ORIGINAL RESEARCH

Stress generation and non-intrusive measurement in virtual environments using eye tracking

Christian Hirt[1] · Marcel Eckard[1] · Andreas Kunz[1]

Abstract

In real life, it is well understood how stress can be induced and how it is measured. While virtual reality (VR) applications can resemble such stress inducers, it is still an open question if and how stress can be measured in a non-intrusive way during VR exposure. Usually, the quality of VR applications is estimated by user acceptance in the form of presence. Presence itself describes the individual's acceptance of a virtual environment as real and is measured by specific questionnaires. Accordingly, it is expected that stress strongly affects this presence and thus also the quality assessment. Consequently, identifying the stress level of a VR user may enable content creators to engage users more immersively by adjusting the virtual environment to the measured stress. In this paper, we thus propose to use a commercially available eye tracking device to detect stress while users are exploring a virtual environment. We describe a user study in which a VR task was implemented to induce stress, while users' pupil diameter and pulse were measured and evaluated against a self-reported stress level. The results show a statistically significant correlation between self-reported stress and users' pupil dilation and pulse, indicating that stress measurements can indeed be conducted during the use of a head-mounted display. If this indication can be successfully proven in a larger scope, it will open up a new era of affective VR applications using individual and dynamic adjustments in the virtual environment.

Keywords Virtual reality · Stress generation · Stress measurement · Virtual stressors · Eye tracking · VR application

1 Introduction

Picard (1995) introduced the concept of "affective computing". Here, software and hardware are used to automatically detect a user's affective state, which in turn should be used to alter a computer's behavior. Accordingly, Picard (1999) further stated that affective computing would generally improve modern human computer interaction (HCI), e.g. due to the reduction of user frustration. According to Greene et al. (2016), nowadays, affective computing has become a large research branch in HCI, which among others is concerned about detection and recognition of human emotional information in the form of affective states. Within these affective states, so-called emotional states like joy, anger or fear can be deduced, which again are further used to compute a qualitative measure for other states such as emotional stress. According to the concept of affective computing, which says that affective states generally influence the performance in cognitive processes, these states can be valid instruments to affect HCI and generally humans' perception of virtual content.

Virtual content or specifically virtual environments (VEs) nowadays are easily accessed and experienced using Head-Mounted Displays (HMDs). In most cases, such virtual reality (VR) applications cause an engulfing feeling for the user, which can be used as a form of quality assessment according to Bowman and McMahan (2007). Slater (2003) already divided this feeling into two subcategories, *immersion* and *presence*. In this case, immersion stands for the technological aspects of an application that can be measured objectively, while presence describes the human reaction to immersion. In order to increase immersion for such VR systems, multiple technological factors need to be considered,

✉ Christian Hirt
hirtc@iwf.mavt.ethz.ch

1 Innovation Center Virtual Reality, ETH Zurich, Leonhardstrasse 21, 8092 Zurich, Switzerland

such as display resolution, field of view, latency, and others. Although these technological factors are simple to specify, they do not allow a clear definition on how they will influence the user. Thus, questionnaires are often used to retrieve users' subjective feedback on a system or an application. Since the answers in these questionnaires are based on the human cognitive process and are thus subject to the individual perception of each person, they may differ strongly among users. However, it was shown in many studies that presence generally correlates positively with immersion (cf. Cummings and Bailenson 2016).

Now, translating the concept of affective computing to state-of-the-art VR, the same positive effects identified previously in affective computing (i.e. improved HCI) should be possible to be achieved for VR. Therefore, we expect that specifically targeting and exploiting affective states and thus generating true "affective VR" will have an immensely positive impact on users' sense of presence. Identifying such a correlation may allow developers to create VEs in the future that adapt to each user differently and thus immerse them based on their individual needs and reactions.

In our case, we chose to address the emotional state stress through a variety of measuring technologies. Besides the aforementioned emotions, Greene et al. (2016) also identify stress as an important aspect in affective computing. Further, stress is already well researched in medicine, e.g. as described by Everly and Lating (1989). It can be seen as a physiological response to external sensations (i.e. stressors) and—in a simplified form—follows the circle shown in Fig. 1. After a personal evaluation and classification of the perceived stressor's relevance, the human body triggers signals in the central nervous system. These signals are transmitted to the sympathetic nervous system, which eventually activates so-called target organ effects like pupil dilation or an increase of the heart rate. If the initial stressor

is then removed, the parasympathetic nervous system resets the body to a calm state. Naturally, if these stressors are not acute, but are active for a longer period of time, long-term consequences may occur.

Since the sympathetic nervous system activates multiple different target organ effects, there is a variety of stress measurement techniques, summarized by Greene et al. (2016). Here, brain and heart activity [i.e. heart rate (HR) and heart rate variability (HRV)], eye activity [i.e. blink rate (BR), blink duration (BD), and pupil diameter (PD)], skin conductivity (SC), and cortisol level in saliva are the most important ones. These methods complement each other, but also have different resolutions, temporal horizons, and error-proneness, making them more or less suitable for an application in VR. Moreover, most of the measurement equipment might not even be applicable in immersive VR settings due to their intrusive nature or excessive preparation time. Looking at modern VR devices, they are equipped with multiple sensors for tracking a user's position, orientation and motion, but also their gaze can be determined using eye trackers. Additionally, in daily life, many people are already accustomed to watches, fitness trackers, phones or similar devices constantly monitoring their movement and body (e.g. their HR). Since eye tracking sensors integrated in HMDs are unobtrusive, even unnoticeable, and HR monitoring seems to be popular, widespread, and easy to access, a considerable step to truly affective VR is to find that a correlation also holds true in a virtual application between subjectively perceived stress, HR and as suggested by Haak et al. (2009) BR or by Pedrotti et al. (2014) PD.

In this paper, we first discuss related work in this research field. Next, we describe the study design including the overall system containing hardware, software, and questionnaires used. For the user study, we developed a VE intended to induce stress to a single user, and measured their eye

Fig. 1 Simplified stress cycle—from perceiving the stressor to target organ effects and back to a calm state
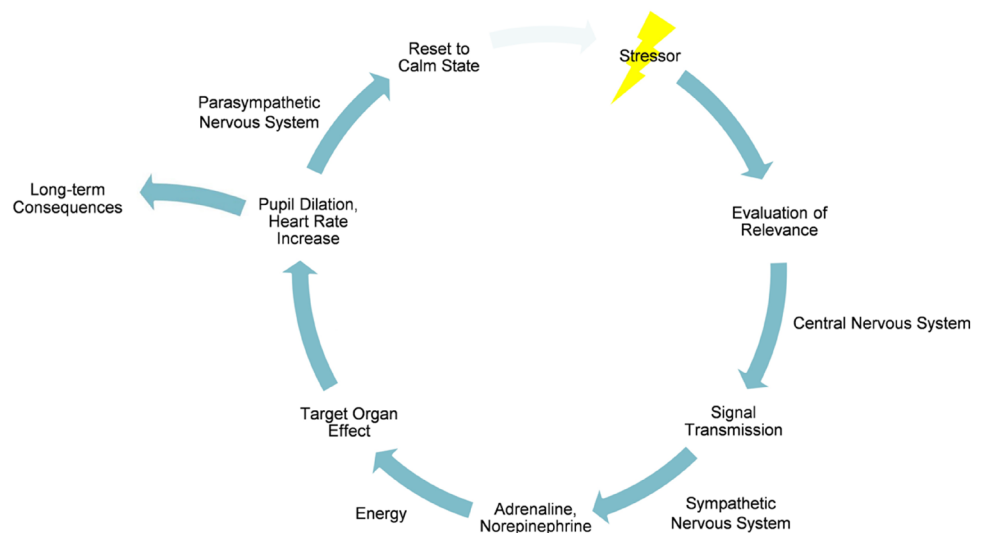
tracking signals from an eye tracker integrated in an HMD. Together with this, we also monitored and recorded heart activity using a pulse chest belt. Finally, we show the measured results and discuss the limitations of the work, before we conclude with a summary and an outlook on future work.

2 Related work

In this chapter, we will focus on various approaches to perform stress measurements, both in real and virtual applications. In accordance with Table 1, first, real applications concerning stress measurements are discussed. Then, we move to studies that were performed with a simple computer screen before presenting the latest user studies conducted in fully immersive VR.

Healey and Picard (2005) describe how they assessed a driver's stress level in a real-world scenario by measuring HR, HRV, muscle strain (electromyography (EMG)) and SC. They showed that HR and SC were best related to the subjectively perceived stress. Kosch et al. (2018) further investigated the PD in a calculation task. They used two initial trial recordings for each participant to train a classifier which was then used in a subsequent calculation task to adjust its complexity to the PD. In their study, participants also filled out Hart and Staveland's (1988) NASA-TLX questionnaire to measure the mental workload. Using this setup, they validated the positive correlation between the task's complexity level and the PD.

While these studies so far were carried out in a real-world environment, the next studies were conducted on a computer screen. The dual-task study by Ryu and Myung (2005) was used to measure HRV and blink intervals (BIs). Compared to a fast-moving tracking task, changing the difficulty of an additional arithmetic task did not significantly influence the BI. Later, Tsai et al. (2007) added such a dual task to a virtual car-driving task. Again, the BR and furthermore the PD were measured, but they found no significant change in the BD. Another study by Haak et al. (2009) was related to a race-car simulator. While driving, they measured participants' BR and brain activity using electroencephalography (EEG), which showed a clear correlation between the BR and the stress level of participants. Contradicting a previous study, Benedetto et al. (2011) showed that the BD decreases with increasing task complexity, while the BR increases. Additionally, the longer test subjects were exposed to the task, the more the BD increased, suggesting that fatigue and BD are strongly correlated. The participants also rated their subjectively perceived mental demand, which significantly correlated with the measured objective values. The GASICA system by Van der Vijgh et al. (2014) was created to use different physiological sensors to engage users in an adaptable game. The game is adapting to a participant's stress level using a simple feedback model. Due to the adjustability of the game, the stress level was kept within a certain range, which could be interesting for specific user studies. In order to measure the impact of stressors on the users, they relied on their previous work (2015), in which they found electrodermal activity (EDA), diastolic and systolic blood pressure (DBP resp. SBP), HRV and HR to be the most suitable measures.

These next studies were conducted in fully immersive VR applications using an HMD. Cho et al. (2017) used an extreme learning machine to classify stress levels that were generated in different VR scenarios. In addition, a dual-task user study was performed (i.e. subtracting numbers), while SC, skin temperature (ST), and HRV were measured. With this setup, they achieved a success rate of 95% of correctly classified stress (levels). In a balancing task by Peterson et al. (2018), subjects were asked to walk on a physical beam in three different viewing conditions. At the same time, they were supposed to listen for an auditory clue, and they should

Table 1 Overview on stress measurements in different setups; *HR* heart rate, *HRV* heart rate variability, *EMG* electromyography, *SC* skin conductivity, *PD* pupil diameter, *BI* blink interval, *BR* blink rate, *BD* blink duration, *EEG* electroencephalogram, *SBP* systolic blood pressure, *DBD* diastolic blood pressure, *EDA* electrodermal activity, *ST* skin temperature

Authors	Year	Measurement setup	Measured variables
Healey J, Picard RW	2005	Real	HR/HRV/EMG/SC
Kosch T, Hassib M, Buschek D, Schmidt A	2018	Real	PD, Task Completion
Ryu K and Myung R	2005	Screen	HRV/BI
Tsai YF, Viirre E, Strychacz C, Chase B, Jung TP	2007	Screen	BR/PD/BD
Haak M, Bos M, Panic S, Rothkrantz LJM	2009	Screen	BR/EEG
Benedetto S, Pedrotti M, Mining L, Baccino T, Re A, Montanari R	2011	Screen	BD/Task Completion/Exposure Time (Fatigue)
Van der Vijgh B, Beun RJ, Van Rood M, Werkhoven P	2014	Screen	HR/HRV/SBP/DBP/EDA
Cho D, Ham J, Oh J, Park J, Kim S, Lee NK, Lee B	2017	VR	SC/ST/HRV
Peterson S, Furuichi E, Ferris D	2018	VR	EEG/EDA/HRV/HR

press a button when hearing it. Participants performed the task first in the real environment and then wearing an HMD, in which two different heights were visualized. During the study, measurements of the participants' stress reaction were recorded using EEG, EDA, HRV, and HR. The results stated that height is an important stressor even in VR, but also that the limited field of view of the HMD impaired the user's balance and motor skills. Table 1 provides an overview on the existing work.

The work described in the above clearly shows that it is beneficial to use combined measures, such as HR and HRV. However, the setups used so far are obtrusive and reduce the overall immersion. It is thus a promising approach to use existing sensors of a VR setup, such as an eye tracker. In some cases, eye trackers are already integrated in the HMD or they can easily be upgraded. This allows a straightforward measurement of BR, BD, and PD. However, it has yet to be proved that BR and BD consistently correlate to stress, as opposed to PD, which was proven to be linked to stress in exemplary applications, but not in VR so far.

3 Study design

3.1 Hardware and tracking space

The setup consists of an HTC Vive with an integrated eye tracker from Pupil Labs (@200 Hz). The HTC Vive natively offers a walkable tracking space of 4×4 m. The VE is generated in Unity 2018, utilizes SteamVR 2.0, and runs on a laptop computer (XMG U507, NVIDIA GeForce GTX 1070, i7-6700 @3.4 GHz, 16 GB RAM). In order to relate the eye tracking data to other biometric data, a pulse chest belt linked to a Garmin Edge 1000 is used to record the HR (@1 Hz). Additionally, Bose QuietComfort 35 II over-ear

noise cancelling headphones complement the VE with sound and block most real world background noise.

3.2 Virtual environment

In this user study, two different VEs were used. The first, the VR Tutorial, aims at giving the participants a brief introduction to VR. The second VE, the VR Main Task, contains a larger environment that covers different stressors and tasks for the participants to fulfil.

3.2.1 VR tutorial

Since we expect many participants to have not much or even no prior experience in VR, we implemented an introducing scene called VR Tutorial. The tutorial has a simplified aesthetic and introduces the user to the required interactions in a simplified context with minimal distraction (see Fig. 2, left). In this introductory VE, participants are mainly instructed using text boxes, but also the study conductor may help if necessary.

Since the full-scale VR Main Task covers an area of about 15×25 m, it is much larger than the walkable space of the HTC Vive. Accordingly, participants need to become acquainted with SteamVR's teleportation functionality. They can change their location using the Vive controller's multifunctional touchpad (see Fig. 2, right). With a simple press on the pad, they point and jump (i.e. teleport) to their preferred next location being selected with the controller. Beside this, the participants' head orientations are completely coupled to their real motions. In the VR Main Task, there are two kinds of active object interactions possible for the participant. First, they can collect certain objects by simple touching them and pulling the trigger of the Vive controller. Further, they can grab certain other objects and utilize them. We mapped the grab functionality on the "grip
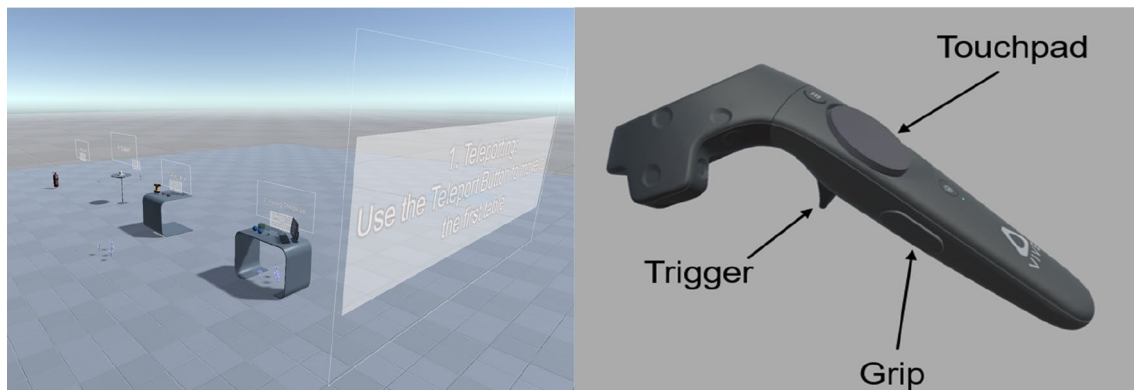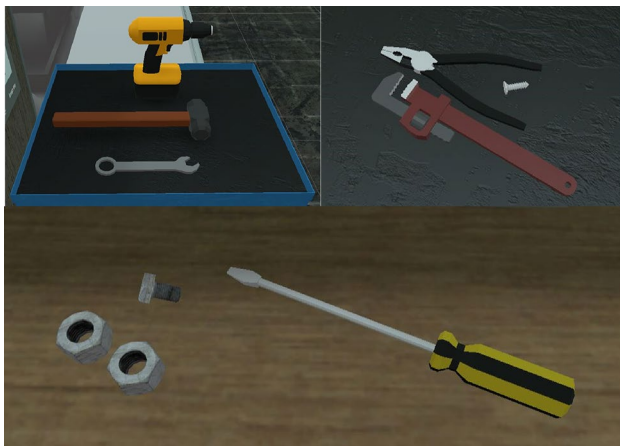

Fig. 2 Left: the VR Tutorial introducing all possible interactions from teleportation, grabbing items, collecting items, pressing a button, to handling a fire extinguisher, right: HTC Vive Controller and its buttons
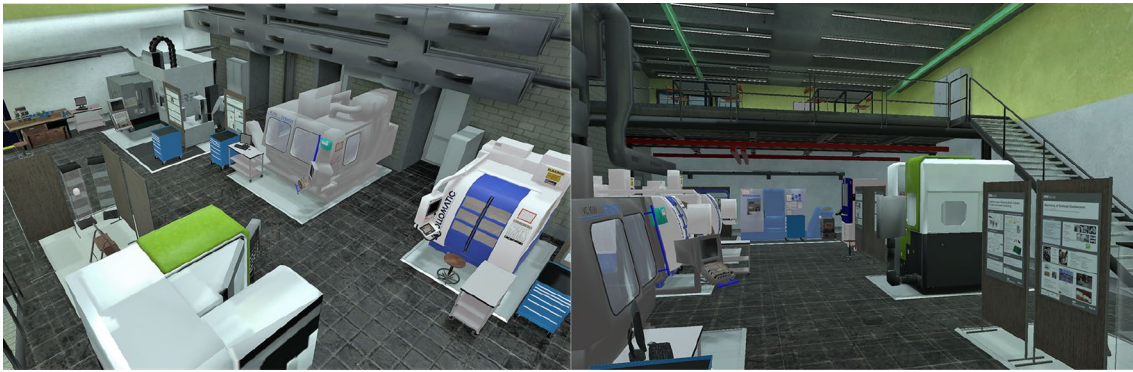
Fig. 3 The virtual shop floor used for the study

Fig. 4 Several tools like screwdrivers, hammers, and more are spread throughout the virtual shop floor
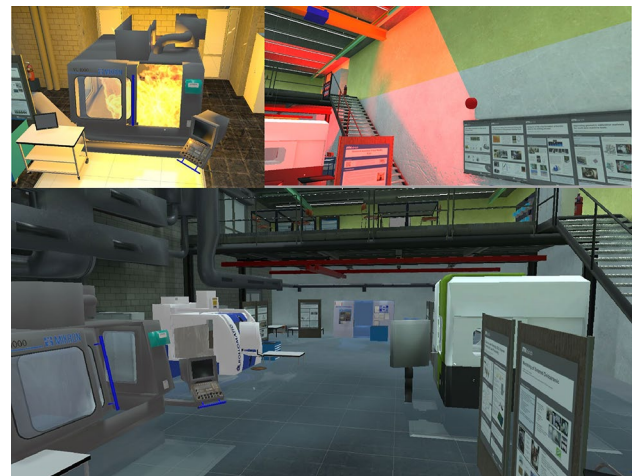

Fig. 5 Three different kinds of stressors integrated in the VE, top left: a machine tool on fire, top right: visual and auditory alarm, bottom: raising water level

button" on the side of the controller, whereas the usage of the grabbed object lies on the trigger. Furthermore, a passive object interaction was chosen to interact with buttons. This most simple interaction only requires the participant to touch the button with the virtual representation of the controller to activate said button.

3.2.2 VR main task

This VE called the VR Main Task consists of a real-scale shop floor with multiple machine tools, several work benches, and a gallery on the first floor with multiple desks (see Fig. 3).

In this shop floor, several other tools such as screwdrivers, hammers, nuts, bolts, etc. are spread (see Fig. 4).

In the virtual shop floor, the study conductor can activate several stressors manually. One of the machine tools can catch fire, which is complemented with an initial explosive sound and afterwards with a crackling fire sound. Further, a visual and acoustic fire alert is activated

in the form of a red strobe light and the sound of a siren. Another stressor is implemented by a rising water level in the room, which from a narrative perspective originates from automatically activated sprinklers combined with a broken water drainage system (see Fig. 5). The water level rises until it reaches the upper floor and can completely submerge participants, which is visualized by applying a blur filter to the image rendered in the HMD. Further, on the upper floor, participants find the control system to drain the water manually (see Fig. 6).

3.3 Data acquisition

In this section, we briefly introduce the different questionnaires used and what physiological data was recorded during the study.

Fig. 6 Left: high water level in the flooded shop floor, right: the control panel for the auxiliary drainage system

3.3.1 Questionnaires

In total, we employ two self-designed questionnaires (i.e. the demographic and the stress questionnaire) and two questionnaires known from literature. In the first self-designed questionnaire, participants mainly state their demographic information and their experience with VR. In the second one, the stress questionnaire, we assess the participant's self-perception of their current stress level and general bodily reaction in stressful situations (e.g. sweating, elevated HR). Further, we ask about special fears like water and fire to avoid panic attacks or similar later during the study. However, to not give away too much of the VR Main Task, we also included some obsolete questions to throw the participants off (e.g. "Are you afraid of spiders?" or "Are you afraid of heights?"). The stress questionnaire is divided into three parts (i.e. Stress 1, Stress 2, and Stress 3). The parts Stress 2 and Stress 3 are filled out after the VR exposures to re-evaluate the participants' current stress level. In Stress 3, another question identifies the stressors, which influenced the participants the most.

The questionnaires used from the literature are Kennedy et al.'s (1993) Simulator Sickness Questionnaire (SSQ) and the Slater-Usoh-Steed Presence Questionnaire (SUS Presence Questionnaire), presented by Slater et al. (1994) and extended by Usoh et al. (2000). We employ the SSQ to ensure that participants do not suffer from simulator sickness, which would considerably bias the physiological stress measurements. The presence questionnaire serves to identify potential weaknesses with the sense of presence during VR exposure in case we would record significant outliers in the physiological measurements.

3.3.2 Physiological measurement

As mentioned before, the physiological measurements consist of an eye tracker recording the PD with 200 Hz and a standard sports chest belt situated below the participants clothing. This chest belt logs the participants' HR with a frequency of 1 Hz.

For the eye tracker, each eye of the participant is illuminated by five infrared LEDs, while one camera per eye, placed in the HMD, takes the recordings. Employing Pupil labs' out of the box software by Kassner et al. (2014), the PD can be extracted from the images in real time, while also blinks and gaze would be possible to analyze. However, to limit the unknown variables in this study, neither blinks nor gaze analysis is used and it the focus lays PD.

3.4 Study overview

In this section, we first discuss the initial pilot study, in which changes in PD are measured that only depend on the change of the environment's illumination. This will serve as a baseline for PD in the two VEs used in the main study. Next, we present the main study, which consists of five parts—parts 1, 3, and 5 are questionnaires, whereas parts 2 and 4 mainly take place in the two VEs (see Fig. 7).

3.4.1 Pilot

The aim of this paper is to show a significant change of the PD correlating to stress-inducing virtual events. However, since the main task and the tutorial are implemented in different VEs, we expect that changing the VE results in a difference of the base light intensity, which also affects the PD. To exclude this potential bias from further test results, we need to quantify this difference in the baseline of the two VEs. In order to find these two baselines, we conduct a pilot study, in which we record the PDs of ten participants in the VEs for roughly 5 min. In both VEs, the sound, animations, and interactions are deactivated such that users are only looking around in the virtual scene. By identifying the difference of the mean PD values of the two VEs, we find the change of the baseline of the PD from the VR Tutorial to the VR Main Task and can consider this ΔPD in the evaluation of the main study.

Fig. 7 Overview of the study
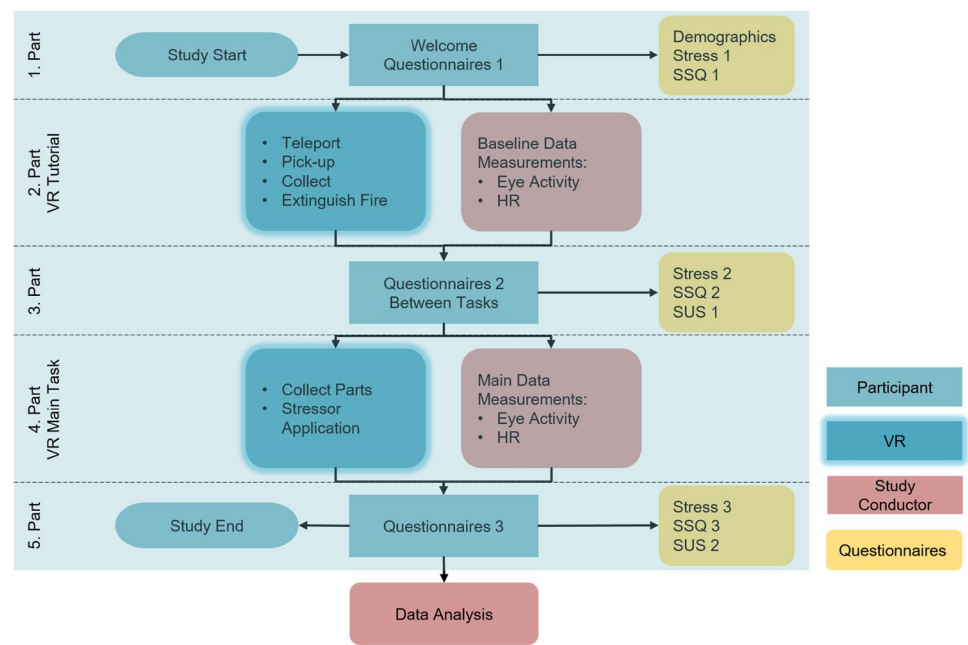
3.4.2 Pre-questionnaire

The demographics questionnaire aims at retrieving information about the participant and their experience with VR applications. The VR experience is further divided into VR experience and non-VR gaming experience. Using these VR and gaming-specific questions, we hope to see a correlation between participants' experience with virtual content and their ability to adapt to our VR study. This demographics questionnaire is further supplemented by the Stress 1 questionnaire, which asks participants for their tolerance for stress and how their personal reaction to stress would be (e.g. heavy breathing, sweating). Additionally, they are asked about their currently perceived stress level on a five point Likert scale. This set of pre-questionnaires is concluded with the SSQ 1, which is used as baseline for the exclusion criteria of simulator sickness.

3.4.3 VR tutorial

Entering the VR Tutorial, we first introduce a calibration sequence, which is necessary for the eye tracker to work more reliably. This calibration sequence is provided by Pupil Labs and can be used out of the box. After the calibration, the VR Tutorial is loaded and the participants can act as they wish. While participants are exposed to the tutorial scene, we collect physiological data from the eye tracker and from the pulse chest belt. These measurements would later serve as the baseline in the comparison since there are no stressors activated in this VE.

3.4.4 Between tasks

After the VR Tutorial, we ask participants to take off the HMD and relax. During this relaxation phase, they answer SSQ 2, SUS 1, and Stress 2 (see Fig. 7). Before moving on to the next task in VR, we explain a model of the virtual shop floor to the participants on a simple computer screen. Additionally, we guide them through this shop floor introducing the relevant points of interest, so participants would not lose their way during the VR Main Task. Further, since the study takes place in a "professional" shop floor, the participants are also instructed about the safety protocol. Accordingly, they are taught about hazards, exits and general emergency behavior in case of an unexpected incident. Consequently, they learn the location of multiple fire extinguishers (see Fig. 8, left), how the sprinklers work, about the issue of a potentially broken drainage system, and how to handle such a situation with a bypassing auxiliary system on the gallery.

3.4.5 VR main task

Entering back into VR, participants experience the VR Main Task. During this task, we record PD and HR in the same way as in the VR Tutorial. In this task, participants experience the following storyline: A machine in the shop floor has broken down and faces a critical problem, which should be addressed. To complete this task, all necessary tools (i.e. screwdrivers, nuts, bolts, etc.) need to be collected which lie somewhere in the shop floor because their colleagues were too lazy to clean up. In addition, an upcoming presentation of the shop floor to external

customers requires the participants to fix the machine within 10 min. In case any unexpected situation would arise, the participants are requested to handle it themselves and present the shop floor in an impeccable state when the customers arrive.

While looking for the tools to complete the storyline, participants hear typical environmental sounds of machine tools coming from drives, compressors or similar. In order to adjust to this new VE, participants are allowed to navigate around and freely explore the shop floor for about 3 min while collecting the tools. Then, the study conductor manually starts the first stressor—an explosion and an animated fire in one of the machine tools. According to the story line, the shop floor needs to be ready for the presentation and thus participants are supposed to fight the fire using any of the fire extinguishers (see Fig. 8, right).

Shortly after the participants managed to extinguish the fire, the second main stressor is triggered. The sprinkler system starts, accompanied by a corresponding acoustic siren and a visual alarm. In accordance with the storyline, the water drainage is broken which subsequently leads to a flooding of the shop floor. This requires the participants to flee to the gallery, where the control of the auxiliary water drainage system is located. Since the water level in the shop floor rises quickly, participants need to rush to the gallery. However, it is possible for the participants to become completely submerged in the water. This is visually displayed by applying a blur filter to the participants' vision. After reaching the control panel, which coincidentally is also defective, participants are required to identify the correctly working button, which drains the water. This acts as an additional stressor, since the participants supposedly act under time pressure due to the quickly rising water level (see Fig. 6). If the participant presses the correct button, a "positive" jingle plays and the water effectively starts to drain. Shortly after, participants can move back downstairs and finish collecting the necessary tools. Upon completion, all physiological measurements are stopped.

3.4.6 After the study: post-questionnaire

After the VR Main Task, participants may take off the HMD and are asked to answer the final set of questionnaires. Again, they start by filling out SSQ 3 and SUS 2, followed by Stress 3. Next, they receive a small compensation and may provide some qualitative feedback on their experience.

4 Results

In this section, we present the results acquired in this user study.

4.1 Questionnaires

In total, 27 participants took part in the user study, 5 female and 22 male. All of them had normal or corrected to normal vision, their mean age was $M = 25.4$ with a standard deviation of $SD = 5.9$ years (i.e. 25.4 ± 5.9 years). In the questionnaires, most questions required the participants to rate their experience on a Likert scale from 1 to 5. For VR experience (i.e. $1 = $"no experience", $5 = $"a lot of experience" respectively), the result was 1.78 ± 1.03. When rating their non-VR gaming experience on the same scale, it resulted in 2.3 ± 1.3. From the initial stress questionnaire, regarding their stress tolerance ($1 = $"high tolerance", $5 = $"low tolerance" respectively) the result is 2.59 ± 0.78.

The participants filled out the SSQ three times, when starting the user study (SSQ 1), and after each of the VR tasks (SSQ 2 after the tutorial, SSQ 3 after the VR Main Task). The SUS Presence questionnaire was filled out only after each VR task (SUS 1 and 2 respectively). According to literature, SUS Presence was then analyzed in two ways: Counting only the answers that are rated "6" or "7" (i.e. *SUS Count*), and by calculating the mean value of all given scores (i.e. *SUS Mean*). For the VR Tutorial, the SUS Count reached a value of 2.04 ± 1.45, and for the VR Main Task it

Table 2 Results of the simulator sickness questionnaires (SSQ) and the SUS Presence questionnaires

	SSQ	SUS count	SUS mean
Pre	8.87 ± 8.71	–	–
Mid	6.37 ± 6.50	2.04 ± 1.45	4.56 ± 0.79
Post	8.03 ± 8.31	2.81 ± 2.00	5.06 ± 1.03

Values are given in $M \pm SD$. While the SSQ was filled out three times (Pre = SSQ 1, Mid = SSQ 2, and Post = SSQ 3), the SUS was filled out only twice (Mid = SUS 1 and Post = SUS 2)

Table 3 Self-reported stress levels on a scale from 1 to 10, Stress 1 (i.e. Pre), Stress 2 (i.e. Mid), and Stress 3 (i.e. Post)

	Stress score
Pre	3.48 ± 1.71
Mid	2.81 ± 1.63
Post	4.93 ± 1.72

Values are given in $M \pm SD$

reached 2.81 ± 2.00. An overview of these results is provided in Table 2.

One of the most important questions was the self-evaluation of the perceived stress in the stress questionnaires. The stress was rated between 1 and 10 (1 = "not stressed", 10 = "very stressed"). The results are given in Table 3.

Most of the participants stated that the VR Tutorial did not induce any stress, while the VR Main Task was perceived to be more stressful. This more stressful VE mainly triggered symptoms like "increased HR" or "sweating". Regarding the stressors in the VR Main Task, we identified the rising water level and the visual and audio alert to be the most stress-inducing events, followed by the fire and the extinguishing task, and finding the right button at the control panel for water drainage.

4.2 Physiological data

In the pilot study, we recorded the PD for both VEs while not engaging the user in any interaction besides looking around in the environment. This allows for an identification of the change in the base PD when changing the VEs. To achieve a comparable value ΔPD, we first average the PDs for each eye for a single participant over the five minutes of exposure in the VR Tutorial ($PD_{L,Tut}$ and $PD_{R,Tut}$ respectively). Then, we take the mean of these two values and thus find the average PD for the VR Tutorial, PD_{Tut}. Repeating the same procedure for the VR Main Task, we identify PD_{Main}. Now, ΔPD for a single participant is given by:

$$\Delta PD_{ID} = PD_{Main} - PD_{Tut}$$

We average these values ΔPD_{ID} over all participants and thus identify the overall $\Delta PD = 3.93px$. Since this ΔPD is

a basic bias introduced by the change of illumination of the different VEs, we consider ΔPD for all evaluations without further mention by subtracting it from the PD measurements in the VR Main Task.

In the main study, HR and PD were recorded during both, the VR Tutorial and the VR Main Task. The HR was recorded in one-second intervals, and the PD was measured for each eye with a sampling frequency of 200 Hz. From the measured eye tracking data, outliers were eliminated using the interquartile range (IQR), introduced by Upton and Cook (1996), which is defined as the difference between lower and upper quartiles (Q_1 resp. Q_3):

$$IQR = Q_3 - Q_1$$

Thus, data points P_i lying outside the following range were eliminated:

$$P_i = \begin{cases} \emptyset, & P_i < Q_1 - 1.5 \cdot IQR \\ \emptyset, & P_i > Q_3 + 1.5 \cdot IQR \\ P_i, & otherwise. \end{cases}$$

The noisy data was then adjusted with a piecewise cubic smoothing spline to improve the visibility of possible stress reactions. For the HR measurements, no data post-processing was required. The measured data was then synchronized with the time of occurrence of stressors in the VE. Vertical black lines indicate when the fire and the water level stressors were introduced. This resulted in the following graphs that are typical for each participant (see Fig. 9).

The data for the HR is displayed in the upper part of the graph, while the eye tracking data is shown in the lower part. The left part of the graph shows data from the VR Tutorial, while the right part shows data from the VR Main Task. In addition, the horizontal lines indicate average values for HR and PD for the two different tasks.

The graph shows that the participant's PD increased immediately when the fire started. The diameter was already decreasing again, when the sprinkler system started. Here, the PD immediately increases again and stays at a higher level. In principle, the HR shows the same characteristics, but is delayed by a couple of seconds. This can be explained by the autoregulation of the pulse that reacts with a certain temporal delay.

To further analyze the data, the mean values of the HR and PD are plotted as a bar diagram (see Fig. 10).

With the help of these mean values, changes were computed in percentage showing the increase or decrease of the physiological measures comparing again VR Tutorial and VR Main Task. For this, the mean of the right (M_r) and left (M_l) eye of the VR Tutorial were averaged (($M_r + M_l$)/2) and the same was done with the values collected during the VR Main Task. Afterwards, a change in percentage was

Fig. 9 Typical data set with smoothened PD and HR for a single participant
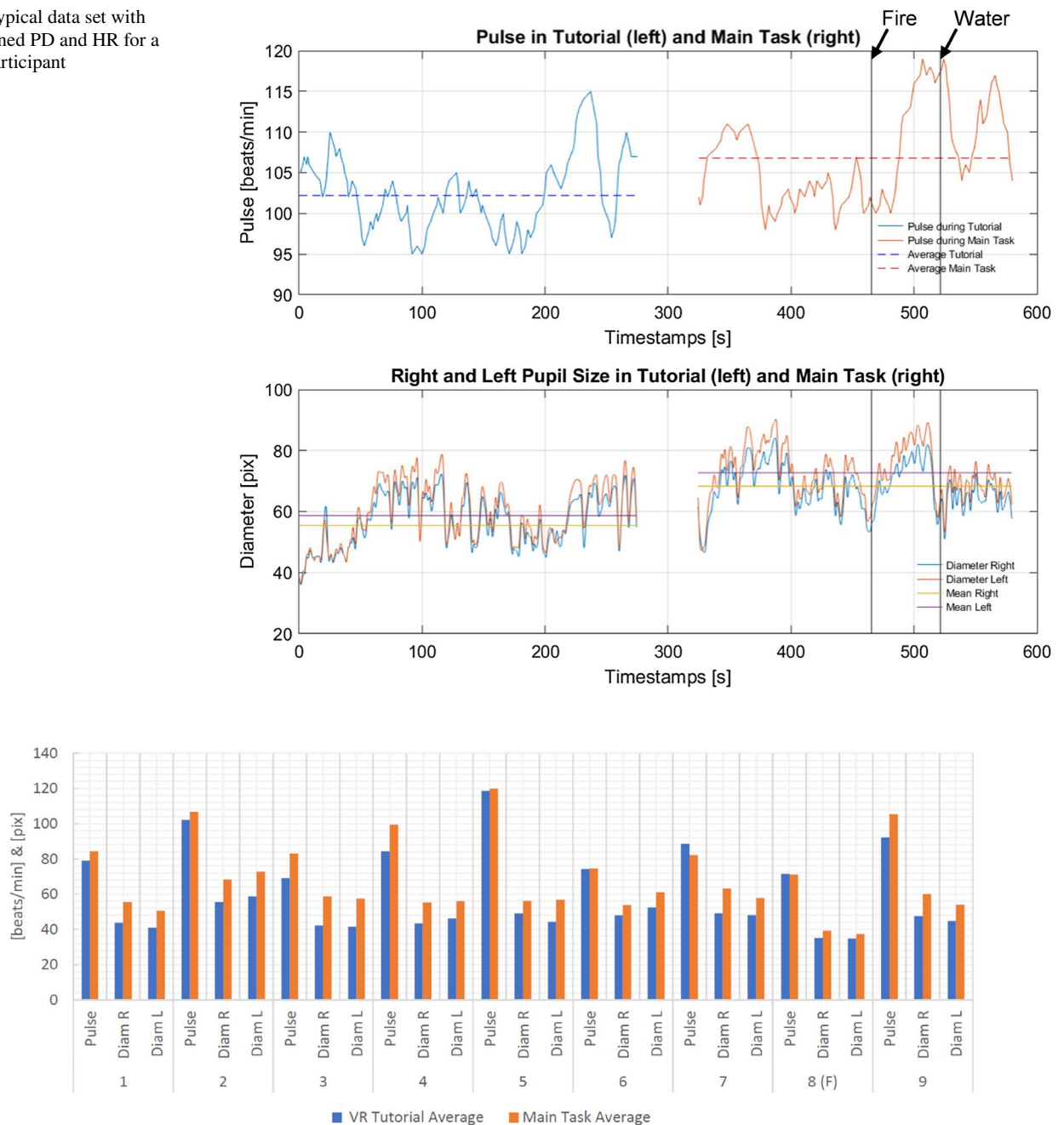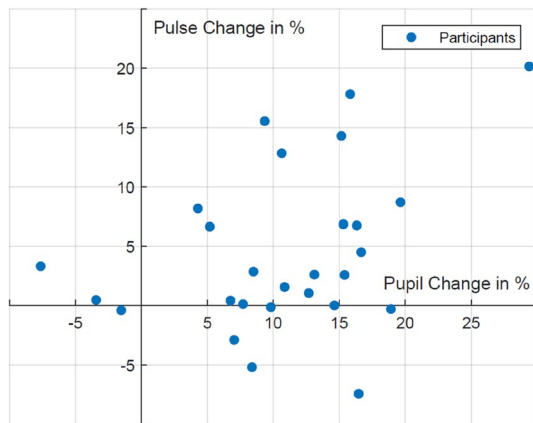

Fig. 10 Mean values of the measured data for the VR Tutorial (blue) and for the VR Main Task (orange) (exemplary for the first nine participants, F = female)

calculated and one notices that most values increased during the main task. To compare the change of HR and PD between the different VR tasks, the differences are shown in a scatter plot (see Fig. 11). In this plot, the horizontal axis represents the relative change of the PD in percentage, while the vertical axis represents the relative change of the HR, also in percentage. It is clearly visible that while average PD, except for three participants, always increased, this was not the case for the average HR recordings. If the HR changes

were negative, they never exceeded − 7.45%. The largest documented positive HR change was 20.16%. The largest recorded increase in PD was 29.40% and the smallest was − 7.64%. Since there are only three negative values for the PD, there is a positive correlation between the occurrence of stressors and the PD in most cases.

In order to determine whether the changes of the mean values of the two measured signals are significant, we performed paired-samples t-tests. Beside the two measured
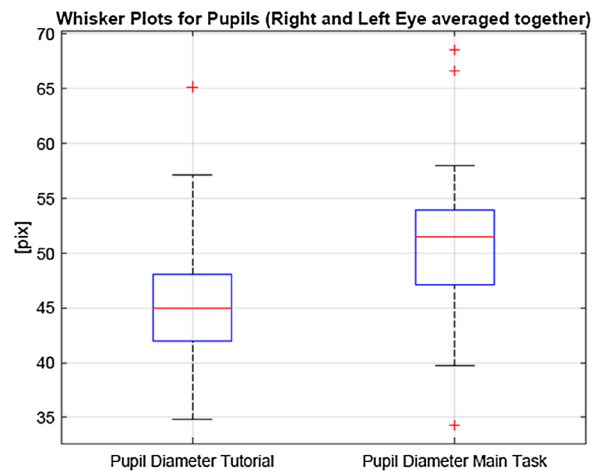
variables (HR and PD), we also included the self-reported stress levels in the t-tests. All t-tests used a significance level of $\alpha = 0.05$, which is a commonly accepted value in literature. In order to apply the t-test, a normal distribution of the pairwise difference of the data was confirmed beforehand using the Anderson–Darling-test, for which we also utilized the same significance level $\alpha = 0.05$.

The t-test of the HR values shows a slight but still statistically significant increase with $p = 0.0027$. The PD showed a higher statistically significant increase with $p = 3.76 \times 10^{-8}$. In addition, the self-reported stress levels show a significant increase: the comparison of the mid- to post-test values have a $p = 3.66 \times 10^{-7}$ (see Fig. 12).

Fig. 11 Scatter plot of changes in HR and PD from the VR Tutorial against the VR Main Task

(a) $p = 0.0027$

(b) $p = 3.76 \times 10^{-8}$

(c) $p = 3.66 \times 10^{-7}$

Fig. 12 Whisker plots showing the statistically significant increase of all the measured variables (a: HR, b: PD, c: subjectively perceived stress). The means are given by a red horizontal bar for all graphs

5 Discussion and limitations

The t-tests show a statistically significant increase of the average PD and the self-reported stress level when switching from the VR Tutorial to the VR Main Task. This shows that a correlation between PD and stress in VR exists and is measurable via eye tracking. This finding is additionally supported by a statistically significant increase of the HR, which is in accordance with Healey and Picard (2005), Taelman et al. (2009), Maarsingh et al. (2017), and Peterson et al. (2018) an already established measure for stress in either real-life applications, or in HCI, as well as in VR.

The eye tracker only recorded PD values if they were above a certain confidence value. Periods in which this confidence value was not met are thus not accounted for the data analysis. Usually, this only occurred for one eye at the time. As the PDs of both eyes tend to behave similarly, this phenomenon therefore is not crucially affecting the results. Additionally, the eye tracker infrequently measured larger jumps in PD, which do not seem realistic. It appears that the measurements of the eye tracker were affected by long eyelashes, which may have obstructed the direct line of sight or even by eyeliner, which may have caused single eyelashes to be misinterpreted for pupils due to their high contrast and curved pattern. Both these incidents occurred only rarely and affected the results insignificantly. Further, pupil data was measured in pixels with a 2D eye model, which does not account for perspective errors of the eye tracker. Future quality improvement could be achieved using a newer 3D-eye model correcting said errors. Furthermore, since the utilized chest pulse belt was sampling with a rate of 1 Hz, there may have occurred smaller time errors during the course of recording.

Additionally, we did not consider the BR and BD for the evaluation. We expected that since the HMD creates a rather closed environment for the eyes with little to no air circulation, participants would blink less frequently because their eyes would dry out less. Additionally, after some time, the enclosed space creates a humid and warm environment, which makes blinking even less essential to keep the eyes hydrated. In accordance with this initial expectation, many participants really changed their blinking behavior significantly while using the HMD. However, we only observed this behavior qualitatively and did not investigate it any further.

6 Conclusion and future work

In this paper, we showed the correlation between eye activity and stress generated in VR. In a study with 27 participants, we collected data on the PD using an eye tracker being integrated in the HMD of the HTC Vive. We also measured the HR, since this was already proved to relate to stress. The data was acquired in two separate VR tasks—the VR Tutorial and the VR Main Task. In order to foster the participants' relationship to the VE and to create a relaxed virtual atmosphere, participants learned how to move and interact with the VE in the VR Tutorial. While collecting some simple tools in the VR Main Task, participants were exposed to several stressors such as fire, an acoustic fire alarm, a rising water level, and malfunctioning buttons to drain the water. The reported stress level, the measured HR, and the PD showed a statistically significant increase between the two VR tasks. Accordingly, we were able to demonstrate that there exists a positive correlation between the generated stress in the VE and the objective reaction of the participants' bodies. On top of that, we further showed that this correlation could be translated to the participants' eye activities, which in terms can be recorded non-intrusively while exploring VEs.

Further work on stress in VR will focus on the integration of other physiological measurements like the number of saccades or the change of fixation points to improve the detection. By introducing these additional physiological measurements, we hope to strengthen and objectively improve the measured stress reaction to quantitatively relate different stressors to their corresponding bodily reaction. It is further envisioned to correlate this data using a deep learning approach, trained using EEG data, which is so far established to be the most reliable source for stress identification.

Beside stress, a different step to achieve true affective VR may be to address other emotional states to gain a more complete understanding of the user's current well-being. For example, would it be possible to identify joy or liking by analyzing user's eye behavior? If this will be the case, completely new fields in marketing and customer service could be created and designed. Furthermore, increasing the overall immersion of the VE will be considered and the effect on a user's emotional state will be tested. Specifically, allowing real walking in the VE and introducing multiple users at the same time will have a large impact.

Thematically, various approaches known to VR research will be tested and evaluated in correlation to affective states. For example, considering real walking, it will be interesting to investigate if and how so-called redirected walking and stress perception in VR are related. A core concept of learning and teaching in VR will also be further evaluated, i.e. how the perception of stress can either improve or deteriorate learning behavior of participants. Specifically, multiple users' teamwork or teacher-student relationship will be investigated.

References

Benedetto S, Pedrotti M, Mining L, Baccino T, Re A, Montanari R (2011) Driver workload and eye blink duration. Transp Res Part F Traffic Psychol Behav 14(3):199–208. https://doi.org/10.1016/j.trf.2010.12.001

Bowman DA, McMahan RP (2007) Virtual reality: how much immersion is enough? Computer 40(7):36–43

Cho D, Ham J, Oh J, Park J, Kim S, Lee NK, Lee B (2017) Detection of stress levels from biosignals measured in virtual reality environments using a kernel-based extreme learning machine. Sensors 17(10):2435. https://doi.org/10.3390/s17102435

Cummings JJ, Bailenson JN (2016) How immersive is enough? A meta-analysis of the effect of immersive technology on user presence. Media Psychol 19(2):272–309

Everly GS, Lating JM (1989) The anatomy and physiology of the human stress response. Clin Guide Treat Hum Stress Response. https://doi.org/10.1007/978-1-4939-9098-6

Greene S, Thapliyal H, Caban-Holt A (2016) A survey of affective computing for stress detection: evaluating technologies in stress detection for better health. IEEE Consum Electron Mag 5(4):44–56. https://doi.org/10.1109/MCE.2016.2590178

Haak M, Bos M, Panic S, Rothkrantz LJM (2009) Detecting stress using eye blinks and brain activity from EEG signals. In: Proceedings of the 1st driver car interaction (DCII), pp 35–60

Hart S, Staveland L (1988) Development of NASA-TLX (Task Load Index): results of empirical and theoretical research. Hancock, P.; Meshkati, N. (eds.): Human Mental Workload 52: 139–183. https://doi.org/10.1016/S0166-4115(08)62386-9

Healey J, Picard RW (2005) Detecting stress during real-world driving tasks using physiological sensors. IEEE Trans Intell Transp Syst 6(2):156–166. https://doi.org/10.1109/TITS.2005.848368

Kassner M, Patera W, Bulling A (2014) Pupil: an open source platform for pervasive eye tracking and mobile gaze-based interaction. In: Proceedings of the 2014 ACM international joint conference on pervasive and ubiquitous computing: adjunct publication. ACM, pp 1151–1160

Kennedy RS, Lane NE, Berbaum KS, Lilienthal MG (1993) Simulator sickness questionnaire: an enhanced method for quantifying simulator sickness. Int J Aviat Psychol 3(3):203–220. https://doi.org/10.1207/s15327108ijap0303_3

Kosch T, Hassib M, Buschek D, Schmidt A (2018) Look into my eyes: using pupil dilation to estimate mental workload for task complexity adaptation. Extended Abstracts of the 2018 CHI conference on human factors in computing systems: LBW617. https://doi.org/10.1145/3170427.3188643

Maarsingh BM, Martin Abello K, Bos J (2017) Stressjam: a mindset game changer? A pilot study. https://irp-cdn.multiscreensite.com/f25f951f/files/uploaded/Whitepaper-Stressjam-a-game-changer-v2-2017.pdf. Accessed 5 Aug 2019

Pedrotti M, Mirzaei MA, Tedesco A, Chardonnet JR, Mérienne F, Benedetto S, Baccino T (2014) Automatic stress classification with pupil diameter analysis. Int J Hum-Comput Interact 30(3):220–236. https://doi.org/10.1080/10447318.2013.848320

Peterson S, Furuichi E, Ferris D (2018) Effects of virtual reality high heights exposure during beam-walking in physiological stress and cognitive loading. PLoS ONE 13(7):e0200306. https://doi.org/10.1371/journal.pone.0200306

Picard RW (1995) Affective computing-MIT media laboratory perceptual computing section Technical Report No. 321. Cambridge, MA, p 2139

Picard RW (1999) Affective computing for HCI. In: HCI (1), pp 829–833

Ryu K, Myung R (2005) Evaluation of mental workload with a combined measure based on physiological indices during a dual task of tracking and mental arithmetic. Int J Ind Ergon 35(11):991–1009. https://doi.org/10.1016/j.ergon.2005.04.005

Slater M (2003) A note on presence terminology. Presence Connect 3(3):1–5

Slater M, Usoh M, Steed A (1994) Depth of presence in virtual environments. Presence Teleoper Virtual Environ 3(2):130–144. https://doi.org/10.1162/pres.1994.3.2.130

Taelman J, Vandeput S, Spaepen A, Van Huffel S (2009) Influence of mental stress on heart rate variability. In: Proceedings of the 4th European conference of the international federation for medical and biological engineering, pp 1366–1369. https://doi.org/10.1007/978-3-540-89208-3_324

Tsai YF, Viirre E, Strychacz C, Chase B, Jung TP (2007) Task performance and eye activity: predicting behavior relating to cognitive workload. Aviat Space Environ Med 78:B176–B185

Upton G, Cook I (1996) Understanding statistics. Oxford University Press, Oxford

Usoh M, Catena E, Arman S, Slater M (2000) Using presence questionnaires in reality. Presence Teleoper Virtual Environ 9(5):497–503. https://doi.org/10.1162/105474600566989

Van der Vijgh B, Beun RJ, Van Rood M, Werkhoven P (2014) GASICA: generic automatic stress induction and control application design of an application for controlling the stress state. Front Neurosci 8:400. https://doi.org/10.3389/fnins.2014.00400

Van der Vijgh B, Beun RJ, Van Rood M, Werkhoven P (2015) Meta-analysis of digital game and study characteristics eliciting physiological stress responses. Psychophysiology 52(8):1080–1098