Journal of Imaging Science and Technology® 63(6): 060413-1–060413-11, 2019.
© Society for Imaging Science and Technology 2019

User Experience Evaluation in Virtual Reality based on Subjective Feelings and Physiological Signals

Yunfang Niu, Danli Wang, and Ziwei Wang
State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing, China
E-mail: danli.wang@ia.ac.cn

Fan Sun
School of Computer and Information Technology, Liaoning Normal University, Dalian, China

Kang Yue and Nan Zheng
State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing, China

Abstract. *At present, the research on emotion in the virtual environment is limited to the subjective materials, and there are very few studies based on objective physiological signals. In this article, the authors conducted a user experiment to study the user emotion experience of virtual reality (VR) by comparing subjective feelings and physiological data in VR and two-dimensional display (2D) environments. First, they analyzed the data of self-report questionnaires, including Self-assessment Manikin (SAM), Positive And Negative Affect Schedule (PANAS) and Simulator Sickness Questionnaire (SSQ). The result indicated that VR causes a higher level of arousal than 2D, and easily evokes positive emotions. Both 2D and VR environments are prone to eye fatigue, but VR is more likely to cause symptoms of dizziness and vertigo. Second, they compared the differences of electrocardiogram (ECG), skin temperature (SKT) and electrodermal activity (EDA) signals in two circumstances. Through mathematical analysis, all three signals had significant differences. Participants in the VR environment had a higher degree of excitement, and the mood fluctuations are more frequent and more intense. In addition, the authors used different machine learning models for emotion detection, and compared the accuracies on VR and 2D datasets. The accuracies of all algorithms in the VR environment are higher than that of 2D, which corroborated that the volunteers in the VR environment have more obvious skin electrical signals, and had a stronger sense of immersion. This article effectively compensated for the inadequacies of existing work. The authors first used objective physiological signals for experience evaluation and used different types of subjective materials to make contrast. They hope their study can provide helpful guidance for the engineering reality of virtual reality.* © *2019 Society for Imaging Science and Technology.*
[DOI: 10.2352/J.ImagingSci.Technol.2019.63.6.060413]

1. INTRODUCTION

Virtual reality (VR) has a bright future for application. It eliminates the limitations of geographic and environmental factors through a computer-generated virtual world, giving users an immersive experience. However, VR technology also has some drawbacks. Users often suffer dizziness and nausea after being in the VR environment for a long time. Therefore, it is important to evaluate different feelings that VR and 2D environments bring to people, whether in research, teaching, training or entertainment. What are the differences between the virtual reality environment and the real environment, and how do people have different emotional and physiological changes? If we had solved these problems well, we would reduce the negative impact of the virtual environment on people's psychology and physiology, so as to make better use of the characteristics and value of virtual reality technology.

There have been some researches on VR and its effects on people. Some scholars have proposed different models for experience evaluation, and provides the foundation for VR study. Increasing works focused on the positive effects of VR, and some researches compared the experience of people before and after using VR products. However, existing works are still not sufficient. Few works about evaluation of VR have been made in general, especially in the field of user experience evaluation. What is more, the technique of VR was not very mature yet, so that VR products may bring discomfort to users. Therefore, it is very important to make a detailed evaluation about user experience of VR products, especially the emotional experience.

In this article, subjective and objective data are collected by a designed experiment, and then analyzed by statistical and categorical methods to study the differences in participants' emotions caused by video in VR and 2D environments. The experiment was based on common models of Affective computing and uses objective physiological signals to analyze emotions. The emotion-inducing dataset was produced by referring to the International Affective Picture System (IAPS) [1], emotion analysis database DEAP [2] and other public video datasets. Based on common models of Affective computing, the experiment used objective physiological signals to analyze emotions. Furthermore, the experiment also focuses on participants' subjective emotional experience and objective physiological responses. We used Positive And Negative Affect Schedule (PANAS) [3], the Simulator

Received July 14, 2019; accepted for publication Nov. 5, 2019; published online Dec. 20, 2019. Associate Editor: Rita Hofmann-Sievert.
1062-3701/2019/63(6)/060413/11/$25.00

Sickness Questionnaire (SSQ) [4] and Self-Assessment Manikin (SAM) [5] for self-evaluation and collected three physiological signals such as electrocardiogram (ECG), skin temperature (SKT) and electrodermal activity (EDA) for objective evaluation. To explore the relationship between physiological signals and subjective feelings, we designed the model to predict subjective categories through physiological signals and compared the effects in VR and 2D environments.

2. RELATED WORKS

Affective computing [6] is an important subject in the research of human–computer interaction, which aims to let computers understand and analyze people's emotions, thus achieving human–machine combination of affection. Researchers try to use various methods to induce emotions, choose appropriate methods to monitor emotional changes, and build emotional models to define emotions. People have been studying the emotional effects of virtual reality technology for more than 30 years. The main advantage of the virtual environment is to eliminate the trade-off between experimental control (precise manipulation of independent variables) and secular realism (the similarity between the experimental environment and the scenes encountered in life) that is difficult to solve in human perception and behavioral experiments [7, 8].

Since the end of the century, researchers have been using virtual reality technology to treat anxiety, eating disorder and other stress-related diseases. Virtual reality technology provides users the opportunity to immerse themselves in a fearful environment, thereby activating the structure of stimuli and stimulating meanings in fear memory, making it possible to treat phobias [9]. In 2007, Riva et al. [10] showed sixty-one college students three virtual park scenes of "anxiety," "relaxation" and "neutrality" environment, and proved that virtual reality is effective as an emotional medium. That is, virtual park scenes can make participants anxious or relaxed. Based on Blaskovich's proposal [11], the virtual environment had a more efficient advantage in psychological experiment research. The Stanford University Virtual Interaction Lab creates a public dataset containing 73 VR video clips, and used the average score of the participants to emotionally label all videos. Despite their efforts to locate and shortlist immersive VR clips for the study, there appears to be an underrepresentation for clips that both induce negative valence and are highly arousing, according to the valence–arousal plane. Generally, this work provides a good platform for studying the connection between virtual reality technology and emotion.

At present, some scholars have begun to study the differences of participants' emotions caused by video in VR and 2D environments. In 2013, Rooney et al. [12] compared the impact of 2D and 3D movies on audience sentiment. The study surveyed the emotions, participation, arousal and satisfaction of 225 cinema patrons who had just watched the 3D version of Thor and 10 viewers who watched the 2D version. The results showed that the audience had no significant difference in emotional arousal, satisfaction, etc. In 2017, Ni et al. [13] first compared the impact of movie clips on participants' emotions in VR and 2D environments. They randomly divided the 40 college students who participated in the experiment into two viewing groups, and arranged them to watch the clips of monkeys and snakes in the VR and 2D version of the Jungle Book. Finally, they came to the following conclusion: the VR environment can evoke a stronger emotional experience. The emotional arousal materials used in these two studies are limited to the same film. The former induced multiple emotions, but the evaluation of emotions is relatively simple. The latter intercepted fragments of monkeys and snakes in the film, which were designed to evoke the negative emotions of the participants. Therefore, the experiment was not representative. Beyond that, previous researches on emotion in the virtual environment was limited to the subjective level, while emotion calculations based on physiological signals are more objective and real.

Emotion detection refers to inferring emotional state through expressions, intonations and physiological signals. Some studies for the affect recognition had implemented supervised classification approaches. Some studies had proposed a variety of algorithms for emotion detection, such as k-Nearest Neighbor (k-NN) [14], Support Vector Machine (SVM) [15, 16] and Neural Networks [17, 18]. Most of these studies focus on improving the accuracy of detection in the same environment with little emphasis on the comparison of detection effects in different environments.

Despite extensive research advancement, there are still shortcomings in existing works. First of all, the research on emotion in the virtual environment is limited to the subjective materials, and there are very few studies based on objective physiological signals. Second, there are few comparative experiments on the emotional impact of different dimensions and experimental environments. Third, previous research focus on designing algorithms to improve the performance of emotion detection, and there are almost no researches on the comparison of emotion detection effects in VR and 2D environments. To remedy these problems, this article proposes experiments to further explore the difference between VR and 2D environment.

3. EXPERIMENT METHODOLOGY

Emotion elicitation is the process that makes the experimenter produce emotional response naturally through specific materials or events. Levenson [19] defined emotional response as the type, intensity and duration of an individual's response to internal or external environmental factors. James Coan's [20] Handbook of emotion elicitation and assessment summarizes eight ways in which human emotions are commonly used in laboratories: video clip, static picture, scene reconstruction, etc. In this article, emotion elicitation modes based on VR and 2D videos could be classified as video clips. Figure 1 shows the experiment process.

The experiment combined subjective and objective data to compare the differences in user emotions elicited by the same material in VR and 2D environments. Thirty

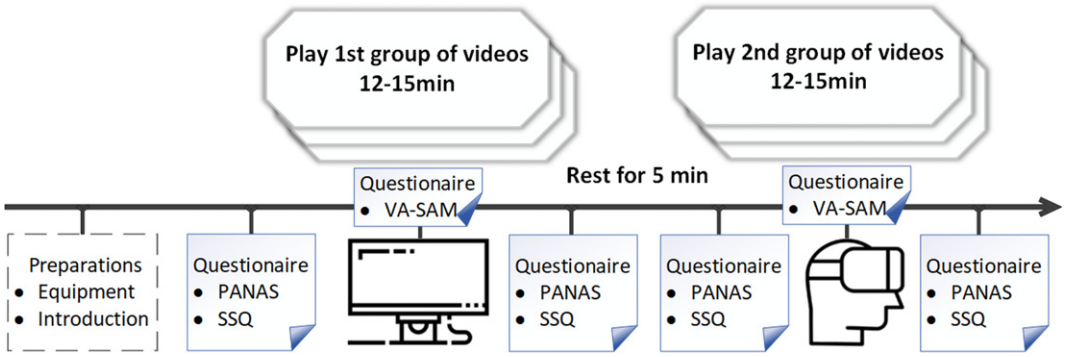

Figure 1. The process of experiment.

volunteers aged 20–26 years, including 18 males and 12 females, participated in this experiment. Each experiment was completed by two interviewers and one subject. The dataset used in the experiment contained 12 videos (3 videos for each valence–arousal degree quadrant). The participants were asked to watch 2 randomly selected videos from each quadrant. After the experiment, each video will be viewed 20 times. There are two reasons for this design: Firstly, previous studies have shown that watching a video for more than 15 minutes can cause fatigue and drowsiness, so a shorter viewing duration—12 minutes—is selected for the purpose of avoiding fatigue. Secondly, there may be program errors, poor contact, etc. during the experiment, resulting in a disruption in viewing. Thus, it is unreasonable for each participant to view the entire video.

This experiment made up for the inadequacies in Ni et al. [13], in which only one elicitation material was used. However, the VR environment has many differences with the 2D environment in some ways and comparing the emotional effects of specific video between two environments is still very risky. For example, the 2D video clips used in the experiment are extracted by VR video. Although we guaranteed the same duration and content, and tried to make the video perceived by the participants in the two environments as clear as possible, the VR environment provides 360° immersive images which could contain more information, and the loss of 2D video information after processing is unavoidable. Therefore, the experiment adopted the following principles to minimize the experimental error: (1) considering the impact of the sequence of experiments on the experimental results, the sequence of experiments performed in the 2D and VR environments was random and half of the subjects were first tested in a 2D environment. (2) The selection and viewing order of the eight videos is also random. Second, randomly select 8 videos and ensure that the playback order is random. (3) The adjacent two videos have different valence (low/high) and arousal (positive/negative) [19]. According to these criteria, 30 sets of video viewing sequences that meet the requirements were generated by the program before the experiment. The following will introduce the dataset and methods data collection about subjective experience and objective physiological signals.

4. MATERIALS

The emotion elicitation dataset consists of a VR video and its corresponding 2D video. The VR video was selected from a public database with valence–arousal scores created by Benjamin [11] from virtual interaction lab at Stanford University. Benjamin's dataset contains 73 360° VR video clips for emotional research. After the tag value of these videos was drawn into a scatter plot such as Figure 2, all the videos in the dataset were better distributed in different quadrants, indicating that the dataset is reasonable. However, through the further understanding of the process and content of the dataset, some problems still existed. First, a significant portion of the video in the dataset is an English short play or documentary. These videos contain a single scene or have extremely high requirements for English listening. Because most of the subjects recruited in this experiment are Chinese students, this type of video will be excluded. Second, since the material source of the public dataset is some mainstream video websites, there are strict restrictions on violence and horror videos, and the dataset lacks representativeness in the high arousal low valence (HALV) quadrant. This topic should supplement videos in this quadrant. Third, the titles and trailers of some videos are too long, and some videos contain long, content-independent portions. These parts may not be beneficial to emotion elicitation, but will increase the fatigue of the participants, and should be removed by editing on the basis of ensuring video consistency.

In order to solve the above three problems, we had screened, supplemented and improved Benjamin's dataset, and established a more appropriate emotional induction dataset for our experiments. First, VR video clips with less English content were selected from the three quadrants of high arousal high valence (HAHV), low arousal high valence (LAHV) and low arousal low valence (LALV). Secondly, according to the existing video content of the original database HALV quadrant, similar videos that are more irritating were found from other websites, and ensured that there are three videos in each quadrant. After cutting off the

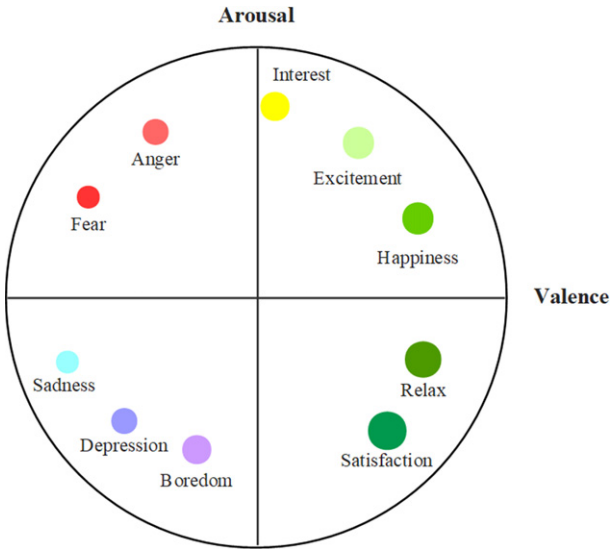

Figure 2. Valence–arousal model.

clip of each video that is not related to the content, a dataset containing 12 VR videos was obtained.

The experiment tried a total of three possible methods to obtain the 2D version from the VR video: the first method is to play directly through the video website, the second method is to use the *Unity 3D* to paste the video inside the ball, and place a camera at the center of the ball to output the camera's picture, the third method is to use the *Go Pro VR* player to output the normal view version with "standard mode." After repeated attempts, the processing result of the first video has the highest resolution, but it is limited by network conditions; the second method has low video resolution after processing; the third method is the most convenient and acceptable effect. After the comprehensive comparison, the third method is used to play the video in the 2D environment.

4.1 *Questionnaire Design*

Participants completed two subjective questionnaires before and after each group of experiments: Positive And Negative Affect Schedule (PANAS) and Simulator Sickness Questionnaire (SSQ). After watching each video, subjects were asked to rate their valence and arousal in the self-assessment manikin (SAM).

PANAS is a self-rating scale that evaluates the emotional state of both positive and negative dimensions proposed by Watson in 1988. The two emotional dimensions contain several English words describing emotions. The words give a score of 1–5, means the degree from "nearly no" to "very intense" [21]. The scale was proved to be of good credibility and has been widely used by diverse groups from different regions in the past ten years [3].

SSQ [4] is a scale proposed by Kennedy in 1993 to evaluate flight simulator and marine ship simulator diseases. If the user spends too much time in the virtual environment, symptoms such as nausea, dizziness and dazzling symptoms may appear. The SSQ assessed the symptoms of motion sickness in the virtual environment due to visual impact, including 16 indicators describing physiological sensations.

SAM is a self-evaluation model based on picture description, so that volunteers can directly measure the valence and arousal caused by emotion elicitation materials. They determine the score by referring to the emotional state of the villain in the picture, and the score range of valence and arousal is between 1 and 9. SAM is a simple and effective way to assess emotions that can be applied to quickly evaluate emotions [5]. In this model, the horizontal axis (arousal) is used to measure the degree of emotional excitement, and the vertical axis (valence) is used to measure the state of the emotion (positive–negative). The valence–arousal model can express multiple emotions on a two-dimensional plane, as shown in Fig. 2.

4.2 *Physiological Signal*

In emotional research, the process of selecting and collecting physiological parameters is particularly important. Usually, people's measurements and definitions of their own emotions are one-sided and single. It is difficult for us to directly judge complex and specific emotional categories and intensity through words or expressions. In order to reflect the emotional state better, we chose electrocardiogram (ECG), skin temperature (SKT), and electrodermal activity (EDA) for research, as shown in Figure 3. These signals are described below:

- ECG signals are the expression of the response of the autonomic nervous system to changes in the environment and physiological systems [32]. It reflects the change of central potential in the process of heart contraction and diastole.
- SKT is thought to be closely related to emotions in the study of physiological signals: higher skin temperature is related to positive emotions while lower skin temperature is related to negative emotions [23–25].
- EDA refers to all the parameters of skin electrical signals, and is the only autonomic physiological parameter that is not affected by parasympathetic activity. Therefore, EDA is often used as an important indicator of emotional cognition [26, 27].

The objective physiological signals were recorded in the experiment using BIOPAC's MP150 multi-channel physiological signal recorder to collect participants' ECG, SKT and EDA signals. After connecting the MP150 to the ECG100C, SKT100C and GSR100C modules, connect the MP150 to the computer via a network cable. The acquired physiological signals are displayed in real time through the software Acqknowledge 4.2.

4.3 *Study I: Analysis of Subjective Data in 2D and VR Environments*

We analyzed the three subjective data of VA-SAM, PANAS, and SSQ, which will be introduced in the following article.

Table I. Prominence of *T* test.

	0	1	2	3	4	5	6	7	8	9	10	11
Valence	0.333	0.038*	0.399	0.055	0.514	0.848	0.741	0.366	0.236	0.366	0.463	0.0005**
Arousal	0.048*	0.002*	0.0001**	0.0001**	0.025*	0.035*	0.011*	0.001*	0.002*	0.0002**	0.0437*	0.005**

Note: "*" indicates significant difference, "**" indicates extremely significant difference.

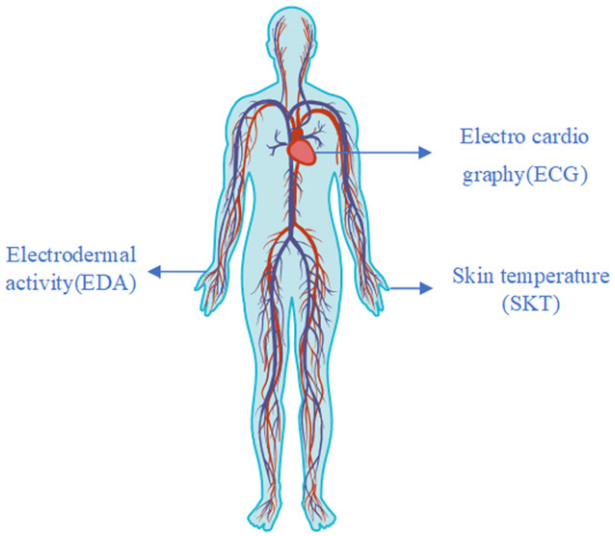

Figure 3. The position of the physiological signals.

4.4 *Analysis of VA-SAM*

After each video trail, participants completed their valence–arousal assessment of their emotional state in a short period of time. The scores of all participants were recorded for each video, and Figure 4 shows the mean and standard deviation of the valence and arousal. It can be seen that after watching video clips, the positive and negative index score of participants' sentiment between the 2D and VR environments is not much different, but the degree of excitement had changed to a large extent: participants were more excited overall in the VR environment than in the 2D environment.

In this experiment, the paired sample T test [22] was used to test whether the valence and arousal induced by the video segment were significantly different between VR and 2D environments. Suppose H1 has significant difference in the subjective data impact of watching video in two environments. For each video segment, the valence–arousal assessment results of the same participant population were paired in two environments. We collected and summarized the valence and arousal scores of the same video, and then got a total of 24 pairs of data from 12 videos in the two emotional dimensions of valence and arousal, and perform paired sample T test. The results are shown in Table I.

The results are consistent with the previous analysis of average values. The test results of almost all video clips show that, in different viewing environments, the positive and negative evaluations of participants' emotions are not significantly different, while the arousal evaluation of the participants in the two environments, as known as the degree of agitation, is of significant difference. The overall analysis can conclude that participants in the VR environment are more agitated. The average of arousal scores of all 12 videos in the VR environment is higher than 2D. And compared to the 2D environment, there is a significant difference in the mean value of arousal scores.

Table II. Results of one-way ANOVA in PANAS and SSQ.

	2D		VR		one-way ANOVA	
	Mean	Sd	Mean	Sd	F	P
Positive emotions	−13.444	13.992	6.222	7.4293	13.872	0.018*
Negative emotions	2.778	3.114	2.667	5.8737	1.895	0.951
Symptoms	0.0357	4.4121	4.250	3.873	8.38	0.007**

Note: "*" indicates significant difference, "**" indicates extremely significant difference.

4.5 *Analysis of PANAS*

First, we analyzed the differences of positive and negative emotions between the two environments. The PANAS mean results after removing the baseline (the difference between the values before and after watching the video) are shown in Figure 5. In all the emotional factors of the positive sentiment dimension, participants gave higher scores in the VR environment. In virtual reality, participants had stronger positive moods after watching a set of video clips. On this basis, the branches of all positive and negative emotion factors were summed and analyzed by variance test. The results of one-way ANOVA (Table II) reveal that positive emotions were significantly different in both environments, $F = 13.872$, $P = 0.018$), consistent with the mean value. However, there was no significant difference in negative emotion ($F = 1.895$, $p > 0.05$).

When comparing the differences of each emotional index, the statistical method of paired sample T test was used to match the emotional scores of the two environments into a pair, and the emotions of the participants before and after the experiment were analyzed separately. In the 2D environment, four positive emotions: enthusiastic ($p = 0.004$), pride ($p = 0.005$), happy ($p = 0.041$), grateful ($p = 0.018$); and a negative emotion: sadness ($p = 0.015$) showed a significant difference. In the VR environment, three positive emotions: active ($p = 0.001$), happy ($p = 0.018$), euphoria ($p = 0.006$)

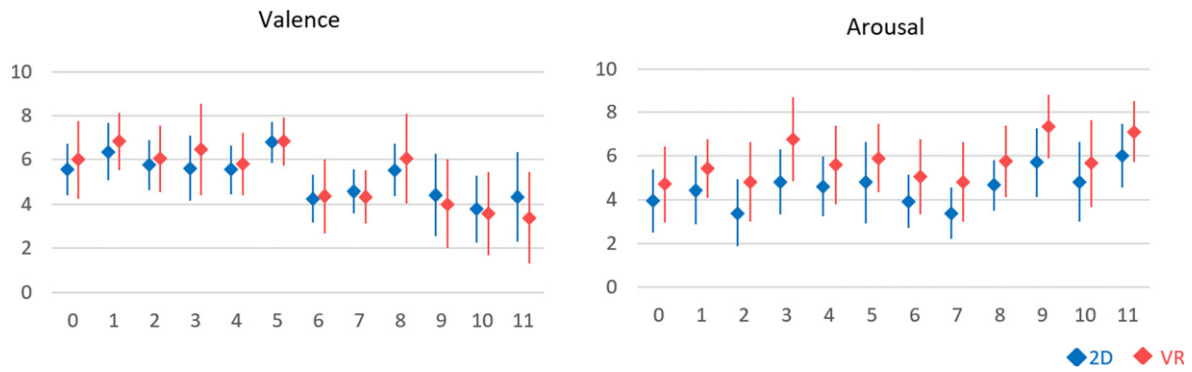

Figure 4. Mean and standard deviation of SAM.

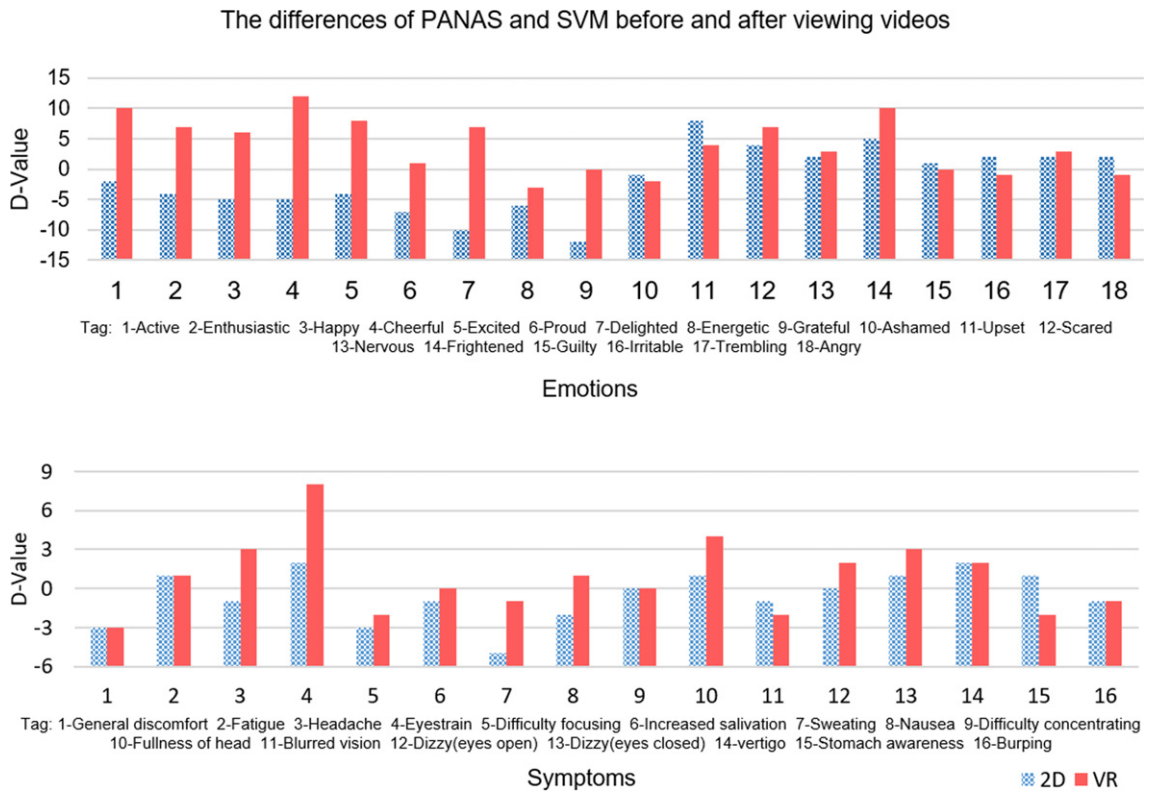

Figure 5. The differences of PANAS and SVM before and after viewing videos.

and three negative emotions: fear ($p = 0.021$), tension ($p = 0.004$), and panic ($p = 0.025$) showed significant differences. The result showed that some positive and negative emotions showed significant differences before and after watching video in two environments. In particular, more negative emotions showed significant differences in the VR environment than in the 2D environment.

4.6 *Analysis of SSQ*

The mean results of the simulator questionnaire after removing the baseline (the difference between the values before and after watching the video) are shown in Fig. 5. Participants in the VR environment generally felt eye fatigue, accompanied by headaches, dizziness and hair swelling, and the sweating situation was more pronounced as well.

The results of one-way ANOVA [23] showed that the participants in the VR environment have indeed realized the discomfort caused by the virtual world, as shown in Table II.

The analysis using the paired sample T test showed that in the 2D environment, the two indicators—eye fatigue ($p = 0.036$) and sweating ($p = 0.0171$)—before and after the experiment were significantly different, and in the VR environment, the three indicators—eye fatigue ($p = 0.028$), dizziness ($p = 0.038$) and vertigo ($p = 0.036$)—showed significant differences. This was in line with the experience of using virtual reality devices in our daily lives.

4.7 *Study II: Analysis of Physiological Signal in 2D and VR Environments*

4.8 *Signal Extraction and Optimization*

The experiment used the multi-channel physiological signal recorder MP150 produced by BIOPAC to record the three-channel signals of each participant's ECG, SKT and EDA signals. In order to facilitate statistical analysis later, there was a need to convert continuous signals into discrete values in certain ways. Recording high-quality ECG data plays a key role in studying heart rate variability (HRV, which is the difference in the continuous heartbeat cycle). Bentson et al.'s [28] findings suggested that even a miscalculation of heart rate variability caused by a single heart artifact in two-minutes ECG signals may have a much larger impact than typical effects in psychological and physiological studies. During the experiment, we set the appropriate sampling rate (1000 Hz) and filter parameters (0.5–35 Hz), and asked the subjects to keep their hands relatively static to reduce the interference caused by the movement of the wires. However, the heart telegrams originally obtained in the experiment also inevitably contain baseline drift and artifacts. Therefore, in this experiment, the sampling record was confirmed by visual observation before the feature value is extracted. If an abnormal signal occurs between the marked times, the signal will be processed by the 0.5–35 Hz bandpass FIR filter of the Blackman window function. The order of the filter is 8000, which was calculated by Eq. (1).

$$N = 4 \times \frac{\text{Sample Rate}}{\text{Lowest Frequency}}. \tag{1}$$

The time-frequency domain processing of ECG signals mainly includes R wave detection and calculation of parameters such as R–R interval, heart rate, and interval standard deviation. We used a flexible dynamic threshold algorithm provided by heart rate detection in Acqknowledge software to detect peaks in a given window that match set thresholds, as shown in Eq. (2).

$$\text{New Peak} = 0.75 \times \text{Old Peak}_{\text{Max}} - \text{Old Peak}_{\text{Min}}. \tag{2}$$

When using dynamic threshold detection, set the heart rate window to 40–120 BMP, and the maximum and minimum values of the signal peaks in the window will be updated continuously. If the input signal exceeds the specified heart rate window range, the heart rate calculation and automatic threshold detection function will be reset and the reset calibration will be output. Noise outside the peak 5% range will also be suppressed. After the threshold detection and noise removal, the interval between the two peaks of the ECG signal is recorded, which is the R–R interval. Through the RR interval, we can calculate the HR (heart rate), SDNN (the RR standard interval of the normal sinus of the human body) Eqs. (4) and (5) and SDSD (the standard deviation of the difference between adjacent RR intervals), which is expressed by the following formula.

$$\text{HR} = \frac{60}{RR\,\text{Interval}} \tag{3}$$

$$\text{SDNN} = \sqrt{\frac{\sum_{t=1}^{N}(RR - \text{meanRR})^2}{N}} \tag{4}$$

$$RR_{n-1} - RR_n = D_{n-1} \tag{5}$$

$$\text{SDNN} = \sqrt{\frac{\sum_{t=1}^{n-1}(D_i - D_{mean})^2}{n-1}}. \tag{6}$$

Skin electrical activity is considered to be a common observation channel for sympathetic nervous system activity, and is the activity associated with skin conductivity level and short-term stimulation events, which is the superposition of skin's conductive response. Throughout the past two decades since 1992, the method of extracting the characteristics of skin electrical signals had undergone a transition from analyzing the basic activity components to the phase activity components [29]. Boucsein believed that the underlying activities of skin electrical impedance were constantly changing for individuals and there are significant differences within individuals as well. Therefore, some researchers believed that it is difficult to obtain valuable information by analyzing the skin level of an individual directly. Due to the interference of phase activity, the result of simply averaging the entire skin electrical signal is much larger than the true value. Such analysis was very unreasonable [26]. The characteristics of the skin electrical signals used in this experiment were extracted according to the method proposed by Greco et al., and the skin electrical signals were regarded as the superposition of the basic reaction, the phase response and the Gaussian white noise, and the cvxEDA algorithm was used for waveform detection to obtain all the skin electrical signals. After that, feature values such as time of peaks and valleys, amplitude, slope and rise time trough were extracted for statistical analysis [27]. The experiment performed analysis of the whole skin electrical signals of each video. Each segment of the signal was sequentially subjected to filtering processing and peak detection. Firstly, the signal was subjected to fourth-order Butterworth low-pass filtering, and the cutoff frequency was set to 1. Secondly, the Filt_filt function was called to input the skin electrical signals into the zero-phase filter function in the positive and negative order, respectively, and the filtered signal was output for the skin electrical reaction detection. The peak and peak time of the skin electrical signal were obtained by the idea of the first-order differential sign before and after the peak occurrence time. Finally, three parameters of amplitude (Eq. (7)), slope (Eq. (8)) and rise time (Eq. (9)) were obtained from the peak detection and performed threshold processing.

$$\text{Amplitude}_{\text{i}} = eda(edr.Peak_i) - eda(edr.\text{Valley}_i) \tag{7}$$

$$\text{Risetime}_{\text{i}} = \frac{edr.Peak_i - edr.\text{Valley}_i}{f_s} \tag{8}$$

$$\text{Slope}_{\text{i}} = \frac{\text{Amplitude}_i}{\text{Risetime}_i}. \tag{9}$$

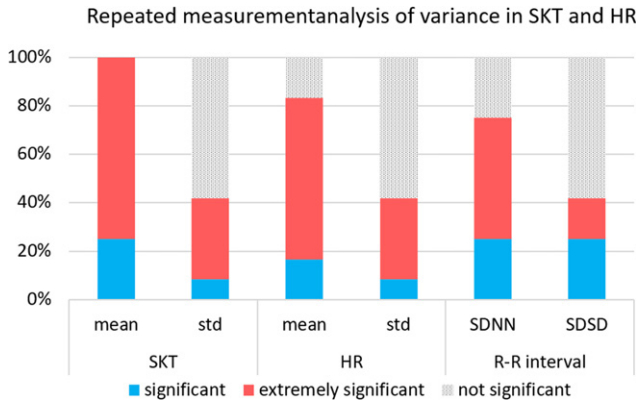

Figure 6. Repeated measurement analysis of variance in SKT and HR.

4.9 *Analysis of SKT and HR*

For SKT and ECG signals, a repeated measures analysis of variance model was used to detect differences in eigenvalues extracted from these two physiological signals in different environments. For skin temperature and ECG signals, the eigenvalues of the same participant in different time segments of a video are a set of variables, and were tested in two different environments, VR and 2D, so we applied repeated measurement analysis of variance to detect differences in eigenvalues extracted from these two physiological signals in different environments. There are six kinds of data analyzed here: mean and standard deviation of SKT and HR, SDNN and SDSD; the distribution of the results is shown in the Figure 6.

The average skin temperature showed a very significant difference between the two environments. The parameter p of 10 videos were less than 0.01. Similarly, mean and interval standard deviations of HR were equally significant in 2D and VR environments, with three quarters of the videos showing significant differences and inconspicuous video distribution at different valence–arousal quadrant.

4.10 *Analysis of EDA*

The electrical signals have been shown to be closely related to the autonomic nervous system and psychological stimulation, and thus this physiological signal is widely used to quantify the level of arousal in emotional and cognitive processes. Different participants have different perceptions of the same video, and the obtained skin electrical signals and peak detection results are also different [30]. Electrodermal response (EDR) is the response of participants to stimuli within a short period of time. After preliminary examination of the peak detection results, it was found that even in the same participant, the skin electrical reaction data in 2D and VR environments differed a lot, and it was difficult to specifically determine the critical time point at which the peak of the skin electrical signal was triggered. Therefore, the view of comparison was translated into the intensity of the overall fluctuation of the electrical signal of the skin. On each video, the mean of the three values of the amplitude, slope and rise time of all skin electrical responses was analyzed.

Table III. Results of T test.

	Peaks	Slope	AMP	Rise time
P	0.0000**	0.0069**	0.0176*	0.0257*
Sd	0.9760	0.1434	0.0597	0.1844

Note: "$*$" indicates significant difference, "$**$" indicates extremely significant difference.

The processed EDA signal was extracted and analyzed as a whole, and the mean results of the four indicators of peak number, amplitude, slope and rise time are obtained as shown in Figure 7. According to the histogram of the mean, it can be seen more intuitively that in the VR viewing environment, the values of the number of peaks, amplitude, and slope are higher than the 2D environment. That is to say, in the VR environment, participants' emotions change more frequently and their moods fluctuate more severely. It is worth mentioning that in the VR environment, all four indicators of skin electrical signals are significantly higher than the 2D environment in HALV quadrant (Video 09–Video 11).

The statistical method of paired sample T test was used to measure the difference between the four indicators. The results showed that in the two viewing environments, the changes in the EDR signals caused by the video clips are significantly different: The number of peaks ($p = 0.000$), amplitude ($p = 0.006$) and slope ($p = 0.017$) and rise time ($p = 0.026$) in the VR environment were significantly higher than the 2D environment, shown in Table III.

4.11 *Emotion Detection based on Dataset of VR and 2D Environments*

In the experiments of Ali [18] et al., different classification methods were used for emotion detection. The ECG, EDA and SKT signals were selected and tested independently. The experimental results showed that the emotion detection based on EDA signals achieves the highest accuracy. In other words, EDA signals are more suitable for emotion detection than ECG and SKT signals. They used the MAHNOB [31] dataset for prediction, which only contains emotional data based on 2D video. For this reason, this work is limited to testing different methods in a single experimental environment.

The peak number, amplitude, slope and rise time characteristics are extracted for testing. According to the conclusions in previous studies, the difference in valence degree between the two experimental environments is little, and the arousal degree is quite different. Therefore, we selected the arousal as the indicator and tested five learning models, Decision Tree, Logistic Regression, KNN and SVM, using 70% of the data as the training set and 30% as the verification set. The experimental results are shown in Table IV.

The results in the above table show that the SVM model has the highest prediction accuracy in both VR and 2D environments. The prediction accuracy of all four

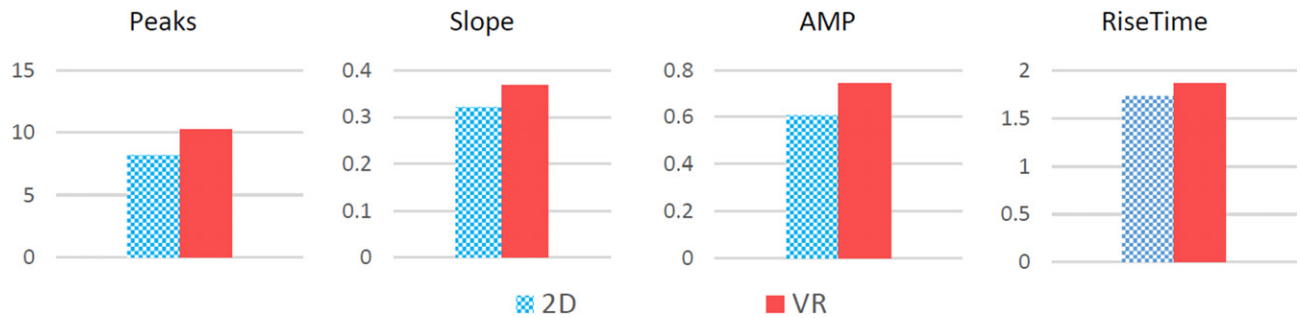

Figure 7. The difference of four indicators in 2D and VR environments.

Table IV. Accuracy of emotion detection.

	Decision tree	Logistic regression	KNN	SVM
VR	0.62	0.77	0.71	0.78
2D	0.51	0.65	0.51	0.62

models in the VR environment is higher than that of the 2D environment, which is also the side proof that the VR environment can trigger more intense emotional expression.

5. DISCUSSION

Through the comparison with the 2D environment, this topic explores the different emotional effects caused by the VR environment from two aspects: subjective emotional experience and objective physiological signals. We created datasets for experiments in VR and 2D environments, and performed emotion elicitation by videos in different valence–arousal quadrants. Based on the establishment process of common emotional datasets, a comparative experiment of emotion elicitation in two environments was designed. We recruited volunteers to participate in the experiment, recorded the entire experiment with a camera, and used the subjective questionnaires and physiological signals obtained for statistical analysis. In the data analysis stage, this topic analyzes the subjective questionnaire and physiological signal obtained by participants through viewing emotion elicitation video clips in multiple aspects, and constructed emotion detection models to analyze and compare the emotion detection results in VR and 2D environments. The comprehensive analysis of all experimental results revealed some important characteristics of participants' emotions in the VR environment.

First, the analysis of the subjective questionnaire showed that the VR environment induced higher arousal and more intense emotions, which was similar to the research of Ding [13]. The results of the valence–arousal table showed that participants in the VR environment were more excited, and the average of the arousal scores of all 12 videos is higher than that of the 2D environment. Compared to the 2D environment, there is a significant difference in the mean value of the arousal of each video. That was different from Rooney's research [12] and the reason may be related to the material and the way of elicitation. The positive and negative emotions felt by the participants are also significantly different in the VR and 2D environments. The positive and negative emotion evaluation results showed that there is a significant difference in positive emotions between VR and 2D environments; that is, the average score of positive emotion factors in the VR environment is significantly higher than that in the 2D environment. Although there is no significant difference in the user's valence scores between the two environments in the valence–arousal table, all the mean values of valence in the VR environment are higher than the 2D environment in the high arousal low valence (HALV) quadrant. The results of valence–arousal table and the positive and negative emotion analysis showed that, participants' positive and negative emotions were more sensitive in the VR environment. In the same video, the user felt more intense in the VR environment than that of the 2D environment overall. In addition, the results of the SSQ questionnaire showed that users experienced symptoms of simulator disease such as fatigue and dizziness in the VR environment.

The eigenvalues of the three physiological signals of ECG, SKT and EDA signals were significantly different under the two environments. The RR interval was extracted from the processed ECG signals, and calculated the HR (heart rate), SDNN (the RR standard interval of the normal sinus of the human body, Eqs. (4) and (5)) and SDSD (the standard deviation of the difference between adjacent RR intervals) from the RR interval. Time-dependent peak numbers, amplitudes, slopes and rise times were obtained from the processed skin electrical signals. The eigenvalues extracted from the ECG and SKT signals every 10 seconds were analyzed in the same way. The results showed that the average values of SKT, HR and SDNN of the participants while viewing the majority of video clips were significantly different in the VR and 2D environments. The EDA signal is thought to be related to the arousal, so experiments had shown that participants in the VR environment had a higher degree of excitement, and the mood fluctuations were more frequent and more intense.

Previous research focuses on designing algorithms to improve the performance of emotion detection, and there is almost no research on the comparison of emotion detection effects in VR and 2D environments. We compared four

models of emotion detection in VR and 2D environments. We selected the EDA signal, the highest correlation with arousal, for follow-up study. The four algorithm models of Decision Tree, Logistic Regression, KNN and SVM were used for emotion detection experiments. The experimental results showed that the SVM has the best detection effect under the two environments. The detection accuracy of the four algorithms in the VR environment is higher than that of the 2D environment, which confirmed that the volunteers in the VR environment have more obvious skin electrical signals, and had a stronger sense of immersion.

6. CONCLUSION

The experiment reveals the specific differences of user's emotional experience between the virtual reality and the traditional 2D environment from both subjective and objective aspects. An emotion elicitation dataset was made before the experiment. For subjective data (VA-SAM, PANAS and SSQ) analysis showed that subjects in the VR environment were more excited and it is easier to elicit their emotions, but valence did not show any significant difference in these two environments from our analysis. At the same time, the VR environment is prone to cause symptoms such as dizziness, nausea and fatigue. The physiological signals ECG, SKT and EDA differed greatly in the two environments, which verified the effectiveness of using them to measure emotions. Studies in emotion detection had shown that the EDA signals in the VR environment had a stronger correlation with subjective perception, indicating that it is easier to elicit emotions.

In the future work, it is possible to mark the plots of the video clips that may cause emotional fluctuations. In this way, we could get more specific changes in the objective physiological signals and obtain more accurate experimental data. In addition, it is also possible to optimize and improve the feature extraction method in emotion detection, which can further improve the accuracy and further analyze the relationship between subjective and objective emotions. In addition, it is also possible to optimize the feature extraction method in emotion detection in order to improve the accuracy and make further analysis on the relationship between subjective and objective emotion data.

Since the video data used in this experiment is in line with people's daily viewing video, the above findings might provide helpful guidance for the design of VR devices and video. We hope that our work can play a positive role in the development of the VR industry.

ACKNOWLEDGMENT

The authors greatly appreciate Kang Yue, Yaoguang Song, Keqiao Zhang and other personnel who had participated in the evaluation experiments and helped them the most.

This research is supported by the National Key Research and Development Program under Grant No. 2016YFB0401202, and the National Natural Science Foundation of China under Grant No. 61872363, 61672507, 61272325, 61501463 and 61562063.

REFERENCES

[1] W. C. Huang and S. K. Chiang, "The international affective picture system (IAPS)—comparison of evaluating method in young adults sample," Adv. Psychol. Res. **4**, 202–209 (2014).
[2] S. Koelstra, C. Mühl, M. Soleymani, A. Yazdani, J.-S. Lee, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "DEAP: a database for emotion analysis using physiological signals," IEEE Trans. Affective Comput. **3**, 18–31 (2012).
[3] L. Qiu, X. Zheng, and Y. F. Wang, "Positive and negative affect schedule (PANAS)," Appl. Psychol. **14**, 249–254 (2008).
[4] R. S. Kennedy, N. E. Lane, K. S. Berbaum, and M. G. Lilienthal, "Simulator sickness questionnaire: An enhanced method for quantifying simulator sickness," Int. J. Aviat. Psychol. **3**, 203–220 (1993).
[5] J. D. Morris, "SAM: the self-assessment manikin. An efficient cross-cultural measurement of emotional response," J. Advertising Res. **35**, 63–68 (1995).
[6] R. W. Picard, *Affective Computing [M]* (MIT Press, Cambridge, MA, 2000).
[7] J. Blascovich, L. Loomis, A. C. Beall, K. R. Swinth, C. L. Hoyt, and J. N. Bailenson, "Immersive virtual environment technology as a methodological tool for social psychology," Psychol. Inquiry **13**, 103–124 (2002).
[8] A. Birenboim, M. Dijst, D. Ettema, J. de Kruijf, G. de Leeuw, and N. Dogterom, "The utilization of immersive virtual environments for the investigation of environmental preferences," Landscape Urban Plan. **189**, 129–138 (2019).
[9] B. Bandelow, M. Reitt, C. Rover, S. Michaelis, Y. Gorlich, and D. Wedekind, "Efficacy of treatments for anxiety disorders: a meta-analysis," Int Clin. Psychopharmacol **30**, 183–192 (2015).
[10] G. Riva, F. Mantovani, C. S. Capideville, A. Preziosa, F. Morganti, D. Villani, A. Gaggioli, C. Botella, and M. L. A. Raya, "Affective interactions using virtual reality: the link between presence and emotions," CyberPsychol. Behav. **10**, 45–56 (2007).
[11] B. J. Li, J. N. Bailenson, A. Pines, W. Greenleaf, and L. M. Williams, "A public database of immersive VR videos with corresponding ratings of arousal, valence, and correlations between head movements and self report measures," Front. Psychol. **8**, 2116 (2017).
[12] B. Rooney and E. Hennessy, "Actually in the cinema: A field study comparing real 3D and 2D movie patrons' attention, emotion, and film satisfaction," Media Psychol. **16**, 441–460 (2013).
[13] N. Ding, W. Zhou, and A. Y. H. Fung, "Emotional effect of cinematic VR compared with traditional 2D film," Telemat. Inform. **35**, 1572–1579 (2018).
[14] L. Zhao, L. Yang, H. Shi, Y. Xia, F. Li, and C. Liu, "Evaluation of consistency of hrv indices change among different emotions," *2017 Chinese Automation Congress (CAC)* (IEEE, Piscataway, NJ, 2017), pp. 4783–4786.
[15] M. Wiem and Z. Lachiri, "Emotion classification in arousal valence model using mahnob-hci database," Int. J. Adv. Computer Sci. Appl. **8** (2017).
[16] M. Matsubara, O. Augereau, C. L. Sanches, and K. Kise, "Emotional arousal estimation while reading comics based on physiological signal analysis," *Proc. 1st Int'l. Workshop on coMics ANalysis, Processing and Understanding, ser. MANPU'16* (ACM, New York, NY, 2016), pp. 7:1–7:4.
[17] L. Santamaria-Granados, M. Munoz-Organero, G. Ramirez-Gonzalez, E. Abdulhay, and N. Arunkumar, "Using Deep Convolutional Neural Network for Emotion Detection on a Physiological Signals Dataset (AMIGOS)," IEEE Access **7**, 1–1 (2018).
[18] M. Ali, F. A. Machot, A. H. Mosa, and K. Kyamakya, "CNN based subject-independent driver emotion recognition system involving physiological signals for ADAS," *Advanced Microsystems for Automotive Applications* (Springer International Publishing, Cham, Switzerland, 2016).
[19] J. J. Gross and R. W. Levenson, "Emotion elicitation using films," Cogn. Emotion **9**, 87–108 (1995).
[20] J. A. Coan and J. J. B. Allen., "Handbook of emotion elicitation and assessment," Neuroimage **37**, 866–875 (2015).
[21] D. Watson, L. A. Clark, and A. Tellegen, "Development and validation of brief measures of positive and negative affect – the panas scales," J. Pers. Soc. Psychol. **54**, 1063–1070 (1988).
[22] L. Sheffield, H. Whitford, and N. A. Cressie, "Use of paired sample T-Test in the real world," Pract. Otol. **3**, 201–205 (1985).

[23] M. T. Alice, S. Brian, K. Geoffrey, S. V. Joseph, M. Joan, B. James, M. Scott, A. Greg, G. Sachin, and S. Jan-Benedict, "Analysis of variance," J. Consum. Psychol. **535** (2001).

[24] S. D. Kreibig, "Autonomic nervous system activity in emotion: A review," Biol. Psychol. **3**, 394–421 (2010).

[25] R. A. McFarland, "Relationship of skin temperature changes to the emotions accompanying music," Biofeedback Self-Regul. **3**, 255–267 (1985).

[26] J. J. Braithwaite, D. G. Watson, R. Jones, and M. Rowe, "A guide for analysing electrodermal activity (EDA) & skin conductance responses (SCRs) for psychological experiments," Psychophysiology **1**, 1017–1034 (2013).

[27] A. Greco, G. Valenza, A. Lanata, E. P. Scilingo, and L. Citi, "cvxEDA: A convex optimization approach to electrodermal activity processing," IEEE Trans. Biomed. Eng. **4**, 797–804 (2015).

[28] G. G. Berntson and J. R. Stowell, "ECG artifacts and heart period variability: Don't miss a beat," Psychophysiology **1**, 127–132 (1998).

[29] L. Shu, J. Xie, M. Yang, Z. Li, Z. Li, D. Liao, X. Xu, and X. Yang, "A review of emotion recognition using physiological signals," Sensors **7**, 2074 (2018).

[30] W. Boucsein, *Electrodermal Activity* (Springer Science & Business Media, US, 2012).

[31] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, "A multimodal database for affect recognition and implicit tagging," IEEE Trans. Affective Comput. **1**, 42–55 (2012).

[32] P. Das, A. Khasnobish, and D. N. Tibarewala, "Emotion recognition employing ECG and GSR signals as markers of ANS," *2016 Conf. on Advances in Signal Processing (CASP)* (IEEE, Piscataway, NJ, 2016), pp. 37–42.