Sensing Interruptions Using Psycho-Physiological Sensors

Anna Scholtz
University of British Columbia
Vancouver, Canada
ascholtz@cs.ubc.ca

Hayley Guillou
University of British Columbia
Vancouver, Canada
guillouh@cs.ubc.ca

ABSTRACT

Interruptions are a key factor in decreasing productivity; however, our current methods of detecting and tracking interruptions can be inaccurate and invasive. By detecting interruptions objectively, we could more accurately study the effects of interruptions. In this paper, we collected psycho-physiological data from sensors using a Muse headband and created a classifier that detects if a segment of time is an interruption or not. We found that our classifier has an accuracy of 87.2% and has the potential to capture that an interruption is happening in real-time. Our contributions are an experimental design that captures interruptions using psycho-physiological sensors and a promising method of analysis that could lead to a real-time, objective identification of interruptions.

CCS CONCEPTS

• **Information Interfaces and Presentation**;

KEYWORDS

Interruptibility, EEG, Psycho-physiological sensors

1 INTRODUCTION

Whether working in a shared office space or alone, interruptions are destined to occur. They are seen to be disruptive and have a negative impact on productivity, but in order to measure the effects of interruptions it is useful to know when they happen. In previous studies, external and self-interruptions have been detected by observations and self-reporting; [6, 10] however, this is invasive, not always exact, and might influence the final results.

With recent advances, psycho-physiological (biometric) sensors have become more practical to use in casual settings. Improvements in form and accuracy has made it possible for these sensors to be worn comfortably and with minimal disruption during controlled experiments and field studies. Using psycho-physiological sensors for detecting interruptions would reduce the need for self-reporting and improve accuracy over observing perceived interruptions. Discovering when interruptions happen would help make studies less invasive. For example, in the context of HCI this is useful when determining which interface design of an application is best to keep the user focused on the task at hand.

In our work, we aim to investigate whether interruptions cause measurable changes using psycho-physiological sensors by examining brain activity and head position. We also examine if different types of interruptions elicit different and distinguishable psycho-physiological responses. The three types of interruptions to be analyzed are *interruptions that require an immediate response*, like a person asking a question or a phone call; *interruptions with a delayable response*, like email notification or a task that can be done at a later time; and *interruptions that require no response*, like loud noises. Due to the limited scope of this project, we focus on the first two mentioned types. The main research questions we address are:

(1) Can interruptions be detected with psycho-physiological sensors?
(2) Can interruptions with delayable or immediate responses be differentiated?
(3) How do different types of interruptions affect the recovery time?

To address these questions, we design a user study in which we have participants continuously work on a main task while being interrupted by another task. We use a sensor headband to measure psycho-physiological responses from nine participants and also collect key-logging data while they complete the tasks. We analyzed this data and trained a classifier, with the best classifier having an accuracy of 87.2% when classifying segments as interruptions and non-interruptions. Our contributions are a proposed study that captures interruptions using psycho-physiological sensors and a promising method of analysis that could lead to a real-time, objective identification of interruptions.

2 RELATED WORK

Different related work looked into the effects [3, 5], disruptiveness [7, 12] and types [2, 6] of external, as well as internal, interruptions. Psycho-physiological sensors have previously been applied to measure the interruptibility of software developers [14] and to determine task difficulties [4]. In [7] pupil dilation was measured during external and internal interruptions and used to determine the disruptiveness. EEG measures the electrical activity in the brain. Spectral power frequency bands can be extracted from the recorded signals using fast fourier transformation and allow conclusions to be drawn on the cognitive state, such as being focused or attentive [1]. These states might also give an indication of whether a person feels interrupted. However, no related work has been found that focuses on detecting interruptions using psycho-physiological sensors.

EEG sensors produce a large amount of data recorded from multiple channels. In [11] an approach for analysing EEG data is presented. It proposes a method where the data gets denoised, then segmented, then for each segment features are extracted followed by a feature selection procedure. The extracted features can be used for machine learning, for instance for training a classifier. Classification has been applied to EEG data in the context of classifying different recorded epileptic seizures [13]. Therefore we assume that it might also be possible to detect and classify different interruptions recorded by EEG sensors and by applying the proposed method of analysis.

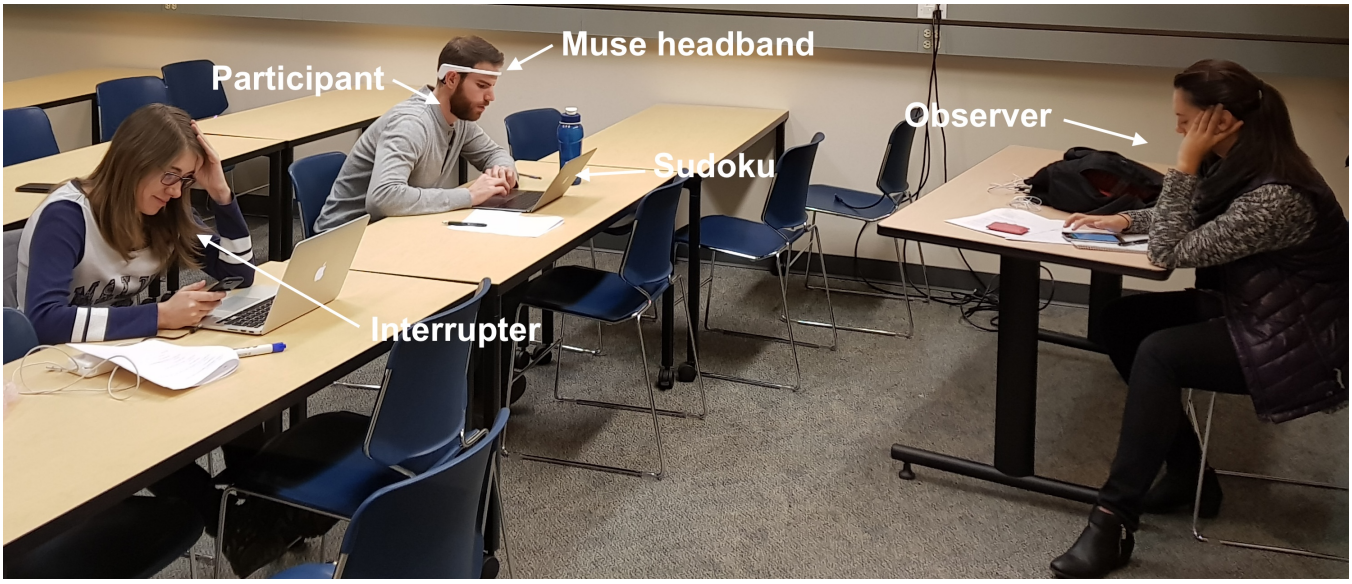

Figure 1: Setup of the experiment with participant solving the Sudoku while wearing the Muse headband. Interrupter interrupts participant by asking questions while observer writes down observations about the movements and behaviour of the participant.

3 EXPERIMENT

We conducted experiments with 9 participants. During the experiments subjects had to solve Sudoku puzzles and were interrupted several times while their psycho-physiological measurements were recorded.

3.1 Experimental Design

The main task was to work on medium difficulty sudoku puzzles. Although we considered using a programming task, ultimately we decided that working on a sudoku would provide more consistent data with less confounding factors. Medium difficulty was chosen so that the puzzles would not be trivial but also not too difficult; skilled participants would not complete them too quickly and novice participants would not be too stuck to continue filling in squares. Participants could also skip a sudoku if they made mistakes or felt stuck.

The interrupting task was similar to the one in [7]: questions from a casual chat conversation and simple arithmetic, such as "What is your favourite movie?" and "Calculate 2 * 19 + 32". The interrupting task is irrelevant to the main task, requires some time, and can be interesting for the participant. We had a collection of 30 different questions, giving each participant a subset of these. We wanted the questions to invoke some thought and since our participants knew each other we switched the questions around for each participant.

The windows used in the experiment are shown in Figure 2. The notifications were built into the same window as the sudoku and a small dialog popped up when answering a question.

3.2 Conditions

One factor was varied into the experiment, the response to the interruption could be delayable or required to be immediate. There were 16 interruptions for each participant, 8 of each kind.

3.3 Apparatus and Setup

Figure 1 shows the experimental setting. Participants were seated at a table with a 12 inch LCD laptop with screen resolution of 2304-by-1440 pixels and screen density of 226 pixels/inch. They were asked to wear a Muse headband that captures brainwave data recorded with four sensors located at different spots on the scalp and ears. The headband also captures eye blinks, head movements and forehead touches.

We recorded the data captured by the Muse headband using MuseLab as shown in Figure 3. MuseLab automatically divides the EEG data into alpha (7.5-13Hz), beta (13-30Hz), gamma (30-44Hz), delta (1-4Hz) and theta (4-8Hz) frequency bands. In addition, we collected all the input data and input times during the Sudoku as well as the times and durations for the immediate interruptions and demographic information. Overall, we collected 4.5 hours of data.

3.4 Participants

9 participants (age range/mean/SD: 22-38/27/5.05, gender: female 33.3% and male 66.6%) including the two authors of this paper, participated in this experiment. All our participants had varying exposure to sudoku before the study. They all gave informed consent and received no compensation for their participation.

3.5 Procedure

The experiment lasted approximately 30 minutes. Before the experiment started, participants sat in meditation with closed eyes

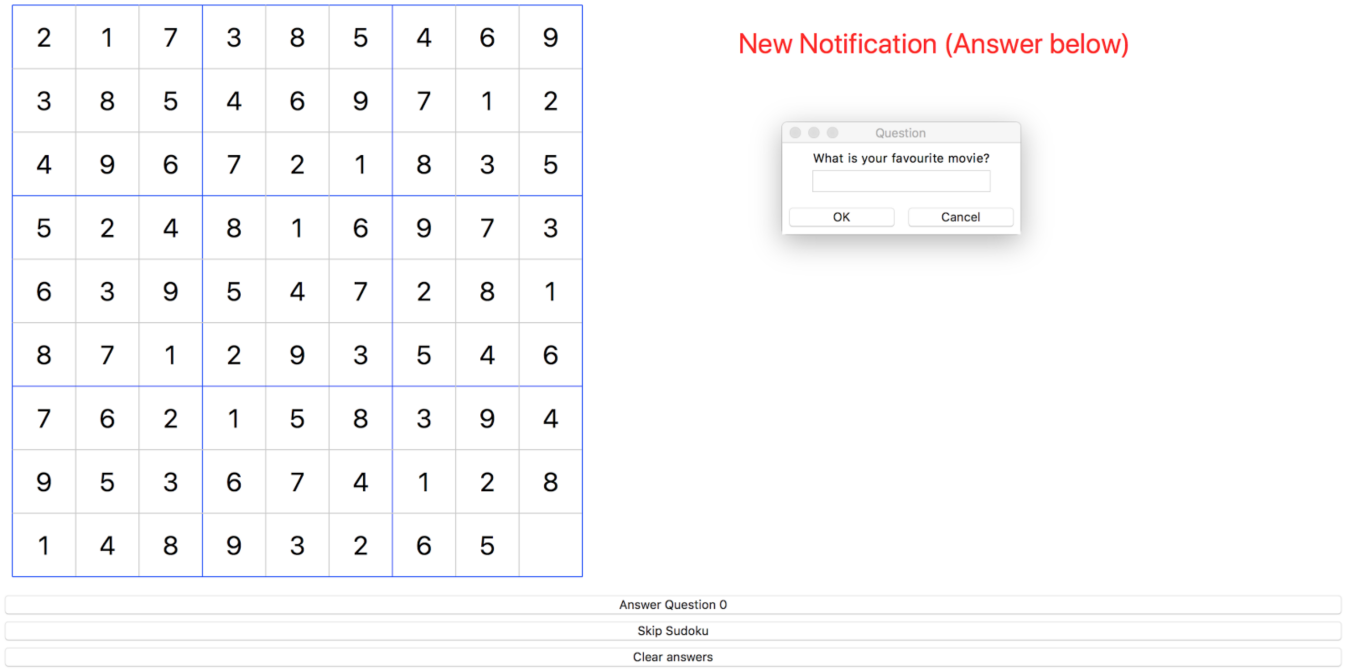

Figure 2: Sudoku screen with new notification and question dialog.

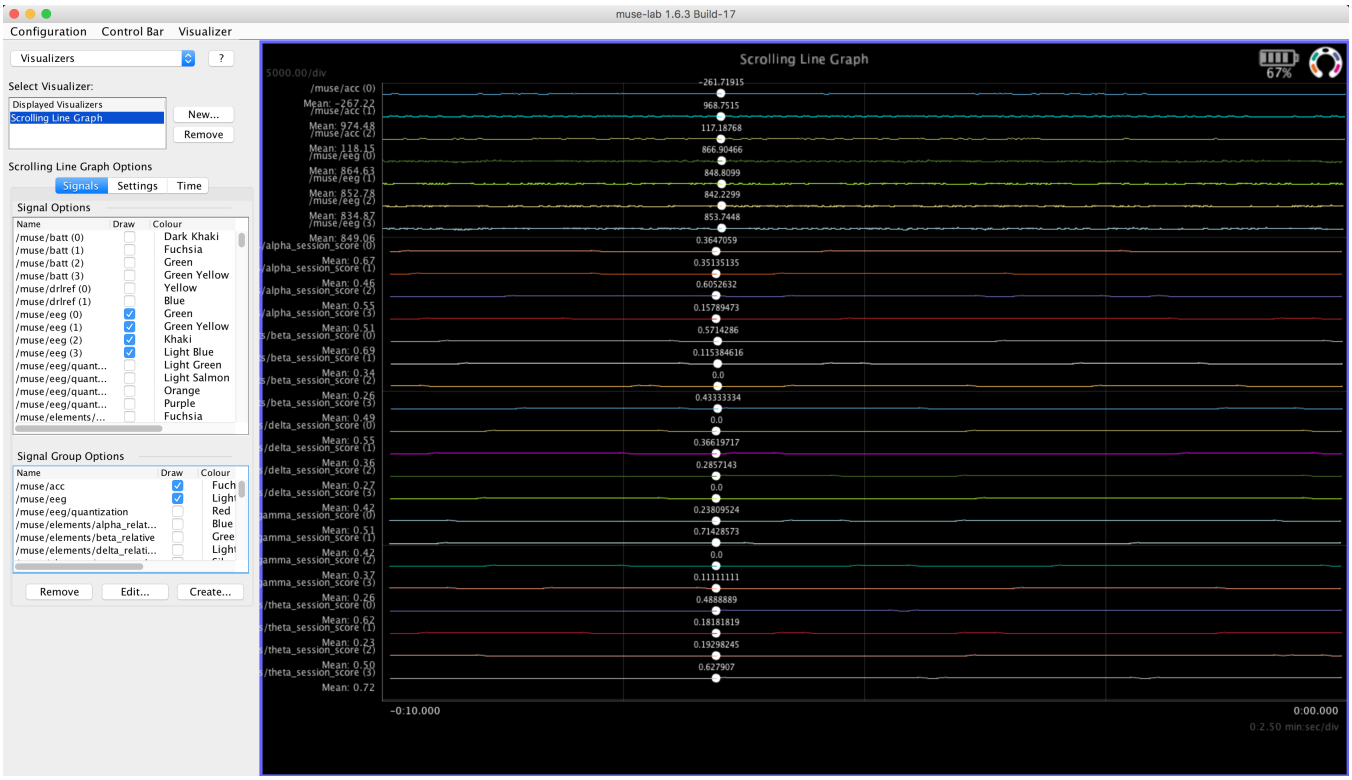

Figure 3: Muselab. Recording of EEG and accelerometer data of the participant.

for 2 minutes in order to capture their baseline EEG recordings. They were instructed that the notifications on the computer could be answered anytime before the end of the experiment and the questions in person should be answered immediately. A new notification would appear in the sudoku program every 3 minutes. One of the experiment runners would ask a question every 3 minutes, alternating with the notifications. The second experiment runner observed the participant and wrote down patterns of movement. Every move the participant made was logged in a text file. After 30 minutes the participants were informed that the experiment was over.

4 DATA ANALYSIS

The recorded data as well as logged interruptions are used to train a machine learning classifier. In the following, we describe the steps of preparing the data and training the classifier.

4.1 Data Cleaning

The Muse headband samples EEG data at a frequency of 220Hz. The accelerometer data is sampled at 50Hz. Eye blinks and forehead touches are inferred from this data. We applied a butterworth bandpass filter with corner frequencies between 0 to 50Hz to the EEG data in order to reduce noise. The accelerometer data was filtered using a butterworth bandpass filter with corner frequencies between 0 to 20Hz.

4.2 Normalization

Since the amount of data is too limited to create machine learning classifiers for individual participants, we decided to train the classifiers across participants. EEG data is particularly individual for each participant; therefore, we decided to normalize the data using the baseline data of each participant. The baseline data was recorded for two minutes before the Sudoku started. We asked participants to close their eyes and relax during this time. For normalization we subtracted the baseline data from the recorded data.

4.3 Data Segmentation

The recorded data consists of parallel time series of various values. In order to transform the collected data into a feature vector suitable for training a classifier, we first applied segmentation. In general, the segmentation is based on sliding windows. Each windows gets assigned one of three labels: *no interruption*, *immediate interruption* or *delayable interruption*. The interruption labels were assigned to windows which overlapped with an interruption task. All other windows were labelled as no interruption.

At first, we tried to segment the entire data into overlapping segments (cf. ① Figure 4). We performed multiple iterations with different window sizes and offsets. On average, an immediate interruption took 16 seconds and a delayable interruptions took 19 seconds, so we chose window sizes between 5 and 10 seconds. However, after generating features and training classifiers none of the classifiers was able to detect interruptions in the test data. Several reasons might have contributed to this poor outcome. Firstly, the segments labelled as no interruption could contain noisy data quite similar to actual interruptions. We observed that sometimes shortly after an immediate interruption ended, participants continued working but seemed less focused and started talking. Other participants started coughing or moving. Secondly, some segments labelled as being an interruption may have only had a few milliseconds of the actual interruption. Those segments do not contain enough data for characterizing the effect of an interruption and the generated features would be more similar to non-interruptions. Lastly, our observations show that the most significant changes, such as movements, happen in the first 1 to 10 seconds of an immediate interruption. The remaining time of the interruption might not present enough changes to detect an interruption.

Based on these findings, we modified our data segmentation approach (cf. ② Figure 4). Non-interruption data that occurred within 60 seconds after an immediate interruption ended was discarded in order to reduce noise. In addition, only data that contained the critical starting interval of the interruption was used for segments labelled as interruptions. These segments also had to exceed a minimum overlap threshold with the interruption. We chose to have non-overlapping segments, otherwise data present in two different segments could get labelled differently and negatively influence the accuracy of the classifier. This is especially crucial for segments close to the start of interruptions.

We ran several iterations with different window sizes (3, 4, 5, 6, 8, 10 seconds) to test which windows size best contained the whole critical starting interval of an interruption.

4.4 Feature Generation

Next, feature extraction was applied to each segment. The following features were used:

- Accelerometer: $\Delta\{\text{Min, Max, Mean, Stdev}\}X, Y, Z$
- Frequencies: $\Delta\{\text{Min, Max, Mean, Stdev}\}\alpha, \beta, \gamma, \delta, \theta$
- Relations of frequencies: $\Delta\frac{\alpha}{\beta}, \Delta\frac{\alpha}{\gamma}, \Delta\frac{\alpha}{\delta}, \Delta\frac{\alpha}{\theta}, \Delta\frac{\beta}{\alpha}, \Delta\frac{\beta}{\gamma}, \Delta\frac{\beta}{\delta}, \Delta\frac{\beta}{\theta}, \Delta\frac{\gamma}{\alpha}, \Delta\frac{\gamma}{\beta}, \Delta\frac{\gamma}{\delta}, \Delta\frac{\gamma}{\theta}, \Delta\frac{\delta}{\alpha}, \Delta\frac{\delta}{\beta}, \Delta\frac{\delta}{\gamma}, \Delta\frac{\delta}{\theta}, \Delta\frac{\theta}{\alpha}, \Delta\frac{\theta}{\beta}, \Delta\frac{\theta}{\gamma}, \Delta\frac{\theta}{\delta}$
- Task difficulty: $\Delta\frac{\theta}{\alpha+\beta}, \Delta\frac{\beta}{\alpha+\theta}$
- Number of eyeblinks

The frequency bands $\alpha, \beta, \gamma, \delta, \theta$ are derived using fast fourier transformation. The Muse SDK automatically calculates these frequencies. Certain frequencies are characteristic for different cognitive states. The ratio of two frequencies is used to allow comparison between participants. We also compute the measure of task difficulty as introduced in [8, 9].

In addition, head movements which are captured by the accelerometer might indicate immediate interruptions. We observed that some participants moved turned their head to look at the interrupter while answering the interruption.

4.5 Feature Selection

We selected the most significant features by running them through a SelectKBest feature selector. To calculate the significance between features and labels, the ANOVA F-value for the provided samples was computed. Features that were not significant were removed from the feature vectors.

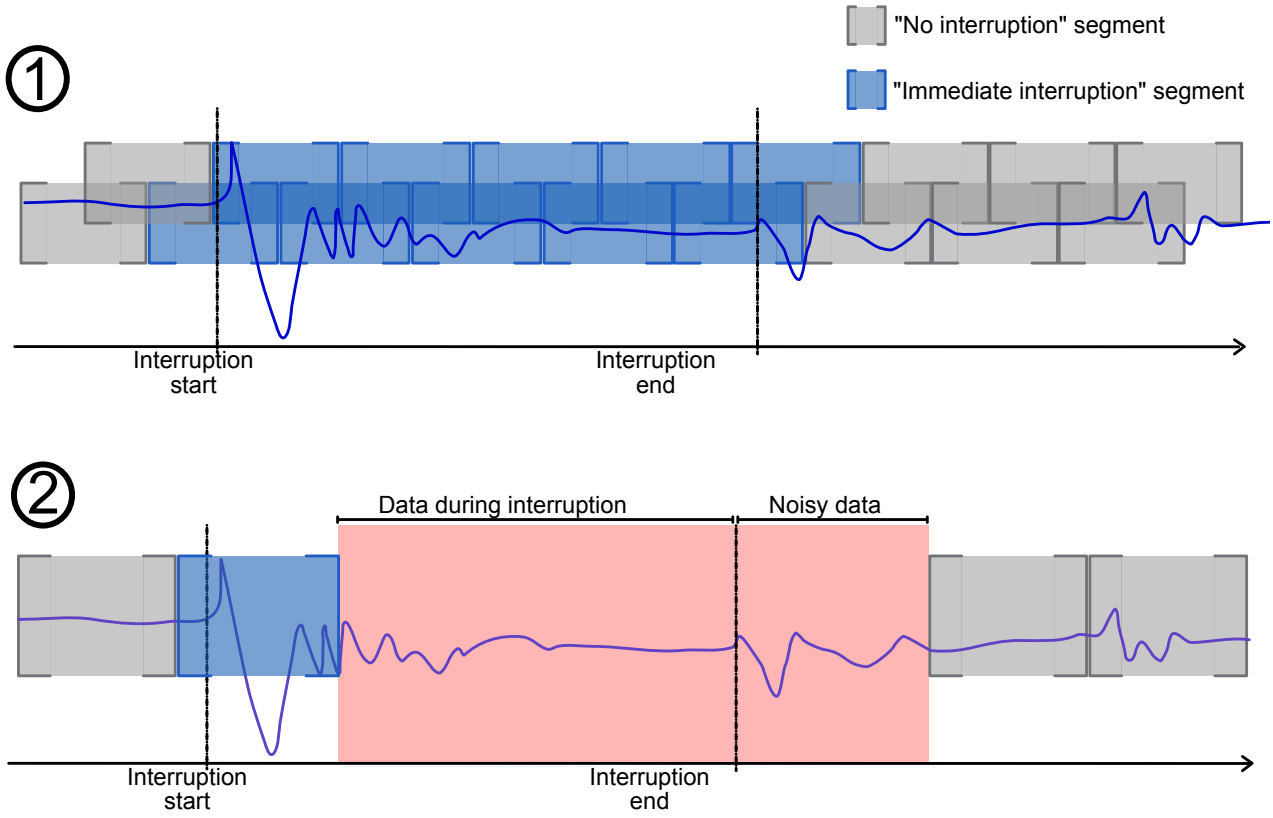

Figure 4: Segmentation of the recorded data. ① First attempt: segmenting and labelling the entire data. ② Second attempt: Segmenting and labelling selected parts of the data.

4.6 Training the Classifier

The feature vectors were used to train various classifiers. We selected different classification algorithms to see which performed best. First, we considered Naive Bayes, however, Naive Bayes does not allow negative data. Next we trained a Support Vector Model (SVM) and a Random Forest Classifier. A few initial tests showed that the Random Forest Classifier had much more promising results compared to SVM, so we chose to train a Random Forest Classifier for detecting interruptions.

To train the classifier we used the data of one participant as test data and the data of the remaining eight participants as training data (leave-one-out strategy). We performed nine training iterations on untrained Random Forest Classifiers in order to ensure that each participant's data is used for training and testing. In addition, we performed iterations with varying window sizes.

5 RESULTS

In the following we present our results of training a classifier to detect interruptions and determine whether those are immediate or delayable interruptions.

5.1 Detecting Interruptions

To detect interruptions, we trained a random tree classifier with the psycho-physiological data collected during the experiments. We are especially interested in detecting when an interruption starts. For training and testing we used the leave-one-out strategy, in each training fold we used the data of one participant for testing and the remaining data for training. The window sizes significantly influence the classification performance, so we determined the window sizes delivering the best results. The window size with the best accuracy was 8 seconds.

We compared the accuracy of our classifier to a naive classifier that simply assumes that there are no interruptions. Overall, the best classifier had an accuracy of 87.2% and was slightly better than the naive classifier with 81.3%. Next, we examined the rate of false positives the classifier predicts. 50% of the interruptions are correctly detected and the false positive rate is 15.7%. This means 15.7% of the non-interruption segments are predicted as being interruptions.

5.2 Classification of Interruptions

We trained our classifier to distinguish between immediate and delayable interruptions. Out of the segments correctly determined to be interruptions, 75% of those were classified correctly. There were

no significant differences between the accuracy of differentiating between immediate and delayable interruptions.

6 DISCUSSION

While we are pleased with the accuracy of our classifier, at this point it is difficult to make a conclusion on whether or not our classifier is statistically significant due to large amounts of noise in our data. We had a relatively high rate of false positives, which may have actually been other interruptions such as the participant asking a question, a loud noise, or someone else in the room causing a disruption.

If we were to continue on with this project, our next step would be to make changes in our experimental design that would hopefully lead to more consistent results. We would have a designated, quiet room with as few other distractions as possible. With a consistent room, we could have the same seating arrangement for all participants. With fewer outside distractions, such as windows, other people in the room, or noises, we would be able to classify our non-interrupted periods more accurately and would potentially have less false positives. We would also have recorded our baseline before explaining the task to avoid the higher levels of excitement participants were experiencing right before starting the task.

Without the time-constraints of a class project, we would have used this initial study as a pilot to tweak our experimental design before running a larger study with more participants. Having higher quality data would give us more confidence in our results. Even with our current data, the average segment length we used for classification was 8 seconds, which means that a real world implementation would react in close to real-time.

In addressing our research questions, we have found that our results show potential in being able to obtain a real-time, objective identification of interruptions in a workflow with psycho-physiological sensors. Our classification of the two different types of interruptions was less accurate; however, we spent less time on this problem than on classifying general interruptions. The next step after properly verifying this would be to conduct a field study to see if our classification could be generalized outside of a controlled environment.

For our third research question, we looked at the recovery time after an interruption, that is, the time it took to add another number to the sudoku after being interrupted. Without an interruption, participants added numbers to the puzzle at an average rate of 15 seconds per number. For interruptions with a delayable response the recovery time was 25 seconds, significantly longer than the average rate. An interpretation of this could be that participants would choose to be interrupted when they were already stuck on the sudoku and saw that as a good time to complete a different task. For interruptions requiring an immediate response, participants had an average recovery time of 15 seconds, the same as the average rate, which we found to be perplexing. A possible explanation is that our interrupting task was not disruptive enough to make the participant lose focus. It would be interesting to see what would happen if we interrupted with a similar task, i.e. interrupt the sudoku-doing participant with another sudoku, which would be similar to getting interrupted with a programming question while programming. Although this research question was originally one that we marked as optional, we found that there are interesting differences in the interruptions that may lead to more accurate classification in the future.

7 THREATS TO VALIDITY

As with any controlled lab study, there are several threats to the validity of our study.

7.1 External Validity

It is difficult to say how generalizable our classifications could be at this point. We only have one kind of interrupting task, so it is still unknown whether this would be applicable to work environments outside of a controlled study. In the future, we could perform a field study to validate potential findings.

7.2 Internal Validity

By using biometrics, a threat to the study is that the data captured with the sensors might be affected by aspects such as the study participants' personality traits or general stress level. To mitigate this risk, we captured a baseline and normalized the data with it. As stated before, in the future we would take this baseline earlier in the session and would incorporate a relaxing video instead of meditation.

Different head sizes and movements of participants prevented ideal contact of sensors, so quality of data is varying for each participant. An alternative for the Muse headband in the future would be the Muse glasses, but at this point the data collection for the glasses would require more time than we had available for this project.

During the experiment, some participants got stuck on some sudokus which caused some frustration. This might have influenced the mental state of the subjects. Alternatively, in future studies a larger set of easier Sudoku puzzles could be used or participants could choose the difficulty based on their proficiency.

8 CONCLUSION

Interruptions are currently detected using self-reporting and observation which can be inaccurate and invasive. We propose the possibility of detecting interruptions using psycho-physiological sensors. As a first step towards this, we designed a controlled study that has participants work on a main task while being intentionally interrupted. We used the Muse headband to take EEG and head movement recordings and used keylogging software to measure the participant's rate of work on the main task. Using this data, we created a classifier that has an accuracy of 87.2% when classifying segments as interruptions and non-interruptions.

If continuing this work, we would have a number of goals. First, we would collect more data with our improvements to the experimental design, which would allow us to collect higher quality data to increase classification accuracy. Second, we would run a field study to confirm that our classification works outside of a controlled environment, making our contribution even more generalizable. Third, we would use this interruption identification in the development of interfaces that help the user remain uninterrupted and improve productivity.

REFERENCES

[1] L. Aftanas and S. Golocheikine. Human anterior and frontal midline theta and lower alpha reflect emotionally positive state and internalized attention: high-resolution eeg investigation of meditation. *Neuroscience letters*, 310(1):57–60, 2001.

[2] D. M. Cades, N. E. Werner, D. A. Boehm-Davis, and Z. Arshad. What makes real-world interruptions disruptive? evidence from an office setting. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 54, pages 448–452. Sage Publications Sage CA: Los Angeles, CA, 2010.

[3] C. D. Fisher. Effects of external and internal interruptions on boredom at work: Two studies. *Journal of Organizational Behavior*, pages 503–522, 1998.

[4] T. Fritz, A. Begel, S. C. Müller, S. Yigit-Elliott, and M. Züger. Using psycho-physiological measures to assess task difficulty in software development. In *Proceedings of the 36th International Conference on Software Engineering*, pages 402–413. ACM, 2014.

[5] R. Harr and V. Kaptelinin. Unpacking the social dimension of external interruptions. In *Proceedings of the 2007 international ACM conference on Supporting group work*, pages 399–408. ACM, 2007.

[6] J. Jin and L. A. Dabbish. Self-interruption on the computer: a typology of discretionary task interleaving. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 1799–1808. ACM, 2009.

[7] I. Katidioti, J. P. Borst, M. K. van Vugt, and N. A. Taatgen. Interrupt me: External interruptions are less disruptive than self-interruptions. *Computers in Human Behavior*, 63:906–915, 2016.

[8] A. F. Kramer. Physiological metrics of mental workload: A review of recent progress. *Multiple-task performance*, pages 279–328, 1991.

[9] J. C. Lee and D. S. Tan. Using a low-cost electroencephalograph for task classification in hci research. In *Proceedings of the 19th annual ACM symposium on User interface software and technology*, pages 81–90. ACM, 2006.

[10] A. N. Meyer, T. Fritz, G. C. Murphy, and T. Zimmermann. Software developers' perceptions of productivity. In *Proceedings of the 22nd ACM SIGSOFT International Symposium on Foundations of Software Engineering*, pages 19–29. ACM, 2014.

[11] A. Procházka, M. Mudrová, O. Vyšata, R. Háva, and C. P. S. Araujo. Multi-channel eeg signal segmentation and feature extraction. In *Intelligent Engineering Systems (INES), 2010 14th International Conference on*, pages 317–320. IEEE, 2010.

[12] J. G. Trafton and C. A. Monk. Task interruptions. *Reviews of human factors and ergonomics*, 3(1):111–126, 2007.

[13] L. Wu and J. Gotman. Segmentation and classification of eeg during epileptic seizures. *Electroencephalography and clinical neurophysiology*, 106(4):344–356, 1998.

[14] M. Züger and T. Fritz. 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*, pages 2981–2990. ACM, 2015.