LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN
INSTITUT FÜR INFORMATIK
LEHRSTUHL FÜR MOBILE UND VERTEILTE SYSTEME

Affective Computing

Studimotion: a Learning Application Adapting Task Difficulty based on Integrated Overload and Underload Detection

Creators: Sophia Sigethy

Julia Pühl

Yara Fanger

Document created
November 21, 2023

Studimotion: a Learning Application Adapting Task Difficulty based on Integrated Overload and Underload Detection

Abstract

Digitalization has revolutionised learning as digital learning apps are becoming more and more popular. However, common learning apps are usually unable to react accordingly to their users' mood or emotions. Here, *Affective Computing* offers a new opportunity. By applying *Affective Computing* approaches, intelligent systems become able to detect, infer or interpret human emotions. We intend to take advantage of this, in order to avoid certain negative emotions while using learning apps. Our goal is to create an enjoyable and enhanced learning experience and thereby increase intrinsic motivation, improve memory recall, and prevent the avoidance of performance situations. To this end, we focused on the avoidance of under- and overload. In this work, we present *Studimotion* - a *Web Application* capable of customising the difficulty level of tasks based on emotion recognition. Based on our insights we have gained through our research on related work, we implemented a *Web App* that is able to detect over- and underload. Subsequently, we conducted a study to evaluate whether our application could indeed correctly detect these emotions. In addition, we also investigated how the concept is perceived and evaluated by the participants. The results showed that underchallenge was detected with a probability of 66.66 % and overchallenge with 62.22 %. Furthermore, the majority of participants welcomed the concept of *Studimotion* and would use it in real life. Some further interesting observations also showed how the concept could be improved in the future.

Contents

1 Motivation

Due to digitisation, learning nowadays no longer takes place only in the classroom, but is increasingly available on digital devices such as computers or mobile phones. In this context, learning has become more flexible and individualised as digital devices greatly expand learning opportunities. Thus, learning can take place anytime and anywhere. As a result many learning applications have made their way onto computers or smartphones.
However, one major disadvantage that computers or smartphones usually suffer from is their inability to detect the mood or emotions of their users as a human teacher would be able to. This is where *Affective Computing* comes to bear and provides essential support. *Affective Computing* is an increasingly important research area that aims to enable intelligent systems to recognise, infer, or interpret human emotions [21]. It is an interdisciplinary field covering psychology, cognitive science, physiology as well as computer science. Thus, if the learning task is too challenging, for example, this leads to stress and may impair the retrieval and updating of information in memory [24]. The opposite - underload - also has a negative effect on learning as underload is usually associated with boredom and reduced motivation [11]. *Affective Computing* offers the possibility to recognise these emotions at an early stage in order to avoid them by individual adaptation of the task. Thus, entirely novel benefits could emerge compared to conventional learning by combining the advantages of digital learning with *Affective Computing*. For this reason, we present *Studimotion* - a *Web Application* that is able to individually adjust the difficulty of tasks based on emotion recognition. This idea also meets the demands for more individualised software as for example over 90 % of students worldwide are interested in personalised support regarding digital learning technologies [5]. We want to make learning a pleasant experience without negative emotions. To achieve this, we focus primarily on avoiding over- and underload.
Therefore, we first dealt with the theoretical basis for the recognition of over- and underload. We also familiarised ourselves with existing approaches. Subsequently, a major part of our project consisted of implementing a *Web Application* with under- and overload detection. For this we mainly used a convolutional neural network and facial landmarks. After the implementation was completed, we conducted a study in order to evaluate whether *Studimotion* actually produces the desired results and feedback we were looking for. On the one hand, we investigated whether the detection of over- and underload matches the participants' self-assessment. On the other hand, we examined whether the concept of automatically detecting emotions and adjusting the level of difficulty was welcomed by the participants. We then analysed and evaluated the results of the study. This revealed that underload was detected with a probability of 66.66 % and overload with 62.22 %. In addition, the majority of participants were enthusiastic about the concept of *Studimotion* and stated that they would also use it in reality. Finally, this paper offers a discussion that evaluates further observations and provides a future outlook.

2 Related Work

2.1 Detection of Stress and Boredom

The functionality to detect whether someone is stressed or bored is not only useful in an educational setup. It has been researched and tested out in many other fields, such as traffic or health. In the following chapter, already existing implementations on recognising under- and overload are described, together with a short specification on how each of these states are defined.

2.1.1 Underload

Underload in the context of this study describes the state of being insufficiently challenged. This is the case, when one's intellectual needs are neglected or one's skills are more advanced than required [23]. As a consequence, the affected person often feels bored which causes their performance to decrease [14]. Boredom can be expressed by various parts of the body, like for example the posture, gesture or facial expressions. Defining boredom is rather difficult since it is a complex emotional state[6]. However, there are characteristics of boredom that are more easily specifiable and therefore also more easily recognisable. One of them is fatigue. There has already been done a lot of research and work on fatigue detection as this is particularly useful for making car driving safer.
In the study by Zhuang et al., driver fatigue is detected based on how wide the driver's eyes are opened [25]. They first identify the eyes using *dlib* face key points and then determine the eyelid closure via the *PERCLOS* value. The *PERCLOS* value is a parameter commonly used in fatigue detection and is calculated by dividing the number of frames with closed eyes by the total number of frames times 100 %. If the *PERCLOS* value is larger than 0.2, in other words, if the eyes are closed 80 % of the time, fatigue is detected. Whether an eye is open or closed is determined from the eye aspect ratio, which results from division of the height of the eye by the width of the eye.
Another indication for fatigue is expressed by a person's mouth, in the form of yawning. Based on this feature, Ochocki and Sawicki developed a fatigue detection algorithm [18]. The challenging part of this approach is to distinguish yawning from other actions, like laughing or singing, that show very similar mouth shapes. Ochocki and Sawicki managed to do so by recording the maximum heights of the mouth over time. From this data it became apparent that the maximum heights for yawning are larger than the ones for singing and smiling.
Fatigue is expressed in many other parts of the body as investigated in the study by Kroes [15]. However,

since for the implementation of this project solely facial data provided by the front camera is tracked, these approaches are not further elaborated on.

2.1.2 Overload

Stress overload, being defined as an excessive amount and type of demands requiring an action, is a human response which is experienced as a problem. The characteristics are an intensive perception of stress, a sense of tension or pressure, difficulty functioning as usual, problems with decision making, increased feelings of anger and impatience, and reports of negative effects of stress such as physical symptoms or psychological problems [17].

In 2014, Gao et al. developed a system for recognising stress of car drivers [9]. They defined stress as anger or disgust or a combination of both. Their experimental results resulted in a correct recognition rate of 90.5 % for the tests in indoor environments and 85 % for the tests in the vehicle. The categorisation of emotions is based on the six basic emotions as defined by Ekman et al. [7]. Besides neutral, these are joy, anger, fear, surprise, sadness, and disgust.

Last year Gordon et al. examined perceived stress, emotions, heart rate, and blood pressure during daily life in a three-week app-based study with over 20.000 participants [10]. They found that perceived stress was associated with higher blood pressure and heart rates. Blood pressure increased with negative emotions with high arousal (such as anger) and decreased with positive emotions with low arousal (such as satisfaction). The high blood pressure is associated with perceived stress, which suggests that negative emotions with high arousal can be stress provoking. The fact that the emotions anger, disgust and fear correspond to this category can be seen by applying the *Circumplex Model of Affect* by Posner et al. [20] to the basic emotions of Ekman [7]. This model arranges affective states in a 2-dimensional space of the dimensions valence and arousal. Figure 1 displays such an application as proposed by Tkalcic et al. [22].

Further research papers support the thesis that the emotions of anger, disgust or fear can be signs of stress. Aseltine et al. conducted a study with 939 American adolescents and showed that a person's perceived stress can predict their anger and was more likely to lead to misbehaviour [1]. In 2005, Lee et al. conducted a study with 185 cancer patients in which perceived stress significantly correlated with perceived anger and depression [16]. A literature review by Davey from 2011 references some research work that identifies disgust as a cause of anxiety and stress, especially since this emotion is also involved in feelings of shame and guilt [3]. One of them is Olatunji and Armstrong's study with 82 participants, which allowed the conclusion that high levels of disgust may be involved in the development of clinical contamination anxiety [19] .

However, stress may also serve as a predictor of anxiety symptoms [8]. The emotion fear is often seen to be related to stress. In 2018, a study by Du et al. was published [12]. A hundred students participated and reported their perceived stress, current grief, and negative emotions (including anxiety, depression, and anger) five times a week. This study confirmed that there is a link between anxiety and stress.

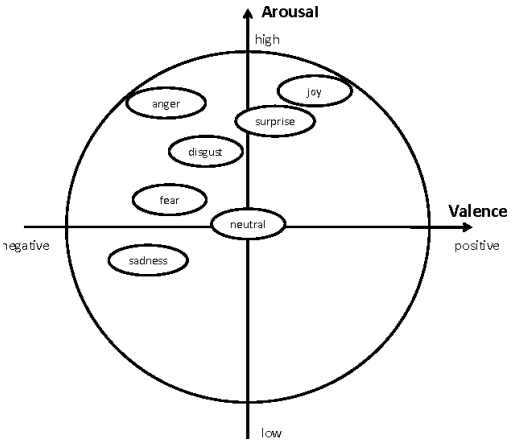

Figure 1: Ekman's seven base emotions mapped to the Valence-Arousal-System [22].

2.2 Learning Applications with Emotion Recognition

Several research projects have already dealt with the topic of integrating under- and overload detection in learning applications. Dingler et al., for instance, developed a language learning application that provides vocabulary tasks as notifications to the user when boredom is detected [4]. Their boredom detection reached an accuracy of 74.5 %. However the detection was not based on emotion recognition, instead the notifications are triggered by an algorithm that uses features such as demographic data, frequency of communication, or intensity of recent phone use.

In 2016, Kadar et al. investigated the use of *Affective Computing* to improve the emotional health of schoolchildren in order to prevent school abortion and to support teachers in recognising and managing the emotional states of schoolchildren during lessons [13]. For this purpose, they considered emotion recognition using a RGB camera that analyses the face, the gait and posture of the students as well as their eye movements. Based on the collected data, they developed a real-time early warning system for the classroom use case, that allowed them to draw conclusions about the engagement of the students.

3 Concept

3.1 Idea

The idea of this project is to make learning more efficient and more individual, which means that the user is no longer forced to struggle with tasks that are too difficult or underchallenging. Since this is often an issue with traditional learning applications, we plan to avoid this problem by using emotion recognition to adapt learning according to the appropriate level of difficulty - to be neither too challenging nor too easy. In this way, we want to make learning a pleasant experience that can be individually customised to the user's needs and skills, thus avoiding negative emotions. The target group is anyone who wants to learn without experiencing unfavourable feelings. We focus on automatic emotion recognition without the user having to initiate the switch to easier or harder tasks himself.
A closer look at the concept shows that overload, causing stress, is detected by identifying stress-realted emotions. These emotions are, as already defined in the chapter related work: anger, disgust and fear. In contrast, underload, which is often connected with boredom and fatigue, is recognised on the basis of eyes and mouth. Here, the degree to which these components are opened plays a decisive role. For example, a user is considered bored if his eyes are open less than a certain threshold. In addition, if the mouth is open more than a certain threshold, it will be considered as yawning.
We finally implemented this idea within a *Web Application*, called *Studimotion*, that contains several tasks with different difficulty levels and can recognise emotions in real time. In order to be able to use this idea for future work and projects, a short study was conducted to determine whether the developed *Web App* really provides the desired results. For this purpose, we first investigated whether the detection of overload and underload works and matches the self-assessment of the participants. In addition, participants were surveyed to obtain feedback, especially on functionality, but also on usability, possible improvements, and preferences.

3.2 Tasks

The tasks shown in *Studimotion* are very similar to those of the *IST 2000 R* intelligence test. They evaluate verbal, numerical, and figural intelligence, as well as memorisation skills. We have divided the tasks into three different levels: easy, normal and difficult. The aim was to induce boredom with the simple tasks and overload with the difficult tasks. The normal tasks were intended to provoke neither of the two.

4 Implementation

For implementation an anaconda environment with the python version 3.7.12 was used. First we will give a background on how overload, then underload and finally the combination of both was implemented in our project.

4.1 Overload

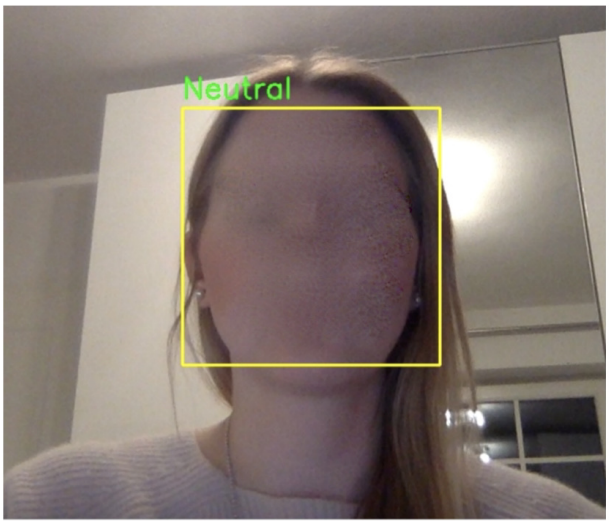

Figure 2: Real time detection of the emotion neutral. The green border around the eyes is for the illustration of the eye opening

To recognise overload, we use a self-trained convolutional neural network that can classify a facial expression into one of Ekman's seven basic emotions [7]. For the model training, the libraries *Tensorflow* (version 2.3.0) and *Keras* (version 2.7.0) are used. The former is an open-source machine learning and artificial intelligence framework from *Google*, while the latter is an open-source neural network library integrated with *Tensorflow*. Finally, *OpenCV*, an open source library for computer vision and machine learning software was used to capture the face on the camera feed and display textual content overlayed on the camera feed. Figure 2 shows an example of the real time detection of emotions in a video feed.
The dataset used to train the model is the *Face expression recognition dataset* from *Kaggle.com* uploaded three years ago by Jonathan Oheix [1]. It is based on *Google*'s dataset *FER 2013* and contains 28.821 images. For Training this includes 3.993 images for the emotion 'anger', 436 for 'disgust', 4.103 for 'fear', 7.164 for 'happy', 4.982 for 'neutral', 4.938 for 'sad' and 3.205 for 'surprise'. The remaining images are used for testing with 960 images for 'anger', 111 for 'disgust', 1.018 for 'fear', 1.825 for 'happy', 1.216 for 'neutral', 1.139 for 'sad' and 797 for 'surprise'. The images are all in gray-scale, sized 48x48 pixel and cropped to the face. It is noticeable that the dataset is somewhat unbalanced

[1] https://www.kaggle.com/jonathanoheix/face-expression-recognition-dataset

(see 'disgust' for example). Nonetheless the trained model seemed to suffice our needs. The first convolutional neural network (CNN) layer of our sequential model uses 64 filters with a kernel size of (3,3). The second layer goes up to 128 filters with a kernel size of (5,5). Layer number three and four both use 512 filters with a kernel size of (3,3). After flattening the model, two fully connected layers are added using *Dense()*. For model compiling the *Adam* optimizer is used with a learning rate of 0.0001. Finally, the model was trained over fifty epochs and resulted in an accuracy of about 0.748 with a loss of about 0.677. The model history regarding accuracy and loss over all epochs can be seen in Figure 3.

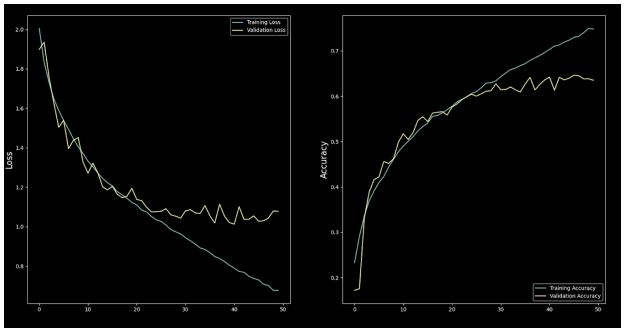

Figure 3: History of our trained model: loss development is shown on the left, while accuracy development is shown on the right

4.2 Underload

As already mentioned, underload can be identified by the symptoms of fatigue. These include squinted eyes and yawning. We use facial landmarks to recognise both.

Opening of the eyes To identify the width of the eye opening, the eye aspect ratio is calculated. This results from the following equation, where the points p1, p2, p3, p4, p5 and p6 are arranged symmetrically along the outline of the eye:

$$EAR = \frac{\|p2-p6\|+\|p3-p5\|}{2\|p1-p4\|}$$

If the EAR falls below a minimum value which was 0.26 in our implementation, we can assume that the eye is closed.

Opening of the mouth To detect a yawn, we calculate the distance between the upper and lower lip of the mouth:

$lip\_distance$ = top_lip_center - bottom_lip_center

As soon as this distance exceeds a certain maximum value over a time period of 5 seconds, we can assume that it is a yawn.

4.3 Combination

Now that some indications for the under- and overload have been explored and their recognition has been technically implemented, we were faced with another challenge. We had to consider how to handle the detection of different components simultaneously. However, combining the individual outputs into one final prediction is not trivial as there are some inconsistencies. For instance, squinted eyes which are intended to be used for detection of underload can also be present in the facial expression when experiencing anger, fear or disgust, all three of which however are indications for overload. On the other hand, an open mouth over longer periods of time is considered a yawn which again is a sign of underload. The problem here is that an open mouth for a longer period of time can also be part of the facial expression of happiness, surprise and sometimes even sadness, all of which do not necessarily indicate underload. The system is supposed to make sense of all these cues and bring them together in a meaningful way in order to predict whether there is an underload, overload or neutral situation. Figure 4 illustrates the structure of the concept we developed in this regard. First we check the current emotion. If that is one of the stress-indicating emotion already mentioned ('fear', 'anger' or 'disgust') and remains for a longer period of time, the system concludes that there is an overload. If the stress-indicating emotion does not last long enough, there is no firm basis for claiming that it is overload, so the system ignores it or rather classifies it as a neutral situation. However, if the current emotion is detected to be 'neutral', we check whether the eyes are squinted. A neutral face expression combined with squinted eyes would result in a facial expression that indicates tiredness, therefore leading the system to believe that there is underload. A neutral expression with eyes open on the other hand shall be ignored as this most likely corresponds to the user being active and in a rather neutral situation. If the emotion that is detected is neutral and the system recognises that the mouth is open for a period of 5 seconds, we can conclude that the user is most likely yawning at the moment. In this case, our system predicts an underload.

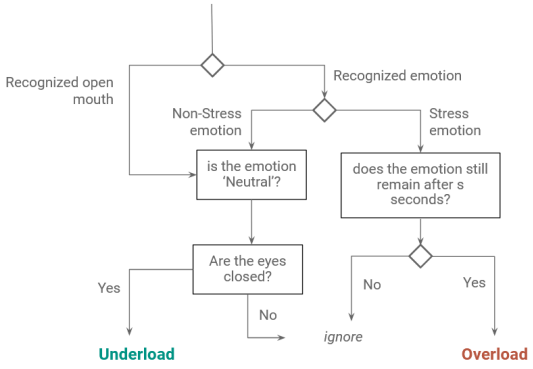

Figure 4: Concept of the integration of the different indications of under- and overload into one system

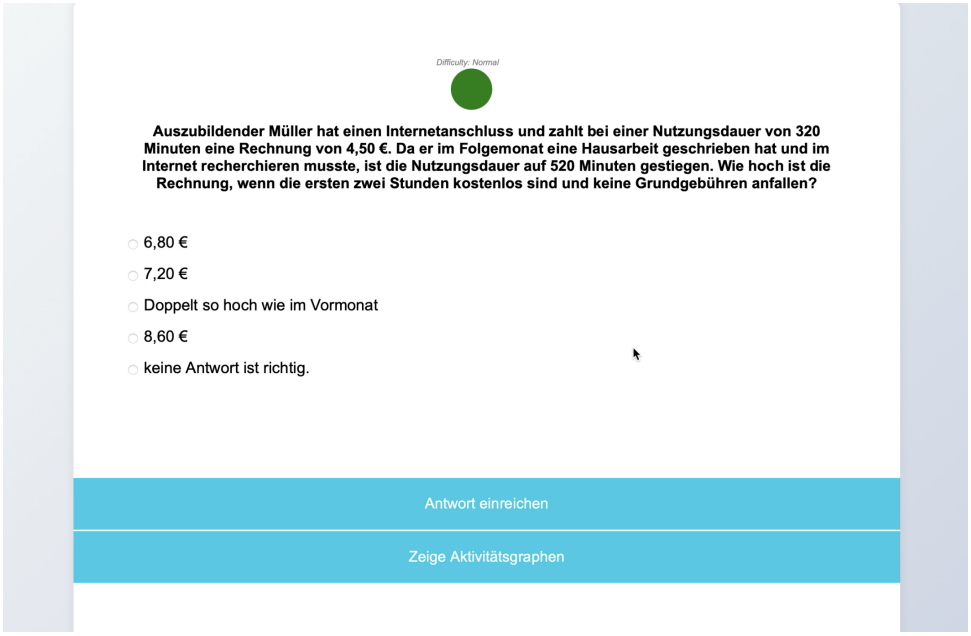

Figure 5: *Studimotion* in the browser. The difficulty level is set to normal.

4.4 Web Application

To embed the emotion recognition part, which was implemented with *Python* as mentioned previously, into a *Web App*, we used *Flask*. *Flask* is a web framework for *Python*. [2] It can be used to simply start a web server available on your computer. This allowed us to connect the *Python* components and the *Web Application*, based on *Javascript* and *HTML*. We have structured the interface by displaying the difficulty level at the top. Beneath this is a question, which can consist of either text or images. Five answer options are displayed below this question, from which the user can select one. Once an answer option is selected, the user can press the *"Submit answer"* button and a new task of the same difficulty level will be displayed. Figure 5 shows the final design of *Studimotion*. A video feed was also embedded in the *Web App* underneath the tasks, where the recognition of emotions can be observed in real time (see Figure 6).
An advantage of *Flask* is that the detected emotions could be passed dynamically to the interface and the website was thus updated accordingly. Every time the emotion anger, disgust or fear was detected, this was passed to the interface. If this occurred several times (e.g. five times), the user was shown a *pop-up* asking *"You seem stressed. Are you overchallenged?"*. The same scheme was followed when fatigue (half-open eyes or yawning) was detected. Here the question was asked *"You seem bored. Are you underchallenged?"*. The user then had the option of agreeing to the question or dismissing the *pop-up window*. With the latter, nothing happened. However, if the user agreed, the task level was directly adjusted. Figures 6 to 9 show the chronological procedure.

[2]https://pythonbasics.org/what-is-flask-python/
[3]https://d3js.org

Furthermore, an activity graph was implemented using *D3.js*, which is a *Javascript* library for manipulating documents based on data. [3] Through this graph, users can check whether they have been underchallenged, overchallenged or active in the past time (see Figure 10).

Figure 6: *Studimotion* detecting underload

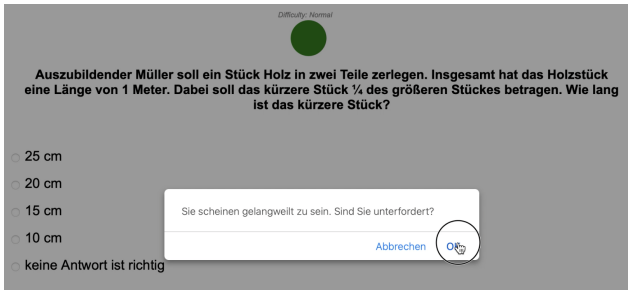

Figure 7: *Studimotion* showing a *pop-up window* asking if the user is underchallenged after underload was detected

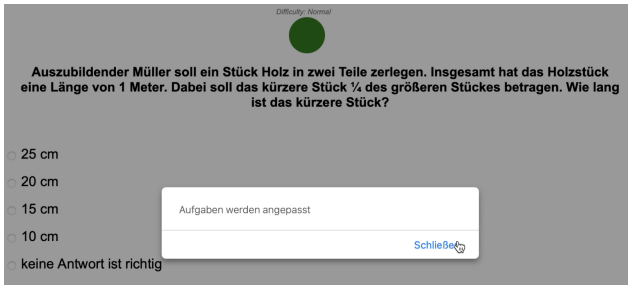

Figure 8: *Studimotion* informing the user that the tasks are going to be adjusted.

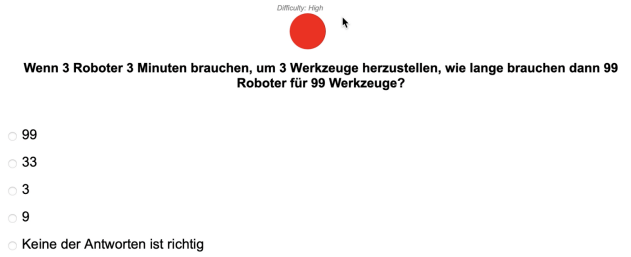

Figure 9: *Studimotion* has just adjusted the task difficulty and now shows tasks of the difficult level. This can be seen by the red circle labeled "Difficulty: High"

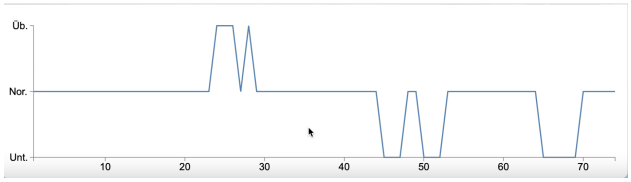

Figure 10: Activity graph below the video feed. It shows when the user was recently, overchallenged (highest value), normal (medium value) or underchallenged (lowest value)

5 Study

In order to ensure that *Studimotion* also correctly detects emotions as well as to evaluate whether the participants would welcome such a concept, a study was conducted. The former was particularly necessary because it was not clear whether the emotions defined as stress and boredom could really be evaluated as over- or underload and thus be recognised correctly.
As shown in chapter 4.3, we made the following hypotheses for our detection system.

H_1 If the emotion anger, disgust or fear is detected and the eye opening is below a certain threshold simultaneously, it is defined as stress.

H_2 If the emotion neutral is detected and the eye opening is below a certain threshold simultaneously, it is defined as boredom.

H_3 If the emotion neutral and an open mouth is detected and the eye opening is below a certain threshold simultaneously, it is defined as boredom.

All other detected emotional states that do not meet the conditions of the hypotheses are considered as *active*.

5.1 Procedure

First of all, the participants' personal data was collected, including gender, age, educational degree and whether they had ever used a learning app in their lives. The data was then documented in writing using a questionnaire.
Subsequently, the *Web App* was briefly explained to the participants, who were then asked to answer a total of nine tasks via the app. The participants had to complete three sets of tasks, each containing three questions. One task set contained three tasks from the easy level, another set consisted of three tasks from the average level and the third set had three tasks from the difficult level. The order of the task sets was varied from participant to participant in order to guarantee randomness and avoid any potential side effects. After each set of tasks, they were then asked to answer two questions. First, to state whether they were underchallenged. Second, whether they were overchallenged. They were asked to rate both questions on a scale from 1 (=Yes) to 5 (=No). It is important to note that while answering the nine questions, the *pop-ups* were disabled in order to avoid influencing the participants. However, emotion recognition continued running in the background and the detected emotional states were logged for each task.
After answering the nine questions, the *pop-up windows* were activated again. Participants were then asked to use the app normally. The starting level was the average level. Participants were also encouraged to be honest when answering the *pops ups*. Afterwards, they were asked to answer the *System Usability Scale* (SUS)

questionnaire [2], which assesses the usability of a system. In addition, they were asked to answer a few more questions in order to gather feedback on the system as a whole. Firstly, they were asked whether they would prefer a learning application with emotion recognition - like the one presented - compared to a learning application without emotion recognition. Secondly, they were asked whether they would use a learning application with individual adaptation of difficulty on the basis of emotion recognition. Lastly, they were encouraged to make suggestions for improvement.

5.2 Participants

The study consisted of fifteen participants. Seven of them were male and eight female. While the youngest participant was 14 years old, the oldest one was 60. The average age was 36.9 years. Among the participants were ten people with an university degree, 4 with a higher education entrance qualification and one with no type of graduation. In addition, 12 participants had already had used a learning app at one point of their life.

5.3 Results

In the following the different results regarding quantitative data, qualitative data and the *system usability score* are presented. The questions of the study questionnaire where tailored to answer the following two research questions:

1. How accurate does our proposed prototype predict under- and overload for users following our hypotheses of chapter 5?

2. How is the usability of the proposed prototype perceived by users and what could be improved?

5.3.1 Quantitative data

First, we investigate the quantitative data that we need to assess whether we have been successful in provoking under- and overload as well as normal situations in the participants. Questions from the IQ-Test catalogue which are considered easy were expected to lead to underload, while questions considered difficult were expected to provoke overload. The 'normal' titled questions are supposed to trigger neither of the two extremes. On the question of whether the participants feel underchallenged with the easy set of questions, we achieved a mean value of 4.6 on a scale from 1 to 5, 1 being "Not at all" and 5 being "Very much so". Normal questions resulted in a mean rating for underload of 3.6, while the difficult set of tasks had a mean of 2.2. On the question of whether the participants feel overchallenged on a scale of 1 ("Not at all") to 5 ("Very much so"), the easy set of questions achieved a mean value of 1.4. The normal-level questions had a mean of 3.8, and the difficult tasks received in their ratings a mean of 4.2.
In order to further evaluate whether the recognition matches the self-assessment of the participants, the logged data was first normalised. This was done for each task set, and the total number of logs for a task set was used as the basis for normalisation. In total, the states stressed, bored or active could be logged, following the conditions of the hypotheses. The data was also considered separately for both questions - whether the participant was overchallenged or underchallenged while answering a set of tasks. After that, the data per task set was considered. The ratio in which the computer recognised active, bored or stressed was calculated. This ratio was then mapped to the scale 1 (=Yes) to 5 (=No). Subsequently, the results were compared with the self-assessment of the participant. To illustrate this, the following example is given:
Participant 1, for example, answered 1 (=Yes) to the question *"Did you feel underchallenged when answering the easy tasks"?* Now the three tasks from participant 1's easy task set are analysed. Therefore we determined how often active, bored or stressed was detected. These numbers are then normalised. Next, the ratio of stressed, bored and active in this easy task set is considered. For example, if bored occurred twice as often as stressed and active, it is scored as 1 (=Yes). For the question *"Was participant 1 underchallenged when answering the easy task set"*, the computer thus outputs 1 = Yes. Hence, the computer's detection matches the participant's self-assessment.
Figure 11 shows the overall results over all task sets and participants for underload. The app's assessment of underload was 66.66 % of the time correct. Here, the presence of the emotion boredom was predicted with almost the same precision as the absence of boredom. The app's assessment of overload was 62.22 % accurate as Figure 12 shows. This time the presence of a stress-indicating emotion was recognised more poorly than the absence of a stress-indicating emotion.

Unterforderung		Predicted (App)	
		Nicht Unterfordert	Unterfordert
Actual (Teilnehmer)	Nicht Unterfordert	20	9
	Unterfordert	6	10

Figure 11: Correlations matrix for the question *"Did you feel underchallenged?"*

Überforderung		Predicted (App)	
		Nicht Überfordert	Überfordert
Actual (Teilnehmer)	Nicht Überfordert	23	7
	Überfordert	10	5

Figure 12: Correlations matrix for the question *"Did you feel overchallenged?"*

5.3.2 Qualitative Data

In order to answer our second research question and gain insight into an user's demands for such a learning application, the participants where asked to state whether they would prefer a learning app with emotion recognition like the one presented to a learning app without. Among 15 participants, 13 reported that they would prefer to use a learning app with emotion recognition. Only one participant said that he would prefer to use it without emotion recognition, and one participant said that he would not care. When asked if participants would use a learning app with individual difficulty adjustment based on emotion recognition, 13 also stated yes and 2 reported no.
Furthermore, the participants were asked to give some suggestions for improvements on our prototype. Four participants [P8, P10, P12, P15] pointed out that there should be a possibility to answer the question before the difficulty level changes. Participant 4, 5 and 7 remarked that the *pop-up* appears to often. Participant 7 further emphasised that this is especially undesired in the beginning of a new task, as there is some time needed to first read the questions. Participant 1 asked to "not repeat the same questions" over and over again in our *pop-ups*. Instead the *pop-up* content could be more varied. Finally, participant 2 and 12 both propose to add the option of choosing the difficulty by themselves.

5.3.3 System Usability Score

To evaluate the usability of our prototype, we used the *System Usability Scale* [2]. This consists of a series of 10 questions about the evaluation of the usability of a system on a scale from 1 to 5 and allows the computation of a *System Usability Score* that ranges from 0 (worst usability) to 100 (best usability). The resulting *System Usability Score* of our study has a value of 89 which is defined as good to excellent usability.

6 Implementation of Improvement Suggestions

Studimotion was designed in such a way that whenever a *pop-up window* appeared and a participant agreed that he or she was over- or underchallenged, the task was immediately adjusted. Regarding suggestions for improvement four participants stated, as already mentioned, that they would like to complete the current task before it changes. Based on this feedback, we modified *Studimotion* to display a second *pop-up window* after agreeing to the first one, asking whether the user would like to switch directly to the harder/easier task (see Figure 14). If one does not agree to this, it is possible to complete the currently displayed task before adjusting the difficulty. Otherwise the procedure is as illustrated from Figure 13 to 14.
Furthermore, two participants stated that they would like to have the additional option of manually adapting the difficulty level. We had already implemented this functionality in order to be able to conduct the first part of the study, in which participants had to answer three sets of questions from different difficulty levels, without interruption. However, this functionality was only used by us developers, but not by the participants. Originally, we wanted to remove this feature again after the study, but with regard to the feedback, we decided to leave this option available (see Figure 15).

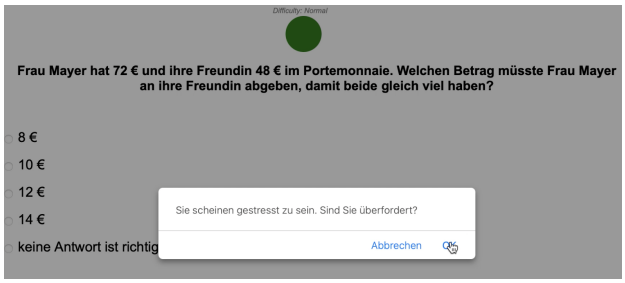

Figure 13: *Studimotion* asking if the user is overchallenged after overload was detected

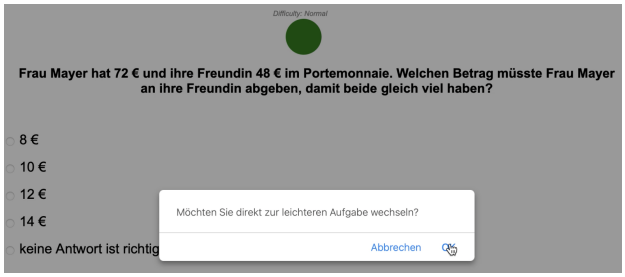

Figure 14: *Studimotion* asking if the user wants to switch directly to the difficult task after the user indicated at the previous *pop-up window* that he was overchallenged

Figure 15: Additional features of *Studimotion* enabling the user to manually change the level of difficulty or to disable the *pop-up windows*

7 Discussion

The results of the mean values related to the question of how bored or stressed the participants feel show that we have achieved our goal of provoking underchallenge in the easy tasks and overchallenge in the difficult tasks. The results for the normal tasks are also in line with our aim, as these mean values are located in the mid-range. As mentioned in chapter 5.3.1 the app's assessment of overload was 62.22% correct. With 66.66% accuracy on underload prediction our system is slightly better at detection of underload than overload. This could be an indication of the better accuracy and precision of the recognition of facial landmarks compared to the emotion recognition of our self-trained model.
There are, however, some further reasons which we believe might have had a significant influence on the prediction of our system. First, we found that the ethnic origin of our participants matters. One Asian participant, for example, received a significantly higher number of bored detections than every other participant. Thus, the predefined threshold that evaluates eye openness did not suit here. We also observed that the seating position plays a role in the detection of eye openness. The reason for this is that the user can look at the screen from below, above or from a certain angle. However, this can affect the correct detection of the eye opening. If the user looks at the screen from the bottom up, for example, boredom will be detected more poorly, since the eyes are wider open when looking up. The exact opposite occurs when the user looks at the screen slightly from above. In this case, it could lead to a smaller eye opening and thus be incorrectly or prematurely interpreted as boredom. The two problems just mentioned could be avoided by prior calibration. For instance, the face could be scanned first when the user is looking neutral at the screen and then parameters such as the threshold value of the eye openness can be adjusted individually.
Moreover, we have found that the use of pencil and paper influences the detection. This is an obvious conclusion, but nevertheless very relevant to avoid incorrect detection of underload. To solve this, the functionality of *Studimotion* should be extended so that, for example, the head tilt is used to determine whether the participant is not looking at the screen.
As mentioned previously we also noticed that overchallenge was recognised somewhat more poorly than underchallenge. A solution for this could be to further improve the accuracy our *CNN*. In addition, the time needed to answer the task could also be taken into account, as this can also be a sign indicating that the participant is not succeeding. However, it should also be noted that people may tend to state that they were underchallenged rather than admitting that they were overchallenged. Since this is more of a psychological question, it is only an assumption and needs to be investigated further. However, this possibility should not be ignored.
Results of the detection of underload could also possibly be better if more than three unchallenging tasks are asked, since some participants might not really be bored after answering only three questions. Therefore, it may be necessary to ask more easy questions over a longer period of time in order to really trigger boredom.
We also made another interesting observation. When we analysed the activity graphs, we noticed that for most participants, underload and overload were not detected immediately one after the other. In other words, if, for example, both underload and overload were detected for a task, this was not detected immediately one after the other, but active was always detected in between. This means that it usually only switched back and forth between active and overchallenged or active and underchallenged, but rarely directly between underchallenged and overchallenged. Figure 16 and 17 show examples of activity graphs from participants 1 and 3.
Lastly, we observed that showing *pop-up windows* too frequent disturbed the participants' concentration. To solve this problem, the parameters that determine how often an emotion must be recognised to trigger a *pop-up window* could be learned gradually in the background. For this purpose, the app could be equipped with an additional neural network that initially uses default values for the parameters, but gradually optimises them individually for the user.
The high system usability rating achieved by our *Studimotion* prototype indicates good to excellent usability. If future applications address the suggestions made by the participants of our study, an even higher score should be achieved. The participants seemed very convinced of the general concept and with 13 out of the 15 participants affirming the question if they would use a learning app with individual difficulty adjustment based on emotion recognition instead of one without, we have a firm ground to have discovered that there is a great demand for such applications. A conduction of this study with a larger number of participants could ascribe greater validity to this conclusion.

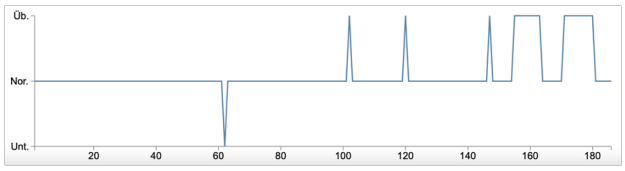

Figure 16: Activity graph of participant 1 during the difficult task set

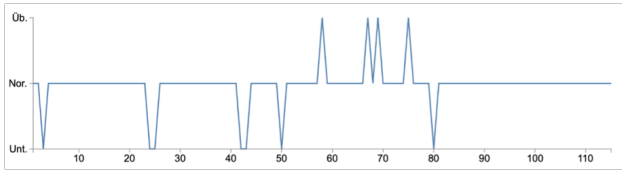

Figure 17: Activity graph of participant 3 during the average task set

8 Conclusion

In order to answer our two research questions as defined in 5.3 we first developed a prototype for a learning application adapting its task difficulty to the user's stress and boredom level. Subsequently, we conducted a user study. We found that the system predicts underload with an accuracy of 66.66% and overload with an accuracy of 62.22%. The prediction of the system could be significantly increased by customising the parameters through prior calibration or even by using machine learning to improve the accuracy of the *CNN*. In addition, taking time needed to answer a question into account, might further improve the predictions. Future work could also be deploying the application on smartphones or testing it in classrooms. The conclusions related to our second research question include the high usability score of our prototype and the aforementioned highlighted suggestions for improvement. Finally, a high demand for a learning application that adapts the task difficulty based on an integrated overload and underload detection was identified.

References

[1] Robert Aseltine, Susan L. Gore, and Jennifer L. Gordon. Life stress, anger and anxiety, and delinquency: an empirical test of general strain theory. *Journal of health and social behavior*, 41 3:256–75, 2000.

[2] John Brooke. Sus: A quick and dirty usability scale. *Usability Eval. Ind.*, 189, 11 1995.

[3] Graham Davey. Disgust: The disease-avoidance emotion and its dysfunctions. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 366:3453–65, 12 2011.

[4] Tilman Dingler, Dominik Weber, Martin Pielot, Jennifer Cooper, Chung-Cheng Chang, and Niels Henze. Language learning on-the-go: Opportune moments and design of mobile microlearning sessions. In *Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services*, MobileHCI '17, New York, NY, USA, 2017. Association for Computing Machinery.

[5] Erin Duffin. E-learning and digital education - statistics facts. *Statista*, 2020.

[6] John Eastwood, Alexandra Frischen, Mark Fenske, and Daniel Smilek. The unengaged mind: Defining boredom in terms of attention. *Perspectives on Psychological Science*, 7:482–495, 09 2012.

[7] Paul Ekman, Robert W Levenson, and W V Friesen. Autonomic nervous system activity distinguishes among emotions. *Science*, 221 4616:1208–10, 1983.

[8] Nancy Fiedler, Robert Laumbach, Kathie Kelly-McNeil, Paul Lioy, Zhi-Hua Fan, Junfeng Zhang, John Ottenweller, Pamela Ohman-Strickland, and Howard Kipen. Health effects of a mixture of indoor air volatile organics, their ozone oxidation products, and stress. *Environmental health perspectives*, 113:1542–8, 11 2005.

[9] Hua Gao, Anıl Yüce, and Jean-Philippe Thiran. Detecting emotional stress from facial expressions for driving safety. *2014 IEEE International Conference on Image Processing (ICIP)*, pages 5961–5965, 2014.

[10] Amie M. Gordon and Wendy Berry Mendes. A large-scale study of stress, emotions, and blood pressure in daily life using a digital platform. *Proceedings of the National Academy of Sciences*, 118(31), 2021.

[11] Arthur C Graesser and Sidney D'Mello. Emotions during the learning of difficult material. In *Psychology of learning and motivation*, volume 57, pages 183–225. Elsevier, 2012.

[12] Du Jiaxuan, Jiali Huang, Yuanyuan An, and Wei Xu. The relationship between stress and negative emotion: The mediating role of rumination. *Clinical Research and Trials*, 4, 01 2018.

[13] Manuella Kadar, Emmanuelle Gutierrez y Restrepo, Fernando Luis-Ferreira, Jorge Calado, Andreia Artifice, João Sarraipa, and Ricardo Jardim-Goncalves. Affective computing to enhance emotional sustainability of students in dropout prevention. pages 85–91, 12 2016.

[14] Maike Krannich, Thomas Goetz, Anastasiya A. Lipnevich, Madeleine Bieg, Ana-Lena Roos, Eva S. Becker, and Vinzenz Morger. Being over- or underchallenged in class: Effects on students' career aspirations via academic self-concept and boredom. *Learning and Individual Differences*, 69:206–218, 2019.

[15] Stefan Kroes. Detecting boredom in meetings.

[16] Pyong Sook Lee, Jung Nam Sohn, Yong Mi Lee, Eun Young Park, and Ji sun Park. [a correlational study among perceived stress, anger expression, and depression in cancer patients]. *Taehan Kanho Hakhoe chi*, 35 1:195–205, 2005.

[17] Margaret Lunney. Stress overload: A new diagnosis. *International Journal of Nursing Terminologies and Classifications*, 17(4):165–175, 2006.

[18] Michal Ochocki and Dariusz Sawicki. Yawning recognition based on dynamic analysis and simple measure. *Proceedings of the International Conference on Computer-Human Interaction Research and Applications (CHIRA 2017)*.

[19] Bunmi O. Olatunji and Thomas Armstrong. Contamination fear and effects of disgust on distress in a public restroom. *Emotion*, 9 4:592–7, 2009.

[20] Jonathan Posner, James A. Russell, and Bradley S. Peterson. The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and Psychopathology*, 17:715 – 734, 2005.

[21] Jianhua Tao and Tieniu Tan. Affective computing: A review. In *International Conference on Affective computing and intelligent interaction*, pages 981–995. Springer, 2005.

[22] Marko Tkalcic, Andrej Koir, and Jurij F. Tasic. Affective recommender systems: The role of emotions in recommender systems. 2011.

[23] Hanna Vock. Permanent frustration because of being underchallenged and facing incomprehension. *IHVO Handbuch*, 2011.

[24] Susanne Vogel and Lars Schwabe. Learning and memory under stress: implications for the classroom. *npj Science of Learning*, 1(1):1–10, 2016.

[25] Qianyang Zhuang, Zhang Kehua, Jiayi Wang, and Qianqian Chen. Driver fatigue detection method based on eye states with pupil and iris segmentation.