BEHAVIORAL DATA

Signal	Extracted Features
ECG (9)	Average and standard deviation of HR, R-R interval, square root of the mean squared differences of successive R–R intervals, R-wave amplitude, T-wave amplitude, HRV power spectrum in the low-frequency band [0.02, 0.06] Hz, midfrequency band [0.07, 0.14] Hz, and high-frequency band [0.15, 0.5] Hz
EOG (5)	Eye blink rate, duration, and latency, average and standard deviation of blink intervals
RSP (11)	Average and standard deviation of RSP, average of its derivative, average and standard deviation of RSP rate, four spectral power in the frequency bands from 0 to 1 Hz, average peak-to-peak interval and amplitude
GSR (12)	Average and standard deviation of SCL, average SCR amplitude and frequency, average peak-to-peak interval and amplitude, crossing rate of skin conductance, slow response in the frequency bands [0, 0.08] Hz and [0, 0.2] Hz, 4 spectral power in the frequency bands from 0 to 1 Hz
PPG (5)	Average and standard deviation of PPG, average peak-to-peak interval and amplitude, relative ratio of two contours in a PPG waveform
RT and Error (3)	Detection reaction time, miss rate, false alarm rate

Linear discriminant analysis (LDA) is performed for feature reduction, which projects high-dimensional features with labels into a low-dimensional space with good class separability by maximizing the Fisher separation criterion. Feature reduction is achieved by finding a subset of features that have the largest Fisher separation value. However, when the number of selected features is still larger than the number of observations, the within-class covariance matrix of the features becomes singular. To solve this problem, we apply the sparse LDA (SLDA) [35], which overcomes this limitation by performing LDA with a sparseness criterion imposed such that classification and feature selection are performed at the same time [36]. The SLDA algorithm will perform the elastic net regression with early stopping at a particular value (i.e., the value of the parameter STOP) of the L1 regularization parameter. Therefore, STOP is an integer that determines the desired number of nonzero variables [36], [37].

After the participant completes each anomaly detection task, he or she has to click the button to continue the next task, which will trigger the processing of physiological signal and feature extraction as described in Section III-B. The selected features are entered as input features to an SVM for classification. In our implementation using MATLAB, we apply the SpaSM toolbox for SLDA and the LIBSVM toolbox for the SVM. We optimize the parameter STOP in the SpaSM toolbox. In the LIBSVM toolbox, we choose the radial basis function kernel function and optimize the cost parameter c and the gamma parameter g using SVMcgForClass [38]. In this study, the SVM classifiers are trained on data from some task difficulty levels of one anomaly detection task, and then, tested on the following data:

Fig. 2. Set of 15 images was randomly arranged in a 5 × 3 matrix. In this example, there was only one anomalous image (snow-covered trees) at the cross section of the second row and third column among 14 distracting images (Snow Mountain).

1) data from the same task difficulty levels (i.e., within-level, offline classification);
2) data from the other task difficulty levels of the same task (i.e., cross-level, offline classification);
3) data from the other anomaly detection task (i.e., cross-task, real-time prediction of mental workload).

Obviously, cross-task classification is more challenging than cross-level classification, which in turn is more difficult than within-level classification.

IV. Methods

A. Participants

We recruited 40 undergraduate and graduate students (20 males and 20 females) whose average age was 22.2 years (range = 19–26, SD = 1.76). Participants were screened to ensure that they were right handed with normal or corrected-to-normal visual acuity and hearing. We recruited novices who lacked sufficient knowledge to make use of top-down expectancies and focused on conspicuous features rather than the features relevant to the task.

B. Task Description

1) Anomalous Image Detection: In Task I, participants were asked to identify a variable number of anomalous images from a set of different distracting images (see Fig. 2). Our interest lies in two dimensions of task demands that influence the investment of mental resources: changing the number of anomalies and distracting stimuli (i.e., set size) and manipulating the time pressure (i.e., changing the information rate in essence). The image set size consisted of four levels: a set of 15 images arranged in a 5 × 3 matrix (i.e., a set of 5 × 3 images), a set of 7 × 3 images, a set of 7 × 4 images, and a set of 9 × 4 images. The time pressure consisted of two levels: self-paced or under time pressure. Take a set of 5 × 3 images, for example. At the beginning of each task, a sample image with a highlighted keyword was displayed for 3 s to reduce individual differences in the comprehension of image contents. Then, a set of 15 different images was arranged in a 5 × 3 matrix. Participants were instructed to select all anomalous images by clicking the left mouse button. Double clicking the left mouse button revoked the selection. After completing ten successive tasks, participants took a 2-min break, and then, started another trial.

The ratio of the number of anomalous images to the number of distracting images in each trial was 0.1. In the last example of a 5 × 3 matrix, there were 1–2 anomalous images in each task, amounting to 15 anomalous images in that trial (i.e., 15 anomalous images/150 images = 0.1). Similarly, there were 1–3, 2–4, and 3–5 anomalous images when a set of 21, 28, and 36 images were arranged in corresponding matrices, respectively. Because the number of anomalous images was unknown, participants had to perform an exhaustive search in each trial (i.e., check the entire set of images one by one). When experiencing time pressure, participants were required to complete the detection task as soon as possible without sacrificing accuracy.

Images (anomalous and distracting images) were selected from different categories of items, such as household equipment, home appliances, and natural scenes, including beaches, city streets, forests, highways, mountains, and offices. The natural scene images were originally used in [39], and the remainder were downloaded from Flicker under the Creative Commons license. Images were visually inspected and selected by three experienced experimenters to control the similarity between anomalous and distracting images (i.e., avoid pop-out anomalies) as well as the similarity between distracting images.

2) Anomaly Activity Detection: In Task II, the same participants were instructed to monitor one or a few video windows in accordance with different task difficulties and to identify all abnormal events that were publicly available in the UCSD Anomaly Detection Dataset [40]. This video dataset involves bidirectional pedestrian traffic from two camera viewpoints. The crowd density in the walkways is variable, ranging from sparse to very crowded. In this dataset, there are 50 video clips containing only normal pedestrian frames and 48 video samples containing at least some anomalous frames, such as the presence of abnormal objects (e.g., bikers, skaters, small carts), anomalous pedestrian motions and spatial abnormalities (e.g., walking across a walkway). Each video clip includes 200 individual frames in TIFF format. Each video was played at 20 frames per second and lasted for 10 s.

The anomaly detection task was manipulated with three levels of difficulty: only one video clip was played at the easy level, two were played simultaneously at the medium level, and four were played simultaneously at the difficult level. Fig. 3 demonstrated an example of the easy level of task difficulty, in which only one video sequence was played. Each trial was constituted by 50% normal and 50% abnormal activities for each level of task difficulty. More specifically, when only one video clip was played (easy level), five normal and five abnormal activities were randomly presented in a trial. When two video clips were played at the same time (medium level), ten normal and ten abnormal activities were randomly presented. For each detection task in a trial, two video clips could be any of the three normal–abnormal combinations (e.g., 2–0, 0–2, or 1–1), in case, participants could predict the presence of normal or abnormal events. The probability of the presence of the 1–1 normal–abnormal combination was higher than that of the other two combinations. Similar to the difficulty level, there were 20 normal and 20 abnormal

Fig. 3. Example of normal pedestrians walking and an abnormal biker in the walkway (marked by a red rectangle). All video clips were downloaded from the UCSD Anomaly Detection Dataset [40].

activities in a trial, three normal–abnormal combinations (e.g., 3–1, 1–3, or 2–2) and a higher chance of the presence of the 2–2 normal–abnormal combination for each detection task.

C. Apparatus

Physiological data were continuously recorded using the MP150 data acquisition system (BIOPAC systems Inc.). The MP150 system recorded multiple channels with variable sample rates to a connected Dell Workstation (OPTIPLEX 380, Intel Pentium E5800 at 3.2 GHz). Videos and images were presented at the center of a 22-in touch screen with a recommended 1680 $\times$ 1050 pixel resolution. The touch screen was located 60 cm from the participants' eyes.

D. Experimental Procedure

As shown in Fig. 4, participants were first asked to sign a consent document before engaging in this study. In Task I, participants went through a 15–20 min practice session to familiarize themselves with the detection task and user interface and to ensure their detection accuracy would maintain a stable level before the formal test.

During the formal test, participants were first instructed to take a 20-min rest while keeping their eyes open (i.e., the baseline condition). In each trial, participants completed ten consecutive tasks. After each trial, participants were asked to assess their mental workload, immediately followed by a 2-min break. There were 16 trials in total (4 set sizes $\times$ with/without time pressure $\times$ 2 replications with different images), and the run order was counter-balanced to avoid a potential confounding effect. All participants experienced Task I first because the data from Task I were collected to train the classifier that was used to recognize the real-time state of the mental workload in Task II. In Task II, participants were provided several video clips in the practice session to familiarize themselves with the detection task, normal pedestrians, and all possible abnormal activities. Similarly, participants first took a 20-min rest while keeping their eyes open (baseline condition). Each trial of the formal test consisted of ten consecutive detection tasks, with a short break (5, 10, or 20 seconds with an increased level of task difficulty) between tasks. During the short break, participants were asked to write down the type of anomaly (e.g., biker, skater, small cart, walking across a walkway) using simple letters or symbols in the corresponding position of the video window. Each difficulty level appeared once, and the run order was counter balanced. The entire experiment lasted 2.5–3 h. Participants were paid USD$10 per hour and an additional USD$10 bonus was provided for the top three participants in terms of their detection performance.

E. Measurement

Physiological signals were continuously recorded during all experimental trials and resting periods. In addition to the physiological measures, two behavioral responses were recorded in Task I. Detection reaction time (RT) per image measured the amount of time (in seconds) that a person spent looking for anomalous images divided by the image set size. Errors occurred when an observer failed to identify an anomalous image (i.e., miss) or chose a distracting stimulus as a target item (i.e., false alarm). Detection error ratio (i.e., the number of errors divided by the image set size) was used to reflect detection accuracy. Self-report assessments of the mental workload were obtained using the NASA-Task Load Index (TLX) rating scale [41]. Self-ratings after each task difficulty level compared to the baseline assessment (i.e., task-rest difference) was used as the perceived mental workload during each task difficulty level.

F. Data Analysis

A repeated measures analysis of variance (ANOVA) was performed with the image set size and time pressure as two within-subjects factors. Significant interactions or main effects were followed-up with *post hoc* pair-wise comparisons (e.g., Bonferroni's test) to assess the effects each factor had on the dependent variables. When the sphericity assumption was not met, the Greenhouse–Geisser correction was applied for the repeated measures ANOVA, which had more than one degree of freedom to control type-I error.

Eight task difficulty levels were involved in Task I: image set size (15, 21, 28, or 36) $\times$ with (W)/without (N) time pressure. The following three classifications were of interest to us (see Table II).

1) *Within level:* Physiological and behavioral data from four task difficulty levels (15 N, 21 W, 28 N, 36 W) were divided into four categories according to subjective ratings of mental workload. The four categories captured the entire range of the mental workload (15/21/28/36 $\times$ W/N): two extremes (15 N and 36 W) and two equidistant levels from these extremes (21 W and 28 N). The SVM classifiers were trained on these physiological and behavioral data to predict the mental workload categories of the same four task difficulty levels in Task I.
2) *Cross level:* Physiological and behavioral data from four task difficulty levels (15 N, 21 W, 28 N, 36 W) were

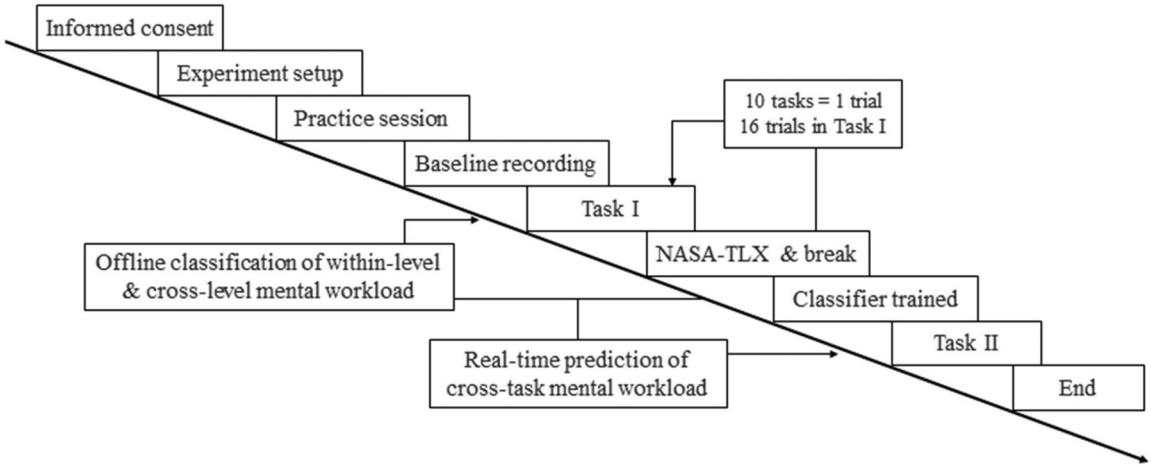

Fig. 4. Framework procedure and flowchart of the experiment.

TABLE II
TASK DIFFICULTY LEVELS AND DATA USED IN THE CLASSIFICATION OF WITHIN- AND CROSS-LEVEL AND CROSS-TASK MENTAL WORKLOAD

	Task I	Task II
Task Description	Detect anomalous images from a set of different distracting images	Detect abnormal activities from surveillance video streams
Task difficulty levels	*Eight Levels:* Image set size (15, 21, 28, or 36) × with (W)/without (N) time pressure	*Three Levels:* One, two, or four video streams were played simultaneously
Within level (offline classification)	*Training and Testing Data:* Data from four task difficulty levels (15 N, 21 W, 28 N, 36 W) were divided into four categories according to subjective ratings of MW	
Cross level (offline classification)	*Training Data*: Data from four task difficulty levels (15 N, 21 W, 28 N, 36 W) were divided into four categories according to subjective ratings of MW *Testing Data:* Data from four task difficulty levels (15 W, 21 N, 28 W, 36 N) were divided into four categories according to subjective ratings of MW	
Cross task (real-time classification)	*Training Data:* Data from eight task difficulty levels (15 N/W, 21 N/W, 28 N/W, 36 N/W) were divided into three categories according to subjective ratings of MW	*Testing Data:* Data from Task II were divided into three categories according to subjective ratings of MW

divided into four categories according to subjective ratings of the mental workload. The SVM classifiers were trained on these physiological and behavioral data to predict the mental workload categories of the other four task difficulty levels (15 W, 21 N, 28 W, 36 N) in Task I. In particular, we used all data from four levels (15 N, 21 W, 28 N, 36 W) for training at the same time, and predicted the mental workload categories of each level in 15 W, 21 N, 28 W, and 36 N.

3) *Cross task:* Physiological signals from eight task difficulty levels (15 N/W, 21 N/W, 28 N/W, 36 N/W) were divided into three categories according to subjective ratings of the mental workload. The SVM classifiers were trained on these physiological data to predict the mental workload categories of the Task II in real time using each individual participant's data.

V. RESULTS

A. Manipulation of Task-I Difficulty Levels

In Task I, the task difficulty levels were manipulated by changing the number of anomalies/distracting stimuli (15, 21, 28, or 36) and the time constraint (with or without). To confirm that such manipulation evoked different levels of the mental workload, we first examined the effects of image set size and time pressure on the subjective ratings of the mental workload, as well as *post hoc* comparisons.

The main effect of the image set size was significant for the subjective ratings of the workload ($F(2.4,\ 93.79) = 20.13,\ p < .001,\ \eta^2 = 0.34$) (see Fig. 5). *Post hoc* pair-wise comparisons (Bonferroni's test) revealed that participants perceived less workload when images were presented in a 5×3 matrix than when images were presented in the 7×3 matrix (mean difference and 95% CI for difference: $-4.1\ (-6.36, -1.84),\ p < .001$), $7 \times 4\ (-4.61\ (-7.38, -1.84),\ p < .001$), and $9 \times 4\ (-6.28\ (-9.05, -3.51),\ p < 0.001$) set sizes. Moreover, the subjective ratings of workload significantly increased with time pressure ($F(1, 39) = 21.29,\ p < 0.001,\ \eta^2 = 0.35$). The image set size × time pressure interaction was not significant for this measure.

According to the subjective ratings of the mental workload, we selected physiological and behavioral data from 15 N (i.e., 5×3 image set size without time pressure), 21 W (i.e., 7×3 image set size with time pressure), 28 N, and 36 W to train the SVM classifiers. Data from the other four conditions (15 W, 21 N, 28 W, and 36 N) were tested to validate the model's cross-level within-task performance.

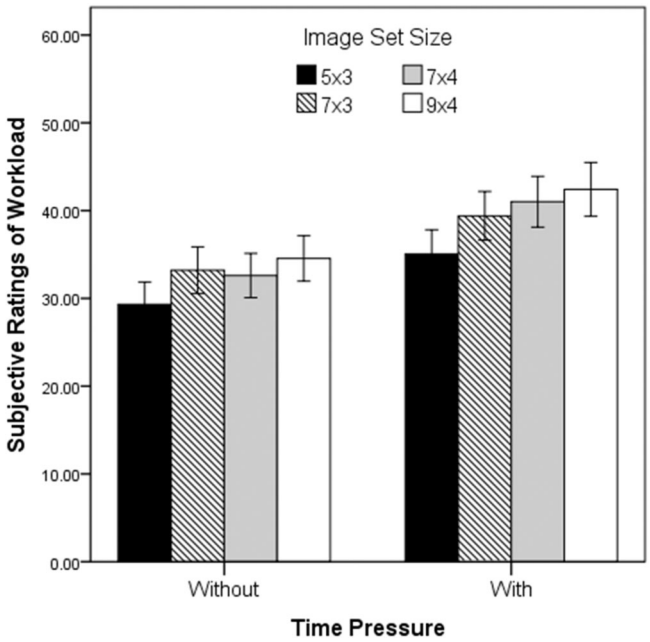

Fig. 5. Subjective ratings of mental workload for the four image set sizes with/without time limit (error bars indicate ±1 standard error).

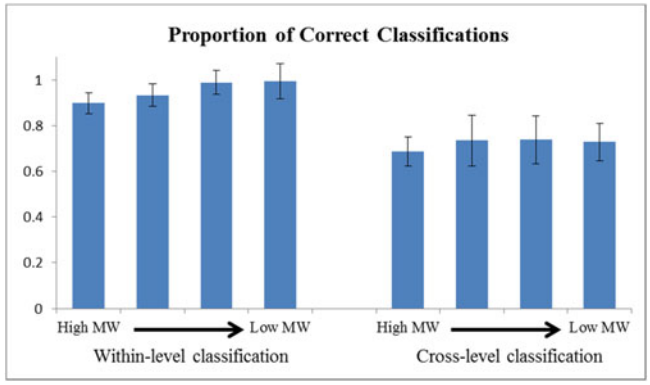

Fig. 6. Proportion of correct classification of mental workload based on subjective ratings at cross-level and within-level within the same anomalous image detection task (error bars indicate ±1 standard deviation).

B. Statistical Analysis of Task-I Performance

A repeated measures ANOVA was performed with image set size and time pressure as two within-subjects variables. Significant findings were followed-up to assess the magnitude of the effects that each independent variable had on the behavioral measures.

A significant image set size × time pressure was revealed for the average detection RT per image ($F(1.97,\ 76.7) = 4.33,\ p = 0.017,\ \eta^2 = 0.10$). *Post hoc* pair-wise comparisons (Bonferroni's test) showed that the 9 × 4 set size, on average, led to a shorter detection RT per image than the 7 × 4 set size, which was better than the 7 × 3 and 5 × 3 set sizes, independent of the manipulation of time pressure ($p < .01$). The average detection RT per image decreased as the number of images increased. There was no significant difference in the average detection RT per image between the 7 × 3 and 5 × 3 set sizes. Additionally, both main effects of the image set size ($F(2.11,\ 82.35) = 59.5,\ p < 0.001,\ \eta^2 = 0.60$) and time pressure ($F(1, 39) = 42.57,\ p < 0.001,\ \eta^2 = 0.52$) were significant for this measure. Larger image set sizes and time pressure shortened the average detection RT per image.

The main effect of the image set size was significant for the number of errors ($F(2.58,\ 100.57) = 90.36,\ p < 0.001,\ \eta^2 = 0.70$). *Post hoc* pair-wise comparisons (Bonferroni's test) indicated that the 5 × 3 set size led to a smaller number of errors than the 7 × 3 set size, which was smaller than the 7 × 4 ($p < 0.001$), showing a linear increase of the number of detection errors as a function of the set size. Alternately, the 9 × 4 set size resulted in a smaller number of errors than the 7 × 4 set size (mean difference and 95% CI for difference: $-0.82\ (-1.43, -0.21),\ p = 0.003$). The main effects of the time pressure and the image set size × time pressure interaction were not significant for this measure.

C. Classification of the Within-Task Mental Workload

Within-task classification accuracy was evaluated from two levels: within- and cross level. Physiological and behavioral data from four task difficulty levels (15 N, 21 W, 28 N, 36 W) were divided into four categories according to subjective ratings of the mental workload. The range of NASA-TLX scores and the number of training data in each category were provided as follows: 8–20 ($n = 80$), 21–31 ($n = 78$), 32–49 ($n = 81$), and 50–83 ($n = 81$). The SVM classifiers were trained on the physiological and behavioral data to predict the mental workload categories of the same four task difficulty levels in Task I (i.e., within-level classification). The leave-one-subject-out cross-validation method was applied for the within-level classification of the mental workload. Specifically, we used 39 participants as the training set, and the rest formed the testing set. The within-level classification accuracy was averaged over 40 results (the number of participants involved in this experiment) since each individual would be the testing set in turn.

Alternately, the SVM classifiers were trained on the same physiological and behavioral data (15 N, 21 W, 28 N, 36 W) combined from all 40 participants to predict the mental workload categories of each individual when he or she experienced the other four task difficulty levels (15 W, 21 N, 28 W, 36 N) in Task I (i.e., cross-level classification). The range of NASA-TLX scores and the number of testing data for the cross-level classification were: 8–20 ($n = 78$), 21–32 ($n = 83$), 33–50 ($n = 80$), and 51–85 ($n = 79$). Each individual would be the testing set in turn, so the cross-level classification accuracy was averaged over 40 results.

As shown in Fig. 6, the within-level proportion of correct classifications (mean = 95.29%, range = 89.76–99.35%) was higher than the cross-level proportion of correct classifications (mean = 72.2%, range = 68.7–73.85%). For the within-level classification, the higher the level of self-reported mental workload was, the lower the classification accuracy was. In contrast,

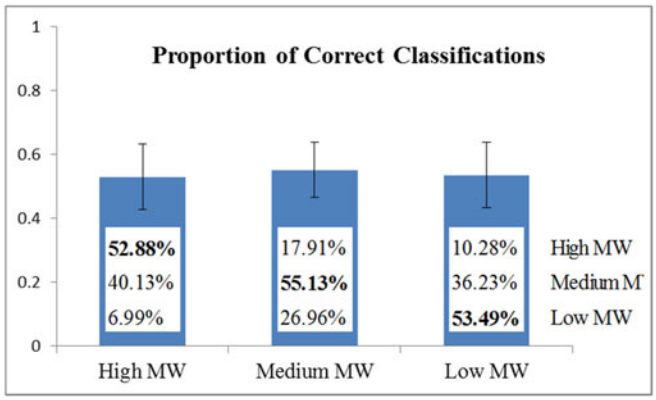

Fig. 7. Proportion of correct classifications of high, medium, and low mental workload categories of the abnormal activity detection task (error bars indicate ±1 standard deviation).

the highest and lowest subjective ratings of the mental workload led to lower cross-level classification accuracies.

Moreover, using only physiological indices, the overall accuracy was very similar to the results using both physiological and behavioral features. Therefore, we used only physiological features to recognize the real-time cross-task mental workload in Task II.

D. Real-Time Recognition of the Cross-Task Mental Workload

To recognize the cross-task mental workload, physiological signals collected from Task I were divided into three categories (high, medium, and low MW) according to participants' subjective ratings of the mental workload to train the SVM classifiers. The range of NASA-TLX scores and the number of training data in each category were: 8–24 ($n = 215$), 25–44 ($n = 211$), and 45–85 ($n = 214$). The classifiers predicted mental workload of the abnormal activity detection task (Task II) at each of the three levels (see Fig. 7). The range of NASA-TLX scores and the number of testing data in each category were: 6–21 ($n = 41$), 22–42 ($n = 43$), and 44–78 ($n = 36$).

Not surprisingly, the cross-task classification accuracy was lower than the within-task accuracy. The proportion of correct classifications was 53.49% (SD = 10.27%) for the low MW, 55.13% (SD = 8.72%) for the medium MW, and 52.88% (SD = 10.28%) for the high MW. The classification accuracy demonstrated an evident decline, while the variability of the model's correct classifications obviously increased, which indicates consistent systematic cross-task misclassification in some individuals. In addition, when the participant responded to each task (i.e., click the button to continue to the next task), our system predicted the level of the mental workload. The average execution time of our proposed system was 1.67 s.

In SLDA, the nonzero values in the outcome sparse vector can be used to select features: the larger the nonzero value in the outcome sparse vector, the more important the feature to the correct classification of the mental workload. As shown in Table III, when removing the least important features (from the full feature set to the top ten features to the top five features),

TABLE III
CLASSIFICATION ACCURACIES WITH DIFFERENT NUMBERS OF FEATURES

Physiological Indices	Classification Accuracy		
	Within Level	Cross Level	Cross Task
Top five features	76.01%	51.12%	52.4%
Top ten features	82.55%	63.35%	52.61%
Full features[a]	95.29%	72.2%	53.83%

[a]45 features (42 physiological indices and 3 behavioral measures) for within- and cross-level classification and 42 physiological features for cross-task classification.

the classification accuracy decreased, but the rates of decline were different among the within-level, cross-level, and cross-task mental workload. Specifically, the prediction of cross-task mental workload relied on fewer physiological features than that of within-level and cross-level mental workload. The difference in the classification accuracy between the top five features and the full feature set was only 1.43%. During the prediction of the cross-task mental workload, the top three physiological features were the power in the 0.1-Hz component of HRV, the average respiration rate, and the average heart rate.

VI. DISCUSSION

This paper proposed a real-time physiological-based system to assess the cross-task mental workload during anomaly detection. Forty participants were recruited to perform an anomalous image detection task (Task I) and an abnormal activity detection task (Task II). Five categories of peripheral physiological signals (ECG, EOG, RSP, GSR, and PPG) were recorded. More than 40 features sensitive to the changes of mental workload suggested in the literature were extracted and entered into the SVM as inputs.

For the offline classification/prediction of the mental workload, physiological and behavioral data from four task difficulty levels were divided into four categories according to subjective ratings of the mental workload. The SVM classifiers were trained on these data to predict the mental workload categories of: 1) the same four task difficulty levels (within-level); and 2) the other four task difficulty levels in Task I (cross level). We found that the within-level classifications (95.29% on average) were more accurate than the cross-level classifications (72.2% on average), which were much higher than the random level (25%). For the within-level classification, the lower the level of self-reported mental workload was, the better classification accuracy was. As a comparison, the medium level of the mental workload led to the highest classification accuracy for the cross-level classification. In the applied environment, recognition of an operator's functional/cognitive states from physiological-based indices largely relies on the classification accuracy and acceptability of the data acquisition and processing methods [29]. This requires that these functional assessment methods be highly accurate. The previous studies have suggested that within-level classification accuracy between two task difficulty levels must approach 95% to be acceptable [29], [42]. In this paper, our system achieved an overall accuracy of

95.29% (ranging from 89.76 to 99.35%) for within-level classifications. Because there were four task difficulty levels (i.e., 25% accuracy at random), our system undoubtedly met the within-level criteria.

A key challenge in the development of real-time and adaptive systems for recognizing the mental workload is the development of accurate machine-learning algorithms for tasks and difficulty levels that the participant has not yet experienced. In this paper, the three-class SVM classifiers were trained on physiological data from Task I and tested on Task II (cross task) using each individual participant's (subject dependent) data. We presented such a generalized system to predict the subjective mental workload in a different task, and we achieved better classification accuracies than the chance level. To the best of our knowledge, the classification accuracies of the cross-task mental workload in the previous studies were below or just around chance levels [31], [43]. Compared to these two studies, our proposed system achieved a higher cross-task classification accuracy of 53.83% than 33.33% accuracy at random. One possible reason was that we used more-portable and practical devices to collect peripheral physiological measures, and all features were selected based on the extensive literature review of mental workload.

In our study, the top three features that made the greatest contribution to the assessment of relatively low-moderate changes in mental workload were the power in the 0.1-Hz component of HRV, the average respiration rate, and the average heart rate. These results are consistent with previous findings: the 0.10-Hz component is sensitive to low or moderate changes in mental workload [19], [22]. Cardiac functions and respiration are sensitive indices of mental workload and are sensitive to changes in task difficulty [14], [44]. Moreover, lower mental workload was observed when people performed a self-paced detection task. From the perspective of energetic resources, the effectiveness of adaptation includes its costs in terms of psychological and physiological energy. This cost is based on the assumption that biological systems seek equilibrium states of minimal energy expenditure. When task demands increase (e.g., the manipulation of time pressure in this study), the human adaptation system becomes instable: Performance is maintained or even enhanced at the cost of compensatory effort manifested in measures of workload. These costs can eventually render people more vulnerable to task failure as a result of depleted energetic resources [45].

Detecting subtle perceptual features can be highly challenging, especially within a time limit. Brooks *et al.* [4] observed a widespread failure of medical expert diagnosticians to notice supposedly obvious prototypical symptoms of diseases such as pancreatitis, and they explained that this stemmed in part from the fact that naturalistic stimuli were typically characterized by ambiguity as to what should be taken as a feature. Our study presented in this paper hints that physiological-based assessment methods have potential application in the development of situation-aware recognition systems of real-time mental workload or adaptive human–computer interaction platforms. According to the predicted levels of the operator's mental workload, our proposed system may have great practical value through sending a warning signal to suggest a work shift or short break or freezing the task to probe whether the individual is suffering from a high level of mental workload. Such responses will help adjust and customize the sensitivity and efficiency of the system according to individual preference. Our proposed system can also adjust the task difficulty level adaptively via a human–machine interface (HMI). Future work includes developing a human–machine interaction system that can both monitor real-time mental workload and provide concurrent feedback to minimize the risk of working with high task demands for a long period of time (i.e., adaptive aiding and training).

In addition, our preliminary success in the classification of cross-task mental workload, which was much higher than chance level, indicated that multimodal physiological indices can be applied in a wide range of visual anomaly detection contexts, such as security, medical diagnosis, monitoring, and quality control. Unlike the manipulation of task difficulty levels under laboratory conditions, it is difficult to identify task demands in the actual working environment (e.g., count the number of images or video streams presented simultaneously). In addition, the same task demands do not result in an equal level of workload for all individuals. Human operators can adapt their behaviors and cope with increasing demands. As a result, it is impossible and inaccurate to predict the mental workload relying solely on task demand. In our study, we not only manipulated different task difficulty levels but also evaluated each individual participant's ratings of mental workload to ensure that both task demands and subjective feelings were consistent. For example, with the aid of wearable and mobile sensors [46], a human screener in a security checkpoint X-raying passenger bags can be monitored in real time. If the predicted level of mental workload exceeds the predefined threshold, the transmission belt can slow down to reduce the amount of luggage that enters the gate.

This study did not measure brain activities, future work might benefit from a combination of the peripheral and EEG indices in the prediction of cross-task mental workload. The high cross-level and cross-task classification accuracies indicate that machine-learning methods have great potential to be developed for predicting the task and difficulty levels that the participant has not yet experienced. It is also interesting to investigate the underlying cognitive mechanism for these cross-level and cross-task implications in our future work.

ACKNOWLEDGMENT

The authors thank the editor and all reviewers for their careful reviews and constructive comments, which help them to improve the quality of this paper.

REFERENCES

[1] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Comput. Surveys*, vol. 41, no. 3, 2009, Art. no. 15.

[2] I. Graves *et al.*, "The role of the human operator in image-based airport security technologies," in *Innovations in Defence Support Systems-2*. Berlin, Germany: Springer, 2011, pp. 147–181.

[3] E. M. Kok *et al.*, "Learning radiological appearances of diseases: Does comparison help?" *Learn. Instruction*, vol. 23, pp. 90–97, 2013.

[4] L. R. Brooks, V. R. LeBlanc, and G. R. Norman, "On the difficulty of noticing obvious features in patient appearance," *Psychological Sci.*, vol. 11, no. 2, pp. 112–117, 2000.

[5] T. S. Horowitz and J. M. Wolfe, "Search for multiple targets: Remember the targets, forget the search," *Attention, Perception, Psychophys.*, vol. 63, no. 2, pp. 272–285, 2001.

[6] B. Cain, "A review of the mental workload literature," Defence Research and Development Canada Toronto, Human System Integration Section, Toronto, ON, Canada, OMB No. 0704-0188, 2007.

[7] J. Zhang, Z. Yin, and R. Wang, "Recognition of mental workload levels under complex human-machine collaboration by using physiological features and adaptive support vector machines," *IEEE Trans. Human-Mach. Syst.*, vol. 45, no. 2, pp. 200–214, Apr. 2015.

[8] S. Wang, J. Gwizdka, and W. A. Chaovalitwongse, "Using wireless EEG signals to assess memory workload in the n-back task," *IEEE Trans. Human-Mach. Syst.*, vol. 46, no. 3, pp. 424–435, Jun. 2016.

[9] J. Harrison *et al.*, "Cognitive workload and learning assessment during the implementation of a next-generation air traffic control technology using functional near-infrared spectroscopy," *IEEE Trans. Human-Mach. Syst.*, vol. 44, no. 4, pp. 429–440, Aug. 2014.

[10] T. O. Zander *et al.*, "Evaluation of a dry EEG System for application of passive brain-computer interfaces in autonomous driving," *Frontiers Human Neurosci.*, vol. 11, 2017, Art. no. 78.

[11] S. Patel *et al.*, "A review of wearable sensors and systems with application in rehabilitation," *J. Neuroeng. Rehabil.*, vol. 9, 2012, Art. no. 21.

[12] D. Gopher and E. Donchin, "Workload: An examination of the concept," in *Handbook of Perception and Human Performance, Vol. 2: Cognitive Processes and Performance*, K. R. Boff, L. Kaufman, and P. Thomas, Eds. Oxford, U.K.: Wiley, 1986, pp. 1–49.

[13] C. D. Wickens, "Multiple resources and performance prediction," *Theoretical Issues Ergon. Sci.*, vol. 3, no. 2, pp. 159–177, 2002.

[14] J. A. Veltman and A. W. K. Gaillard, "Physiological workload reactions to increasing levels of task difficulty," *Ergonomics*, vol. 41, no. 5, pp. 656–669, 1998.

[15] J. G. Casali and W. W. Wierwille, "On the measurement of pilot perceptual workload: A comparison of assessment techniques addressing sensitivity and intrusion issues," *Ergonomics*, vol. 27, no. 10, pp. 1033–1050, 1984.

[16] R. Parasuraman and M. Rizzo, *Neuroergonomics: The Brain at Work*. New York, NY, USA: Oxford Univ. Press, 2006.

[17] L. J. M. Mulder, "Measurement and analysis methods of heart rate and respiration for use in applied environments," *Biol. Psychol.*, vol. 34, no. 2-3, pp. 205–236, 1992.

[18] S. W. Porges and E. A. Byrne, "Research methods for measurement of heart rate and respiration," *Biol. Psychol.*, vol. 34, no. 2-3, pp. 93–130, 1992.

[19] P. Nickel and F. Nachreiner, "Sensitivity and diagnosticity of the 0.1-Hz component of heart rate variability as an indicator of mental workload," *Human Factors*, vol. 45, no. 4, pp. 575–590, 2003.

[20] R. W. Backs and L. C. Walrath, "Eye movement and pupillary response indices of mental workload during visual search of symbolic displays," *Appl. Ergon.*, vol. 23, no. 4, pp. 243–254, 1992.

[21] G. Borghini *et al.*, "Measuring neurophysiological signals in aircraft pilots and car drivers for the assessment of mental workload, fatigue and drowsiness," *Neurosci. Biobehavioral Rev.*, vol. 44, pp. 58–75, Jul. 2014.

[22] R. W. Backs and K. A. Seljos, "Metabolic and cardiorespiratory measures of mental effort: The effects of level of difficulty in a working memory task," *Int. J. Psychophysiol.*, vol. 16, no. 1, pp. 57–68, 1994.

[23] A. F. Kramer, "Physiological metrics of mental workload: A review of recent progress," in *Multiple-Task Performance*, D. L. Damos, Ed. London, U.K.: Taylor & Francis, 1990, pp. 279–328.

[24] D. Novak *et al.*, "Psychophysiological responses to robotic rehabilitation tasks in stroke," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 18, no. 4, pp. 351–361, Aug. 2010.

[25] Y. Lyu *et al.*, "Measuring photoplethysmogram-based stress-induced vascular response index to assess cognitive load and stress," in *Proc. 33rd Annu. ACM Conf. Human Factors Comput. Syst.*, Seoul, Korea, 2015, pp. 857–866.

[26] G. F. Wilson and C. A. Russell, "Operator functional state classification using multiple psychophysiological features in an air traffic control task," *Human Factors*, vol. 45, no. 3, pp. 381–389, 2003.

[27] D. B. Kaber *et al.*, "Workload state classification with automation during simulated air traffic control," *Int. J. Aviation Psychol.*, vol. 17, no. 4, pp. 371–390, 2007.

[28] Z. Wang *et al.*, "Cross-subject workload classification with a hierarchical Bayes model," *Neuroimage*, vol. 59, no. 1, pp. 64–69, Jan. 2, 2012.

[29] G. F. Wilson and C. A. Russell, "Real-time assessment of mental workload using psychophysiological measures and artificial neural networks," *Human Factors*, vol. 45, no. 4, pp. 635–644, 2003.

[30] J. A. Cannon *et al.*, "An algorithm for online detection of temporal changes in operator cognitive state using real-time psychophysiological data," *Biomed. Signal Process. Control*, vol. 5, no. 3, pp. 229–236, Jul. 2010.

[31] C. L. Baldwin and B. N. Penaranda, "Adaptive training using an artificial neural network and EEG metrics for within- and cross-task workload classification," *Neuroimage*, vol. 59, no. 1, pp. 48–56, 2012.

[32] M. A. Hogervorst, A.-M. Brouwer, and J. B. van Erp, "Combining and comparing EEG, peripheral physiology and eye-related measures for the assessment of mental workload," *Frontiers Neurosci.*, vol. 8, 2014, Art. no. 322.

[33] D. De Waard and V. Studiecentrum, *The Measurement of Drivers' Mental Workload*, Traffic Research Center, University of Groningen, Groningen, The Netherlands, 1996.

[34] T. F. Meijman *et al.*, "Psychological aspects of workload," *A Handbook of Work and Organizational Psychology: Volume 2: Work Psychology*, P. J. D. Drenth, H. Thierry, and C. J. de Wolff, Eds. Hove, U.K.: Psychology Press, 1998.

[35] L. Clemmensen *et al.*, "Sparse discriminant analysis," *Technometrics*, vol. 53, no. 4, pp. 406–413, 2011.

[36] Y.-J. Liu *et al.*, "Real-time movie-induced discrete emotion recognition from EEG signals," *IEEE Trans. Affect. Comput.*, 2018, doi: 10.1109/TAFFC.2017.2660485.

[37] G. Zhao *et al.*, "Emotion analysis for personality inference from EEG signals," *IEEE Trans. Affect. Comput.*, 2018, doi: 10.1109/TAFFC.2017.2786207.

[38] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, 2011, Art. no. 27.

[39] D. B. Walther *et al.*, "Simple line drawings suffice for functional MRI decoding of natural scene categories," *Proc. Nat. Acad. Sci.*, vol. 108, no. 23, pp. 9661–9666, 2011.

[40] V. Mahadevan *et al.*, "Anomaly detection in crowded scenes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, San Francisco, CA, USA, 2010, pp. 1975–1981.

[41] S. G. Hart and L. E. Staveland, "Development of NASA-TLX: Results of empirical and theoretical research," in *Human Mental Workload*, P. A. Hancock and P. Meshkati, Eds. Amsterdam, The Netherlands: Elsevier, 1988, pp. 139–183.

[42] G. F. Wilson and C. A. Russell, "Performance enhancement in an uninhabited air vehicle task using psychophysiologically determined adaptive aiding," *Human Factors*, vol. 49, no. 6, pp. 1005–1018, 2007.

[43] Y. Ke *et al.*, "Towards an effective cross-task mental workload recognition model using electroencephalography based on feature selection and support vector machine regression," *Int. J. Psychophysiol.*, vol. 98, no. 2, pp. 157–166, 2015.

[44] L. R. Fournier, G. F. Wilson, and C. R. Swain, "Electrophysiological, behavioral, and subjective indexes of workload when performing multiple tasks: manipulations of task difficulty and training," *Int. J. Psychophysiol.*, vol. 31, no. 2, pp. 129–145, 1999.

[45] J. L. Szalma and G. W. L. Teo, "Spatial and temporal task characteristics as stress: A test of the dynamic adaptability theory of stress, workload, and performance," *Acta Psychologica*, vol. 139, no. 3, pp. 471–485, 2012.

[46] M. Rabbi *et al.*, "Passive and in-situ assessment of mental and physical well-being using mobile sensors," in *Proc. 13th Int. Conf. Ubiquitous Comput.*, Beijing, China, 2011, pp. 385–394.

Guozhen Zhao received the B.S. degree in industrial engineering from Tianjin University, Tianjin, China, in 2007, and the M.S. and Ph.D. degrees in industrial and systems engineering from the State University of New York, Buffalo, NY, USA, in 2009 and 2011, respectively.

He is an Associate Professor with the Institute of Psychology, Chinese Academy of Sciences. His current research interests include the mathematical modeling of human cognition and performance, transportation safety, human computer interaction, emotion recognition, and augmented cognition.

Yong-Jin Liu (SM'16) received the B.Eng. degree in mechanoelectronic engineering from Tianjin University, Tianjin, China, in 1998, and the Ph.D. degree in mechanical engineering from the Hong Kong University of Science and Technology, Hong Kong, in 2003.

He is an Associate Professor with the Department of Computer Science and Technology, Tsinghua University, Beijing, China. His research interests include pattern analysis, human–computer interaction, and computer graphics.

Yuanchun Shi (SM'07) received the B.S., M.S., and Ph.D. degrees in computer science from Tsinghua University, Beijing, China.

She is a Changjiang Distinguished Professor with the Department of Computer Science, Tsinghua University. She was a Senior Visiting Scholar with MIT AI Lab during 2001–2002. She has authored and coauthored more than one hundred papers in *International Journal of Human-Computer Studies*, IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, *ACM Transactions on Computer-Human Interaction*, *ACM Multimedia*, *ACM User Interface Software and Technology*, etc. Her research interests include human–computer interaction, pervasive computing, and multimedia communication.

Dr. Shi had chaired several conferences including ACM Ubicomp2011. She serves as the Area Editor of the *Pervasive and Mobile Computing* (Elsevier), an editor of the *Interacting With Computer* (Oxford University Press), and the Vice Editor-in-Chief of the *Communications of China Computer Federation*.