Employing Consumer Wearables to Detect Office Workers' Cognitive Load for Interruption Management

FLORIAN SCHAULE, Technical University of Munich, Germany
JAN OLE JOHANSSEN, Technical University of Munich, Germany
BERND BRUEGGE, Technical University of Munich, Germany
VIVIAN LOFTNESS, Carnegie Mellon University, USA

Office workers' productivity and well-being are reduced by interruptions, especially if they occur during an inconvenient moment. Interruptions in phases of high cognitive load are more disruptive than in phases of low cognitive load. Based on an explorative study, we suppose the presence of social codes that signal office workers' interruptibility. We propose a system that utilizes the cognitive load of an office worker to indicate situations suitable for interruptions. The cognitive load is inferred from office workers' physiological state measured by a consumer smartwatch. The system adapts an externally mounted smart device to indicate if the office worker is interruptible. To predict the cognitive load, we trained a classifier with ten office workers and achieved an accuracy between 66% and 86%. In order to validate the classifier's accurateness in an office setting, we performed a verification study with five office workers: We systematically triggered interruptions for each subject over an interval of half a day of office work. The classifier was able to infer the level of cognitive load for three office workers. This result supports our hypotheses that inferring cognitive load using a consumer smartwatch is a viable concept.

CCS Concepts: • **Human-centered computing** → **User models**; **User studies**; **Ubiquitous and mobile computing**; *HCI theory, concepts and models*; *Interactive systems and tools*; • **Computing methodologies** → *Machine learning approaches*;

Additional Key Words and Phrases: Cognitive Load, Interruption Management System, Social Code, Wearable Devices, Office Environment, Productivity, Automatic Interruptibility Measurement, Explorative Study

ACM Reference Format:
Florian Schaule, Jan Ole Johanssen, Bernd Bruegge, and Vivian Loftness. 2018. Employing Consumer Wearables to Detect Office Workers' Cognitive Load for Interruption Management. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 1, Article 32 (March 2018), 20 pages. https://doi.org/10.1145/3191764

1 INTRODUCTION

Interruptions during work are a necessary part of most working environments. In particular, during their daily work, office workers face a variety of interruptions, such as emails, questions of co-workers, or instant messages.

Humans need on average about twenty-three minutes to resume their task after they got interrupted [23]. Multiple studies reveal that interruptions make humans slower and more error-prone in performing tasks [8, 13]. Other research argues that interruptions occurring at an inopportune moment have a negative effect on the well-being of humans since they cause stress [2, 24]. To address this problem, social codes were established.

Authors' addresses: Florian Schaule, Jan Ole Johanssen, and Bernd Bruegge, Technical University of Munich, Department of Informatics, Boltzmannstraße 3, 85748 Garching b. München, Germany, florian.schaule@tum.de, jan.johanssen@in.tum.de, bruegge@in.tum.de; Vivian Loftness, Carnegie Mellon University, School of Architecture, 5000 Forbes Avenue, Pittsburgh, PA 15213, USA, loftness@cmu.edu.

For instance, a closed office door can indicate that an office worker does not want to get interrupted, whereas an open door may indicate the opposite. In general, we consider social codes as a human habit expressed in a physical change of the environment, which are commonly understood to have a defined meaning within a larger social group such as office employees. However, these changes have to be actively evoked by the office worker and consequently requires their constant attention—which decreases their effectiveness. This research aims for a solution that does not require any active management by an office worker to signal their state of interruptibility. Based on the acceptance of social codes, we intend to automatically reflect office workers' current state of interruptibility on a smart device, such as a tablet computer mounted on the front door of an office.

The project *PrefMiner* [25] tries to reduce the number of notifications by classifying them into important and unimportant and consequently only presents the important notifications. The system learns from the users' behavior. It classifies notifications by recording those that are dismissed and those that the user interacts with. Hereafter, it applies these patterns on future notifications. The drawback of this solution is that it only takes the importance of the notifications into account, while the timing of the notification is not incorporated. As a result, a notification could be interrupting the user in a phase in which they are focused on a task.

Research reveals that interruptions during phases of high cognitive load are perceived as more disruptive than in phases of low cognitive load [13, 16, 20]. The cognitive load can be inferred by applying machine learning methods on physiological data such as the heart rate variability (HRV) or the galvanic skin response (GSR) [6, 7].

This research strives to improve office workers' satisfaction in office environments by adapting the timing of interruptions in response to the office workers' cognitive load. By scheduling external unexpected interruptions to phases of low cognitive load they will be less disruptive. External interruptions are caused by changes in the environment, e.g., a ringing phone [27]. Unexpected interruptions are not scheduled interruptions, e.g., someone entering the office without an appointment [28]. The cognitive load signature helps to indicate the interruptibility to co-office workers in order to signal a suitable point in time to interrupt the office worker. This process should run without any active interaction, be unobtrusive for the office worker, and improve the productivity of the office worker. We made the following contributions in this paper:

(1) Exploring the presence of social codes to manage interruption. We focused on the effectiveness of social codes regarding open and closed doors to transfer this code to externally controlled devices.
(2) Detecting the level of cognitive load using consumer wearable devices instead of more expensive and less common wearables which are typically used for scientific studies.
(3) Designing and evaluating a process to train a system which adapts to the personal traits of different office workers. The process also includes the evaluation of the system with regard to the accuracy.
(4) Building an interruption management system based on an unobtrusive smartwatch, a light, and an externally mounted tablet computer.

We outline background information on cognitive load theory in Section 2, while Section 3 provides a literature review of related research. Section 4 explores the usage of social codes for managing interruptions. In Section 5, we describe our approach as well as the implementation of the prototype. An evaluation, related results, and interpretations are presented in Section 6. In Section 7, we summarize future work while Section 8 concludes.

2 COGNITIVE LOAD

We hypothesize that interruptions in workplaces should occur during periods of low cognitive load to improve complex task performance. The psychological concept of cognitive load refers to the total mental effort required to perform a task puts on the human working memory [33]. The working memory is limited to 7 ± 2 *chunks* of information [26]. Cognitive load is the pressure put on human working memory while performing a task. Cognitive load is a variable trying to quantify the amount of mental demand a task puts on the mental resources.

Therefore, a mental high workload is reflected in a high cognitive load. Cognitive load is a dynamic property that changes quickly. There are four different ways to quantify the cognitive load of a human [6]:

- **Subjective (self-reported) Measure**: Quantifying the cognitive load by asking the subjects to rate their experienced cognitive load on a rating scale immediately after finishing a task [15].
- **Performance Measure**: Quantifying the cognitive load by tracking the task performance of the subject. The underlying assumption is that an overloaded working memory, which is an indicator of high cognitive load, will result in lower task performance [30].
- **Physiological Measure**: Quantifying the cognitive load by recording changes in physiological data that are closely linked to changes in the cognitive load. By continuously measuring these changes, the cognitive load can be inferred [7].
- **Behavioral Measure**: Quantifying the cognitive load by recording behavioral patterns, such as eye-gaze or mouse movements, which correspond to changes in the cognitive load. By detecting changes in these patterns the cognitive load can be inferred [19].

The amount of cognitive load generated by performing the same task can vary between humans since the cognitive load is influenced by age, gender, personal traits, and their experience in solving the task [32]. The physiological and the behavioral measurement-based methods use the reaction of a human to different levels of cognitive load. Due to the highly personal nature of these reactions, the method needs to be individually adapted to each subject studied [6].

3 RELATED WORK

In this section we describe research that focuses on the measurement of the cognitive load and the management of the interruptions depending on it. We structure these research efforts into three categories depending on the method of measurement of the cognitive load.

3.1 Physiological Measurement Based Systems

Six precedent systems for physiological measurement and response to cognitive loading are important background for this research project: Chen and Vertegaal developed a system that automatically manages mobile phone notifications depending on the movement state and the cognitive load of the user [5]. Their system detects the cognitive load using an electrocardiography (ECG) and an electroencephalography (EEG) sensor. Cinaz *et al.* propose a monitoring system to measure the cognitive load of an office worker using an ECG sensor [7]. They calibrate the system by instructing the participants to perform three different performance tasks. They train machine learning classifiers using this training data and infer in a second step the workload of an office worker during a typical office day. Ferreira *et al.* measure the cognitive load using an EEG, an ECG, and a GSR sensor [9]. They train their machine learning classifier using training data generated by instructing the participants to do various performance tasks. They attain an accuracy between 64% and 73% using the data sliced into frames of 10 seconds (i.e., the latency for the classification is 10 seconds) for two levels of cognitive load.

Compared to our system, all of the above presented systems share the disadvantage that the user has to wear cumbersome scientific devices with electrodes attached to their chest and scalp. Our setup, in contrast, solely relies on a smartwatch worn by users, which we expect to have a higher acceptance in an office environment.

Krause *et al.* present a system that applies unsupervised machine learning to detect the context-dependent personal preferences of the user using the state, context, and user interactions with the system [22]. It focuses on the adaption of a notification setting on a mobile phone. Goyal and Fussell identify suitable moments of interruptions to reduce their negative effects on the task performance [14]. They utilize a GSR sensor integrated into a smartwatch worn by the users. Abdelrahman *et al.* measure the level of cognitive load utilizing the skin temperature of the nose tip and the forehead of the user [1]. The temperature is determined using an infrared

camera. They are able to detect changes in the level of cognitive load after 3.8 seconds. These three systems focus on identifying appropriate moments for interruptions, but not on managing the interruption actively. We add the functionality to manage the timing of the interruption in our system.

3.2 Behavioral Patterns Based Systems

The following three systems for behavioral measurement and response to cognitive loading are important background for our research project.

Züger *et al.* present a system that uses the behavioral patterns of an office worker's interaction with their mouse and keyboard to detect their interruptibility linked to cognitive loading [35]. They conduct a large-scale study to evaluate the performance of the system. Their system indicates the state of interruptibility using a light mounted next to the office worker, which changes its hue accordingly. The study participants indicated that the system improves their productivity and self-motivation and most of them keep on using the system in the months-long study. This system can only be applied to tasks which include the active usage of the mouse or keyboard. Instead, our solution is able to also manage interruption for tasks scenarios in which the user is not actively engaging with a computer, e.g., when watching a lecture video or completing paperwork.

Katidioti *et al.* propose a system that uses the pupil dilution to detect the point in time to interrupt an office worker, with an assumption of better times linked to lower cognitive load [19]. The system detects the pupil dilution using an eye-tracker. Tanaka *et. al* develop an interruption management system that uses head movements and the interaction with the computer to identify phases of low cognitive load [34]. The two before cited interruption management systems require the user to face an eye-tracker or a camera. In case users look into a different direction, the systems will not be able to work correctly. We overcome this limitation by only using the data gathered by a smartwatch which delivers constant data independent of the users' point of view.

3.3 Mixed Measurements

Two precedent systems of combined measurement and response to cognitive loading form important background in the area of mixed measurements.

Züger and Fritz present a system that predicts the interruptibility of professional software developers using the data of wearable physiological sensors [36]. They used the data of an EEG sensor, heart rate sensor, eye-tracking system, blood volume pulse sensor, and GSR sensor. They were able to identify the software developers' state of interruptibility by means of machine learning classifiers with a high accuracy. The system has the disadvantage compared to our proposed solution that the user has to wear cumbersome electrodes for the EEG sensor.

In their research, Nourbakhsh *et al.* present a system that classifies the level of cognitive load of a human using the frequency of their eye blinks and their GSR [29]. The participants performed different performance tasks to generate physiological data, which was used to train two machine learning classifiers. They achieved a 75% accuracy for two levels and a 50% for four levels of cognitive load. While this research also requires that the user faces the eye-tracking device, we overcome the limitation using only a smartwatch.

4 EXPLORING THE SOCIAL CODE OF INTERRUPTIBILITY

We approached our investigations on *social codes* by applying an explorative study. We intended to learn about the presence of social codes, i.e., the behavior of subjects, when confronted with an open or closed door. Therefore, we carefully designed two study scenarios as depicted in Figure 1 to stimulate and prompt social codes.

In *Study A*, we explore if a closed door implies that an office worker prefers not to get interrupted while an open door implies they are willing to participate in any kind of interaction. In a second step, as part of *Study B*, we want to investigate whether we can transfer the social codes of open and closed doors to an automatically controlled environment adaption strategy. A system consisting of a smartphone mounted at each door frame is

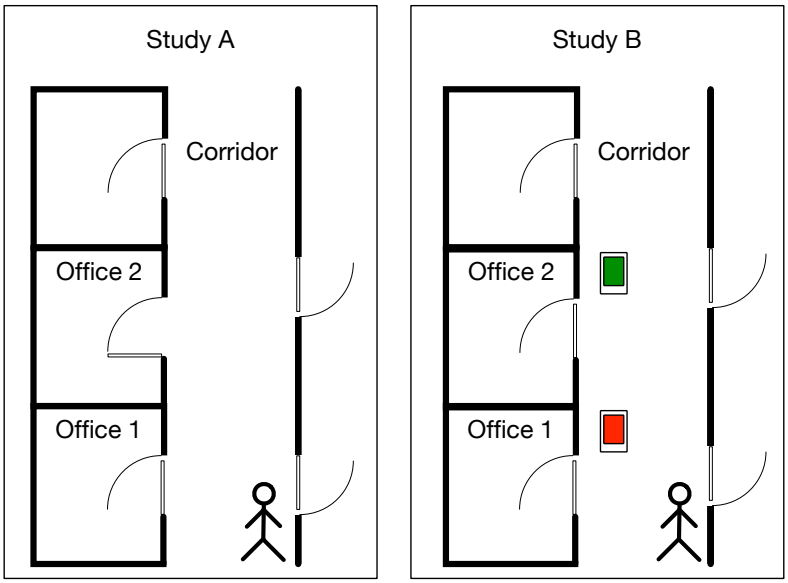

Fig. 1. We explored the social codes of interruptibility in office environments with two studies, in each of which ten participants were asked to interrupt office workers to a find out the room number of a particular member of staff. In *Study A*, the participants faced a closed door of *Office 1* and an open door of *Office 2*. In *Study B*, both office doors were closed, yet equipped with red and green signals displayed on a smartphone mounted at each door frame.

used as the prototype of such an environment adaption. In front of the first office, a smartphone displays the text *Do not disturb* on a red background to imply the social code of a closed door. The second office was equipped with a smartphone that displays the text *Welcome* on a green background to simulate the social code of an open door.

We randomly selected 20 participants in the corridor of an academic facility. We refrained from recording any person-related information about them to avoid any privacy concerns. The 20 participants were divided into two groups of the same size, one subgroup each for *Study A* and *Study B*. The study environment consisted of the corridor that connects two offices—*Office 1* and *Office 2*—located next to each other as shown in Figure 1. Beside these two offices there are further offices on both sides of the corridor while we explicitly excluded the offices on the right side of the corridor. This environment was not fully controlled; therefore, each participant faced a slightly different situation, e.g., people were in the corridor or people entering and leaving offices, when performing the subsequently described task. Both offices in question created the impression that they were occupied, which could be inferred from artificial light shining either through the frosted glass window in the doors in case they were closed or given the fact that the door stood open.

One author of this paper was sitting in *Office 2*, while another author was conducting the study, i.e., recruiting and introducing participants. The task required participants to ask the staff sitting in the offices on the left side of the corridor for the room number of a particular member of staff. We choose this task, because it is a typical question you would ask an unknown person in an office environment. We guaranteed that they do not know anybody in the offices nor the person they were asked to look for. The participant starts the study at the beginning of the corridor which guaranteed both offices within sight. We depicted the start position with a stick figure in Figure 1. To avoid the bias that the participants choose the closer door (Office 1), we defined this door as the door associated with the social code for not-interrupting. The study was considered completed if the participant

either knocks or opens one of the doors or if they are walking further down the corridor. After the participant completed the study, we consulted them for the following questions which differed depending on the study:

- *Study A* and *Study B*: Why did you choose the door of *Office 1* or *Office 2*?
- *Study A* and *Study B*: Why did you not choose the door of *Office 1* or *Office 2*?
- *Study B*: What meaning do you relate to smartphones with green and red background?

Figure 2 shows which door the study participants choose first to ask for the staff member. In *Study A* and *Study B*, the majority choose the second door.

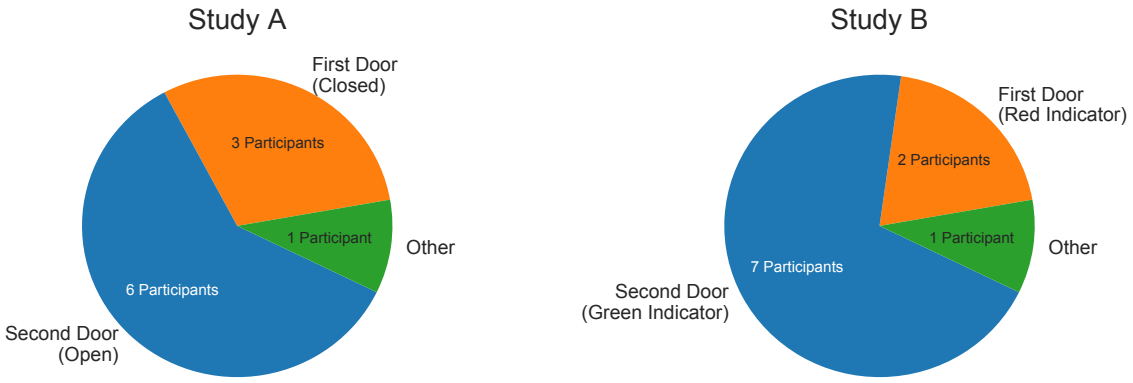

Fig. 2. Distribution of the participants which choose either the first or second door in the explorative study in which we intended to learn about social codes in office environments. One participant of each study chooses an office further down the corridor, which accounts for the *Other* category in the pie chart.

In *Study A*, the majority choose the door of *Office 2*, which is a signal for the presence of a social code implied by closed doors. The reason participants provided in *Study A* for their decision to open the second door was they interpreted the open door as an indicator that they can interrupt the office worker. Participants who opened the first door justified their behavior in the fact that the first door was closer, and they wanted to complete the task both quickly and efficiently. To summarize the results of the *Study A*, we can assume that an open and closed door are recognized by most of the participants as an indicator for an office worker interruptibility.

The results of the *Study B* point out that most of the participants understood that the smartphones introduce a social code to signal who to interrupt for a request. Of the two participants who entered the first door, the first participant stated that they did not recognize the smartphone and wanted to complete the task while the second participant explained they wanted to complete the task and did not care about the smartphones at all. The participant who just walked pass the two doors reported to be confused by the two smartphones mounted at the doors and tried to find another way to solve the task, i.e., other than asking office workers of one of the two offices. To condense the answers of the *Study B*, most of the participants recognized that the smartphone with a red background implies that the office worker is *not-interruptible* while the green background implies the opposite. Green and red were recognized as the signal colors for *interruptible* and *not-interruptible*. The second finding is that the social code of the open and closed door can be transferred to the smartphones mounted on the door frame. This finding supports the assumption that the way we adapt the environment to avoid interruption in inconvenient moments is intuitive to the majority of the participants without any initial explanation.

After elaborating on the results, we enumerate the following threats to validity of the study. They highlight its explorative character. At the same time, they raise the need for a more detailed and controlled execution of the

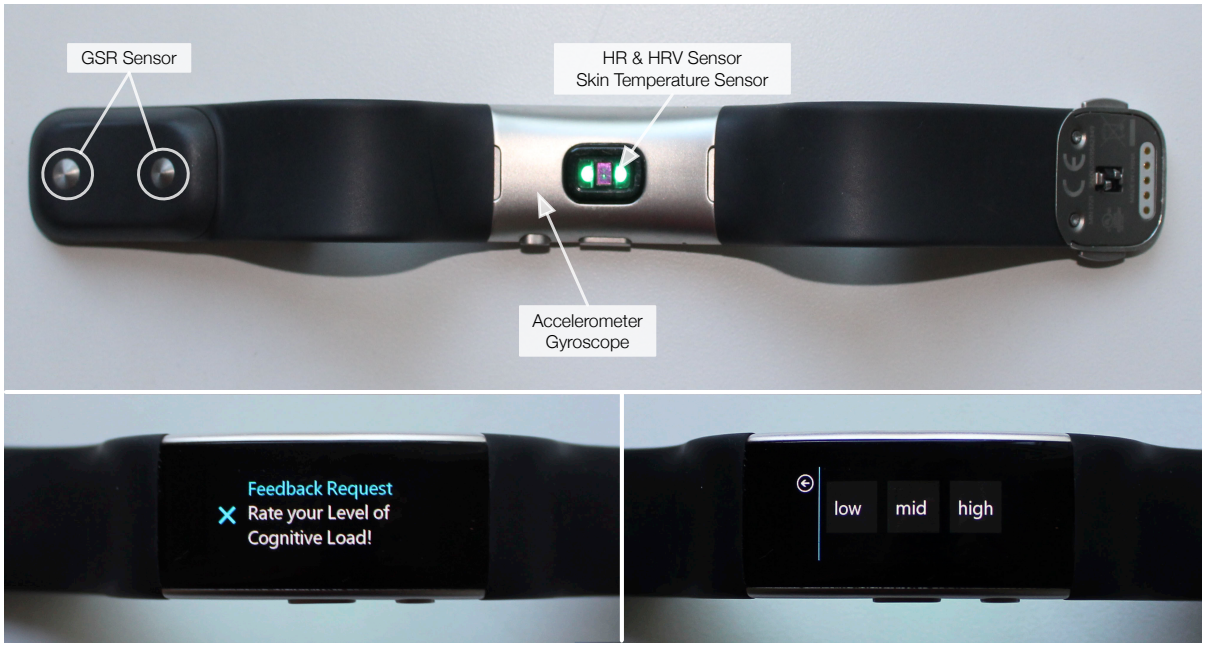

Fig. 3. Overview of the sensors of the Microsoft Band 2 (top). Screenshot of the notification of the feedback request (bottom left). The possible answers for the feedback request (bottom right).

study. The behavior of the participants is not fully influenced by the open or closed doors nor by the smartphones, but by a variety of other factors. Further, a person's principal character traits influence how careful they are about interrupting other persons. The task we gave the participants is artificial and not typical for interruptions in real working environments, in which the importance of a task might influence the decision on whom to interrupt. Furthermore, in an office setting most of the office workers know each other which poses a different attitude and inhibition threshold for the interruption of their peers. Consequently, the effect of both the mounted smartphone as well as the open or closed door tend to vary rather than being clearly determinable upfront.

For the introduction of the COLLINS system the users would be briefed, and we assume the users would understand the meaning of the displayed content of a smartphone. Thereby, the problem of participants' confusion about the smartphones that we discovered would be prevented.

5 COLLINS - AN INTERRUPTION MANAGEMENT SYSTEM

We suggest that interruptions in the workplace should be timed to periods of low cognitive load to improve performance at complex tasks. We developed the COLLINS (**Co**gnitive **L**oad C**l**assification to prevent **In**terruption**s**) system that infers the cognitive load of an office worker using their physiological state to adapt their environment.

The goal of this adaption is to minimize interruptions during phases of high cognitive load and postpone them to phases of low cognitive load. The system measures the physiological state using a Microsoft Band 2

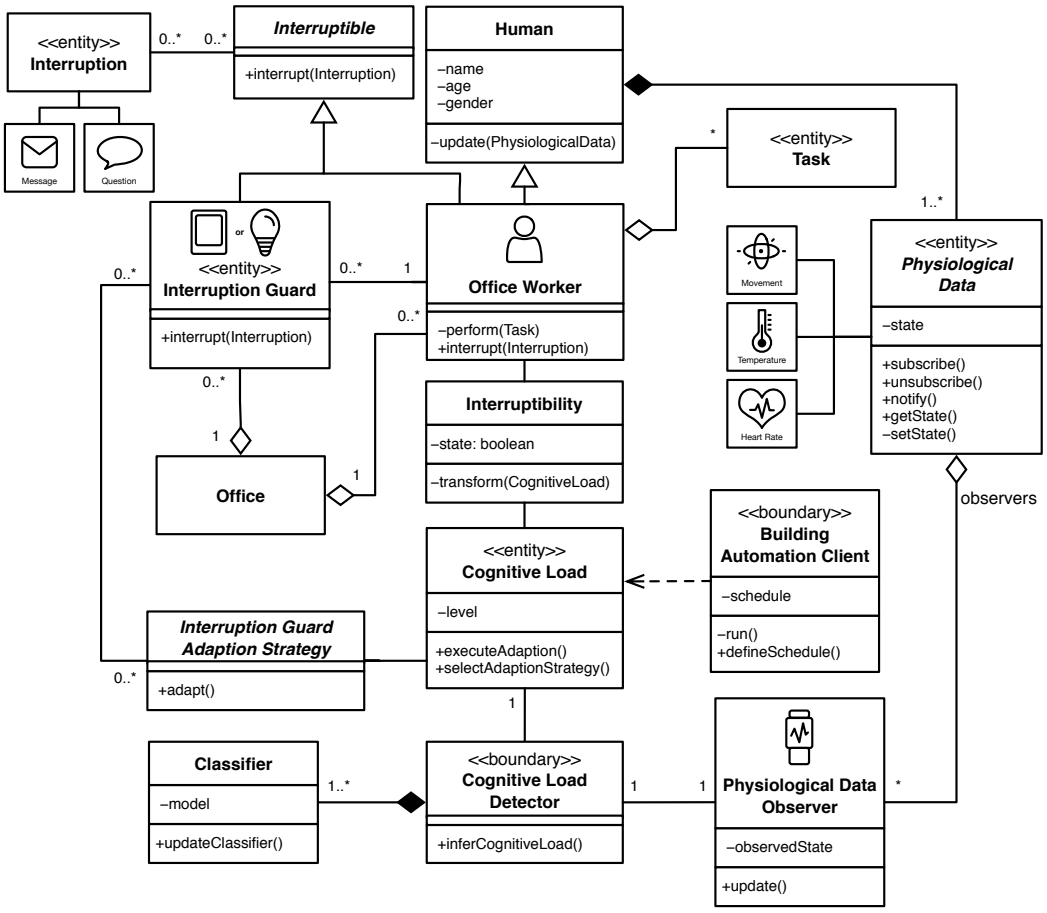

Fig. 4. The class diagram shows the office worker in the center, which is protected from interruptions during phases of high cognitive load. The cognitive load is inferred using the physiological data and—depending on it—interruptions are postponed.

smartwatch worn by the office worker. As depicted in the top part of Figure 3, the Microsoft Band 2 provides all the required sensors. Furthermore, the Microsoft Band 2 is considered to be a consumer smartwatch[1].

The smartwatch measures the heart rate, HRV, GSR, skin temperature, and wrist movements. A smartphone bridges between the Bluetooth connection of the smartwatch and the Wi-Fi of the dedicated server hosting the application domain. In case of a connection loss between the smartphone and the smartwatch, the smartphone application tries to reconnect to the smartwatch every 15 second. We do not handle other error, e.g., running out of battery. The reason for this simple approach for error handling is our implementation serves as a prototype. Using machine learning techniques on the data provided by the smartwatch, the cognitive load can be classified. The inferred cognitive load is then mapped to two states of interruptibility: *interruptible* and *not-interruptible*. The state is visually indicated to the co-office workers in form of a light and an externally mounted tablet computer.

[1]As of today, the future of the Microsoft Band is uncertain: recent reports indicate it is discontinued (http://www.zdnet.com/article/microsoft-pulls-band-listings-from-its-store-admits-no-band-3-this-year/, accessed June 26, 2017). The Microsoft Band SDK is no longer available (http://developer.microsoftband.com/bandSDK, accessed June 26, 2017), requiring future version of COLLINS to use other smartwatches.

5.1 Application Domain

Figure 4 depicts the *class diagram* used to analyze and describe the application domain of the COLLINS system. Furthermore, it serves as the foundation to develop the system architecture.

The *Office Worker*, the human component in the system, performs a *Task* and is protected from an *Interruption* during inopportune points in times. The *Physiological Data* from the *Office Worker* unconsciously adapts depending on the *Cognitive Load* a task induces. The *Physiological State Observer* quantifies the changes in the *Physiological Data*. The *Physiological State Observer* implements an subscriber and the *Physiological Data* implements the publisher in *Observer Pattern* [11]. The quantified data is used by the *Cognitive Load Detector* to infer the *Cognitive Load* using a pre-trained *Classifier*. The inferred *Cognitive Load* is mapped to a state of *Interruptibility*. The *Interruption Guard Adaption Strategy* adapts the *Interruption Guard* depending on the *Cognitive Load* to prevent *Interruption* in phases of high cognitive load. We use the *Strategy Pattern* for *Interruption Guard Adaption Strategy* to adapt different *Interruption Guard* [11]. In this pattern the *Cognitive Load* is the context, the *Building Automation Client* is the client, and the *Cognitive Load Detector* is the policy.

5.2 Cognitive Load Inference

A multi-step approach is used to infer the cognitive load of an office worker illustrated in Figure 5.

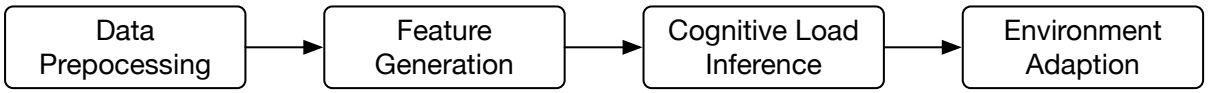

Fig. 5. Block diagram showing the data flow in the COLLINS systems for the inference of the cognitive load.

Firstly, we generate frames of the data stream with a length of 30 seconds. Each frame overlaps 25% with the previous and 25% with the following one. Secondly, the physiological data is cleaned of artifacts such as wrist movements, which distort the measured HRV [10]. Therefore, all RR intervals are removed which differ more than 40% from the average of the frame. We use the same parameter as in [7]. The threshold for the removal of RR frame is doubled to 40%, since our hardware is less accurate than the one used in [7]. A RR interval is the time between two consecutive heart beats.

After the data preprocessing, the feature vector is generated with the following time and frequency features:

- **Galvanic Skin Response**: The average and standard deviation of the time frame.
- **Skin Temperature**: The average and standard deviation of the time frame.
- **Heart Rate**: The average and standard deviation of the time frame.
- **Heart Rate Variability**: The average and standard deviation of the time frame; the root mean square of the successive difference of the RR intervals (RMSSD) and the percentage of the number of successive RR intervals varying more than 50 ms and 20 ms from the previous RR interval (pNN50 and pNN20).

The feature vectors are used in the classification process to infer the current cognitive load. The COLLINS system incorporates several classification algorithms and compares the performance: Support Vector Machine (SVM) with a radial basis function kernel, Random Forest, and the Naïve Bayes Classifier. The *scikit-learn* library was used for the classification of the discrete levels of cognitive load [31]. The classification results in a discrete level of cognitive load.

5.3 Environment Adaption

The inferred level of cognitive load is used to adapt the environment to indicate the state of interruptibility of the co-office workers as *not-interruptible* or *interruptible* (to simplify the interaction of the co-office worker with

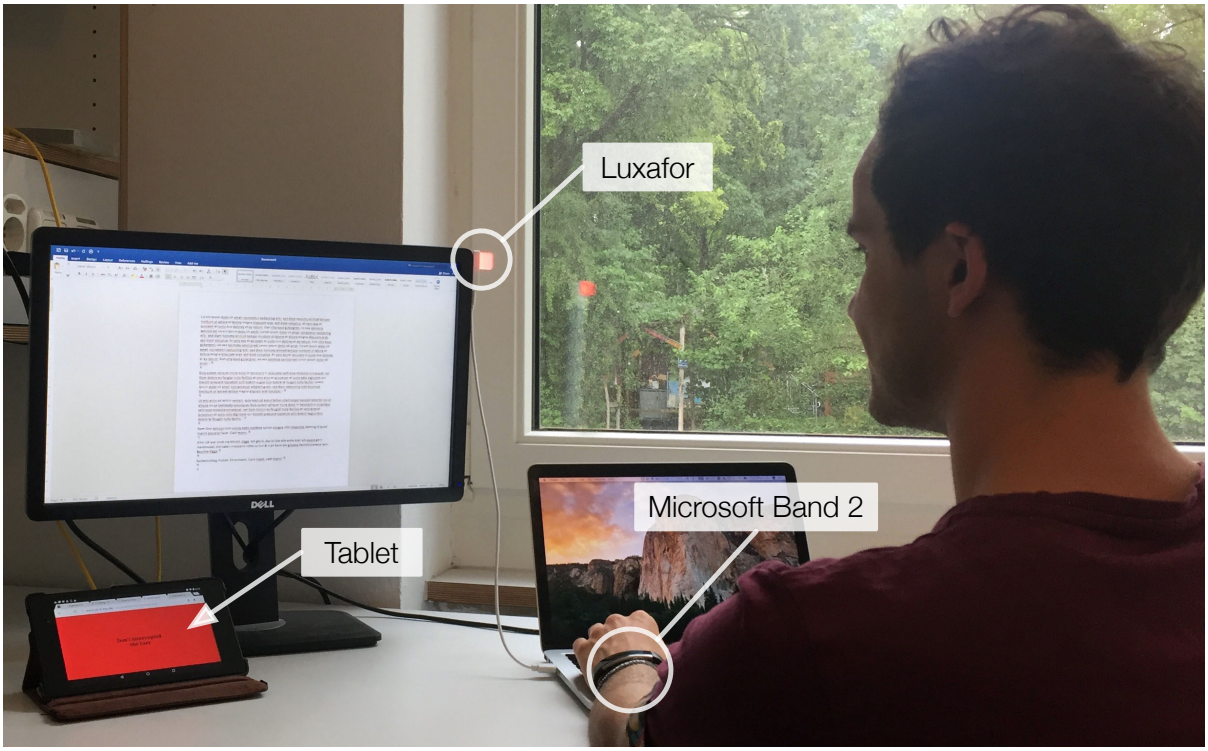

Fig. 6. The COLLINS system indicating that the office worker should not be interrupted.

the system). In Section 4, we showed a system using a smartphone displaying a red or green background is an effective and intuitive method to indicate the interruptibility. Taking this finding as the foundation, we employed in the prototype of COLLINS system two different devices to communicate to co-workers: an externally mounted tablet computer and a light. The tablet computer background color is set to red when a worker should not be interrupted and to green when they can. A Luxafor[2] light is also set to red if the office worker should not be interrupted and to green if they can. This light can be mounted on top of the screen of the computer of the office worker. The setup of the devices placed in a test work space is shown in Figure 6.

For a closed office environment with only one office worker per room we would use a tablet computer mounted next to the door of the office. In other cases, an open office environment or a closed office environment with multiple occupants, a Luxafor mounted on each computer screen would be employed.

Publicly displaying and constantly measuring the current level of interruptibility might raises ethical questions. An office worker whose state of interruptibility is mostly *interruptible* might be labeled by their co-workers as 'lazy', since this state is linked to a low level of the office worker's cognitive load. Züger *et al.* do not report on such correlation in their long-term study with 449 participants [35]. In their study, a system publicly displays the state of interruptibility of each office worker next to their workplace. Their research allows us the conclusion that there is no such correlation. The second ethical questions the system rises is that it could be used for the surveillance and assessment of the employees' performance. Following the same approach as in Züger *et al.* research, we solved this question by not centrally collecting and storing the state of the office workers.

[2] http://www.luxafor.com

6 EVALUATION

In this section, we evaluate COLLIONS with two consecutive studies: a *training study* and a *verification study*. We further include the results and interpretations of each study.

6.1 Evaluation Approach

As a first step, consent forms were solicited from each participant to ensure the ethical use of the generated data for evaluation. Then, using a general survey, we wanted to determine how the participants perceive interruptions during work. The results are described in Section 6.2.

The training study focused on the machine learning classifier for each individual's physiological reaction to different levels of cognitive load. We present the outcomes of this study in Section 6.3. The verification study validated the accuracy of the classification in a normal office setting; in Section 6.4 its findings are described. The structure of the evaluation is shown in Figure 7.

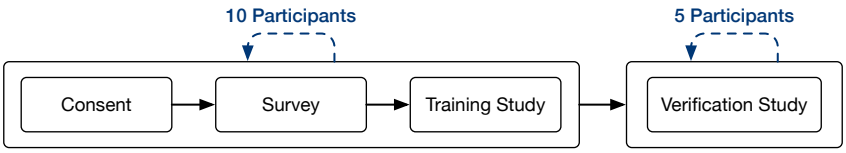

Fig. 7. The evaluation was conducted in two parts. First, ten participants were asked to fill in a consent and survey before participating in the training study. Hereafter, five of them participated in the verification study.

6.2 Survey

Ten participants with an average age of 29.3 (±7.6) were recruited from a pool consisting of faculty staff, Master, and Ph.D. students. The participants were both male and female. The ten participants provided general information on how they perceive interruption during their work. The result of each of the questions is depicted in Figure 8.

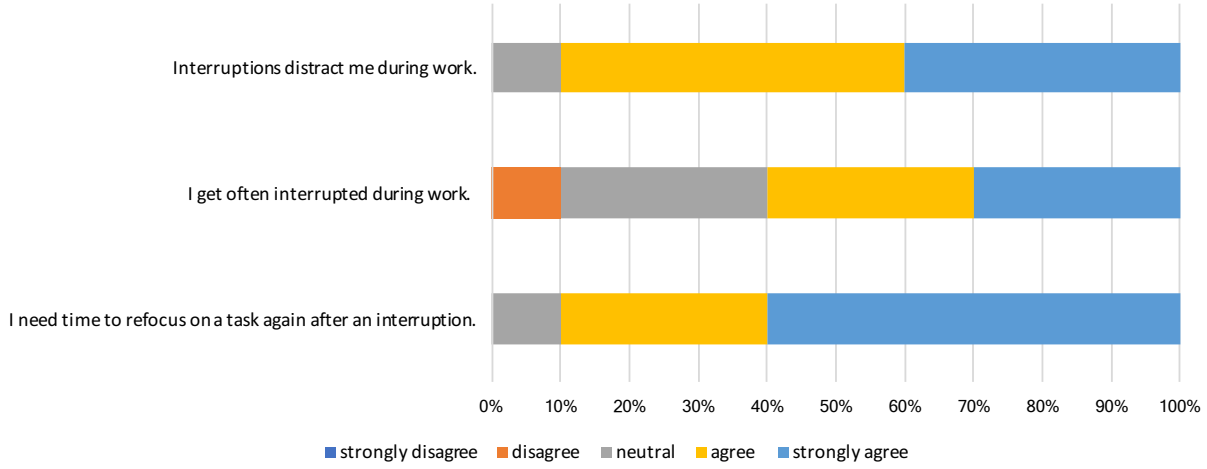

Fig. 8. The survey's results of the three questions reflect the way interruptions are perceived by the participants.

Overall, the responses of the participants can be concluded as follows:

- **Interruptions distract me during work**: 90% of the participants agree or even strongly agree with this statement. Therefore, interruptions are a reason for distractions for our group of participants.
- **I get often interrupted during work**: With this statement only 10% of the participants disagree. Participants perceive that interruptions during work happen frequently.
- **I need time to refocus on a task again after an interruption**: 90% of the participants either agree or strongly agree that they need time to refocus after an interruption as significant.

6.3 Training Study

The *training study* is used to adapt the system to recognize individual reactions of participants experiencing different levels of cognitive load when performing a task. The participants performed three different performance tasks to trigger three different levels of cognitive load (low, medium, and high). Figure 9 depicts the process.

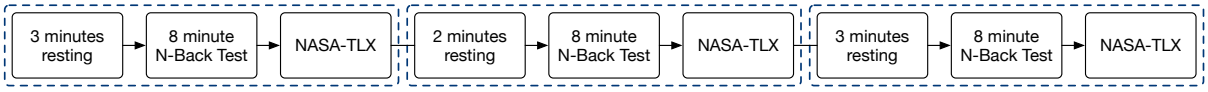

Fig. 9. The training study consists of three parts, while each part includes a resting phase, a phase for inducing the cognitive load, and a NASA-TLX survey for the self-evaluation of the cognitive load.

Before each performance task, a participant passes a resting phase to acquire a baseline as a reference value. Each resting phase is followed by a workload phase. During these workload phases, the participant performs different variants of the Dual N-Back Test to induce low, medium, and high levels of cognitive load [17]. The Brain Workshop[3] software was selected with the following variations of the Dual N-Back Test

- **Task 1**: In the *Position 1-Back Test*, a square appears every three seconds in a three by three grid. The participant has to indicate if the position of the current square is the same as the one shown before by pressing the 'A' key. This should simulate simple repetitive tasks, which induces a low workload.
- **Task 2**: In the *Arithmetic 1-Back Test*, a number between zero and twelve appears every three seconds on the screen. A math operator (plus, minus, times, divide) is presented via the headphones. The participant has to apply the operator to the current number and the one presented before and enter the result. The task represents a medium workload.
- **Task 3**: The *Dual 2-Back Test* is a combination of the two tests mentioned before. A number between zero and twelve is presented in a three by three grid. The participant has to respond if the number appears at the same position in the grid as two steps before by pressing the 'A' key. Furthermore, a math operator is presented via the headphones, which they have to apply to the current number and the number, which was shown two steps before and to enter the result. The task represents a high workload, since the participant has to keep the last two numbers and the last two positions in mind.

After finishing a task, the participant fills in the NASA-TLX questionnaire [15] to state a self-evaluation of the perceived workload for each task. The results of the NASA-TLX survey indicate that the participants perceive Task 1 the lowest and Task 3 the highest cognitive load. The difficulty of Task 2 ranks between Task 1 and Task 3. The results of the NASA-TLX are displayed in Figure 11.

The task performance of the three different Dual N-Back Test, shown in Figure 10, suggests that the difficulty of the tasks is perceived in the same order as designed. Task performance is defined as the percentage of tasks which are correctly solved. The hypotheses that Task 1 is the easiest, Task 3 is the most difficult, and Task 2 is between the two is supported by the results.

[3] http://brainworkshop.sourceforge.net/

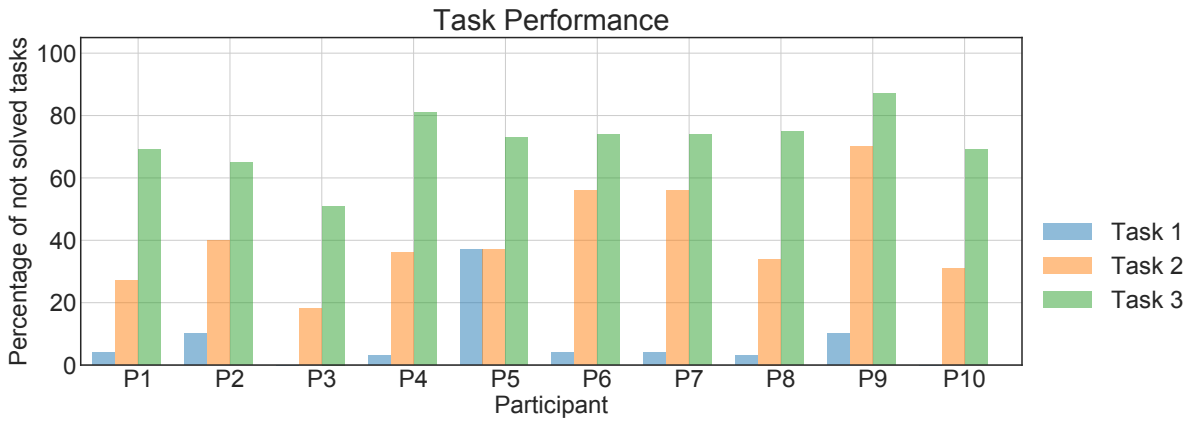

Fig. 10. Percentage of wrong answers of the three levels of the N-Back Test.

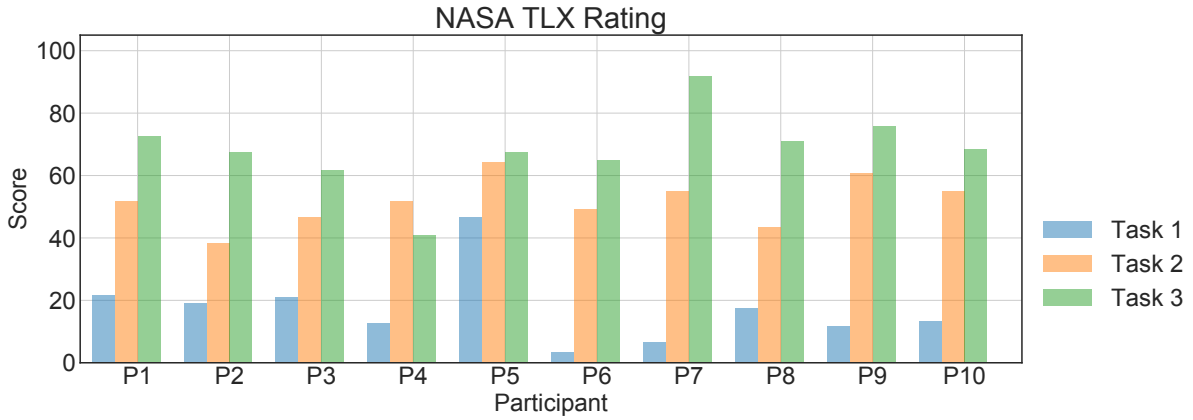

Fig. 11. Result of the NASA-TLX survey for the three levels of the N-Back Test for the self-reported level of cognitive load.

Using the data from the training study, different machine learning classifiers were trained to evaluate their accuracy for inferring the correct level of cognitive load taking the physiological data into account. In Figure 12, the accuracy for the five-fold cross-validation of trained three machine learning classifiers for each of the participants with four classes are displayed. The four classes are the four levels of cognitive load.

Figure 13 reveals the accuracy of the classification for two classes of cognitive load merging the resting and low level as one and medium and high as the other. The Random Forest algorithm has the highest accuracy with one exception for the four cognitive loads and the highest accuracy on average for the consolidation into two classes of cognitive loading.

For a general classifier, the model was trained with the leave-one-subject-out method. The classifier is trained using all data except for one subject, applied in succession to the ten participants. For the ten-fold cross-validation, the accuracy for SVM was 35.5 ± 0.3%, for Random Forest was 31.9% ± 0.5%, and for Naïve Bayes 35.5% ± 0.4%. For the two consolidated classes of cognitive loading, the same method of ten-fold cross-validation revealed an accuracy for SVM of 53.1% ± 0.3%, for Random Forest of 53.1% ± 0.3%, and for Naïve Bayes 53.1% ± 0.3%.

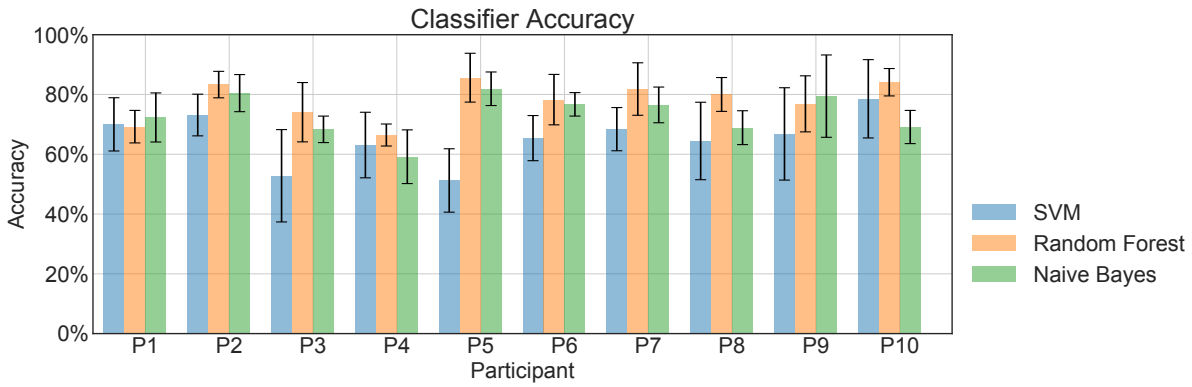

Fig. 12. Classification accuracy for four classes using the different classifiers for the participants.

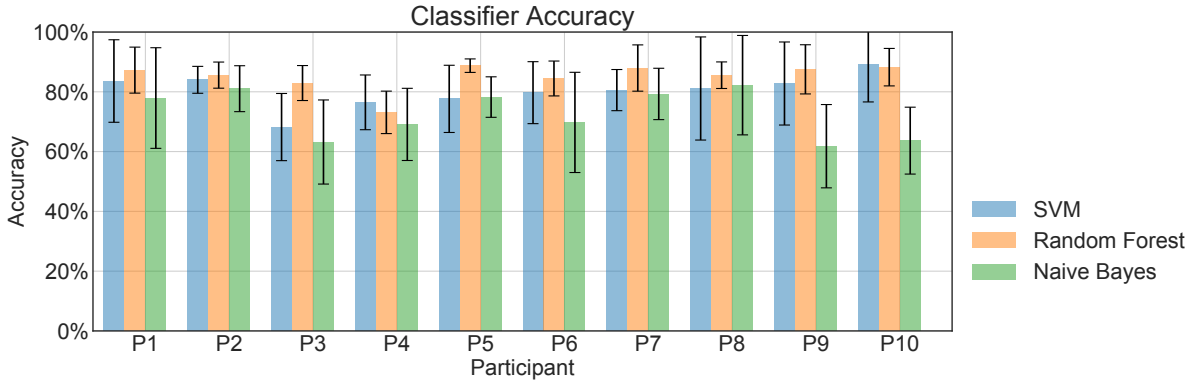

Fig. 13. Classification accuracy for two classes using the different classifiers for the participants.

6.4 Verification Study

The goal of the *verification study* is to validate the classifier trained with the data from the training study and to determine the right threshold in cognitive load for the system to decide to change the state of an office worker from *interruptible* to *not-interruptible* and vice versa. In addition to the cognitive load of an office worker derived from their physiological state during a typical office day, direct feedback is added.

If a certain level of physiologically recorded cognitive load was ongoing for a long-time period, the office worker is interrupted by a systematically triggered notification to ask them to rate the current cognitive load on their smartwatch at three levels: low, medium, or high.

The participants are working on their regular office tasks while the system is continuously inferring their cognitive load. In case the system detects a sufficiently high level of cognitive load, the participant gets interrupted by a vibration alert of the smartwatch. The smartwatch displays a dialog to ask the participant for feedback about the currently perceived level of cognitive load. Both the notification and the answer options displayed to the participant are shown in the bottom part of Figure 3.

A multi-step approach was used to determine if the detected level of cognitive load is long enough or high enough for the individual input. To determine if the pre-trained classifier is accurate, interruptions are triggered

when an office worker is in a phase of high cognitive load for longer periods. This strategy includes a potential uncertainty: If the classifier is inaccurate or the office worker never reaches a high level of cognitive load, the system might never infer a high level of cognitive load. To avoid this gap, the required level of cognitive load or the time is lowered after some time has elapsed, following the five steps as shown in Table 1.

We define as the level of cognitive load of a time frame the most commonly inferred level in case we do not infer the same level for all segments of the time frame.

Table 1. Trigger levels for the verification study. After the time frame stated in the first column, the required level of inferred cognitive load is lowered in case no high enough level was inferred during the time frame. The observed time frame is the duration considered for taking a decision.

Time after level is lowered in minutes	Length of observed time frame in minutes	Minimum level of cognitive load
30	4	high
5	2	high
5	4	medium
5	2	medium
-	1	low

Five participants from the initial pool of ten subjects in the training study were observed for four hours, the time defined to send a maximum of ten interruptions. Interruptions were not sent when they were walking, returned to their desk less than five minutes ago, or if they received an interruption in the last twenty minutes. These decisions were made to improve the user-friendliness of the system during the verification study.

Figure 14 shows the levels of inferred cognitive load from physiological data of each participant. For most participants, almost all time segments were classified into one or two levels. The expected distribution of levels of cognitive load, from the lowest level of *Resting* to the highest level *High Load*, did not occur.

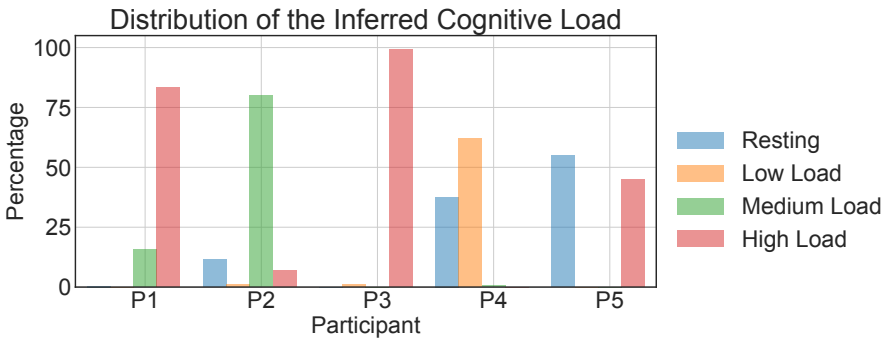

Fig. 14. Distribution of inferred cognitive load in the verification study for the participants.

Figure 15 shows the distribution of the self-reported evaluations of the cognitive load, when the same subjects were interrupted for feedback. For participants P1, P4, and P5, there is a correlation of inferred and self-reported high cognitive loading. For P2, however, the inferred high cognitive load did not match the participant's indication of a low level of cognitive load. Participant P3 answers the feedback with all the three possible response types, even though their level of cognitive load is always inferred with the highest level.

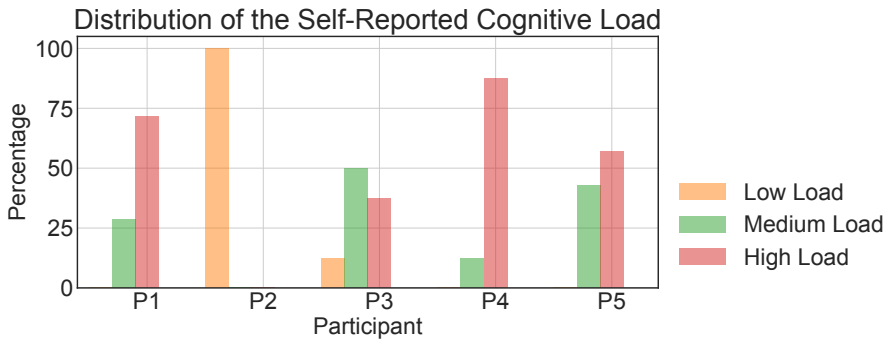

Fig. 15. Distribution of the feedback of the subjective self-reported measurement of the level of cognitive load.

6.5 Interpretations on Cognitive Load

Given the results of the studies described, the accuracy in the classification of the level of cognitive load for the training study is high. A five-fold cross-validation of the individual classifier for each participant has an accuracy between 66% and 86%. This is comparable accuracy to similar research [29] in the classification of the level of cognitive load. We achieve an accuracy for the general classifier between 31% and 35% compared to an accuracy between 35% and 40% using a GSR sensor [29]. We use the same machine learning algorithm Naïve Bayes and SVM. The low accuracy of the consumer wearable used in our research could explain the lower accuracy in classification of the cognitive load. It offers the advantage that it is cheaper and more convenient to wear.

The individual perception of the different tasks' induced level of cognitive load matches the way we designed them. The task performance of the participants and their rating of the task on the NASA-TLX survey supports the hypothesis that Task 1 is the easiest, Task 3 the most difficult one, and Task's 2 difficulty is between both of the other tasks. The survey only considers the scores of the tasks for each participant relative to each other and can not compare the scores of the different participants, since the perception of cognitive load is highly individual.

In the verification study, a lower accuracy in the classification of the level of cognitive load is achieved with two participants not classified correctly. The physiological reaction to different levels of cognitive load, however, is highly personal. The accuracy of a general classifier trained with the data from all the participants shows a lower accuracy than the personalized classifiers. This is an indicator that changes in the physiological state due to changes in the level of cognitive load are highly personal and we need to train a classifier for each office worker.

A second interpretation is that the participant perceived the levels of cognitive load differently compared to the training study. The mapping between the self-reported evaluation level and the inferred level of cognitive load is shifted. For instance, a low level of classified cognitive load could match a self-reported high level of cognitive load for certain participants. For some participants, the tipping point between *interruptible* and *not-interruptible* may be set between the resting level and the low classified level of cognitive load rather than between low and high. This could explain the lower accuracy in classification of the level of cognitive load in the verification study.

6.6 Threats to Validity

The threats to validity of the results and interpretations include:

6.6.1 Correlation between Physiological State and the Cognitive Load. Even though previous research shows that there is a correlation between the physiological measurements and the cognitive load, these measurements are also influenced by other factors. Stress influences the HRV [21], GSR, and the heart rate [18]. Furthermore, there are other factors we cannot control during the evaluation.

Changing health conditions, such as the flu, can alter the skin temperature and the GSR; similar alterations can be observed by a change in the room temperature. We performed the evaluation during winter which is why these conditions applied. Furthermore, the movement of an office worker influences their physiological state.

6.6.2 Accuracy of the Measurement. The Microsoft Band 2 is a consumer product and therefore not made to have the accuracy of a scientific device. Nevertheless, research shows that consumer wearables (Microsoft Band 2, Fitbit Surge, Apple Watch) are reliable for measuring the heart rate in resting condition (sitting) [3].

6.6.3 Dual N-Back Test as Benchmark. In the scientific context, the N-Back Test is used to simulate different levels of cognitive load [7]. However, one question remains: how comparable are these tasks to office work tasks regarding the effect on cognitive load? By using the underlying assumption of our system that interruptions are more disturbing in phases of high cognitive load, it is sufficient that the N-Back Test generates similar levels of cognitive load to office work.

6.6.4 Biased Feedback of the Participants. The participants are interrupted to rate their level of cognitive load. People can have a general bias in their answers. They could tend to lean towards two directions: either they always consider to be in a phase of high cognitive load or the opposite depending on their understanding of the nature of their work. Furthermore, the question always depends on the understanding of the office worker's concept of cognitive load. We do not explain the concept of cognitive load in detail to the participants to prevent to bias them by any explanation. Therefore, they do not have a scale to reference the three answer options to.

6.6.5 Low Number of Participants. The data set we generate is rather small. The training study has only ten participants and the verification study only five. Small data sets tend to have the problem of over-fitting [4]. Furthermore, the composition of them can have an influence on the results: all the participants have a background in academia and the majority is male.

6.6.6 Design of Experimental Setup. The setup of the studies can have an influence on the results. The order of the training study and the way we interrupt people in the verification study is defined and not alternated during the study. However, we shuffle the order of the different Dual N-Back Tests to avoid training effects of the task.

7 FUTURE WORK

In this section, we present an additional study that should be part of our evaluation to test the system in real world environment Furthermore, we describe future extension for the COLLINS system.

7.1 In the Wild Study

With the results of the field study as a foundation, the usability of the COLLINS system should be tested 'in the wild'. This would be a long-term field study in an office environment, in order to quantify the improvements the COLLINS system offers for the productivity and well-being of the office workers as a result of reduced interruptions at inconvenient times. Beside this factor on well-being, we also want to take into consideration if the changes in the hue of light and the knowledge that the office worker is protected from unwanted interruption effect their well-being. This factor could have further effects on the attention span of an office worker.

The study could be implemented by the following scenario: An office worker starts their day by putting on their smartwatch and starting the COLLINS system, which was trained to their personal traits in a training study. The office worker takes care of their typical tasks in the office, while the system is constantly measuring their physiological state. Using this data, the cognitive load is inferred. The cognitive load is mapped to a state of interruptibility. Depending on this state the hue of the smart light mounted on the screen of the office worker is either set to green if they are interruptible or to red if not. The office co-worker knows if the office worker is ready for an interruption or not. The new social code in the office is that a red light implies that the office worker

is *not-interruptible*. The COLLINS system provides the current level of interruptibility in form of an Application Programming Interface (API) to other applications. Consequently, the COLLINS system sets the office worker's status in their instant messaging and email client to *Do not disturb* in order to avoid interruptions of emails and instant messages using COLLINS' API. Consequently, interruptions of the office worker are rescheduled to phases of low cognitive load and are less disruptive. The office workers productivity and satisfaction with time management soar.

The classifier will be continuously adjusted given the feedback of the office worker to improve the accuracy of the classification over time. To quantify the value, the office worker provides feedback on their perception of how well the COLLINS system infers their state of interruptibility. Whenever possible, the improvements in productivity and satisfaction of the worker office by using the COLLINS system will be captured.

7.2 System Extensions

Besides the *in the wild* study, there are possible extensions to the COLLINS system for consideration: A first step is to apply the paradigm of edge computing to the system [12]. This approach in the development of the system would move all the functionality encapsulated in the server to the smartphone. The functionalities are the cognitive load inference and the adaption of the environment. The paradigm suggests moving the processing of the data away from a central authority back to the edges of the network where the data is generated. The source of the data should include also the functionality for analytic and knowledge generation.

The second extension is to test the system with different kinds of smartwatches. There is the *Simband*[4] developed by Samsung, which offers an outlook into future developments in the area of wearable computing. Beside the accelerometer, gyroscope, photoplethysmogram (PPG), GSR sensor, and the skin temperature sensor the Microsoft Band 2 is already equipped with, the *Simband* has additionally an ECG sensor. The ECG combined with the multiple PPG sensors offers a more accurate measurement of the HRV as well as a measurement of the blood pressure. The test with different smartwatches would allow testing the result for reproducibility. Furthermore, it would allow for evaluating the influence of different kinds of sensors and more accurate sensors on the accuracy of the results.

Beside the two before mentioned extension, the third extension is the automation of the training of the system. We want to adapt the COLLINS system to multiple users quickly and without an expert performing the calibration. The approach would be to personalize a general classifier to the individual reactions of a user to cognitive load during the usage of the system. Therefore, the system would ask the user occasionally to rate their level of cognitive load to generate data points of verified levels of cognitive load. Taking these self-evaluated levels into account the classifier could get adjusted during time thus the general becomes an individual classifier.

8 CONCLUSION

The COLLINS system manages disruption for an office worker by scheduling interruptions to times of low cognitive load. The level of cognitive load, which implies a state of interruptibility, is inferred using individual physiological data from a consumer smartwatch and a pre-trained machine learning classifier. To manage interruptions the current level of interruptibility is indicated to co-workers using lights and messages, which show the current state of interruptibility of the office worker. As a result, interruptions can be postponed to a point in time when the office worker is at a level of low cognitive load and interruptions are less disruptive. We evaluated the effectiveness and intuitiveness of social codes of the open and closed door and of smartphones mounted on the door frame for interruption management. The results suggest most participants understood our proposed social codes.

[4] https://www.simband.io/

The COLLINS system has a high accuracy for classification of the level of cognitive load using consumer wearable, between 66% and 86% for a five-fold cross-validation with the individually trained classifier. For the general classifier, a lower accuracy was achieved between 32% and 36% for a ten-fold cross-validation. The level of consensus between self-reported and the inferred level of cognitive load in the verification study had a lower accuracy than in the training study indicating that classification of cognitive loading is highly personal.

Taking these findings as the foundation, we found the following interpretations: A classifier to infer the level of cognitive load based on physiological data for each office worker must be individually trained with classification input from that office worker. Otherwise, the accuracy of the classifier is too low. Also, the perception of cognitive load is highly personal as the findings of the verification study indicate. This implies that the scales of cognitive load need to get adapted to each office worker individually.

ACKNOWLEDGMENTS

This work was supported by a fellowship within the FITweltweit program of the German Academic Exchange Service (DAAD) and the Lothar and Sigrid Rohde Foundation. We thank the participants of the evaluation for their time and helpful feedback.

REFERENCES

[1] Yomna Abdelrahman, Eduardo Velloso, Tilman Dingler, Albrecht Schmidt, and Frank Vetere. 2017. Cognitive Heat: Exploring the Usage of Thermal Imaging to Unobtrusively Estimate Cognitive Load. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 3, Article 33 (2017).

[2] Brian P. Bailey, Joseph A. Konstan, and John V. Carlis. 2001. The Effects of Interruptions on Task Performance, Annoyance, and Anxiety in the User Interface. In *Proceedings INTERACT '01*. IOS Press, 593–601.

[3] Olaf Binsch, Thymen Wabeke, and Pierre Valk. 2016. Comparison of three different physiological wristband sensor systems and their applicability for resilience- and work load monitoring. In *2016 IEEE 13th International Conference on Wearable and Implantable Body Sensor Networks (BSN)*. 272–276.

[4] Christopher M. Bishop. 2006. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA.

[5] Daniel Chen and Roel Vertegaal. 2004. Using mental load for managing interruptions in physiologically attentive user interfaces. *Extended abstracts of the 2004 conference on Human factors and computing systems - CHI '04* 2004 (2004), 1513.

[6] Fang Chen, Jianlong Zhou, Yang Wang, Kun Yu, Syed Z. Arshad, Ahmad Khawaji, and Dan Conway. 2016. *Robust Multimodal Cognitive Load Measurement*.

[7] Burcu Cinaz, Bert Arnrich, Roberto La Marca, and Gerhard Tröster. 2013. Monitoring of mental workload levels during an everyday life office-work scenario. *Personal and Ubiquitous Computing* 17, 2 (2013), 229–239.

[8] Mark B. Edwards and Scott D. Gronlund. 1998. Task Interruption and its Effects on Memory. *Memory* 6, 6 (1998), 665–687.

[9] Eija Ferreira, Denzil Ferreira, SeungJun Kim, Pekka Siirtola, Juha Röning, Jodi F. Forlizzi, and Anind K. Dey. 2014. Assessing real-time cognitive load based on psycho-physiological measures for younger and older adults. In *2014 IEEE Symposium on Computational Intelligence, Cognitive Algorithms, Mind, and Brain (CCMB)*. 39–48.

[10] Hayato Fukushima, Haruki Kawanaka, Shoaib Bhuiyan, and Koji Oguri. 2012. Estimating heart rate using wrist-type Photoplethysmography and acceleration sensor while running. In *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. 2901–2904.

[11] Erich Gamma, Richard Helm, Ralph Johnson, and John Vlissides. 1995. *Design Patterns: Elements of Reusable Object-oriented Software*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.

[12] Pedro Garcia Lopez, Alberto Montresor, Dick Epema, Anwitaman Datta, Teruo Higashino, Adriana Iamnitchi, Marinho Barcellos, Pascal Felber, and Etienne Riviere. 2015. Edge-centric Computing: Vision and Challenges. *SIGCOMM Comput. Commun. Rev.* 45, 5 (2015), 37–42.

[13] Sandy J. J. Gould, Duncan P. Brumby, and Anna L. Cox. 2013. What does it mean for an interruption to be relevant? An investigation of relevance as a memory effect. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 57, 1 (2013), 149–153.

[14] Nitesh Goyal and Susan R. Fussell. 2017. Intelligent Interruption Management Using Electro Dermal Activity Based Physiological Sensor for Collaborative Sensemaking. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 3, Article 52 (2017).

[15] Sandra G. Hart and Lowell E. Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In *Human Mental Workload*, Peter A. Hancock and Najmedin Meshkati (Eds.). Advances in Psychology, Vol. 52. North-Holland,

139 – 183.
[16] Shamsi T. Iqbal and Brian P. Bailey. 2005. Investigating the Effectiveness of Mental Workload As a Predictor of Opportune Moments for Interruption. In *CHI '05 Extended Abstracts on Human Factors in Computing Systems (CHI EA '05)*. ACM, New York, NY, USA, 1489–1492.
[17] Susanne M. Jaeggi, Martin Buschkuehl, John Jonides, and Walter J. Perrig. 2008. Improving fluid intelligence with training on working memory. *Proceedings of the National Academy of Sciences* 105, 19 (2008), 6829–6833.
[18] Palanisamy Karthikeyan, Murugappan Murugappan, and Sazali Yaacob. 2011. A review on stress inducement stimuli for assessing human stress using physiological signals. In *2011 IEEE 7th International Colloquium on Signal Processing and its Applications*. 420–425.
[19] Ioanna Katidioti, Jelmer P. Borst, Douwe J. Bierens de Haan, Tamara Pepping, Marieke K. van Vugt, and Niels A. Taatgen. 2016. Interrupted by Your Pupil: An Interruption Management System Based on Pupil Dilation. *International Journal of Human–Computer Interaction* 32, 10 (2016), 791–801.
[20] Ioanna Katidioti, Jelmer P Borst, and Niels A Taatgen. 2014. What happens when we switch tasks: Pupil dilation in multitasking. *Journal of experimental psychology: applied* 20, 4 (2014), 380.
[21] Desok Kim, Yunhwan Seo, and L. Salahuddin. 2008. Decreased long term variations of heart rate variability in subjects with higher self reporting stress scores. In *2008 Second International Conference on Pervasive Computing Technologies for Healthcare*. 289–292.
[22] Andreas Krause, Asim Smailagic, and Daniel P. Siewiorek. 2006. Context-aware mobile computing: learning context- dependent personal preferences from a wearable sensor array. *IEEE Transactions on Mobile Computing* 5, 2 (2006), 113–127.
[23] Gloria Mark. 2015. *Multitasking in the digital age*. Morgan & Claypool, San Rafael, CA, USA.
[24] Gloria Mark, Daniela Gudith, and Ulrich Klocke. 2008. The Cost of Interrupted Work: More Speed and Stress. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '08)*. ACM, New York, NY, USA, 107–110.
[25] Abhinav Mehrotra, Robert Hendley, and Mirco Musolesi. 2016. PrefMiner: Mining User's Preferences for Intelligent Mobile Notification Management. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '16)*. ACM, New York, NY, USA, 1223–1234.
[26] George A. Miller. 1956. The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological review* 63, 2 (1956), 81.
[27] Yoshiro Miyata and Donald A. Norman. 1986. Psychological issues in support of multiple activities. *User centered system design: New perspectives on human-computer interaction* (1986), 265–284.
[28] Stacey F. Nagata. 2003. Multitasking and Interruptions during Mobile Web Tasks. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 47, 11 (2003), 1341–1345.
[29] Nargess Nourbakhsh, Yang Wang, and Fang Chen. 2013. *GSR and Blink Features for Cognitive Load Classification*. Springer Berlin Heidelberg, Berlin, Heidelberg, 159–166.
[30] Fred Paas and Jeroen J. G. Van Merriënboer. 1994. Instructional control of cognitive load in the training of complex cognitive tasks. *Educational Psychology Review* 6, 4 (1994), 351–371.
[31] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake VanderPlas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Edouard Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.
[32] Jan L. Plass, Roxana Moreno, and Roland Brünken. 2010. *Cognitive load theory*. Cambridge University Press.
[33] John Sweller. 1988. Cognitive Load During Problem Solving: Effects on Learning. *Cognitive Science* 12, 2 (1988), 257–285.
[34] Takahiro Tanaka, Ryosuke Abe, Kazuaki Aoki, and Kinya Fujita. 2015. Interruptibility Estimation Based on Head Motion and PC Operation. *International Journal of Human–Computer Interaction* 31, 3 (2015), 167–179.
[35] Manuela Züger, Christopher Corley, André N. Meyer, Boyang Li, Thomas Fritz, David Shepherd, Vinay Augustine, Patrick Francis, Nicholas Kraft, and Will Snipes. 2017. Reducing Interruptions at Work: A Large-Scale Field Study of FlowLight. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. ACM, New York, NY, USA, 61–72.
[36] Manuela Züger and Thomas Fritz. 2015. Interruptibility of Software Developers and Its Prediction Using Psycho-Physiological Sensors. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. ACM, New York, NY, USA, 2981–2990.

Received August 2017; revised November 2017; accepted January 2018