CEAP-360VR

A Continuous Physiological and Behavioral Emotion Annotation Dataset for 360 VR Videos

Xue, Tong; El Ali, Abdallah; Zhang, Tianyi; Ding, Gangyi; Cesar, Pablo

DOI
10.1109/TMM.2021.3124080

Publication date
2021

Document Version
Final published version

Published in
IEEE Transactions on Multimedia

Citation (APA)
Xue, T., El Ali, A., Zhang, T., Ding, G., & Cesar, P. (2021). CEAP-360VR: A Continuous Physiological and Behavioral Emotion Annotation Dataset for 360 VR Videos. *IEEE Transactions on Multimedia*, *25*, 243-255. https://doi.org/10.1109/TMM.2021.3124080

CEAP-360VR: A Continuous Physiological and Behavioral Emotion Annotation Dataset for 360° VR Videos

Tong Xue, *Student Member, IEEE*, Abdallah El Ali, *Member, IEEE*, Tianyi Zhang, *Member, IEEE*, Gangyi Ding, *Member, IEEE*, and Pablo Cesar, *Senior Member, IEEE*

***Abstract*—Watching 360° videos using Virtual Reality (VR) head-mounted displays (HMDs) provides interactive and immersive experiences, where videos can evoke different emotions. Existing emotion self-report techniques within VR however are either retrospective or interrupt the immersive experience. To address this, we introduce the *Continuous Physiological and Behavioral Emotion Annotation Dataset for 360° Videos (CEAP-360VR)*. We conducted a controlled study (N=32) where participants used a Vive Pro Eye HMD to watch eight validated affective 360° video clips, and annotated their valence and arousal (V-A) continuously. We collected (a) behavioral (head and eye movements; pupillometry) signals (b) physiological (heart rate, skin temperature, electrodermal activity) responses (c) momentary emotion self-reports (d) within-VR discrete emotion ratings (e) motion sickness, presence, and workload. We show the consistency of continuous annotation trajectories and verify their mean V-A annotations. We find high consistency between viewed 360° video regions across subjects, with higher consistency for eye than head movements. We furthermore run baseline classification experiments, where Random Forest classifiers with 2s segments show good accuracies for subject-independent models: 66.80% (V) and 64.26% (A) for binary classification; 49.92% (V) and 52.20% (A) for 3-class classification. Our open dataset allows further experiments with continuous emotion self-reports collected in 360° VR environments, which can enable automatic assessment of immersive Quality of Experience (QoE) and momentary affective states.**

***Index Terms*—360° video, virtual reality, emotion, dataset, HMD, physiological signals, head and eye movement, continuous annotation.**

Manuscript received 27 May 2021; revised 22 October 2021; accepted 25 October 2021. Date of publication 2 November 2021; date of current version 13 January 2023. This work was supported in part by the National Key Research and Development Program of China under Grant 2020YFF0305200 and in part by the National Natural Science Foundation of China under Grant 62177005. The Associate Editor coordinating the review of this manuscript and approving it for publication was Prof. Mohammed Daoudi. *(Corresponding author: Tong Xue.)*

Tong Xue is with the School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081, China and also with the Distributed and Interactive Systems, Centrum Wiskunde & Informatica (CWI), 1089XG Amsterdam, Netherlands (e-mail: xuetong@bit.edu.cn).

Abdallah El Ali is with the Distributed and Interactive Systems, Centrum Wiskunde & Informatica (CWI), Amsterdam 1089XG, Netherlands (e-mail: abdallah.elali@gmail.com).

Tianyi Zhang and Pablo Cesar are with the Distributed and Interactive Systems, Centrum Wiskunde & Informatica (CWI), 1089XG Amsterdam, Netherlands and also with the Multimedia Computing Group, Delft University of Technology, 2600AA Delft, Netherlands (e-mail: tianyi.zhang@cwi.nl; p.s.cesar@cwi.nl).

Gangyi Ding is with the School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081, China (e-mail: dgy@bit.edu.cn).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TMM.2021.3124080.

Digital Object Identifier 10.1109/TMM.2021.3124080

I. Introduction

With the rapid development of VR technologies and increasing availability of commercial HMDs, 360° video has been flooding into our daily life and drawing great attention [3], [4]. As a new multimedia type, 360° video can provide virtual and immersive experiences by occupying the entire vision of the viewer. While watching 360° videos, viewers are allowed to freely rotate their head and focus on objects and regions of interest, which enables more immersive and interactive experiences [3], by contrast to desktop video. One key aspect is the capacity of VR to evoke a wide range of emotions in users [5], [6]. Example research areas include inducing emotional responses for educational purposes [7], tourism experiences [8], or for developing emotion recognition and adaptive systems [6] within immersive experiences. For such research, it is necessary to not only measure user experiences using a wide range of behavioral and physiological sensing devices, but also to collect accurate and precise emotion labels (i.e., ground truth).

To better understand users' emotion in virtual environments, recent research has measured user emotion states by collecting quantifiable user behavioral and physiological signals [9]–[11]. Common physiological measurements include Electroencephalography (EEG), Heart Rate Variability (HRV), and Electrodermal Activity (EDA). These are used in Quality of Experience (QoE) studies [12], Affective Virtual Reality Systems (AVRS) aimed at immersive emotion induction [13], and sensor-based affect data collection [6]. An important aspect of such virtual experiences is that individuals interact differently across emotion induction scenarios. In this respect, prior work has revealed a significant relationship between viewing behavior such as head movement (HM) and eye movement (EM) and dimensional emotion aspects of valence and arousal [14], [15]. However, emotions can be subjective and constructed (cf., facial emotion expressions [16]), where user behavior within VR can exhibit high variance across individuals. This means that for some emotional states, we do not always observe a clear overt behavioral manifestation, or what we observe may not represent the users' true emotion state. Given this, user self-reports

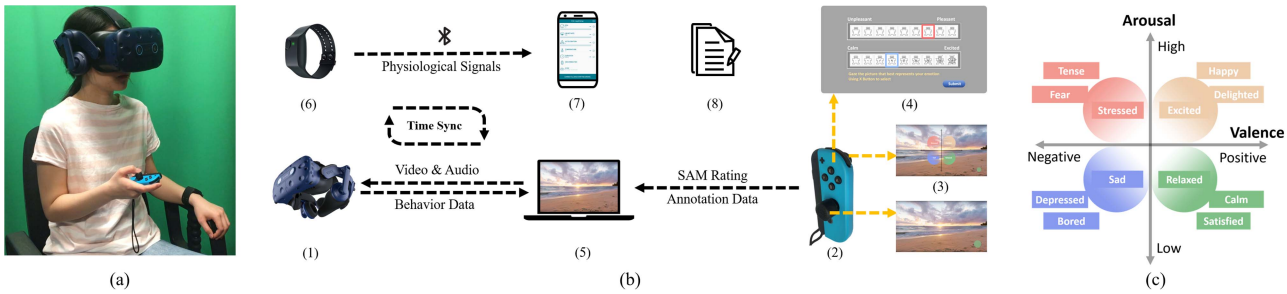


Fig. 1. (a) A participant in our experiment watching a 360° video using the HTC VIVE Pro Eye HMD and annotating her emotional state using a Joy-Con controller, while wearing an Empatica E4 Wristband on the non-dominant hand. (b) The system schematic shows various aspects of the experiment set-up and data acquisition. (c) Valence-Arousal model space based on Russell's Circumplex model [1]. In our annotation system, four distinct colors are selected across quadrants (HEX values = #eecdac, #7fc087, #879af0, #f4978e for quadrants one to four clock-wise, respectively, which has been shown to be intuitive and easy for users to understand [2]).

are essential to assessing whether a VR experience results in a dominant emotion.

Widely used emotion annotation methods are typically done post-stimuli (i.e., retrospectively after the experience), like Self-Assessment Manikin (SAM) [17], which divide human emotions into discrete basic emotion categories. Considering the time-varying nature of emotion [18], [19], there has been work on continuous emotion annotation systems which enable collecting more precise ground truth labels, and continuously through the duration of an experience [20]–[22]. As Toet *et al.* [23] pointed out, existing methods of collecting emotion data for 360° videos are either time consuming, require significant cognitive effort and task explanations, or are performed outside the VR environment [23], which interrupts the immersive experience. This requires new techniques and approaches for collecting continuous emotion ground truth data within a VR environment, with minimal interruption of user engagement while immersed in a VR experience.

Within the 360° video research, there have been several public datasets focused on the study of visual attention patterns [24], [25], visual quality assessment [26], or user viewing behavior [14]. For viewing behavior, this includes HM data and post-stimuli SAM ratings to explore the possible links between HM and valence/arousal. To further enable advances in emotion within VR environments, there is a need to create a high quality multi-modal dataset that contains HM/EM, physiological signals and corresponding continuous and precise ground truth emotion labels collected during immersive, virtual experiences. Our work offers two primary contributions:

1) Dataset: We conducted a controlled user study with 32 participants where each watched eight one-minute 360° video clips (as shown in Fig. 1(a)), and publicly make available the Continuous Physiological and Behavioral Annotation Dataset for 360° VR Videos *CEAP-360VR Dataset*. Our multi-modal 360° video dataset features precise and continuous emotion annotations alongside measured behavioral and physiological signals. Our dataset is publicly available at https://github.com/cwi-dis/CEAP-360VR-Dataset and https://www.dis.cwi.nl/ceap-360vr-dataset.

2) Analysis: We performed statistical analyses to validate our collected data, better understand affective states in 360° VR videos, and enable reproducibility and usage of our data by subsequent work. By automatically classifying self-reported affective states, we provide a means to assess the relationship between physiological and behavioral measures, and the moment-by-moment affective states during immersive 360° VR video watching experiences. We tested our dataset with common baseline classification methods, including both classical machine learning (ML) and deep learning (DL) classifiers. Results with a Random Forest classifier using a 2s segment length show good classification accuracies: for a subject-dependent model, 68.45% (V) and 71.33% (A) for binary classification, and 60.42% (V) and 62.38% (A) for 3-class classification; for a subject-independent model, 66.80% (V) and 64.26% (A) for binary classification, and 49.92% (V) and 52.20% (A) for 3-class classification. Furthermore, results from an ablation study shows that using only behavioral signals or only physiological data can yield reasonable recognition accuracies, however using both modalities improves classification performance.

Our dataset can be used for building more temporally precise emotion recognition models for 360° VR video watching. This can additionally be used for further analysis on visual attention modelling on 360° videos [4], [27], with considerations of momentary emotion self-report states. Researchers can also explore the relationship between HM/EM features and discrete self-reported affective states based on our dataset [15], [28]. Also, the diverse set of physiological signals collected can be used to conduct implicit perceptual experience analyses in HMD-based VR environments [29]. To summarize, our dataset can further advance the HMD-based 360° video community's understanding of momentary (self-reported) emotion states, and physiological and behavioral responses.

II. Related Work

In this section, we provide a review of datasets related to emotion recognition and 360° videos.

A. Datasets for Emotion Recognition in 2D Videos

There have been various datasets based on both explicit and implicit modalities evoked by 2D video stimuli. Soleymani *et al.* [30] presented work on emotion recognition where they analyzed the physiological responses (Electrocardiograph (ECG), EDA, EEG, Respiration (RESP), SKT) of 27 participants who watched various stimuli including 34 videos and some images. The proposed MAHNOB-HCI dataset contains face video, eye

gaze data and discrete scale of valence dominance, predictability as well as emotional keywords. The DEAP dataset [31] consists of implicit tagging from EEG and peripheral physiological signals (Electrooculography (EOG), EDA, RESP, Blood Volume Pulse (BVP), ECG, SKT) of 32 participants while watching 40 video clips. It includes a continuous scale of arousal, valence, liking, dominance and discrete scale of familiarity. Similarly, Abadi *et al.* [32] added Magnetoencephalogram (MEG) and presented the DECAF dataset. It contains a discrete scale of valence, arousal and dominance of 30 participants while watching 40 videos and 36 movie clips. The dataset AMIGOS [33] is compiled to model multi-class emotional data including EEG, ECG, EDA from 40 participants during the viewing of 20 short and long videos. It includes annotations of both internal self-assessment (scale questionnaires) and external assessment (frontal and full body videos) of affective levels. In the ASCERTAIN dataset presented by Subramanian *et al.* [34], the data recordings consist of physiological modalities (ECG, EDA, EEG) and facial activity. Discrete scale of valence, arousal, liking, engagement, familiarity and Big Five personality are also included. More recently, Sharma *et al.* [20] collected the CASE dataset of 30 participants in responses to eight validated videos. It includes synchronized recordings of physiological signals (ECG, BVP, EDA, EMG, SKT, RESP) and continuous reporting of valence and arousal. However, these datasets did not consider studying participants' emotions in virtual environments.

B. Datasets for 360° Videos

Previous studies [3], [12] have presented comparisons of QoE factors such as presence, engagement, usability and sickness while watching 360° videos among HMD, CAVE-based and 2D-based display screen. The results indicated that users can experience higher QoE ratings with an HMD. Recently, Qiao *et al.* [35] proposed a novel visual saliency model to predict viewport-dependent saliency on 360 videos considering both head movements and eye fixations. Several datasets report HM traces of users while watching 360° videos for visual attention research. Corbillon *et al.* [24] captured viewport traces of 59 participants watching five 70s videos. In [36], six videos were shown to 17 participants and the results of recorded scanpaths and fixation points suggest that users' attention is guided by moving objects. In another study [37], the PVS-HM dataset is created based on HM data of 58 subjects watching 76 videos. Analysis of the dataset indicates that there is similarity and a strong center bias across subjects. For 360° video, HM indicates the position of the subjects' viewport, while EM could reflect where the subject fixates on [4]. The Salient360 dataset constructed by David *et al.* [25] contains 19 immersive videos and 57 subjects' HM/EM data. The head+eye and head-only saliency maps and scan-paths are also included. Li *et al.* [26] proposed the VQA-OV, a 360° video dataset with HM, EM data and subjective quality scores of the sequences to study the links between user behavior and subjective evaluation on visual quality. Zhang *et al.* [38] presented a dataset including head and eye fixations of 104 videos watched by 20+ subjects for better modelling dynamic saliency. To explore HM/EM saliency prediction in dynamic 360° immersive videos, Xu *et al.* [39] presented a large-scale VR dataset including both HM and EM data of 31 participants watching 208 videos. Nguyen *et al.* [40] built a saliency dataset and proposed PanoSalNet, a saliency detection model.

Although human behavior in VR has been thoroughly investigated, few datasets have been developed using 360° videos for emotion induction research. One of the first datasets is gathered by Li *et al.* [14]. It contains HM data and corresponding ratings of arousal and valence captured with 93 participants watching 73 videos, given the purpose of exploring links between HM and emotions when viewing VR content. More recently, Tang *et al.* [28] reported an eye tracking dataset with valence and arousal scores from 19 participants watching 360° images to study the influence of emotions on eye behavior in a virtual setting. Their analysis showed that negative emotions have a significant impact on fixation and saccade features, while positive and neutral content do not. In our prior work, we additionally analyzed HM/EM features across fine-grained emotion labels from 360° video segments with varying lengths (5-60s) [15]. Our exploratory work showed that standard deviation of HM yaw negatively correlated with valence, HM pitch positively correlated with arousal, while standard deviation of EM yaw negatively correlated with valence, and EM pitch negatively correlated with arousal. Furthermore, recent studies in 360° videos took advantage of the relationship between physiological signals and users' emotions. Egan *et al.* [12] first took EDA and HR together to assess QoE in VR content. Marã-n-Morales *et al.* [6] recognized subjects' valence and arousal perceptions from EEG, HRV features and embedded SAM ratings in virtual environments. The findings validate that VR has the capacity to elicit emotional states and allow emotion recognition from physiological responses as with 2D videos. However, most of the existing research are based on authors' own data collection, which leads to limited accessibility for other researchers to reproduce results [41]. In addition, these studies pay attention to the user experience of VR and ignore the viewing behavior, as well as continuous emotion reports. To bridge these gaps, we propose the public CEAP-360VR dataset for emotion recognition in virtual environments watching 360° videos, containing both physiological signals and the corresponding viewing behavior data, as well as continuous self-report emotion ratings.

III. Experiment Protocol

In this section, we present our experiment protocol. This study was carried out in accordance with the recommendations of the Ethics Committee of our institute. Data collection was approved by the board and all participants. Below we describe our experiment setup and procedure.

A. Experiment Setup

We show the experiment architecture in Fig. 1(b), and each part is described in detail below.

(1) Participants viewed the 360° video clips through HTC Vive Pro Eye[1] HMD (in Fig. 1(b1)), with a reported 0.5° accuracy and frequency of 120 Hz Tobii Pro eye tracker integrated. The HMD

[1] [Online]. Available: https://enterprise.vive.com/us/product/vive-pro-eye/

provides a resolution of 2880 × 1600 pixels, a 110° field of view and a refresh rate of 90 Hz. In parallel, the audio signal is sent to the HMD. During the experiment, participants sat on a swivel chair and were free to look in any direction. Correspondingly, head rotation and eye gaze data from the headset were recorded at 120 Hz.

(2) The joystick used was a generic wireless digital gaming peripheral, called Joy-Con,[2] as shown in Fig. 1(b2). With a return spring, the proprioceptive feedback could aid realigning to center position under no force, which makes it suitable for continuous annotation while wearing an HMD. Also, we added a 11-mm heighten cap to extend the length of the joystick, thereby helping to increase flexibility of operation. The movement of the joystick head maps into a 2D Valence-Arousal space, in which the x axis indicates valence while the y axis indicates arousal, as shown in Fig. 1(b). Participants were instructed to annotate their emotion experience by moving the joystick head into one of the four quadrants. To increase the emotion intensity, participants could move the joystick head further. The annotation data was sampled at 10 Hz, in accordance with research on human motor control [42].

(3) We also developed an on-demand helper function, so that participants who forget what color maps to a quadrant with corresponding emotions could use it for easy lookup. This on-demand reference functionality is activated through a joystick button press event. We show the helper function in Fig. 1(b3), where we just include the most representative emotion keyword (by contrast to several keywords in Fig. 1(c)).

(4) After each video, participants were asked to report their emotional experience using a within-VR SAM rating. A SAM rating [17] panel was embedded in VR to visualize the 9-point scales of valence and arousal, which allows users to stay closer to the context of an ongoing exposure than outside of the VR [43]. Arousal scale ranges from "calm" (1) to "excited" (9), while valence ranges from "unhappy" (1) to "happy" (9), as shown in Fig. 1(b4). Participants could gaze at one picture and use the X button on the Joy-Con controller to indicate their self-assessment level.

(5) We constructed a custom scene in Unity Engine[3] (version 2018.4.1f1) to display 360° videos and audio and show the annotation feedback based on users' continuous ratings. Equirectangular content was projected onto the skybox while the camera was fixed into the center of the sphere. We integrated the Tobii Pro SDK[4] to collect data from HMD and eye tracker, along with the SteamVR SDK[5] which provides virtual reality support. The project ran on a 2.2G Hz Intel i7 Alienware laptop with an Nvidia RTX 2070 graphics card.

(6) We captured participants' physiological signals through the Empatica E4 wristband[6] worn on the non-dominant hand [44], as shown in Fig. 1(b6). This device can measure BVP, EDA and SKT. It also contains a 3-axis accelerometer, and a built-in application which calculates HR and IBI from BVP.

(7) A mobile device (Nexus 5, 32GB, 5 inches, 1920-1080) was used to collect data from the E4 band via Bluetooth. Timestamp of this device was set according to the clock of the experiment laptop, synchronized via an NTP server.[7]

(8) Validated questionnaires for sense of presence, workload, and level of motion sickness are used as subjective measures. We chose a standardized Simulator Sickness Questionnaire (SSQ) [45] to measure the level of motion sickness, and use the Igroup Presence Questionnaire (IPQ) [46] to evaluate perceptions of VR videos. For perceived workload, we used the NASA Task Load Index (NASA-TLX) questionnaire [47].

B. Independent Variables

Drawing on the Circumplex model of emotion (shown in Fig. 1(c)), there are four types of videos depending on valence and arousal scores, namely high valence / high arousal (HVHA), high valence / low arousal (HVLA), low valence / low arousal (LVLA), low valence / high arousal (LVHA). We follow a 4 (Video Type: HVHA, HVLA, LVHA, LVLA) X 2 (Peripheral Feedback: HaloLight vs DotSize) study design approach.

1) Stimuli Selection: We selected two sample 360° videos to represent each emotion type (as listed in Table I) from the database provided by Li *et al.* [14], which contains mean valence and arousal ratings (mean V-A ratings) from 95 subjects. We used youtube-dl[8] to download the contents from YouTube with 4K in resolution (3840 × 1920 pixels), equirectangular format. The videos come in different lengths and most are longer than 2 minutes, so we extracted a 60s segment from each of them with no scene cuts. A pilot study with 12 researchers from our institute indicated that clipped 60s videos still provided the same V-A ratings, and valence and arousal were rated similarly across participants, as shown in previous work [27].

In addition, we computed the Spatial Perceptual Information (SpI) and Temporal Perceptual Information (TpI) for eight selected videos in equirectangular format [48] to depict spatial and temporal complexity. SpI indicates the amount of spatial detail and is higher for more spatially complex scenes. TpI indicates the amount of temporal changes and is higher for high motion sequences. We did a two-way consistency intra-class correlation (ICC) analysis between valence/arousal labels from original dataset and SpI / TpI and the results show that there is no correlation ($p > 0.05$). This is not surprising, as our videos were selected on the basis of their emotion ratings, rather than other features such as spatial and temporal complexity. However, the low correlations do suggest that these features do not provide a confound with our emotion labels. Furthermore, the video attributes indicate some high-level semantic attributes such as indoor/outdoor, video category and objects of interests. The audio categories including background music (bgm), ambient sound (ambience), dialog and voice-over. Links and start time offset as well as valence and arousal scores are also presented in Table I.

[2][Online]. Available: https://www.nintendo.com/switch/choose-your-joy-con-color/

[3][Online]. Available: https://unity.com/

[4][Online]. Available: http://developer.tobiipro.com/unity/unity-getting-started.html

[5][Online]. Available: https://store.steampowered.com/app/250820/SteamVR/

[6][Online]. Available: https://www.empatica.com/en-int/research/e4/

[7][Online]. Available: android.pool.ntp.org/

[8][Online]. Available: https://github.com/ytdl-org/youtube-dl

TABLE I
DESCRIPTION OF 360° VIDEOS USED IN OUR EXPERIMENT. V = MEAN VALENCE RATING; A = MEAN AROUSAL RATING

VideoID	Type	DatasetID (V, A)	PilotStudy (V, A)	Name	YoutubeID	Start Offset	SpI	TpI	Audio Categories	Video Attributes	Description
V0	Training	63 (6.36, 5.93)	/	NASA - Encapsulation & Launch of OSIRIS Rex	D7-AmamuJEA	7s	51.91	0.93	voice-over, bgm	indoor, docu-mentary	Documentary film on planning and execution of rocket launches
V1	HVHA	50 (7.47, 5.35)	(7.08, 6.08)	Puppies host SourceFed for a day	c7sA3EdXSUQ	0s	61.41	9.14	bgm	indoor, action, dogs	Viewers get up close with some puppies
V5	HVHA	52 (6.75, 7.42)	(6.83, 7.42)	Speed Flying	g6w6xkQeSHg	0s	65.04	12.88	dialog, bgm	outdoor, sport, pilot	Viewer follows a speed wing pilot as he glides past mountain
V3	LVHA	21 (3.20, 5.60)	(2.58, 6.83)	Zombie Apocalypse Horror	pHX3U4B6BCk	65s	55.98	2.61	dialog, ambience, bgm	indoor, film, zombies	Film following some soldiers defending against zombie attack
V7	LVHA	68 (4.40, 6.70)	(4.42, 7.17)	Jailbreak 360	vNLDRSdAj1U	127s	46.78	2.25	dialog, ambience, bgm	indoor, action, criminal	Short film depicting a jailbreak from closed-circuit cameras
V2	HVLA	38 (6.13, 1.80)	(8.08, 1.91)	Mountain Stillness	aePXpV8Z10Y	10s	39.42	0.97	bgm	outdoor, tour, mountain	Atmospheric shots of Canadian snowy mountains
V6	HVLA	32 (6.57, 1.57)	(7.67, 1.50)	Malaekahana Sunrise	-bIrUYM-GjU	0s	47.34	0.36	ambience	outdoor, tour, sunrise	Viewer sees the sun rising over the horizon at a beach
V4	LVLA	14 (2.53, 3.82)	(2.42, 4.17)	War Zone	Nxxb_7wzvJI	3s	62.99	1.54	voice-over, ambience, bgm	outdoor, film, people	Journalistic clip of a war torn city
V8	LVLA	19 (2.73, 3.80)	(2.17, 3.17)	The Nepal Earthquake Aftermath	5tasUGQ1898	41s	76.11	2.07	voice-over, ambience, bgm	outdoor, film, buildings	Short film on the effects of an earthquake in Nepal
abandon	HVHA	69 (6.46, 6.91)	(4.17, 7.00)	Walk the tight rope	JtAzMFcUQ90	10s	/	/	/	/	Viewer experiences walking a tight rope over a canyon
abandon	HVHA	73 (6.27, 6.18)	(5.50, 6.58)	Through Mowgli's Eyes	bUiP-iGN6oI	13s	/	/	/	/	Short film with a conversation between an ape and a boy

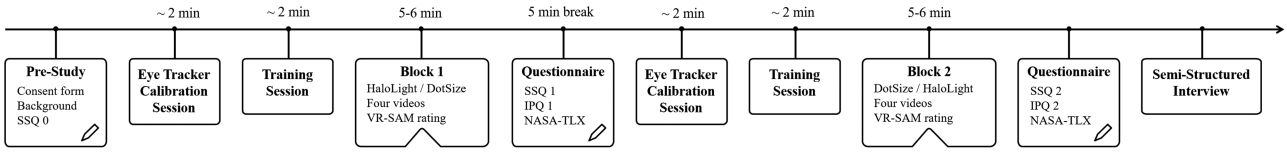
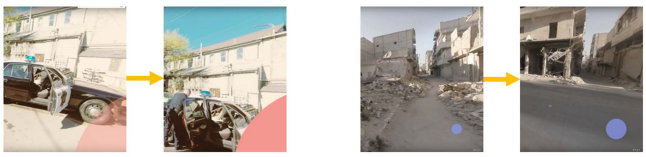


Fig. 2. The experiment procedure.



(a) HaloLight (b) DotSize

Fig. 3. Two peripheral visualization feedback methods.

2) Peripheral Feedback: Since users need to annotate their emotions in real-time while watching 360° videos, this will lead to divided attention. We contributed HaloLight and DotSize methods to provide peripheral feedback and minimize workload [49]. As shown in Fig. 3, HaloLight is a shaded halo arc in bottom-right viewport, which varies in transparency with emotion intensity. DotSize is a circle dot in bottom-right viewport, which varies in size with emotion intensity.

C. Participants

32 participants between the ages of 18 and 33 (M=25, SD=4.0) from different culture backgrounds participated in our data collection experiment. They were recruited by posters from near universities. All participants reported normal or corrected-to-normal vision and were not color-blind. They received monetary compensation for their participation. 50% of the participants are female and 27 participants have used VR devices less than five times before.

D. Experiment Procedure

We show the experiment procedure in Fig. 2. Duration of the entire session lasted around 50 minutes.

(1) Prior to commencing the experiment, we asked the participant to carefully read and sign the consent form and fill in a background information sheet. Then we gave a general explanation about the experiment steps and tasks, including the 2D Circumplex model (Fig. 1(c)) and how to annotate with the joystick. After all the questions about the experiment were addressed, we asked the participant to finish a pre-study SSQ.

(2) During the eye-tracker calibration session, we first helped the participant measure their Inter-Pupillary Distance (IPD). Then the participant sat in a swivel chair and put on E4 wristband

and HMD. The embedded eye tracker was calibrated following the VIVE Pro Eye instruction.[9] The calibration of eye tracker was performed every time the user put on the HMD, namely, before Block 1 and before Block 2.

(3) During the training session, we showed a documentary 360° video. The participant was orally instructed to get familiar with continuous emotion annotation method and visualization feedback, as well as 360° video viewing experience by moving their head and rotating the chair. This session took place before each block.

(4) Our main experiment consists of two blocks. In each block we fixed the peripheral feedback, and let participants watch four representative videos from each of the four quadrants. To counterbalance the effect of HaloLight and DotSize, half participants experienced HaloLight in the first block and then DotSize in the second block. For the other 16 participants, we showed DotSize in the first block and then HaloLight in the second block. Furthermore, we applied fractional factorial design [50] to counterbalance the effect of different videos within each block. To unify participants' starting position, before each video played, there was a black scene displayed in the HMD. We asked participants to find a white cube placed in the scene and then gaze at it. The cube would be highlighted in red while the participant gazed at it. If the cube is highlighted for five seconds, the cube disappears and the video immediately starts playing. In our early tests, we tried other mechanisms like marking the position of the swivel chair, however the advantage of showing a cube is that we can unify users' fixation consistency in the HMD. We introduced this step to participants during the pre-study session.

(5) While a participant viewed a 360° video, they rated emotional states (valence and arousal) continuously using the joystick. The HMD recorded the HM and EM data continuously, as well as the E4 wristband logged the physiological data continuously during the study period. To avoid carry over effects of one emotion to another and reduce the fatigue of viewing 360° video, a delay of 15 seconds was enforced between two videos. We also ensured a time gap of 5 minutes between two blocks following prior work [14], [51].

(6) At the end of each video, the participant submitted a SAM rating using the Within-VR SAM rating panel. At the end of each block, we helped the participant remove the HMD and fill in the SSQ, IPQ, NASA-TLX forms and then a semi-structured interview with five questions about user experience after the two blocks.

IV. Data Validation and Discussion

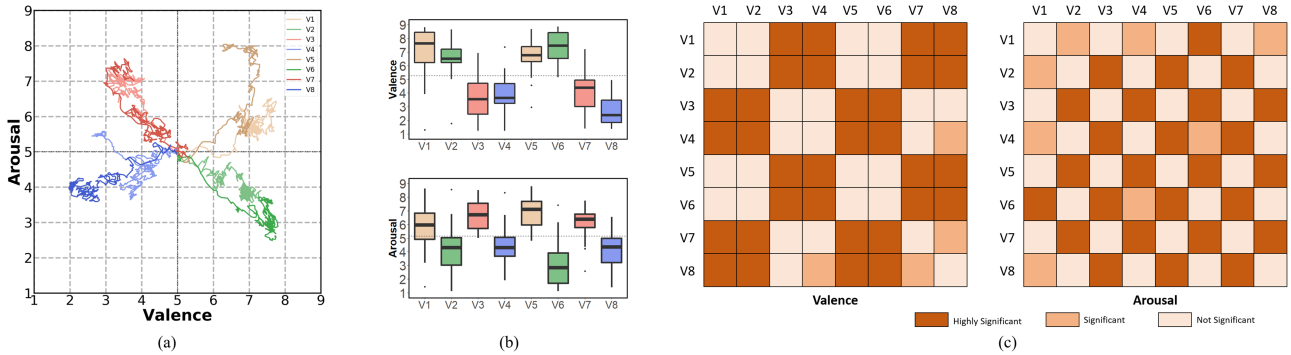
In a previous study [27], we conducted a controlled usability evaluation and found no significant differences between HaloLight and DotSize concerning motion sickness, presence or mental workload, and both techniques do not result in high sickness, workload, nor break presence. Thus in this section, we combined the collected annotations and behavioral data from HaloLight and DotSize and show the results of descriptive statistics.

[9][Online]. Available: https://www.vive.com/us/support/vive-pro-eye/category_howto/calibrating-eye-tracking.html

A. Continuous Annotation Analysis

We combined 32 participants' annotation data by calculating the mean V-A ratings at each frame for each video. The generated eight trajectories are shown in Fig. 4(a). It can be seen that the results of continuous annotations are consistent with the intended emotional experiences of the stimuli videos. Two videos pertaining to the same emotion type span the same quadrant, thus exhibiting agreement in subjective ratings. For different videos annotated across participants, 68.4% of annotation sequences appear in more than two emotion quadrants. The mean of difference between the maximum and minimum values from eight videos annotated by all participants for valence ranged from $[2.475, 6.157](M = 4.637, SD = 0.859)$, for arousal ranged from $[2.678, 6.532](M = 4.831, SD = 1.074)$, indicating that for certain video types, participants used a wide range for annotating, and were not limited to annotating one dimension only.

The mean V-A ratings across 32 participants for eight videos spanning four quadrants are shown in Fig. 4(b). We can find that the mean V-A ratings of eight videos are consistent with the video categories listed in Table I. For example, V1 and V5 belong to HVHA and the mean V-A ratings are >5. To further test the differences among these videos, we run inferential statistics. A Shapiro-Wilk test showed both the mean of valence and arousal are not normally distributed ($p < 0.05$). As we are comparing eight groups within-subjects, we performed a Friedman rank sum test on the mean of valence ($\chi^2(7) = 146.44, p < 0.01$) and then on the mean of arousal ($\chi^2(7) = 120.48, p < 0.01$). The results show significant effects of video emotions on V-A ratings. A post-hoc test using Bonferroni pairwise comparisons was performed to precisely determine whether the ratings of any two videos are similar or different [20], [42], where the results of these comparisons are presented in form of symmetric matrix plots in Fig. 4(b) and (c). Effect sizes for significant post-hoc pairwise comparisons between each video on valence ranged from $[0.943, 1.675]$, while for arousal ranged from $[0.815, 1.655]$. Most of the cases are in line with our expectations, with no significant differences ($p > 0.05$) among videos with the same emotion type, and high significant differences ($p < 0.001$) among videos with the opposite emotion type. However, in some cases, as reported in the literature [52], [53], this was not the case. We could find high significant differences ($p < 0.001$) between HV (V1, V2, V5 and V6) and LV (V3, V4, V7 and V8) videos. Beyond that, there are also significant differences ($0.001 < p < 0.05$) between V4 and V8, as well as V7 and V8, probably because V8 immersed users in the aftermath of the Nepal earthquake with lower valence value than others. For arousal ratings, there are high significant differences ($p < 0.001$) between HA (V3, V5 and V7) and LA (V2, V4, V6 and V8) videos. The significant differences ($0.001 < p < 0.05$) between V1, a cute puppy video with V2, V4 and V8 are not expected to be high. One reason is that more than 50% of participants during the interview said they liked dogs very much, so they were relatively relaxed while watching V1. In addition, V4 and V6 are also significantly different ($p < 0.05$), which may be due to the very low arousal value of V6, with the theme of Hawaiian sunrise.



(a) (b) (c)

Fig. 4. (a) Combined annotation trajectories for eight selected videos from 32 participants. (b) Boxplots for mean ratings of valence and arousal. (c) Pairwise comparisons of mean valence (left) and arousal (right) ratings across eight videos, with colors depicting different significance levels ($p < 0.001$, highly significant; $0.001 < p < 0.05$, significant; $p > 0.05$, not significant).

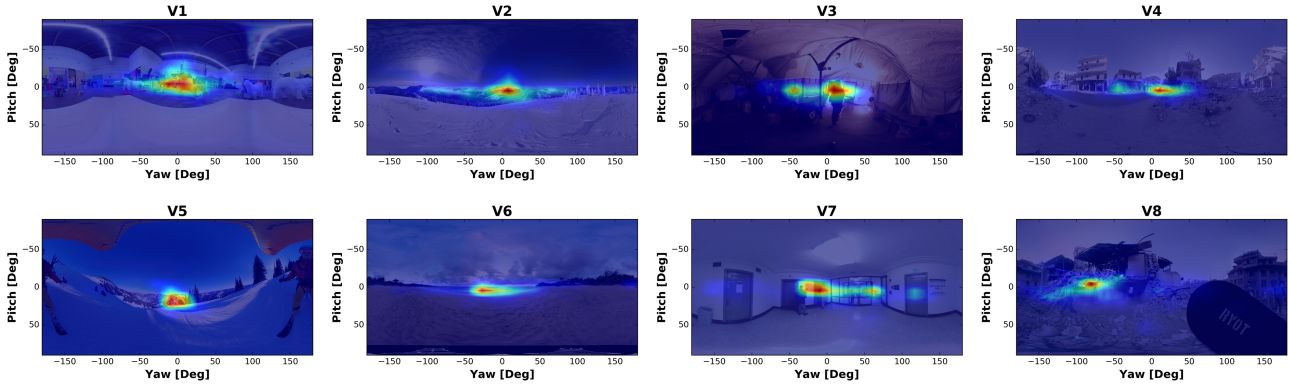


Fig. 5. A sample thumbnail frame with its saliency map for each video from 32 participants.

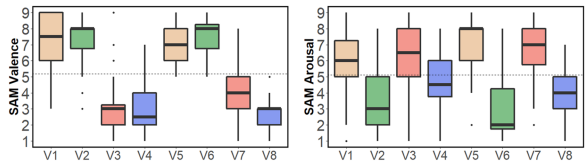


Fig. 6. Boxplots for SAM ratings of valence and arousal.

B. Within-VR SAM Analysis

We show the results of within-VR SAM rating in Fig. 6, which are consistent with expectations. By a two-way random, absolute agreement, average-measures ICC, the results show excellent reliability for the SAM valence ($ICC = 0.984, p < 0.05$) and arousal ($ICC = 0.951, p < 0.05$) ratings, indicating that the SAM valence and arousal were rated similarly across participants [54]. Moreover, to assess the agreement of the two rating methods (within-VR SAM rating and continuous annotation), we performed a two-way mixed, absolute agreement, average-measures ICC. The average resulting ICCs regarding the eight videos suggest excellent reliability for the valence score, total average $ICC = 0.882, p < 0.05$, and good reliability for the arousal score, total average $ICC = 0.714, p < 0.05$. Together they indicate that: (1) the within-VR SAM rating and the continuous annotation methods have a high degree of agreement and (2) valence and arousal are rated similarly across the two rating methods.

TABLE II
THE MEAN AND STANDARD DEVIATIONS VALUES OF CC FOR HM AND EM SALIENCY MAPS BETWEEN $Group1$ AND $Group2$ FOR EACH VIDEO

VID	CC (HM)	CC (EM)
V1	0.881 ± 0.016	0.913 ± 0.012
V2	0.843 ± 0.010	0.952 ± 0.042
V3	0.862 ± 0.047	0.956 ± 0.023
V4	0.917 ± 0.050	0.967 ± 0.032
V5	0.883 ± 0.064	0.971 ± 0.013
V6	0.915 ± 0.046	0.960 ± 0.025
V7	0.861 ± 0.064	0.970 ± 0.012
V8	0.854 ± 0.042	0.926 ± 0.021

C. HM and EM Data Analysis

We first analyze whether viewing behavior among participants is similar, which is an essential indicator of how robust our behavior data is [4]. To test the consistency among participants while watching 360° videos, we follow Qiao and Xu *et al.*'s work [35], [55] in our experiment. We divided participants into two groups $Group1$ and $Group2$ randomly and equally and then generated the HM and EM saliency maps of the two groups for each frame. Then Pearson's linear correlation coefficient (CC) score [56], [57] is calculated to evaluate similarity of saliency maps, which ranges from -1 (perfectly inversely correlated) to 1 (perfectly correlated). Mean and standard deviations of CC are reported in Table II, which show the correlations are sufficiently

high (> 0.8) across different videos. This indicates that the visual attention behavior are highly consistent among participants while watching the eight selected videos.

In Fig. 5, we show the EM saliency maps as equirectangular representations for each video, as obtained from all collected combined eye gaze sample points of 32 subjects, where the Y-axis refers to the pitch and the X-axis the yaw values. Much research [25], [37] has argued that there exists a strong equator and front bias for human attention while viewing 360° videos. In our study, the viewing directions of all participants were initialized at the center of the video. We can see from the Fig. 5 that most viewing attention falls into small regions in the front and center region of the equator. In addition, note that other than the center region, there still exists potential regions attracting human attention depending on the video content [55], [58]. In V3, zombies constantly appear from different places, while V7's perspective is to follow a prisoner's escape route. Thus the participants' long-term focus regions are not unique. For V8 we could find an obvious left bias, one plausible reason is that an embedded logo from the video creators is placed in the right-bottom corner.

D. PD Data Analysis

Prior work indicated that PD changes can be used as an indicator of arousal states [59], but also are largely affected by the lighting conditions [60]. Recently, Pfleging *et al.* [61] and Tarnowski *et al.* [62] modelled PD as the sum of two contributing factors: (1) PD given lighting conditions, (2) PD given experiences from task. In our study, since 360° videos were played around and near to the eyes, there was no light source except for the presentation of 360° videos. Thus for each participant p, PD values affected by video v are calculated from:

$$PD_{p,v} = PD_{p,average} - PD_{p,light} \tag{1}$$

$PD_{p,average}$ is the average PD of both eyes recorded for participant p, while $PD_{p,light}$ is the PD given luminance condition of video v. Following Tarnowski *et al.*'s work [62], we used linear regression method (coefficients k, b) to model the relationship between PD and luminance of video v for participant p:

$$\begin{bmatrix} PD_{p,1} \\ PD_{p,2} \\ \vdots \\ PD_{p,n} \end{bmatrix} = \begin{bmatrix} Light_{v,1} \\ Light_{v,2} \\ \vdots \\ Light_{v,n} \end{bmatrix} * \begin{bmatrix} k \\ b \end{bmatrix} \tag{2}$$

where PD is the average PD values and $Light$ is the luminance values calculated by the V component in the HSV color space for each frame from video v. Then, the estimated value of PD was calculated from:

$$PD_{p,est} = k_p * Light_v + b_p \tag{3}$$

The $PD_{p,est}$ is used to estimate the $PD_{p,light}$ in (1).

We calculated the mean and standard deviation of video affected PD values ($PD_{p,v}$) across each video, in which the Z-score standardization across each participant was performed to eliminate different inter-personal baselines. The results are presented in Fig. 7(left). According to a Shapiro-Wilk normality test, the gathered data was not normally distributed ($p < 0.001$). A Friedman rank sum test revealed a significant effect of video types on PD values ($\chi^2(7) = 155.98, p < 0.001$). Then we performed a post-hoc test using Mann-Whitney tests with Bonferroni correction and show the results in Fig. 7(right). The effect sizes for significant post-hoc pairwise comparisons between each video ranged from $[0.475, 0.859]$. The influence of arousal states on PD values are evident in our data. For instance, there are significant differences ($p < 0.05$) between HA videos (V1, V3, V7) and LA videos (V2, V4, V6) for $PD_{p,v}$ values. It is worth noting that the results of LVLA videos were not as low as expected, which indicates that compared with watching sad videos, the arousal is not as low as watching relaxed videos.

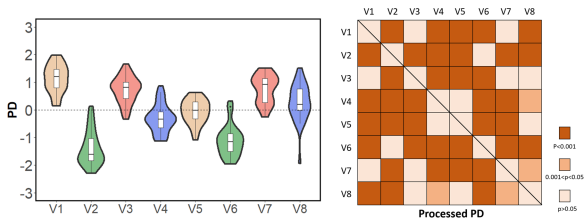


Fig. 7. Violin plot of the distribution for mean processed PD across eight videos (left). Pairwise comparisons of mean processed PD across eight videos, with colors depicting different significance levels ($p < 0.001$, highly significant; $0.001 < p < 0.05$, significant; $p > 0.05$, not significant) (right).

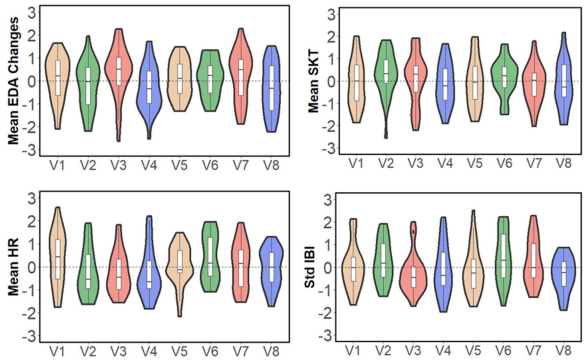


Fig. 8. Violin plot of the distribution for the physiological features across eight selected videos.

E. Physiological Data Analysis

We first normalized the values of each physiological signal after filtering out noise following previous work [63] for each participant viewing each video and then selected one predominantly used feature for each signal. The mean of the first-order differential of EDA signals during video playback was calculated as EDA changes, following previous research [11], [42]. We used the mean of SKT and HR values during each video to describe the time domain variation [63], as mean SKT and mean HR, respectively. The standard deviation of the duration of the detected inter-beat interval was acquired for each video as IBI changes [20]. Due to the lack of IBI data from $P2$ and $P12$, we removed the two subjects and performed Z-score standardization for other participants.

The violin plots in Fig. 8 report the distributions of the selected features across eight different videos. Similar to Sharma

TABLE III
COMPARISON OF THE PERFORMANCE USING RF CLASSIFIER, 1D-CNN, AND LSTM, FOR BOTH SD AND SI MODELS FOR 2S SEGMENT LENGTHS. ACC AND WEIGHTED-F1 SCORES ARE FOR BINARY V-A, 3-CLASS, AND 5-CLASS CLASSIFICATION. HIGHEST ACCURACY IS SHOWN IN BOLD

Evaluation	Classifier	Valence-2		Arousal-2		Valence-3		Arousal-3		5-class	
		acc	w-f1	acc	w-f1	acc	w-f1	acc	w-f1	acc	w-f1
SD (10-fold)	RF	68.45%	0.6315	71.33%	0.6487	**60.42%**	**0.5354**	**62.38%**	**0.5457**	**51.89%**	**0.4340**
	1D-CNN	**68.46%**	**0.5739**	71.37%	0.6121	51.17%	0.3867	56.85%	0.4502	40.49%	0.2828
	LSTM	65.51%	0.6013	**71.39%**	**0.6560**	52.91%	0.4755	56.36%	0.5002	44.48%	0.3735
SI (LOSOCV)	RF	**66.80%**	**0.6238**	64.26%	0.5298	**49.92%**	**0.4419**	**52.20%**	**0.4341**	**31.47%**	**0.3001**
	1D-CNN	64.27%	0.5828	**67.64%**	**0.5808**	45.51%	0.4191	47.17%	0.3923	29.87%	0.2598
	LSTM	65.00%	0.6349	66.30%	0.5934	44.62%	0.4269	43.79%	0.4085	30.12%	0.2768

TABLE IV
ABLATION STUDY ACROSS PHYSIOLOGICAL (EDA, IBI, HR, SKT, BVP) AND BEHAVIORAL SIGNALS (HM//EM) PLUS PD USING RF CLASSIFIER FOR BOTH SD AND SI MODELS UNDER 2S SEGMENTS. ACC AND WEIGHTED-F1 SCORES ARE FOR BINARY V-A, 3-CLASS AND 5-CLASS CLASSIFICATION. HIGHEST ACCURACY IS SHOWN IN BOLD

Evaluation	Component	Valence-2		Arousal-2		Valence-3		Arousal-3		5-class	
		acc	w-f1	acc	w-f1	acc	w-f1	acc	w-f1	acc	w-f1
SD (10-fold)	Physio	65.93%	0.6081	68.32%	0.6124	54.83%	0.4794	59.96% 2	0.5244	45.20%	0.3647
	HM/EM + PD	67.41%	0.6214	69.24%	0.6268	58.29%	0.5147	61.46%	0.5326	50.57%	0.4215
	Physio + HM/EM + PD	**68.45%**	**0.6315**	**71.33%**	**0.6487**	**60.42%**	**0.5354**	**62.38%**	**0.5457**	**51.89%**	**0.4340**
SI (LOSOCV)	Physio	65.69%	0.6126	62.17%	0.5007	44.66%	0.4027	51.45%	0.4200	30.56%	0.2439
	HM/EM + PD	62.68%	0.5344	62.43%	0.5022	47.90%	0.4124	50.00%	0.3555	26.33%	0.2440
	Physio + HM/EM + PD	**66.80%**	**0.6238**	**64.26%**	**0.5298**	**49.92%**	**0.4419**	**52.20%**	**0.4341**	**31.47%**	**0.3001**

et al.'s [20] results, we did not find significant differences for the selected physiological features across different types of videos. One potential reason is that the length of our video stimuli is restricted to one minute, which may be short in duration for clear effects of physiological signals.[10] One consideration is that it is difficult to perform standardized data analysis for videos with inconsistent lengths [31]. Furthermore, longer duration 360° videos can lead to higher motion sickness and workload [14], [64] which can also influence physiological markers, so there is a trade-off in what can be done. On the other hand, our findings from Fig. 8 show that some features can characterize a specific type of video. Prior work [65] indicated that EDA is known to be highly correlated with user arousal. In our work, V3 (V=3.20, A=5.60) and V7 (V=4.40, A=6.70) with high arousal labels result in higher values of EDA changes than other videos. For V4 (V=2.53, A=3.82) and V8 (V=2.73, A=3.80), the sad videos (LVLA), the values of all four features are lower than others. We provide these raw physiological time-series data in our dataset that change over time, in which the peaks and drops are associated with video events [66]. Our data can further help the community to study the relationship between physiological signals and 360° video content.

V. CLASSIFICATION EVALUATION

To further analyze the validity and reliability of our dataset, in this section we provide baseline classification experiments using common machine learning techniques.

A. Baseline Experiments

We draw on prior work [67], where we test three classification tasks on our dataset: (1) Binary classification for low / high levels of Valence and Arousal (V-A). (2) 3-class classification for low / neutral / high levels of V-A. (3) 5-class classification for the four quadrants of V-A space and neutral level. Mapping between continuous V-A values and discretized classes is listed in Table V.

TABLE V
THE MAPPING BETWEEN CONTINUOUS V-A RATINGS AND DISCRETIZED CLASSES.

Class	V-A Ratings (Binary)	V-A Ratings (3-Class)
Low	[1, 5)	[1, 3)
Neutral	-	[3, 6)
High	[5, 9]	[6, 9]
5-class	Valence Ratings	Arousal Ratings
High-High (HH)	[5, 9]	(5, 9]
High-Low (HL)	(5, 9]	[1, 5]
Low-Low (LL)	[1, 5]	[1, 5)
Low-High (LH)	[1, 5)	[5, 9]
Neutral	5	5

Both classic ML and DL methods are proposed to classify and predict the value of valence and arousal [67]. For ML methods, we tested the following: Support Vector Machine (SVM) [68], Random Forest (RF) [69], Gaussian Naive Bayes (GaussianNB) [70], and k-Nearest Neighbor (k-NN) [71]. For DL methods, we tested 1D-Convolutional Neural Network (1D-CNN) [72] and sequential learning approach, Long Short-Term Memory (LSTM) [73]. These are the two most basic and commonly used algorithms in affective computing [74]. Training and evaluation were run on an NVIDIA 2080Ti GPU server.

1) Feature and Model Selection: We first pre-processed HM/EM, PD and peripheral physiological signals (EDA, IBI, HR, SKT, BVP) and then segmented them into 2-s length (sample size: $32 \times 8 \times 30$ segments) for fine-grained emotion recognition, following prior work [67]. To train ML methods, we extracted mean, median, standard deviation for the pitch / yaw

[10][Online]. Available: https://support.empatica.com/hc/en-us/sections/200582445-E4-wristband-data

of HM/EM, PD and original, first and second differential of physiological signals, as well as fixation number, mean, median, standard deviation for fixation and saccade duration, and lastly saccade amplitude. These are widely used features for behavioral and physiological signals in the task of emotion recognition [28], [75], [76]. Aside from RF, we leave the default parameter settings for all classic ML classifiers. For subject-dependent (SD) models, we kept the default parameters ($max\_depth = 2$) for RF. However for subject-independent (SI) models, given that the amount and complexity of the training data are larger for SD models, we increase the maximum depth of the tree ($max\_depth = 4$) to better learn the latent representation.

For DL methods, we tested 1D-CNN and LSTM on the processed original data. The 1D-CNN model employs five 1D-CNN layers whose filter numbers n and sizes s, (n, s) are $(4, 64), (16, 32), (64, 16), (128, 8), (128, 32)$, respectively. All the five 1D-CNN layers are activated by a rectified linear activation function (ReLU). Then a 1D global max pooling layer is followed to select the most salient features from the 1D-CNN layers. A dense layer activated by the softmax function is put as the last layer for classification. The LSTM model consists of one LSTM layer with 100 units where we put the same dense layer as 1D-CNN for classification. The two models are built with keras and trained with *RMSprop* [77] optimizer.

2) Evaluation Metrics: We chose two widely used metrics in machine learning [78] to evaluate classification performance: (1) Accuracy (acc) for the percentage of correct predictions, (2) Weighted F1-score (w-f1) for the harmonic mean of precision and recall for each label. We trained and tested each classification method using both subject-dependent (SD) and subject-independent (SI) models. SD models were tested using 10-fold cross validation and SI models were tested using Leave-One-Subject-Out Cross Validation (LOSOCV). The results we show are the mean accuracy and w-f1 of each fold/subject used as testing data.

B. Results and Discussion

1) Classification Results: Among the classification methods using default architectures, we found that RF outperforms SVM, NB, and KNN methods, thus we only show RF results here and use these results for subsequent analysis. However, we include the results from the other classifiers in our dataset. We ran experiments to investigate model performance using RF, 1D-CNN and LSTM methods for both SD and SI models under 2s instances. As shown in Table III, the accuracies for 3-class classification are lower than binary classification but higher than 5-class classification. Given that many instances (43.32% for 3-class and 27.06% for 5-class) are classified as neutral, the data imbalance can pose problems when recognizing emotions using fine-grained emotion labels (cf., [67]).

Compared with SI models, SD models achieve higher accuracies and w-f1 scores, especially for 3-class and 5-class classification on our dataset. The comparable recognition accuracy of the two models demonstrates that the data volume from one user is sufficient to train a machine learning model for emotion recognition. This also lends support that the number of videos and video lengths we chose for an individual user are sufficient for running classification experiments. These results provide support that our models can generalize across behavioral and physiological data collected in 360° VR environments.

2) Ablation Study: To further analyze the effectiveness of single modality in our dataset, we conducted an ablation study to inspect the effects of: behavioral data (HM/EM and PD[11]) and physiological data (EDA, BVP, HR, and SKT). The results of binary classification evaluated using both SD and SI models are shown in Table IV. We found that only behavioral data or only physiological signals in our dataset can yield good recognition accuracies. Additionally, the accuracies from combining physiological signals with behavioral data are slightly higher than using single modality.

VI. Limitations

First, we are limited in the selection of 360° video stimuli. The different emotion types of videos used in our experiment have perceptual differences, for example color or camera movement, which could affect the user's viewing experience. Participants' personal preferences of the video content may also affect their emotional assessment [79]. However, due to the lack of publicly available 360° video databases with validated emotion labels, these could not be explored further in this work. Furthermore, the age of our participants ranges from 18-33 (M=25, SD=4.0), recruited from our institute or nearby institutes, which may not be well spread to other age groups like seniors. One consideration is that since users need to report their emotional states while watching 360° videos, we do not want age to be a dominant factor. However, it is interesting to consider other population groups which may have greater difficulty in reporting their emotion (e.g., Autism Spectrum Disorder [80]). Fourth, we did not collect EEG because collecting stable, high-quality EEG data is still a challenge [81], especially for immersive virtual environments where users wear an HMD [82]. Finally, the performance of the SI model can still be improved if data imbalance is addressed (e.g., through data synthesis using Generative Adversarial Networks [83]).

VII. Conclusion

The contributions of this paper focus on the provision of a public multi-modal 360° dataset and statistical analyses and baseline classification experiments. We first designed a protocol to collect fine-grained, continuous emotion labels of valence and arousal while users watching 360° videos in a VR setting. In our experiment with 32 participants viewing eight videos, we gathered continuous emotion annotation data, HM and EM behavior data, as well as PD and peripheral physiological data (EDA, IBI, HR, SKT, BVP).[12]

The primary insights of our analyses are: (1) Mean V-A ratings from our dataset are reasonably consistent with the intended attributes of the videos, and there exists high agreement between continuous ratings and post-stimuli SAM ratings, indicating the

[11] PD can be considered as a physiological response, however since we extract data directly from the HMD, we keep PD as part of the EM data.

[12] key steps in the stage of data acquisition and pre-processing are reported in our dataset.

reliability of our data following Sharma *et al.* [20]. (2) For all eight videos, we find high correlations on viewing behavior (HM and EM) among participants, and a center and front bias on saliency maps, in line with [55], [58]. (3) Similar to [60], [61], we found PD values are positively correlated with arousal levels, where the ambient light of the video has an impact on arousal. (4) Preliminary results for RF under 2s segments show good performance on our dataset, and an ablation study shows using only behavior data or only physiological signals can yield reasonable recognition accuracies, however using both modalities is better.

Furthermore, collecting continuous annotations can be used to evaluate the performance of fine-grained emotion recognition algorithms (e.g., weakly supervised learning or regression). As mentioned by Romeo *et al.* [84], the lack of continuous annotations is the reason why they failed to validate their weakly-supervised algorithm for fine-grained emotion recognition. Moreover, if only discrete annotations are available, ML algorithms can overfit because the discrete labels represent only the most salient or recent emotion rather than the dynamic emotional changes that may occur within video watching (cf., *peak-end theory* [85]). This can be reduced if training with continuous emotion labels, since the continuous labels allow the algorithms to learn the precise mappings between the dynamic emotional changes and input signals. To summarize, having continuous annotations becomes essential for developing and validating continuous or fine-grained emotion recognition algorithms.

Our future work comprises different facets: First, it is important to conduct further research on saliency models and attention using our dataset [35], [86], and explore how these correlate with continuous emotion labels. Second, our dataset could be used in various application scenarios such as helping 360° video makers to understand the emotion and behavior of people watching 360° videos. It can also serve as representative moment-by-moment ground truth for developing machine learning algorithms to automatically recognize the user's emotions in 360° VR environments. Third, our work leaves room for future research to design new methods to capture real-time emotions while watching volumetric videos. Fourth, we aim on further investigating automatic content analysis techniques to investigate further how e.g., HM and EM vary specifically with respect to content of video. To conclude, CEAP-360VR is the first public, multi-modal dataset with continuous emotion annotation data, behavior and physiological data, which can enable future research on emotion understanding and prediction within 360° VR environments.

VIII. Supplementary Material

Our dataset includes the raw and processed data from all 32 participants and eight selected videos, the processing and validation scripts, along with dataset description and key steps in the stage of data acquisition and pre-processing. All data were saved in JavaScript Object Notation (JSON) [87], a well-known file format that has native support by most programming languages. This makes the data accessible and easy to process. Also, the scripts to prepare data and features for running ML experiments are reported in our dataset.

The dataset and processing scripts are publicly available on GitHub (https://github.com/cwi-dis/CEAP-360VR-Dataset), under the following license: Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) license. Our dataset can be additionally retrieved on a dedicate webpage (https://www.dis.cwi.nl/ceap-360vr-dataset).

Acknowledgment

The authors would like to thank all participants of our annotation study and the main experiment, as well as our reviewers.

References

[1] J. A. Russell, "A circumplex model of affect," *J. Pers. Social Psychol.*, vol. 39, no. 6, pp. 1161–1178, 1980.
[2] D. Handayani, A. Wahab, and H. Yaacob, "Recognition of emotions in video clips: The self-assessment manikin validation," *Telkomnika*, vol. 13, no. 4, pp. 1343–1351, 2015.
[3] A. MacQuarrie and A. Steed, "Cinematic virtual reality: Evaluating the effect of display type on the viewing experience for panoramic video," in *Proc. IEEE Virtual Reality*, 2017, pp. 45–54.
[4] M. Xu, C. Li, S. Zhang, and P. L. Callet, "State-of-the-art in 360° video/image processing: Perception, assessment and compression," *IEEE J. Sel. Topics Signal Process.*, vol. 14, no. 1, pp. 5–26, Jan. 2020.
[5] M. Alcañiz, R. Baños, C. Botella, and B. Rey, "The EMMA project: Emotions as a determinant of presence," *PsychNol. J.*, vol. 1, no. 2, pp. 141–150, 2003.
[6] J. Marín-Morales *et al.*, "Affective computing in virtual reality: Emotion recognition from brain and heartbeat dynamics using wearable sensors," *Sci. Rep.*, vol. 8, no. 1, pp. 1–15, 2018.
[7] F. Assilmia, Y. S. Pai, K. Okawa, and K. Kunze, "IN360: A 360-degree-video platform to change students preconceived notions on their career," in *Proc. CHI Conf. Extended Abstr. Hum. Factors Comput. Syst.*, 2017, pp. 2359–2365.
[8] J. Beck, M. Rainoldi, and R. Egger, "Virtual reality in tourism: A state-of-the-art review," *Tourism Rev.*, pp. 586–612, 2019.
[9] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "AffectNet: A database for facial expression, valence, and arousal computing in the wild," *IEEE Trans. Affect. Comput.*, vol. 10, no. 1, pp. 18–31, Jan.–Mar. 2019.
[10] F. Nasoz, K. Alvarez, C. L. Lisetti, and N. Finkelstein, "Emotion recognition from physiological signals using wireless sensors for presence technologies," *Cogn., Technol. Work*, vol. 6, no. 1, pp. 4–14, 2004.
[11] J. Fleureau, P. Guillotel, and I. Orlac, "Affective benchmarking of movies based on the physiological responses of a real audience," in *Proc. Humaine Assoc. Conf. Affect. Comput. Intell. Interact.*, 2013, pp. 73–78.
[12] D. Egan, S. Brennan, J. Barrett, Y. Qiao, C. Timmerer, and N. Murray, "An evaluation of heart rate and electrodermal activity as an objective qoe evaluation method for immersive virtual reality environments," in *Proc. 8th Int. Conf. Qual. Multimedia Experience*, 2016, pp. 1–6.
[13] D. Liao *et al.*, "Design and evaluation of affective virtual reality system based on multimodal physiological signals and self-assessment manikin," *IEEE J. Electromagn., RF, Microw. Med. Biol.*, vol. 4, no. 3, pp. 216–224, Sep. 2020.
[14] B. J. Li, J. N. Bailenson, A. Pines, W. J. 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.*, vol. 8, pp. 1–10, 2017.
[15] T. Xue, A. E. Ali, G. Ding, and P. Cesar, "Investigating the relationship between momentary emotion self-reports and head and eye movements in HMD-based 360° VR video watching," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, New York, NY, USA: Association for Computing Machinery, 2021, pp. 1–6. [Online]. Available: https://doi.org/10.1145/3411763.3451627
[16] L. F. Barrett, R. Adolphs, S. Marsella, A. M. Martinez, and S. D. Pollak, "Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements," *Psychol. Sci. Public Int.*, vol. 20, no. 1, pp. 1–68, 2019, PMID: 31313636. [Online]. Available: https://doi.org/10.1177/1529100619832930

[17] M. M. Bradley and P. J. Lang, "Measuring emotion: The self-assessment manikin and the semantic differential," *J. Behav. Ther. Exp. Psychiatry*, vol. 25, no. 1, pp. 49–59, 1994.

[18] F. Nagel, R. Kopiez, O. Grewe, and E. Altenmüller, "EMuJoy: Software for continuous measurement of perceived emotions in music," *Behav. Res. Methods*, vol. 39, no. 2, pp. 283–290, 2007.

[19] M. Soleymani, S. Asghari-Esfeden, Y. Fu, and M. Pantic, "Analysis of EEG signals and facial expressions for continuous emotion detection," *IEEE Trans. Affect. Comput.*, vol. 7, no. 1, pp. 17–28, Jan.–Mar. 2015.

[20] K. Sharma, C. Castellini, E. L. van den Broek, A. Albu-Schaeffer, and F. Schwenker, "A dataset of continuous affect annotations and physiological signals for emotion analysis," *Sci. Data*, vol. 6, no. 1, pp. 1–13, 2019.

[21] J. M. Girard and A. G. Wright, "DARMA: Software for dual axis rating and media annotation," *Behav. Res. Methods*, vol. 50, no. 3, pp. 902–909, 2018.

[22] P. Lopes, G. N. Yannakakis, and A. Liapis, "RankTrace: Relative and unbounded affect annotation," in *Proc. 7th Int. Conf. Affect. Comput. Intell. Interact.*, 2017, pp. 158–163.

[23] A. Toet, F. Heijn, A.-M. Brouwer, T. Mioch, and J. B. van Erp, "The emojiGrid as an immersive self-report tool for the affective assessment of 360 VR videos," in *Proc. Int. Conf. Virtual Reality Augmented Reality*, Springer, 2019, pp. 330–335.

[24] X. Corbillon, F. De Simone, and G. Simon, "360-degree video head movement dataset," in *Proc. 8th ACM Multimedia Syst. Conf.*, 2017, pp. 199–204.

[25] E. J. David, J. Gutiérrez, A. Coutrot, M. P. D. Silva, and P. L. Callet, "A dataset of head and eye movements for 360° videos," in *Proc. 9th ACM Multimedia Syst. Conf.*, 2018, pp. 432–437.

[26] C. Li, M. Xu, X. Du, and Z. Wang, "Bridge the gap between VQA and human behavior on omnidirectional video: A large-scale dataset and a deep learning model," in *Proc. 26th ACM Int. Conf. Multimedia*, 2018, pp. 932–940.

[27] T. Xue, A. El Ali, T. Zhang, G. Ding, and P. Cesar, "RCEA-360VR: Real-time, continuous emotion annotation in 360° VR videos for collecting precise viewport-dependent ground truth labels," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, New York, NY, USA: Association for Computing Machinery, 2021. [Online]. Available: https://doi.org/10.1145/3411764.3445487

[28] W. Tang, S. Wu, T. Vigier, and M. P. D. Silva, "Influence of emotions on eye behavior in omnidirectional content," in *Proc. 12th Int. Conf. Qual. Multimedia Experience*, 2020, pp. 1–6.

[29] S. Moon and J. S. Lee, "Implicit analysis of perceptual multimedia experience based on physiological response," *IEEE Trans. Multimedia*, vol. 19, pp. 340–353, 2017.

[30] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, "A multimodal database for affect recognition and implicit tagging," *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 42–55, Jan.–Mar. 2012.

[31] S. Koelstra *et al.*, "DEAP: A database for emotion analysis; using physiological signals," *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 18–31, Jan.–Mar. 2012.

[32] M. K. Abadi, R. Subramanian, S. M. Kia, P. Avesani, I. Patras, and N. Sebe, "DECAF: MEG-based multimodal database for decoding affective physiological responses," *IEEE Trans. Affect. Comput.*, vol. 6, no. 3, pp. 209–222, Jul.–Sep. 2015.

[33] J. A. M. Correa, M. K. Abadi, N. Sebe, and I. Patras, "AMIGOS: A dataset for affect, personality and mood research on individuals and groups," *IEEE Trans. Affect. Comput.*, vol. 12, no. 2, pp. 479–493, Apr.–Jun. 2021.

[34] R. Subramanian, J. Wache, M. K. Abadi, R. L. Vieriu, S. Winkler, and N. Sebe, "ASCERTAIN: Emotion and personality recognition using commercial sensors," *IEEE Trans. Affect. Comput.*, vol. 9, no. 2, pp. 147–160, Apr.–Jun. 2018.

[35] M. Qiao, M. Xu, Z. Wang, and A. Borji, "Viewport-dependent saliency prediction in 360° video," *IEEE Trans. Multimedia*, vol. 23, pp. 748–760, 2021, doi: 10.1109/TMM.2020.2987682.

[36] C. Ozcinar and A. Smolic, "Visual attention in omnidirectional video for virtual reality applications," in *Proc. 10th Int. Conf. Qual. Multimedia Experience*, 2018, pp. 1–6.

[37] M. Xu, Y. Song, J. Wang, M. Qiao, L. Huo, and Z. Wang, "Predicting head movement in panoramic video: A deep reinforcement learning approach," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 11, pp. 2693–2708, Nov. 2019.

[38] Z. Zhang, Y. Xu, J. Yu, and S. Gao, "Saliency detection in 360° videos," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 488–503.

[39] Y. Xu *et al.*, "Gaze prediction in dynamic 360° immersive videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 5333–5342.

[40] A. Nguyen, Z. Yan, and K. Nahrstedt, "Your attention is unique: Detecting 360-degree video saliency in head-mounted display for head movement prediction," in *Proc. 26th ACM Int. Conf. Multimedia*, 2018, pp. 1190–1198.

[41] J. Marín-Morales, C. Llinares, J. Guixeres, and M. Alcañiz, "Emotion recognition in immersive virtual reality: From statistics to affective computing," *Sensors*, vol. 20, no. 18, pp. 1–25, 2020.

[42] T. Zhang, A. El Ali, C. Wang, A. Hanjalic, and P. Cesar, "RCEA: Real-time, continuous emotion annotation for collecting precise mobile video ground truth labels," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, 2020, pp. 1–15.

[43] S. Putze, D. Alexandrovsky, F. Putze, S. Höffner, J. D. Smeddinck, and R. Malaka, "Breaking the experience: Effects of questionnaires in VR user studies," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, 2020, pp. 1–15.

[44] N. Milstein and I. Gordon, "Validating measures of electrodermal activity and heart rate variability derived from the Empatica E4 utilized in research settings that involve interactive dyadic states," *Front. Behav. Neurosci.*, pp. 1–13, 2020.

[45] 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. aviation Psychol.*, vol. 3, no. 3, pp. 203–220, 1993.

[46] T. Schubert, F. Friedmann, and H. Regenbrecht, "The experience of presence: Factor analytic insights," *Presence: Teleoperators Virtual Environ.*, vol. 10, no. 3, pp. 266–281, 2001.

[47] S. G. Hart, "Nasa-task load index (NASA-TLX); 20 years later," in *Proc. Hum. Factors Ergonom. Soc. Annu. Meeting*, Los Angeles, CA, USA: Sage Publications Sage CA, 2006, vol. 50, no. 9, pp. 904–908.

[48] ITU-T Recommendation, "Subjective video quality assessment methods for multimedia applications," pp. 4–5, 1999.

[49] T. Xue, S. Ghosh, G. Ding, A. El Ali, and P. Cesar, "Designing real-time, continuous emotion annotation techniques for 360° VR videos," in *Proc. Extended Abstr. CHI Conf. Hum. Factors Comput. Syst. Extended Abstr.*, 2020, pp. 1–9.

[50] R. F. Gunst and R. L. Mason, "Fractional factorial design," *Wiley Interdiscipl. Reviews: Comput. Statist.*, vol. 1, no. 2, pp. 234–244, 2009.

[51] A. Lutz, J. Brefczynski-Lewis, T. Johnstone, and R. J. Davidson, "Regulation of the neural circuitry of emotion by compassion meditation: Effects of meditative expertise," *PLoS One*, vol. 3, no. 3, pp. 1–10, 2008.

[52] E. L. Van Den *et al.*, "Affective man-machine interface: Unveiling human emotions through biosignals," in *Proc. Int. Joint Conf. Biomed. Eng. Syst. Technol.*, Springer, 2009, pp. 21–47.

[53] E. van den Broek, "Affective signal processing (ASP): Unraveling the mystery of emotions," Ph.D. dissertation, Univ. Twente, 2011, doi: 10.3990/1.9789036532433.

[54] D. V. Cicchetti, "Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology," *Psychol. Assessment*, vol. 6, no. 4, pp. 284–290, 1994.

[55] M. Xu, C. Li, Z. Chen, Z. Wang, and Z. Guan, "Assessing visual quality of omnidirectional videos," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 12, pp. 3516–3530, Dec. 2019.

[56] J. Li, C. Xia, Y. Song, S. Fang, and X. Chen, "A data-driven metric for comprehensive evaluation of saliency models," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 190–198.

[57] Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, and F. Durand, "What do different evaluation metrics tell us about saliency models?" *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 3, pp. 740–757, Mar. 2019.

[58] V. Sitzmann *et al.*, "Saliency in VR: How do people explore virtual environments?," *IEEE Trans. Visual. Comput. Graph.*, vol. 24, no. 4, pp. 1633–1642, Apr. 2018.

[59] M. M. Bradley, L. Miccoli, M. A. Escrig, and P. J. Lang, "The pupil as a measure of emotional arousal and autonomic activation," *Psychophysiology*, vol. 45, no. 4, pp. 602–607, 2008.

[60] Z. Zhu, K. Fujimura, and Q. Ji, "Real-time eye detection and tracking under various light conditions," in *Proc. Symp. Eye Tracking Res. Appl.*, 2002, pp. 139–144.

[61] B. Pfleging, D. K. Fekety, A. Schmidt, and A. L. Kun, "A model relating pupil diameter to mental workload and lighting conditions," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, 2016, pp. 5776–5788.

[62] P. Tarnowski, M. Kołodziej, A. Majkowski, and R. J. Rak, "Eye-tracking analysis for emotion recognition," *Comput. Intell. Neurosci.*, vol. 2020, pp. 2909267–2909267, 2020.

[63] B. Zhao, Z. Wang, Z. Yu, and B. Guo, "EmotionSense: Emotion recognition based on wearable wristband," in *Proc. IEEE SmartWorld, Ubiquitous Intell. Comput., Adv. Trusted Comput., Scalable Comput. Commun., Cloud Big Data Comput., Internet People Smart City Innov.*, 2018, pp. 346–355.

[64] M. V. d. Broeck, F. Kawsar, and J. Schöning, "It's all around you: Exploring 360° video viewing experiences on mobile devices," in *Proc. 25th ACM Int. Conf. Multimedia*, 2017, pp. 762–768.
[65] W. Boucsein, *Electrodermal Activity*. Berlin, Germany: Springer, 2012.
[66] D. R. Bach, G. Flandin, K. J. Friston, and R. J. Dolan, "Modelling event-related skin conductance responses," *Int. J. Psychophysiol.*, vol. 75, no. 3, pp. 349–356, 2010.
[67] T. Zhang, A. El Ali, C. Wang, A. Hanjalic, and P. Cesar, "CorrNet: Fine-grained emotion recognition for video watching using wearable physiological sensors," *Sensors*, vol. 21, no. 1, pp. 1–25, 2021.
[68] C. He, Y.-j. Yao, and X.-s. Ye, "An emotion recognition system based on physiological signals obtained by wearable sensors," *Wearable Sensors and Robots*. Berlin, Germany: Springer, 2017, pp. 15–25.
[69] G. Rigas, C. D. Katsis, G. Ganiatsas, and D. I. Fotiadis, "A user independent, biosignal based, emotion recognition method," in *Proc. Int. Conf. User Model.*, Springer, 2007, pp. 314–318.
[70] D. S. Wickramasuriya and R. T. Faghih, "Online and offline anger detection via electromyography analysis," in *Proc. IEEE Healthcare Innov. Point Care Technol.*, 2017, pp. 52–55.
[71] L. Chen, M. Li, W. Su, M. Wu, K. Hirota, and W. Pedrycz, "Adaptive feature selection-based AdaBoost-KNN with direct optimization for dynamic emotion recognition in human-robot interaction," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 5, no. 2, pp. 205–213, Apr. 2021.
[72] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Netw.*, vol. 61, pp. 85–117, 2015.
[73] J. Ma, H. Tang, W.-L. Zheng, and B.-L. Lu, "Emotion recognition using multimodal residual LSTM network," in *Proc. 27th ACM Int. Conf. Multimedia*, 2019, pp. 176–183.
[74] S. Zhao, S. Wang, M. Soleymani, D. Joshi, and Q. Ji, "Affective computing for large-scale heterogeneous multimedia data: A survey," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 15, no. 3s, pp. 1–32, 2019.
[75] L. Shu *et al.*, "A review of emotion recognition using physiological signals," *Sensors*, vol. 18, no. 7, pp. 1–10, 2018.
[76] O. Kardan, M. G. Berman, G. Yourganov, J. Schmidt, and J. M. Henderson, "Classifying mental states from eye movements during scene viewing," *J. Exp. Psychol. Hum. Percep. Perform.*, vol. 41, no. 6, pp. 1502–1514, 2015.
[77] F. Zou, L. Shen, Z. Jie, W. Zhang, and W. Liu, "A sufficient condition for convergences of Adam and RMSProp," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 11 127–11135.
[78] M. Fatourechi, R. K. Ward, S. G. Mason, J. Huggins, A. Schlögl, and G. E. Birch, "Comparison of evaluation metrics in classification applications with imbalanced datasets," in *Proc. 7th Int. Conf. Mach. Learn. Appl.*, 2008, pp. 777–782.
[79] D. Z. Rodríguez, R. L. Rosa, E. A. Costa, J. Abrahão, and G. Bressan, "Video quality assessment in video streaming services considering user preference for video content," *IEEE Trans. Consum. Electron.*, vol. 60, no. 3, pp. 436–444, Aug. 2014.
[80] M. Maskey, F. Warnell, J. R. Parr, A. L. Couteur, and H. McConachie, "Emotional and behavioural problems in children with autism spectrum disorder," *J. Autism Devlop. Disord.*, vol. 43, no. 4, pp. 851–859, 2013.
[81] J. Birjandtalab, D. Cogan, M. B. Pouyan, and M. Nourani, "A non-EEG biosignals dataset for assessment and visualization of neurological status," in *Proc. IEEE Int. Workshop Signal Process. Syst.*, 2016, pp. 110–114.
[82] L. He, H. Li, T. Xue, D. Sun, S. Zhu, and G. Ding, "Am I in the theater? Usability study of live performance based virtual reality," in *Proc. 24th ACM Symp. Virtual Reality Softw. Technol.*, 2018, pp. 1–11.
[83] M. Mirza and S. Osindero, "Conditional generative adversarial nets," 2014, *arXiv:1411.1784*.
[84] L. Romeo, A. Cavallo, L. Pepa, N. Berthouze, and M. Pontil, "Multiple instance learning for emotion recognition using physiological signals," *IEEE Trans. Affect. Comput.*, to be published, doi: 10.1109/TAFFC.2019.2954118.
[85] B. L. Fredrickson and D. Kahneman, "Duration neglect in retrospective evaluations of affective episodes." *J. Pers. Social Psychol.*, vol. 65, no. 1, p. 45, 1993.
[86] J. Wang, M. Xu, L. Jiang, and Y. Song, "Attention-based deep reinforcement learning for virtual cinematography of 360° videos," *IEEE Trans. Multimedia*, vol. 23, pp. 3227–3238, 2021, doi: 10.1109/TMM.2020.3021984.
[87] E. T. Bray, "The Javascript object notation (JSON) data interchange format," RFC 7159, Mar. 2014.

Tong Xue (Student Member, IEEE) received the B.E. degree from the Communication University of China, Beijing, China, in 2016. She is currently working toward the Ph.D. degree with the School of Computer Science and Technology, Beijing Institute of Technology, Beijing, China. She is a joint Ph.D. student with Distributed and Interactive Systems, Centrum Wiskunde & Informatica, Amsterdam, The Netherlands. Her research interests include human-computer interaction and affective computing.

Abdallah El Ali (Member, IEEE) received the Ph.D. degree from the University of Amsterdam, Amsterdam, The Netherlands, in 2013. He is currently a tenure-track Researcher with Centrum Wiskunde & Informatica, Amsterdam, The Netherlands, with Distributed & Interactive Systems (DIS) Group. He is leading human-computer interaction (HCI) research with Affective Interactive Systems Research Area. His interests include ground truth label acquisition techniques, affective state visualization across environments (mobile, wearable, XR), and bio-responsive interactive prototypes.

Tianyi Zhang (Member, IEEE) is currently working toward the Ph.D. degree with the Faculty of Electrical Engineering, Mathematics & Computer Science, Delft University of Technology, Delft, The Netherlands. He is associated with Distributed & Interactive Systems (DIS) Group, Centrum Wiskunde & Informatica, Amsterdam, The Netherlands, The National Research Institute for Mathematics and Computer Science, The Netherlands. His research interests include human-computer interaction and machine learning based affective computing.

Gangyi Ding (Member, IEEE) received the B.E. degree from Peking University, Beijing, China, in 1988 and the Ph.D. degree from the Beijing Institute of Technology, Beijing, China, in 1993. He is currently a Professor with the School of Computer Science and Technology, Beijing Institute of Technology. In 1993, he joined the Faculty of the Beijing Institute of Technology. His research interests include computer simulation, software engineering, and digital performance.

Pablo Cesar (Senior Member, IEEE) leads the Distributed and Interactive Systems Group, Centrum Wiskunde & Infomartica (CWI), Amsterdam, The Netherlands, and is a Professor with the Delft University of Technology, Delft, The Netherlands. His research interests include human-computer interaction and multimedia systems, and focuses on modeling and controlling complex collections of media objects, including real-time media and sensor data that are distributed in time and space. He was recently the recipient of the Prestigious 2020 Netherlands Prize for ICT Research, because of his work on human-centered multimedia systems. He is also the Principal Investigator from CWI on a number of projects on social virtual reality and affective computing. He is a member of the Editorial Board of the IEEE MULTIMEDIA, *ACM Transactions on Multimedia*, and IEEE TRANSACTIONS OF MULTIMEDIA, among others. He has acted as an Invited Expert at the European Commission's Future Media Internet Architecture Think Tank.