RCEA-360VR: Real-time, Continuous Emotion Annotation in 360° VR Videos for Collecting Precise Viewport-dependent Ground Truth Labels

TONG XUE, Beijing Institute of Technology Centrum Wiskunde & Informatica, China
ABDALLAH EL ALI, Centrum Wiskunde & Informatica, The Netherlands
TIANYI ZHANG, Centrum Wiskunde & Informatica Delft University of Technology, The Netherlands
GANGYI DING, Beijing Institute of Technology, China
PABLO CESAR, Centrum Wiskunde & Informatica Delft University of Technology, China

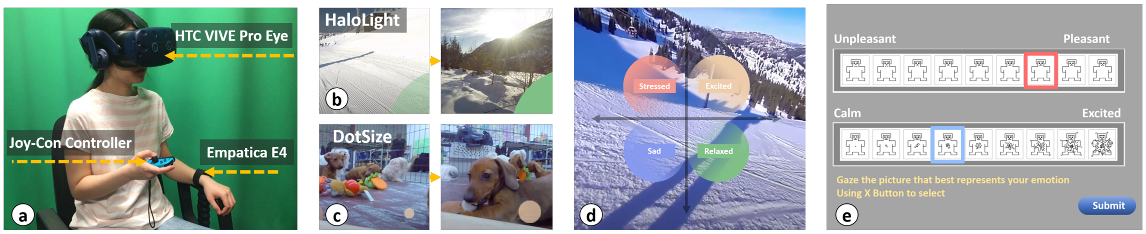

Fig. 1. (a) RCEA-360VR system components. (b) HaloLight: shaded halo arc in bottom-right viewport. (c) DotSize: circle dot in bottom-right viewport. (d) A screen-shot of helper function. (e) Within-VR SAM Rating Panel.

Precise emotion ground truth labels for 360° virtual reality (VR) video watching are essential for fine-grained predictions under varying viewing behavior. However, current annotation techniques either rely on post-stimulus discrete self-reports, or real-time, continuous emotion annotations (RCEA) but only for desktop and mobile settings. We present RCEA for 360° VR videos (RCEA-360VR), where we evaluate in a controlled study (N=32) the usability of two peripheral visualization techniques: HaloLight and DotSize. We furthermore develop a method that considers head movements when fusing labels. Using physiological, behavioral, and subjective measures, we show that (1) both techniques do not increase users' workload, sickness, nor break presence (2) our continuous valence and arousal annotations are consistent with discrete within-VR and original stimuli ratings (3) users exhibit high similarity in viewing behavior, where fused ratings perfectly align with intended labels. Our work contributes usable and effective techniques for collecting fine-grained viewport-dependent emotion labels in 360°VR.

CCS Concepts: • **Human-centered computing** → **Human computer interaction (HCI)**; • **Virtual Reality**;

Additional Key Words and Phrases: Emotion, annotation, 360° video, viewport-dependent, real-time, continuous

ACM Reference Format:
Tong Xue, Abdallah El Ali, Tianyi Zhang, Gangyi Ding, and Pablo Cesar. 2021. RCEA-360VR: Real-time, Continuous Emotion Annotation in 360° VR Videos for Collecting Precise Viewport-dependent Ground Truth Labels. In *CHI Conference on Human Factors in Computing Systems (CHI '21), May 8–13, 2021, Yokohama, Japan.* ACM, New York, NY, USA, 23 pages. https://doi.org/10.1145/3411764.3445487

1 INTRODUCTION

Watching 360° videos using head-mounted displays (HMDs) can provide interactive and immersive Virtual Reality (VR) experiences. Unlike desktop or mobile videos, 360° videos viewed through HMDs allow users to freely rotate their heads and focus on a portion of the scene [67]. Within such experiences, several works have established that such immersive VR environments have the capacity to evoke a wide range of emotions in humans [28, 31, 76, 78], and through sensing of physiological and behavioral markers (e.g., brain and heartbeat dynamics), can enable automatic emotion recognition of valence and arousal during such experiences [68]. Whether the goal is to induce, track, or recognize emotion for educational purposes [1], embodied virtual tourism [7], news engagement [104, 106], or develop emotion recognition and adaptive systems [68] within immersive VR experiencess, it is important to collect accurate and precise ground truth emotion labels. However, collecting emotional responses to 360° VR videos can be time consuming, demand considerable cognitive effort and interpretation [103], or carried out outside the VR experience (cf., [18, 76]) which may break the sense of immersion and presence [54, 87]. Furthermore, by allowing users to dynamically adjust their viewport freely and construct their own viewing experience [67], we can no longer be sure the annotations pertain to that specific scene at any given point in time. This necessitates the development and evaluation of new tools for continuous annotation of affective reactions of users while they watch 360° videos, whereby such viewport-dependent annotations can only be generated in such a setting, so must be provided in real-time.

Typically emotion data collection takes place via post-interaction or post-stimuli self-reports of valence and arousal (cf., [84]), which are retrospective and discrete in nature (e.g., Self-Assessment Manikin (SAM) [10]). However, such self-reports are temporally imprecise, especially for video content, since one can experience multiple emotions throughout [73, 94, 114] (e.g., experiencing >1 emotion when entire video is labeled 'happy'). Moreover, retrospective evaluations rely on episodic memory (cf., self-report construal in HCI [30]), which can introduce episodic memory biases (e.g., peak-and-end effects) [21]. While there have been several works on real-time and continuous emotion annotation, however only for desktop (e.g., CASE [89]) or mobile contexts (e.g., RCEA [114]). For immersive VR experiences, Xue et al. [112] designed two peripheral visualization techniques for continuous annotation: HaloLight and DotSize. However they did not perform any usability tests, nor study their effectiveness in producing meaningful precise ground truth labels considering users' changing head movement behavior, and their consistency with within-VR retrospective emotion (e.g., SAM) ratings. This necessitates the need for creating precise viewport-dependent ground truth labels by leveraging head movement patterns in 360° video watching.

This paper presents the **Real-time and Continuous Emotion Annotation for 360° VR (RCEA-360VR)** system. We ask: **(RQ1a)** How does RCEA-360VR usability compare with discrete and retrospective emotion assessment methods (within-VR emotion ratings), specifically with respect to mental workload, motion sickness, and presence? **(RQ1b)** Which of RCEA-360VR's peripheral visual feedback techniques, **HaloLight** and **DotSize**, provides better usability and user experience? We conducted a controlled, indoor experiment (N=32) (Figure 1(a)) and compared mental workload, presence, and motion sickness between HaloLight, DotSize, and discrete emotion assessments (within-VR SAM ratings) by measuring subjective and physiological measures[1]. To verify if RCEA-360VR's annotations are effective, we ask: **(RQ2)** How can we build precise emotion ground truth labels using RCEA-360VR considering user head movement behavior? We develop a method that considers head movements when fusing labels. It comprises three steps: continuous annotation time-alignment, segment-based viewport clustering, and lastly viewport-dependent annotation fusion.

[1] Raw data, processing scripts, and basic analyses of user physiological and behavioral data will be made publicly available in a separate, dataset paper.

Our exploratory work offers two primary contributions: **(1)** We evaluate RCEA-360VR using subjective and physiological measures, and show that its two peripheral visualization methods, HaloLight and DotSize, are both usable for collecting precise labels while users are immersed in VR. In other words, such techniques are suitable for collecting fine-grained emotion annotations of valence and arousal in real-time while users are watching 360° videos. **(2)** We contribute a method (continuous annotation time-alignment, segment-based viewport clustering, viewport-dependent annotation fusion) with associated algorithms that enables researchers to aggregate continuous ratings while considering varying head movement behavior during 360° video watching. Our method can be used to build accurate and precise ground truth emotion labels by combining viewing traces with annotation fusion methods. For human-computer interaction and emotion computing researchers, this provides greater insight into the temporal nature of reported humans emotion during immersive, viewport-dependent viewing experiences. At the same time, it enables machine learning researchers and practitioners to construct more temporally precise labels for training emotion recognition systems. Below, we start with a survey of related work.

2 RELATED WORK

Several research strands influenced our approach (emotion annotation, VR-based annotation techniques, and viewing behavior in 360° videos), which we describe below.

2.1 Discrete vs. Continuous Emotion Annotation Techniques

Given our task of simultaneously watching 360° videos using HMDs and annotating in real-time continuously, we follow prior work on continuous annotation [23, 34, 89]. Here, we draw on Russell's Circumplex model [84] using the two dimensions of valence and arousal (V-A) and to capture the finer granularity of emotion annotations through the user's immersive experience. Emotion assessments however are typically obtained through post-stimuli measurement instruments. For example, the Self-Assessment Manikin (SAM) [10], Pick-A-Mood (PAM)[27] and AffectButton [14] tools allow users to give detailed emotional feedback about their feelings after experiencing stimuli. However these post-stimulus, discrete annotation techniques cannot capture the temporal nature of emotions that can occur within temporal media (e.g., 360° video) [94, 114]. This led researchers to develop real-time, continuous emotion annotation techniques to obtain finer-grained emotion ground truth labels. With a computer-mouse interface, the FEELCARE [23], EmuJoy [73] and Gtrace [24] software packages require users to annotate emotions in a two-dimensional space by clicking the mouse button continuously, which increases users' physical and cognitive load [72, 113].

Several researchers consider the usage of auxiliary devices to lower mental workload while annotating. Girard et al. [33] developed CARMA which provides users a one-dimensional emotion slider to report basic emotion (positive or negative) by pushing it up and down, and the RankTrace tool [62] was implemented by a physical radial controller to specify a single, continuous dimension such as emotional intensity. DARMA [34] and CASE [89] enable users to input their emotions using a joystick in the V-A space, and display annotation feedback on a coordinate system that is either next to the video player, or superimposed in the upper-right corner of the player. More recently, Zhang et al. [114] designed the RCEA method for mobile settings, where users use a virtual joystick to annotate emotions in real time while watching videos on mobile device. Given the small screen display and distracting nature of mobile environments, they leverage peripheral visual feedback to show emotion states, which further motivates our approach of drawing on users' peripheral visual attention [3, 71]. While some of the foregoing tools allow real-time and continuous annotation, there are currently no such tools developed for VR environments.

2.2 Emotion Annotation in Virtual Reality

During VR experiences, the users' field of view is commonly constrained by the HMD. As Putze et al. stated [80], administering questionnaires in VR is becoming more common, which can ease participation, reduce the Break in Presence (BIP) and avoid biases. Toet et al. introduced the EmojiGrid [102], a smiley grid for emotion assessment in the virtual scenarios. Krüger et al. [54] proposed Morph A Mood (MAM), that provides a set of 3D characters with facial expressions for users to choose, aiming to be more intuitive. Both self-report techniques however occur after the experience. Voigt-Antons et al. [105] designed a stationary V-A grid interface in VR, with the background of 360° video, and users evaluated each video by clicking on a point in the grid continuously. However, they do not address the usability of this technique, nor how to fuse the resulting annotations. Moreover, it appears likely that a static 2D grid superimposed on the video can pose distractions. To address this, Xue et al. [112] considered peripheral visualization techniques to minimize workload and distraction, where they propose the design of HaloLight and DotSize for use in VR. While both techniques aimed at unobtrusively presenting emotion state on the users' periphery while immersed in VR, it is still unknown how usable and effective such techniques are. In this work, we provide a systematic usability evaluation of the RCEA-360VR system and associated peripheral visualization techniques, and provide a comparison with discrete and retrospective within-VR emotion rating methods.

2.3 Head Movements and Viewport-based Clustering of 360° Viewing Behavior

Unlike 2D videos, users can direct their field of view to any part of the scene while watching 360° videos. Thus it is important to understand how users observe and explore VR content [83]. Marmitt et al. [70] conducted a precursory study to analyze visual scanpaths in VR settings, and found that Head Movement (HM) and Eye Movement (EM) data are commonly used to analyze 360° viewing behavior. Wu et al. [109] established a head tracking dataset using HTC Vive across various categories of 360° videos and found that users share common patterns while watching VR videos. David et al. [25] presented a dataset with HM and EM data from 57 participants watching ten 360° videos, and provided guidance on how to generate saliency map and scanpaths from raw behavior data. Xu et al. [111] investigated users' viewing behavior and linked it with evaluation of visual quality of 360° videos. They found a high consistency in viewing direction among subjects, and that users' attention highly correlates with video content. Furthermore, Rossi et al. [81] proposed a graph-based method to identify clusters of users who are attending to the same portion of spherical content, and Nasrabadi et al. [74] presented a viewport-based prediction method based on clustering. While these works aim to model users' viewing behavior and predict visual attention, they do not consider users' annotated emotions with varying viewports. The novelty in our contribution is the viewport-dependent emotion annotation fusion method.

3 EVALUATING USABILITY OF RCEA-360VR

To answer (**RQ1a**) and (**RQ1b**), we evaluate the potential of real-time, continuous emotion annotation for 360° VR videos **(RCEA-360VR)**. Specifically, we conducted a controlled, indoor laboratory experiment (N=32) (Figure 1a) and compared mental workload, presence, and motion sickness between HaloLight, DotSize, and discrete emotion assessments (within-VR SAM ratings) across physiological and subjective measures. Below we describe our study design, usability and annotation consistency results, and discuss how they feed into our viewport-dependent fusion method.

Table 1. Description of 360° videos. Type code: H (high), L (low), Valence (V), Arousal (A).

Video	Type	OriginalDB Mean (V, A)	AnnotationStudy Mean (V, A)	Name	YoutubeID	Start Offset	Description
V0	Training	(6.36, 5.93)	/	NASA - Encapsulation & Launch...	D7-AmamuJEA	7s	Documentary film rocket launches
V1	HVHA	(7.47, 5.35)	(7.08, 6.08)	Puppies host SourceFed for a day	c7sA3EdXSUQ	0s	Viewers get up close with some puppies
V5	HVHA	(6.75, 7.42)	(6.83, 7.42)	Speed Flying	g6w6xkQeSHg	0s	Viewer follows a speed wing pilot
V3	LVHA	(3.20, 5.60)	(2.58, 6.83)	Zombie Apocalypse Horror	pHX3U4B6BCk	65s	Action film on soldiers and zombie attack
V7	LVHA	(4.40, 6.70)	(4.42, 7.17)	Jailbreak 360	vNLDRSdAj1U	127s	Action film depicting closed-circuit jailbreak scene
V2	HVLA	(6.13, 1.80)	(8.08, 1.91)	Mountain Stillness	aePXpV8Z10Y	10s	Atmospheric shots of Canadian snowy mountains
V6	HVLA	(6.57, 1.57)	(7.67, 1.50)	Malaekahana Sunrise	-bIrUYM-GjU	0s	Sun rising over the horizon at a beach
V4	LVLA	(2.53, 3.82)	(2.42, 4.17)	War Zone	Nxxb_7wzvJI	3s	Journalistic clip of a war torn city
V8	LVLA	(2.73, 3.80)	(2.17, 3.17)	The Nepal Earthquake Aftermath	5tasUGQ1898	41s	Short film on effects of an earthquake in Nepal

3.1 HaloLight and DotSize Techniques

Following a user-centric approach [75] with iterative design rounds based on an expert co-design session [112], we designed in earlier work two techniques, HaloLight and DotSize (as shown in Figure 1b&c). These were deemed suitable to indicate annotation state feedback. These visualization techniques are based on three design principles: **P1** - Design for HMD-based 360° VR video, **P2** - Design for input device ergonomics, and **P3** - Design for divided attention. These served as heuristics to narrow down the design space, and based on VR HMD-based interaction design guidelines [47]. Both techniques leverage joystick-based input (cf., Sec 3.2.4), where visual feedback is presented in the periphery of users' visual attention, fixed to the bottom right corner of the HMD viewport. Design attributes including position, size and transparency [43, 59] were considered.

For annotating emotions, we used the 2D V-A model based on Russell's Circumplex model [84]. Each quadrant in our 2D model (Figure 1d) has a distinct color, and represents emotion keywords such as excited, sad, etc. These four colors (HEX = #eecdac, #7fc087, #879af0, #f4978e for quadrants one to four, respectively) are used to provide peripheral feedback to users on which emotion quadrant they are currently annotating [41] in while watching a 360° video. Colors were selected based on a simplified version of Itten's color system [96], which has been shown to be intuitive and easy to understand [41]. Whereas HaloLight uses color opacity to indicate emotion intensity, DotSize uses the size of the filled circle to indicate intensity. How each technique works is shown as a video in **Supplementary Material**.

3.2 Study Design

Drawing on the Circumplex model [84, 90] of emotion (Figure 1d), there are four types of videos shown depending on V-A video ratings. These are: high valence / high arousal (HVHA), high valence / low arousal (HVLA), low valence / low arousal (LVLA), low valence / high arousal (LVHA). Our experiment is a 2 (Annotation Method: HaloLight vs DotSize) x 4 (IV2: Video Emotion: HVHA, HVLA, LVHA, LVLA) within-subjects design, tested in a controlled, indoor environment. We evaluated two videos per Video Emotion, paired with each annotation method, resulting in eight videos (2 x HVHA, 2 x HVLA, 2 x LVHA, 2 x LVLA). Participants annotated four of them using HaloLight and the other four using DotSize. At the end of each video, participants were asked to report their emotional experience using a within-VR SAM rating scale. A SAM rating [10] panel was embedded in VR to visualize the scales of V-A, which allows users to stay closer to the context of an ongoing exposure than outside of the VR [80]. We chose the 9-point scale given that prior work found that 5-points was limited in expressivity (cf., [101]). Arousal scale ranges from calm (1) to excited (9), while valence ranges from unpleasant (1) to pleasant (9), as shown in Figure 1e. Throughout the study, subjective and physiological measures from participants were taken. Our study followed strict guidelines from our institute's ethics and data protection committee. Experiment details are explained below.

3.2.1 Video Stimuli. We selected two 360° videos to represent each emotion type (Table 1) from the database provided by Li et al. [56]. This database contains mean V-A ratings from 95 subjects. We used youtube-dl[2] to download the contents from YouTube with 4K in resolution (3840 x 1920px), equirectangular format. The videos are of different lengths where most are longer than 2 minutes, and this can result in motion sickness and fatigue [13, 56]. To avoid such issues and following Lo et al. [61] and Koelstra et al. [52] work, we extracted 60s segments from each video with no scene cuts. All video stimuli contained audio. An annotation study was conducted to test if the clipped 60s videos still provided the same original V-A ratings. Since prior work on emotion research has shown that affective states can be elicited using film stimuli with lower range lengths of 8s [35] (Western films) or 58s [26] (Asian films), we considered that such clipping should not pose issues for elicitation. 12 researchers from our institute viewed these clips (Table 1) and used the within-VR SAM rating panel to report V-A scores after each video. Agreement of the ratings (N=12) across eight selected videos were assessed by inter-rater reliability (IRR) using a two-way random, absolute agreement, average-measures intra-class correlation (ICC) [40]. Average resulting ICCs regarding the eight videos suggest excellent reliability [19] for valence scores, total average $ICC = 0.972, p < 0.05$, and for arousal scores, the total average $ICC = 0.976, p < 0.05$, indicating that V-A were rated similarly across participants. Results are shown in Table 1, where url links, start time offset, and V-A scores are indicated.

3.2.2 Subjective Measures. To evaluate VR experiences, motion sickness and the sense of presence are two widely considered human factors [8, 17]. We chose a standardized Simulator Sickness Questionnaire (SSQ) [50] to measure the level of motion sickness on a scale from 1 (none) to 4 (severe). Igroup Presence Questionnaire (IPQ) on a scale from 1 (fully disagree) to 7 (totally agree) [86] was used to assess the level of presence experienced in the virtual setting, which is used in our work to evaluate users' perceptions of VR videos. To assess perceived workload, we chose the commonly used NASA Task Load Index (NASA-TLX) questionnaire [44]. Finally, we also measured the usage count of our helper function, which can aid in assessing how familiar or confused users feel when using our RCEA-360VR system.

3.2.3 Physiological Measures. We employed three physiological measures that were shown to correlate with mental workload [16]: Pupil Dilation (PD), Electrodermal Activity (EDA), and Inter-beat Interval (IBI). PD has been shown to be an accurate marker of mental workload [11, 77], where the pupil dilation decreases as the workload increases. However, previous works [79, 116] reported that ambient light will also greatly affect the PD values. Since users' vision is engulfed by the HMD, the illumination of 360° scenes should be considered (cf., Section 3.5). EDA is also quite sensitive to users' arousal level, which reflects activity within the sympathetic axis of the automatic nervous system [32]. Previous works [12, 20] have shown that physiological arousal will increase if the users' mental workload is increasing. Finally, IBI is another sensitive indicator associated with mental workload [46, 114]. We draw on these three physiological measures as objective measures of mental workload.

3.2.4 Hardware and Software Setup. Participants viewed the 360° video clips through an HTC Vive Pro Eye[3] HMD, with a reported 0.5° accuracy and frequency of 120Hz Tobii Pro eye tracker integrated. The HMD provides a resolution of 2880 x 1600 pixels, a 110° field of view and a refresh rate of 90Hz. In parallel, the audio signal was sent to the headset equipped in the HMD. Correspondingly, head rotation and eye gaze data from the HMD were recorded at 120Hz. For annotation input, we used a wireless digital gaming joystick, called Joy-Con[4]. With a return spring, the proprioceptive feedback could aid in realigning to center position under no force, which makes it suitable for continuous annotation

[2] https://github.com/ytdl-org/youtube-dl; last retrieved: 22.12.2020
[3] https://enterprise.vive.com/us/product/vive-pro-eye/; last retrieved: 22.12.2020
[4] https://www.nintendo.com/switch/choose-your-joy-con-color/; last retrieved: 22.12.2020

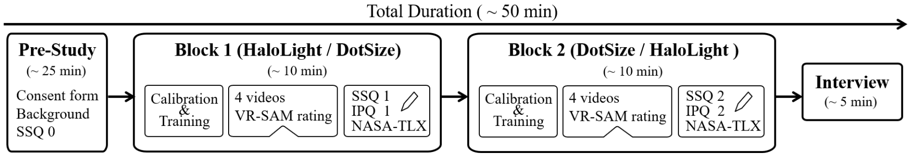

Fig. 2. Our experiment procedure.

(cf., [88]) while wearing an HMD. We also added an 11mm heightening 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 V-A space, where the x axis indicates valence, while the y axis indicates arousal, as shown in Figure 1d. Participants were instructed to annotate their emotion state by moving the joystick head into one of the four quadrants. To increase the emotion intensity, the participant could move the joystick head further. Annotated data was sampling at 10Hz (similarly to [89, 114]), because according to Loram et al. [63] the upper frequency limit of human joystick control is 5Hz and doubling this ensures robustness.

We also developed an on-demand helper function, so that participants who forget what color corresponds to which emotion quadrant could use it for easy lookup. This function is activated through a joystick button press event. The helper function is shown in Figure 1d, where we include the most representative emotion keyword. All keywords however were explained to participants prior to the study. At the end of each video, participants were asked to report their emotion state using the within-VR SAM rating panel. Participants could gaze at a single SAM icon, and use the X button on the Joy-Con controller to confirm their choice. The helper function and within-VR SAM rating panel interaction are shown in video in **Supplementary Material A**.

We constructed a custom scene in Unity Engine[5] to display 360° videos and corresponding 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[6] to collect HM and EM data from the HMD, along with the SteamVR SDK[7] which provides virtual reality support. The project ran on a 2.2 GHz Intel i7 Alienware laptop with an Nvidia RTX 2070 graphics card. We captured participants' physiological signals through the Empatica E4 band[8] worn on the non-dominant hand. This wearable device can measure BVP and EDA, and a built-in application which calculates HR and IBI from BVP. Processing of these signals are described in Sec. 3.4.5. A mobile device (Nexus 5, 32GB, 5", 1920-1080) was used to collect data from Empatica E4 band via Bluetooth. Timestamp of this device was set according to the clock of the experimental laptop, synchronized via an NTP server[9].

3.2.5 Procedure. Our experiment procedure is shown in Figure 2, lasted approximately 50 min. Before the experiment, participants carefully read and signed the data privacy and consent form and filled in demographic details. We explained the study tasks, including the 2D Circumplex model and how to annotate with the Joy-Con controller. They then filled in a pre-study SSQ. We then moved on to a calibration session, where we measured participants' Inter-pupillary Distance (IPD) to set the distance between the lenses. Each participant was equipped with an Empatica E4 wristband, Vive Pro Eye HMD, and sat in a swivel chair. Experiment room was air conditioned (21° Celsius), which results in low humidity. For Empatica E4 measures, we followed the official guide[10], where we ensured participants relaxed their arm on the

[5] https://unity.com/; last retrieved: 22.12.2020
[6] http://developer.tobiipro.com/unity/unity-getting-started.html; last retrieved: 22.12.2020
[7] https://store.steampowered.com/app/250820/SteamVR/; last retrieved: 22.12.2020
[8] https://www.empatica.com/en-int/research/e4/; last retrieved: 22.12.2020
[9] android.pool.ntp.org/
[10] https://support.empatica.com/hc/en-us/articles/206374015-Wear-your-E4-wristband-; last retrieved: 22.12.2020

swivel chair side, and wore the wristband on their non-dominant hand to minimize motion artifacts. Furthermore, we slightly tightened the E4 wristband to avoid electrode movement on users' wrists, where the experimenter checked this before each session. The embedded HMD eye tracker was calibrated following the HMD instructions[11]. During the training session, we showed a 360° video documentary with neutral emotion. Each participant was given a demo on using RCEA-360VR, the helper function, and the peripheral feedback techniques. Either HaloLight or DotSize was provided during training depending on the counterbalance condition. Finally, participants were given time to get familiar with viewing 360° videos by moving their head and rotating their chair.

Our experiment consists of two blocks. In each block we show the respective technique depending on the starting condition, where then participants watch four representative videos from each of the four quadrants. We counterbalanced the effect of peripheral feedback type by showing half participants HaloLight first and other half DotSize first. Further, we applied fractional factorial design [38] to counterbalance the effect of different videos within each block. Importantly, a small cube object was placed at the center of the video before playing, to ensure the participants start watching the videos at the same position. While watching a 360° video, participants rated their emotional states (as V-A) continuously using the joystick. Following prior work [56, 65], we wanted to avoid carry over effects (so-called Halo effects) of one emotion to another and reduce fatigue of viewing 360° video. Therefore, we enforced a delay of 15s between videos, where we also ensured a time gap of 5 minutes between each experimental block. At the end of a video, participants submitted a SAM rating using the Within-VR SAM rating panel. At the end of each block, we helped the participant remove the HMD. They then filled in the SSQ, IPQ, and NASA-TLX questionnaires. Finally, participants were given a brief semi-structured interview about their overall experience with RCEA-360VR, and using HaloLight and DotSize.

3.2.6 Participants. 32[12] participants (16f, 16m) aged between 18-33 years old ($M = 25, SD = 4.0$) were recruited. Participants were recruited from our institute and nearby institutes, and spanned varied nationalities. 37.5% had never experienced 360° VR using an HMD, where the rest had experienced VR at least once. However all were familiar with 360° videos, and none reported visual (including color blindness), auditory or motor impairments. Participants were compensated with a monetary reward for participating, commensurate with policies on user recruitment.

3.3 Results

Below we analyze the consistency of annotations between HaloLight and DotSize and compare with within-VR SAM ratings, and thereafter analyze the subjective (NASA-TLX, IPQ, SSQ) and physiological measures (PD, IBI, and EDA). While our work is exploratory, we expected that workload from low to high is: None < SAM < HaloLight/DotSize. In the None condition, users did not perform actions, so we expect workload to be low, and used this as a baseline. For inputting SAM ratings, users need to give a (retrospective) V-A rating after watching a video. For HaloLight and DotSize however, users need to annotate their emotions continuously while watching. This may incur higher workload, even though peripheral visualization techniques were designed for divided attention (cf., Sec. 3.1).

3.4 Emotion Ratings

3.4.1 Mean Valence-Arousal Ratings of Videos for HaloLight and DotSize. Mean V-A rating distributions across 32 participants for videos spanning four quadrants are shown as boxplots in Figure 3a. We first calculate the mean of continuous V-A ratings annotated by 32 participants watching eight videos. Then for each type of video, e.g., V1 and

[11] https://www.vive.com/us/support/vive-pro-eye/category_howto/calibrating-eye-tracking.html; last retrieved: 22.12.2020

[12] For effect size f=0.25 under $\alpha = 0.05$ and power (1-β) = 0.95, with 8 repeated measurements within factors, we need 24 participants.

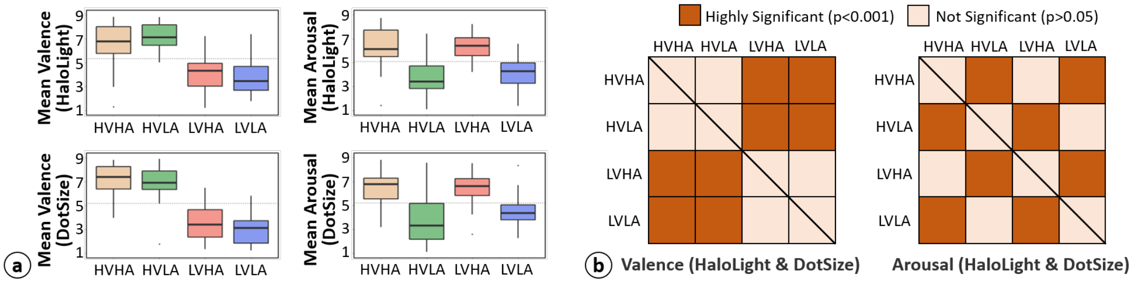

Fig. 3. (a) Boxplots for mean V-A ratings for HaloLight and DotSize. (b) Pairwise comparisons of mean V-A for HaloLight and DotSize ($p > 0.05$, not significant; $p < 0.001$, highly significant).

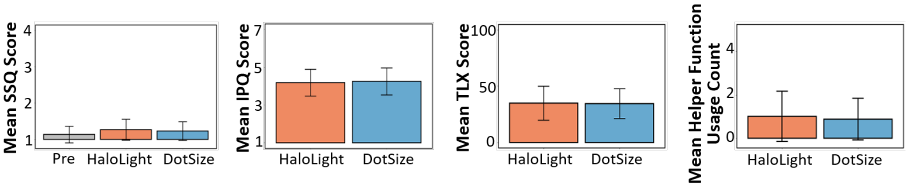

Fig. 4. Barplots for mean SSQ, IPQ, NASA-TLX, and mean helper function usage count.

V5 belong to high V-A, we average the mean of continuous annotations from the two videos across all participants. We run inferential statistics to test differences among the video types. A Shapiro-Wilk test showed that both the mean V-A ratings from HaloLight and DotSize are not normally distributed ($p < 0.05$). We therefore performed a Friedman rank sum test on the mean of valence ($\chi^2(3) = 57.94, p < 0.001$) and arousal ($\chi^2(3) = 56.96, p < 0.001$) for HaloLight, then valence ($\chi^2(3) = 71.44, p < 0.001$) and arousal ($\chi^2(3) = 43.39, p < 0.001$) for DotSize. The results show significant effects of video emotions on V-A ratings. Post-hoc Bonferroni pairwise comparisons using Wilcoxon rank sum tests were performed to precisely determine whether the ratings of any two video types are different [89, 114] and the results of these comparisons are presented in form of symmetric matrix plots in Figure 3b. Effect sizes for significant post-hoc pairwise comparisons between each video type ranged from $[0.600, 0.824]$.

3.4.2 HaloLight and DotSize Consistency across Mean Continuous V-A and Within-SAM Ratings. To assess the agreement of the two peripheral annotation visualization techniques (HaloLight and DotSize) with the mean of continuous V-A ratings, we performed a two-way mixed, absolute agreement, average-measures ICC. The average resulting ICCs suggest excellent reliability for the valence score, total average $ICC = 0.792, p < 0.05$, and of good reliability for the arousal score, total average $ICC = 0.606, p < 0.05$. Similarly, we assessed consistency between HaloLight and DotSize with our within-VR SAM ratings. The average resulting ICCs for HaloLight suggest excellent reliability for the valence score, total average $ICC = 0.855, p < 0.05$, and of good reliability for the arousal score, total average $ICC = 0.731, p < 0.05$. The average resulting ICCs for DotSize suggest excellent reliability for the valence score, total average $ICC = 0.909, p < 0.05$, and of good reliability for the arousal score, total average $ICC = 0.706, p < 0.05$.

3.4.3 SSQ, IPQ Questionnaires & Helper Function Usage. A Shapiro-Wilk normality test showed that participants' average SSQ ratings are not normally distributed ($p < 0.001$). As we compare three matched groups within subjects, then we directly performed a Friedman rank sum test. Here however, we did not find a significant effect regarding pre-study ($M = 1.139, SD = 0.229$), HaloLight ($M = 1.268, SD = 0.290$), and DotSize ($M = 1.234, SD = 0.257$) on SSQ ratings

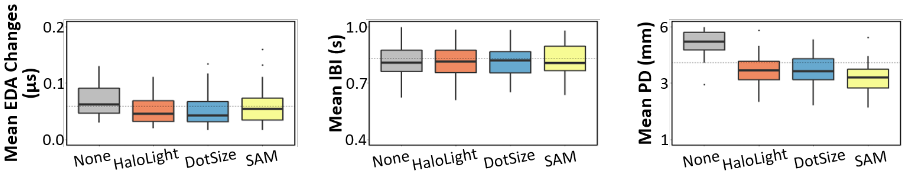

Fig. 5. Boxplots for mean EDA changes, IBI, and PD values for our four different conditions.

($\chi^2(2) = 0.777), p = 0.106$). With respect to IPQ, a Shapiro-Wilk test showed that the average IPQ scores was normally distributed ($p > 0.05$). A paired sample t-test was applied to check the differences in terms of peripheral feedback types. We found no significant differences ($t(31) = 0.397, p = 0.694$) between HaloLight ($M = 4.181, SD = 0.710$) and DotSize ($M = 4.250, SD = 0.711$) for IPQ responses. A Shapiro-Wilk test showed that the usage count of helper function is not normally distributed ($p < 0.05$). Then we performed a Wilcoxon Signed-rank test and did not find significant differences ($Z = 0.801, p = 0.429$) between HaloLight ($M = 1.008, SD = 1.153$) and DotSize ($M = 0.875, SD = 0.963$). For these measures. HaloLight and Dotsize were perceived to be similar. Results across participants are shown in Figure 4.

3.4.4 NASA-TLX Workload Scores. Subjective workload scores (Figure 4) were computed for modified NASA-TLX[13] [44] responses, and analyzed within groups per visualization method (HaloLight, DotSize). A Shapiro-Wilk test showed that the overall workload scores were normally distributed ($p > 0.05$). We therefore run a paired samples t-test, however do not find significant effects of workload ($t(31) = 0.105, p = 0.917$) between HaloLight ($MD = 33.750, IQR = 20.417$) and DotSize ($MD = 38.333, IQR = 19.791$).

3.4.5 PD, EDA & IBI. PD, EDA changes and IBI are compared for each of four conditions: without annotation (None), HaloLight, DotSize and within-VR SAM rating (SAM). The results are shown in Figure 5. For PD, we acquired and used raw values (in mm) from the HMD Tobii eye tracker, sampled at 120Hz. Means and standard deviations of PD values for the four conditions are: $None = 4.777(0.687)$, $HaloLight = 3.473(0.572)$, $DotSize = 3.444(0.574)$, $SAM = 3.209(0.526)$.). A Shapiro-Wilk test showed that the PD values are not normally distributed ($p < 0.05$). As we compare four matched groups within participants, we performed a Friedman rank sum test and found a significant effect of condition on the PD values ($\chi^2(3) = 73.95, p < 0.001$). Post-hoc pairwise comparisons using Bonferroni correction shows significant differences between None and HaloLight ($Z = 5.841, p < 0.01, r = 0.730$), between None and DotSize ($Z = 5.975, p < 0.01, r = 0.747$) and between None and SAM ($Z = 6.297, p < 0.01, r = 0.787$), however did not show significance between HaloLight and DotSize ($Z = 0.081, p > 0.05$), between HaloLight and SAM ($Z = 1.947, p > 0.05$), nor between DotSize and SAM ($Z = 1.846, p > 0.05$).

With embedded sensors, the Empatica E4 collects BVP data from PPG (64Hz), and EDA data from an EDA/GSR sensor in μS (4Hz). For EDA changes, we used the first-order differential of the EDA signal to represent arousal changes following previous work [32]. A third-order low-pass filter with a cut-off frequency of 2Hz was used to remove the artifacts in EDA. Then we calculated EDA changes by the non-negative first-order differential of filtered EDA signals following [114]. EDA changes means and standard deviations for the four conditions are: $None = 0.065(0.030)$, $HaloLight = 0.054(0.027)$, $DotSize = 0.053(0.029)$, $SAM = 0.060(0.034)$. A Shapiro-Wilk test showed that EDA changes

[13] We omit Annoyance and Preference.

is not normally distributed ($p < 0.05$). Then we performed a Friedman rank sum test to compare four matched groups within participants. Here we did not find a significant difference on EDA changes ($\chi^2(4) = 7.609, p = 0.055$).

IBI data measures the interval between individual heart beats and is computed from BVP in seconds. For IBI, we obtained the IBI sequence from the processing of the PPG/BVP signal, where Empatica's processing algorithm[14] already removes incorrect peaks due to BVP signal noise. The mean and standard deviations of IBI values for the four conditions are: $None = 0.825(0.097)$, $HaloLight = 0.838(0.099)$, $DotSize = 0.832(0.101)$, $SAM = 0.839(0.103)$.). A Shapiro-Wilk test showed that the IBI values is not normally distributed ($p < 0.05$). As we compare four matched groups within participants, we performed a Friedman rank sum test and found no significant differences on the IBI values ($\chi^2(3) = 3.902, p = 0.272$).

3.4.6 Subjective Feedback. For real-time annotation while watching 360° videos, most participants (88%) stated they could easily manage both annotating and watching simultaneously. When asked about their technique preference, 13 participants (41%) preferred HaloLight, while the rest (47%) preferred DotSize. Eight participants (53%) felt that HaloLight took up too much space and interfered with their viewing experience. P4 and P14 reflected that the preference of video content affects their preference for the visualization. P4 stated *"the first video I annotated with light is skiing, causing heavy sickness, and I don't like such sports. But for the dot, the first video is doggy, which is very cute. It makes me happy. So I prefer DotSize"*. Also, P2 mentioned that she liked DotSize because she was more familiar with the annotation task in the second block, so had a better impression due to order effects. Four participants (12%) did not have preference, among which P32 expressed *"it would be better if it was a combination of circle dot and transparency change"*.

3.5 Discussion: Usability of RCEA-360VR

We found average SSQ scores to be quite low compared to previous studies on watching 360° video studies [92]. Specifically, we found no significant differences among our three conditions: pre-study, HaloLight and DotSize. This leads us to conclude that RCEA-360VR in general does not lead to heavy motion sickness. However, as we later mention in our limitations (Sec 5.1), this could be due to the overall shorter video, non-rapid camera movement, and swivel chair seating. Furthermore, we found no significant differences between HaloLight and DotSize regarding participants' perceived sense of presence, where scores are comparable with prior work that show good IPQ scores for 3DoF media [99]. Krüger et al. [54] found that filling in discrete ratings inside VR is faster than doing so outside (which also reduces breaks in presence), while Schwind et al. [87] found that while presence did not significantly differ in or outside VR, the consistency of variance did. Furthermore, this lends support to Putze et al.'s [80] findings that administering questionnaires in VR can reduce breaks in presence and avoid biases. In our case, we collected users' SAM rating elicited by 360° videos through a custom-developed within-VR SAM rating method, which allows users to give responses within the virtual environment. Different from EmojiGrid [102], we embed the entire SAM images in the panel and allow gaze-based selection, which we find more intuitive (though this can be tested further).

For mental workload, we do not find significant differences neither from NASA-TLX nor physiological measures (PD, EDA and IBI) between HaloLight, DotSize and the within-VR SAM questionnaire. The NASA-TLX scores were lower than what Zhang et al. [114] found for annotating on mobile devices. This indicates that compared with post-stimuli SAM ratings, our RCEA-360VR techniques HaloLight and DotSize do not increase mental workload. A cautionary note with respect to resulting PD values, is that we find HaloLight, DotSize and Within-VR SAM are significantly different from None. As Pfleging et al. [79] and Zhu et al. [116] stated, people's PD values are also affected by the intensity of

[14] https://support.empatica.com/hc/en-us/articles/360030058011-E4-data-IBI-expected-signal; last retrieved: 22.12.2020

ambient light and values will be higher in a darker environment. Given that in the None condition the scene presented in the HMD was black, it is not surprising that PD values in None are higher than the other conditions. Furthermore, 78% of participants used the helper function on average once per video, where there was no significant differences between HaloLight and DotSize. This indicates that this helper function is hardly necessary. Finally, our qualitative feedback reports lend support that both techniques were easy to use for the dual task of watching and annotating. Together, our findings indicate that our RCEA-360VR method is usable within immersive 360° video environments when compared with discrete within-VR methods **(RQ1a)**, where both the HaloLight and DotSize peripheral visual feedback variants are effective in allowing the collection of precise continuous emotion annotations **(RQ1b)**.

Our resulting mean V-A values from continuous annotations are in line with the labeled V-A ratings from the original Li et al. dataset [56], which are additionally similar to our Annotation Study ratings (Sec. 3.2.1). We found no significant differences ($p > 0.05$) among videos with the same valence/arousal type, and highly significant differences ($p < 0.001$) among videos with the opposite valence/arousal type. This provides a strong initial indication that the continuous annotation of videos are similar to the original labels. We also find that V-A are rated similarly across HaloLight and DotSize. The within-VR SAM ratings and the continuous annotation methods have a high degree of agreement, as well as the within-VR SAM ratings with the original Li et al. labels (V: $ICC = 0.982, p < 0.05$; A: $ICC = 0.941, p < 0.05$). Furthermore, agreement between our within-VR SAM ratings and continuous annotations for both V-A in our experiment were higher than Voigt-Antons et al.'s [105] work on a continuous emotion rating method involving clicking on a point in a two-dimensional orthogonal grid in VR. These indicate the reliability of our annotations, and therefore, in our subsequent step (Sec. 4) of fusing annotations that consider users' head movement behavior, we consider all the data across both peripheral visualization methods.

4 GENERATING VIEWPORT-DEPENDENT EMOTION GROUND TRUTH LABELS

To answer **RQ2**, and to ensure our annotations are effective for building precise ground truth labels based on continuous ratings, we develop a segment-level viewport-dependent annotation fusion method. We aggregate multiple annotators' decision to compute the emotion ground truth [72]. In 360° videos however, users can choose what content to view through head movement, so users' continuous emotion annotations are necessarily driven by their viewport. To address this, below we describe our method , which comprises (1) continuous annotation time-alignment, (2) segment-based viewport clustering, and finally (3) the viewport-dependent annotation fusion. Source code for our method is available online at **https://github.com/cwi-dis/RCEA360VR-CHI2021**.

4.1 Continuous Annotation Time-alignment

As Metallinou et al. [72] stated, there are time delays (e.g., due to gender, age, distraction levels) between the occurrence of an emotional event and its annotation, since continuous annotations are performed in real-time. Thus, we follow Mariooryad et al.'s [69] EvalDep method where they find the evaluator-dependent lag by maximizing the mutual information between a reference feature and the annotations. We follow a similar approach taking into account the dynamic viewports. This involves three key steps: (a) pick dominant feature as reference, (b) calculate time delay for each annotation sequence, (c) shift the sequences to align. Previous work has shown that reaction delays vary from one to six seconds [69]. In our RCEA-360VR experiment (Section 3), we ensured that all participants started watching the video from a fixed center position. To verify how behavior differed during those early seconds, we look at the pitch angle distribution of Head Movements (HM) for all 32 participants. We found that the watching areas lie between -30° and 30° for more than 98% of the first six seconds for all videos, and more than 90% of the time for yaw areas falling

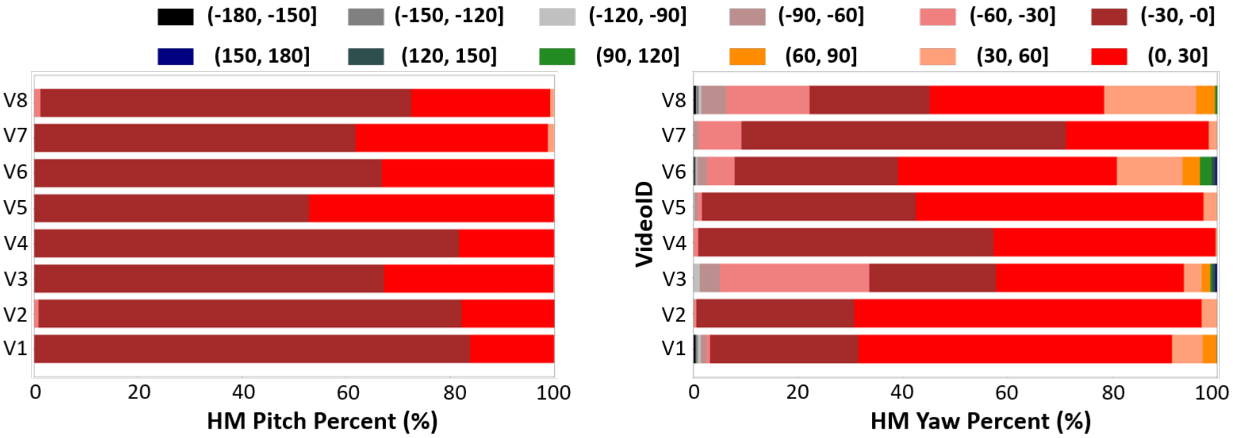

Fig. 6. Percentage of HM points in pitch (left) and yaw (right) bins per video during the first six seconds.

Algorithm 1 Annotation Time delay

```
Input: The V-A ratings $P \in R^{I \times J}$, $P_{ij} = [P_{ij}^1, P_{ij}^2 \cdots, P_{ij}^M]$; The color feature $CF \in R^{1 \times i}$, $CF_i = [CF_i^1, CF_i^2 \cdots, CF_i^N]$
Output: Each participant's annotation time delay for each video $D \in R^{i \times j}$
for $j = 1$ to J do
    for $i = 1$ to I do
        for $\tau = 1$ to 6, step is 0.1 do
            $S\_P_{ij} = [[P_{ij}^{1+\tau * fps_i}, P_{ij}^{2+\tau * fps_i} \cdots, P_{ij}^{N+\tau * fps_i}]$
            $Dis_\tau = DTW(S\_P_{ij}, CF_i)$
        end for
        $D_{ij} = argmin(Dis)$
    end for
end for
```

within -60° and 60° (shown in Figure 6). Thus, to select a suitable reference, we considered commonly used methods to extract visual features related to color, texture and edge [115] from the specific region ([-30°, 30°] for pitch, [-60°, 60°] for yaw) of the first six seconds of each video frame.

Suppose CF_i is the color feature extracted by color moment [97] from video $i \in [1, I]$, $CF_i = [CF_i^1, CF_i^2 \cdots CF_i^N]$, I is the number of videos and N is the the number of frames in the first six seconds of video i. Similarly, TF_i is the texture feature extracted by Gray-Level Co-occurrence Matrix (GLCM) [42], and EF_i is the edge feature extracted by Canny Operator [15]. Since visual feature sequences are not normally distributed ($p < 0.05$), we calculated the spearman correlation between CF_i, TF_i, EF_i and each participant's original continuous valence/arousal sequence separately. The results show that the color features across all videos have the highest Spearman correlations with both valence (ρ range: 0.2-0.6) and arousal sequences (ρ range: 0.2-0.6). As a result, we select color features as reference for subsequent time alignment. The pseudocode to calculate the annotation delay time is shown in Algorithm 1.

Suppose P_{ij} is the annotation (valence or arousal) from participant $j \in [1, J]$ watching video $i \in [1, I]$, $P_{ij} = [P_{ij}^1, P_{ij}^2 \cdots P_{ij}^M]$, D_{ij} is the annotation delay time from participant $j \in [1, J]$ watching video $i \in [1, I]$ and fps_i is the number of frames per second for video i. J and M are the number of participants and frames for video i respectively. We shift a sliding window with a duration of six seconds and step size of 0.1s (same as our joystick sampling rate;

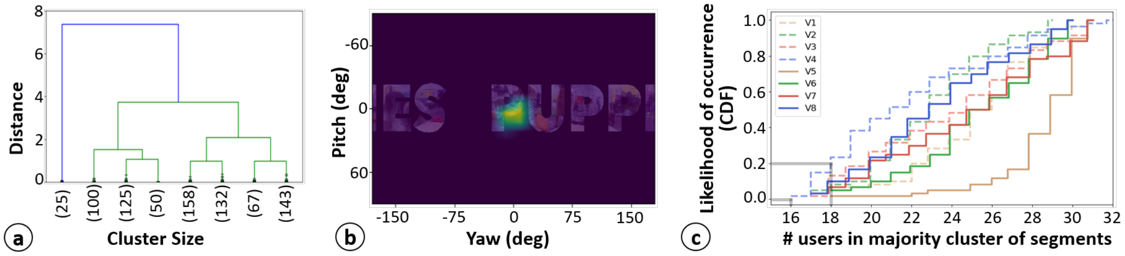

Fig. 7. (a) Hierarchical clustering for segment 1 in V1. (b) Heat saliency map for segment 1 in V1. (c) CDF for hierarchical clustering of 32 participants' viewing behavior.

Section 3.2.4) on P_{ij}, denoted as S_P_{ij}, and the starting position is from the first to the sixth second. Lastly, Dynamic time warping (DTW) [85], one of the most prominent methods in similarity measures for time series data [29], is used to calculate similarity between S_P_{ij} and CF_i. We get the shifted sequence with the highest similarity and the corresponding τ is recorded as the annotation time delay D_{ij}. Finally, we shift annotations from participant i watching video j based on D_{ij} and obtain the aligned annotation sequences. Time shifts (in seconds) across videos for valence rating ranges were [1.478, 3.944] ($M = 3.186, SD = 0.809$), and for arousal [1.969, 3.506] ($M = 2.909, SD = 0.614$). Computing time shifts across participants results in valence rating series ranges of [1.975, 4.638] ($M = 3.186, SD = 0.751$) and for arousal [1.625, 3.85] ($M = 2.909, SD = 0.553$). Our findings lend support to prior work [69] that showed ranges between 1-6s.

4.2 Segment-based Viewport Clustering

In the second step, we clustered users based on similarities in their viewing behavior. The HM data from every participant while watching 360° videos was used as input to find a group of similar users. We segment every video at 1s intervals according to test settings in [110], then the HM data in each segment from all the participants are collected to run hierarchical clustering [48]. We use dynamic hierarchical clustering to be able dynamically adjust the number of clusters, by contrast with methods (e.g., k-means) that require pre-specifying clusters in advance [98]. We dynamically adjust the convergence distance of hierarchical clustering to guarantee that the biggest cluster includes more than 80% of the HM points in one segmentation. Figure 7a shows the dendrogram of the hierarchical clustering of V1's segment 1 and Figure 7b presents a saliency map for this segment. From Figure 7c, we can see that (a) more than 50% of participants are in the selected cluster, (b) 80% of the segments' majority cluster contain at least 18 users, and (c) the number of users in the majority cluster of the segments are identical for all the videos except V5.

4.3 Annotation Fusion

Lastly, we develop a fusion method to robustly fuse multiple viewport-dependent annotations into a single set of continuous emotion ground truth labels. This involves two steps: (a) frame-level fusion, and (b) segment-level fusion. This two-step fusion approach is necessary to discard the annotation outliers at the frame level, and fuse annotations of each frame according to the percentage of viewpoints clustered in each segment. Pseudocode for our annotation fusion method is shown in Algorithm 2, where our fusion results for 8 videos with four emotion types are shown in Figure 8.

Algorithm 2 Viewport-dependent Annotation Fusion

```
Input: The V-A ratings $P \in R^{I \times J}$ for one video $P_{ij} = [P^1_{ij}, P^2_{ij} \cdots, P^N_{ij}]$
Output: Fused V-A ratings for one video $F \in R^{1 \times j}$
for j = 1 to J do
    for n = 1 to N do
        for i = 1 to I do
            $D^n$ of $P^n_{ij}$ using Eq. 1
        end for
        $X_j \leftarrow$ delete $P^n_{ij}$ in $P_j$ which $d^n_{lm} > T$
        $f_n \leftarrow$ fuse $X_j$ using 2
    end for
    $F_j = \sum_{n=1}^{N} \frac{H_n}{\sum_{p=1}^{N} H_p} f_n$
end for
```

Suppose P_{ij} is the annotation (valence or arousal) from participant $i \in [1, I]$ at segment $j \in [1, J], P_{ij} = [P^1_{ij}, P^2_{ij} \cdots P^N_{ij}]$. I and J are the number of participants and segments, respectively. N is the number of sampling points in one segment of annotation. The annotation from multiple participants is first fused in each frame using Bayesian fusion [114]. Following [64, 66], the confidence measure matrix is $D^i j$, where $d^n_{lm} \in D^n$ for frame $n \in [1, N]$ by:

$$d^n_{lm} = erf(\frac{x_l - x_m}{\sqrt{2}\sigma_l}), d^n_{ml} = erf(\frac{x_m - x_l}{\sqrt{2}\sigma_m}) \tag{1}$$

where x_m and x_l are annotations for participant m and l respectively. σ_m and σ_l are the standard deviation of the annotation for participant m and l respectively in one segment. $erf(\theta) = \frac{2}{\pi} \int_\theta^0 e^{-u^2}$ is the error function. Then the outliers for the annotations of frame n are removed by setting a threshold ($T = 0.2$) of d_{lm}. Suppose the annotation after outlier elimination is $X_j = [x1, x2, \ldots, x_K], K \leq 20$, the fusion results of frame n can be calculated as follows:

$$f_n = \sum_{k=1}^{K} (1 - \frac{\sum D^j_k}{\sum D^j}) \cdot x_k \tag{2}$$

where D^j_k represents the k column of D^j. We calculate the frame level fusion result $f = [f_1, f_2, \ldots, f_N]$ for all frame $n \in [1, N]$ in one segment. Then the annotation for segment j is obtained by using the weighted average of f:

$$F_j = \sum_{n=1}^{N} \frac{H_n}{\sum_{p=1}^{N} H_p} f_n \tag{3}$$

where H_n is the number of viewpoints at frame n. We then calculate the fusion for all segments $j \in [1, J]$ to get the fused annotation of a video. Ratings are fused independently here since V-A are orthogonal (independent) variables.

4.4 Analysis: Viewport-dependent Fused Emotion Annotations

To test the consistency of fused continuous V-A ratings, essentially how effective they are, we implement a temporal analysis of each video annotation result. Suppose A_{ij} is the fused arousal value of video $i, i \in [1, I]$, segmentation $j, j \in [1, J]$. If 50% of the $[A_{i1}, A_{i2} \cdots A_{iJ}]$ have low (1-5) or high [5-9) arousal value (cf., [114]), the overall predicted (i.e., classified) arousal for video i equals to the corresponding low/high label. The predicted valence for all eight videos are similarly calculated. We find that our fused V-A ratings can classify/predict both the original Li et al. [56] labels as well as our within-VR SAM ratings each with 100% classification accuracy.

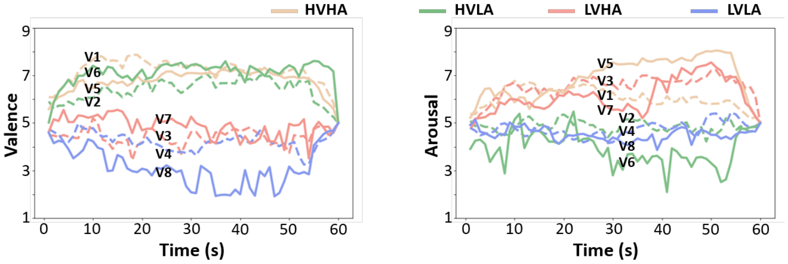

Fig. 8. Fused valence (left) and arousal (right) annotations across eight videos with different intended V-A labels.

4.4.1 Temporal Lens into Emotion States. The strength of our method lies in enabling a more fine-grained temporal lens by which to understand emotion states and specific scenes of an immersive 360° viewing experience. Here, we discuss two examples: (1) V8 is a short film on the effects of an earthquake. Around the 10th second of the video, it depicts the general scene after the earthquake, and participants' valence level changes smoothly. While at the 46th second, there is a big box suddenly dropping from the roof with a lot of dust, and the valence level of the fusion result is apparently lower from 3.15 to 1.92 from Figure 8 (left), despite the overall valence rating. (2) Around the 36th of V7, a prison guard was leading a criminal to the door, and suddenly, the suspect broke free and turned to escape. We could see the arousal level increased rapidly from 5.02 to 7.02 from Figure 8 (right). We see similar fluctuations in arousal for V1 and V5. These examples show how our fused annotations can provide temporal details into the peaks, valleys, and trends of viewers' experience. Short example videos and processing scripts to overlay viewport-dependent V-A labels on 360° videos are available at **https://github.com/cwi-dis/RCEA360VR-CHI2021**.

5 DISCUSSION

5.1 Limitations and Future Work

There were some limitations to our work. First, we did not test longer video durations (>1 min), as they result in higher motion sickness and workload, even though longer videos may be more immersive [39, 51]. Second, we do not test RCEA-360VR in scenarios where users can walk in world-scale virtual worlds [5, 55]. If users walk freely, it may be difficult using an auxiliary joystick to report emotions in real time. Third, we did not look at eye movement patterns, even though prior work [91, 100] has shown correlations between emotions and eye movements. This was beyond the current scope, and we aim to investigate this in future work. Fourth, we focus strictly on the Circumplex emotion model [45, 89] and within-VR SAM ratings [10], and do not test different dimensional models (e.g. vector models [9] or PANAS [107]) nor other discrete methods (e.g. AffectButton [14]). This was done since both Circumplex and SAM are widely used methods [37], and have exhibited good usability in prior studies. Fifth, our peripheral feedback used Itten's color system [96] due to ease of use and standardization based on prior work, however this excludes color-impaired individuals, where future work should consider accessibility to ensure widespread adoption. Furthemore, we fixed the feedback visualization on the right-bottom corner as this was deemed suitable in prior work [112], however future RCEA-360VR versions would benefit users when ensuring a more accessible and customizable design. Finally, we did not specifically measure humidity levels of our experiment room, nor caffeine and medication usage of our participants, so one should interpret our observed EDA changes cautiously.

5.2 Collecting Momentary Emotion Self-Reports in 360° VR Environments

It is now widely agreed that VR environments have the capacity to evoke a range of emotions in humans [28, 31, 76, 78]. Considering prior work in emotion sensing and recognition, it can be asked why we need to resort to seemingly cumbersome self-report collections, especially while immersed in virtual space. Specifically, why not simply track facial expressions and speech [2, 69, 94] alongside continuous annotations? In such virtual settings where currently users wear relatively bulky HMDs, nearly half the users' face is covered, and even if for example a smile can be captured, the rotation in head movements would pose issues for camera-based tracking. Furthermore, while previous work has tracked valence and arousal from speech signals during social VR experiences [57], in our case tracking speech would not be feasible as it requires users to be speaking throughout an otherwise private experience. Furthermore, when it comes to emotion research, as Barrett et al. [4] states, in the absence of an objective, external way to measure emotional experience (especially when facial expressions can indicate more than one emotion, or be altogether misleading about emotional state), we can only examine emotions through self-reports. In this respect, irrespective of automatic affect sensing, we still need self-reports as ground-truth, and ideally in the moment of the experience, rather than retrospectively (whether inside or outside VR).

However collecting self-reports in a momentary and precise manner poses challenges for users' divided attention (cf., Wicken's Multiple Resource theory [108]). This required us to leverage easy to use auxiliary devices and peripheral annotation feedback that to lower demand on users' attentional resources. As a result, we considered certain design measures (ergonomics design principle P3): for input, we used a Joy-Con wireless controller, which is lightweight and highly sensitive to positional shifts. The return spring on the joystick provides proprioceptive feedback which facilitates realigning to center position under no force, making it suitable for continuous annotation (cf., [88]) while wearing an HMD and immersed in video content. For output or peripheral feedback, we drew on peripheral visual interaction techniques, where research has shown that information presented to the periphery of users' visual attention (peripheral displays) can help participants quickly and effectively understand information while performing other primary tasks [3, 71]. This leads us to consider both HaloLight and DotSize as visualization methods. Drawing on physiological (PD, EDA, IBI) and subjective measures of workload (NASA-TLX), presence (IPQ), and motion sickness (SSQ), we collected what we believe to be sufficient evidence to enable a class of annotation techniques that leverage user peripheral attention under immersive 360° VR experiences, without drastically disrupting the user experience or creating discomfort. Despite the foregoing, since our work focused on 360° video, we further consider the need for a new class of emotion annotation techniques, given interactive (incl. locomotion) and highly immersive qualities of virtual worlds, as well as interactive 360° videos. In such interactive and world-scale scenarios, it may be difficult to simultaneously annotate one's emotion and interact with a virtual environment or with video content, and poses challenges for capturing cross-user viewport regularities.

5.3 Viewport-dependency and Fusing Fine-Grained Emotion Labels

Unlike 2D video watching, if user annotations are performed under continuously changing viewports, this creates uncertainty that the annotations pertain to that specific scene at any given point in time. This necessitates methods that consider similarities in viewing behavior. While existing techniques enable greater uniformity in viewing behavior (e.g., looping video textures under a gazed at region of interest [60]), or provide on-display guidance cues for where to look (e.g., Halo- and WedgeVR [36]), our goal was to allow as much viewing freedom as possible without manipulating video content. In this respect, our showed how RCEA-360VR takes advantage of regularities in head movement patterns (cf.,

[82]) to ensure effective fused annotations (**RQ2**). For human-computer interaction and emotion computing researchers, this unlocks greater insight into the temporal nature of reported emotion across videos (cf., Sec 4.4.1) during immersive 360° VR experiences. Similarly, it enables building more temporally precise labels for training emotion recognition systems [6, 94, 114] that can perform predictions at a more fine-grained level.

However, what if we do not witness regularity in head movement behavior across viewed 360° video content? In our study, we fixed users' video watching start position from the same central position. However, calibration of starting point in real-world settings can be more complex (cf., [93]), which may cause too much divergence for navigation patterns. In this respect, our viewport-dependent fusion method can be heavily influenced by the type of content. This raises an issue: if the viewport-dependent clustering result contains two or more dominant clusters, then we may end up with more than one set of fused annotations per cluster. In the extreme case of too many clusters, then perhaps we should go towards personalized individual viewing patterns of emotion analysis and explore the relationship between different starting points and viewport clustering results. Essentially, this impacts whether we are able to develop subject-independent emotion recognition models [22, 53]. An implication of this is that content creators may need to define visual saliency cues [58, 95] to help guide users towards focal points, which would improve segment clustering and allow meaningful viewport-dependent annotations. Interestingly, recent work by Jun et al. [49] showed in a large scale study (N=511), that the preferable 360° videos, which were likely to have attention grabbing focal points, were overall less explored by participants. This lends credence to the effectiveness of our method, should it be be used for in the wild data collection.

6 CONCLUSION

We presented a real-time, continuous emotion annotation system for 360° VR videos (RCEA-360VR). Our system comprises two peripheral visualization techniques, HaloLight and DotSize, that allow annotators to see in their visual periphery which emotion state (as valence and arousal) they are annotating. Our system enables researchers to collect fine-grained emotion annotations of valence and arousal while watching 360° videos, as well as within-VR SAM ratings. Through our controlled usability evaluation, we found no significant differences between HaloLight and DotSize concerning motion sickness, presence, or mental workload. Furthermore, both techniques do not result in high sickness, workload, nor break presence. RCEA-360VR also performs as well as retrospective, discrete rating methods, where we verified the reliability of our continuous annotations. Finally, we proposed a viewport-dependent fusion method to aggregate annotations based on 360° viewing behavior. Our work enables further research on capturing momentary emotion annotations in 360° VR, which is essential for collecting precise viewport-dependent ground truth labels.

7 ACKNOWLEDGEMENTS

This work is supported by the National Key Research and Development Program of China (No. 2020YFF0305200) and the China Scholarship Council (Scholarship No. 201906030015). We thank our participants for taking part in our study, and we thank S. Ghosh for participating in early analysis discussions.

REFERENCES

[1] Fathima Assilmia, Yun Suen Pai, Keiko Okawa, and Kai Kunze. 2017. IN360: A 360-degree-video platform to change students preconceived notions on their career. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. 2359–2365.

[2] Jeremy N. Bailenson, Emmanuel D. Pontikakis, Iris B. Mauss, James J. Gross, Maria E. Jabon, Cendri A.C. Hutcherson, Clifford Nass, and Oliver John. 2008. Real-time classification of evoked emotions using facial feature tracking and physiological responses. *International Journal of Human-Computer Studies* 66, 5 (2008), 303 – 317. https://doi.org/10.1016/j.ijhcs.2007.10.011

[3] Saskia Bakker, Doris Hausen, and Ted Selker. 2016. *Peripheral Interaction: Challenges and Opportunities for HCI in the Periphery of Attention*. Springer.

[4] Lisa Feldman Barrett. 2004. Feelings or words? Understanding the content in self-report ratings of experienced emotion. *Journal of personality and social psychology* 87, 2 (2004), 266.

[5] Chiara Bassano, Giorgio Ballestin, Eleonora Ceccaldi, Fanny Isabelle Larradet, Maurizio Mancini, Erica Volta, and Radoslaw Niewiadomski. 2019. A VR game-based system for multimodal emotion data collection. In *Motion, Interaction and Games*. 1–3.

[6] Yoann Baveye, Christel Chamaret, Emmanuel Dellandréa, and Liming Chen. 2017. Affective video content analysis: A multidisciplinary insight. *IEEE Transactions on Affective Computing* 9, 4 (2017), 396–409.

[7] Julia Beck, Mattia Rainoldi, and Roman Egger. 2019. Virtual reality in tourism: a state-of-the-art review. *Tourism Review* (2019).

[8] Maximino Bessa, Miguel Melo, David Narciso, Luís Barbosa, and José Vasconcelos-Raposo. 2016. Does 3D 360 Video Enhance User's VR Experience? An Evaluation Study. In *Proceedings of the XVII International Conference on Human Computer Interaction*. Association for Computing Machinery, New York, NY, USA, 4.

[9] Margaret M Bradley, Mark K Greenwald, Margaret C Petry, and Peter J Lang. 1992. Remembering pictures: pleasure and arousal in memory. *Journal of experimental psychology: Learning, Memory, and Cognition* 18, 2 (1992), 379.

[10] Margaret M Bradley and Peter J Lang. 1994. Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of behavior therapy and experimental psychiatry* 25, 1 (1994), 49–59.

[11] Margaret M Bradley, Laura Miccoli, Miguel A Escrig, and Peter J Lang. 2008. The pupil as a measure of emotional arousal and autonomic activation. *Psychophysiology* 45, 4 (2008), 602–607.

[12] Iuliia Brishtel, Shoya Ishimaru, Olivier Augereau, Koichi Kise, and Andreas Dengel. 2018. Assessing cognitive workload on printed and electronic media using eye-tracker and EDA wristband. In *Proceedings of the 23rd International Conference on Intelligent User Interfaces Companion*. 1–2.

[13] Marc Van den Broeck, Fahim Kawsar, and Johannes Schöning. 2017. It's all around you: Exploring 360 video viewing experiences on mobile devices. In *Proceedings of the 25th ACM international conference on Multimedia*. 762–768.

[14] Joost Broekens and Willem-Paul Brinkman. 2013. AffectButton: A method for reliable and valid affective self-report. *International Journal of Human-Computer Studies* 71, 6 (2013), 641–667.

[15] John Canny. 1986. A computational approach to edge detection. *IEEE Transactions on pattern analysis and machine intelligence* 6 (1986), 679–698.

[16] Rebecca L Charles and Jim Nixon. 2019. Measuring mental workload using physiological measures: a systematic review. *Applied ergonomics* 74 (2019), 221–232.

[17] Umer Asghar Chattha, Uzair Iqbal Janjua, Fozia Anwar, Tahir Mustafa Madni, Muhammad Faisal Cheema, and Sana Iqbal Janjua. 2020. Motion Sickness in Virtual Reality: An Empirical Evaluation. *IEEE Access* 8 (2020), 130486–130499.

[18] Alice Chirico and Andrea Gaggioli. 2019. When Virtual Feels Real: Comparing Emotional Responses and Presence in Virtual and Natural Environments. *Cyberpsychology, behavior and social networking* 22 3 (2019), 220–226.

[19] Domenic V Cicchetti. 1994. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological assessment* 6, 4 (1994), 284.

[20] C Collet, E Salvia, and C Petit-Boulanger. 2014. Measuring workload with electrodermal activity during common braking actions. *Ergonomics* 57, 6 (2014), 886–896.

[21] Tamlin S Conner and Lisa Feldman Barrett. 2012. Trends in ambulatory self-report: The role of momentary experience in psychosomatic medicine. *Psychosomatic medicine* 74, 4 (2012), 327.

[22] Layale Constantine and Hazem Hajj. 2012. A survey of ground-truth in emotion data annotation. In *2012 IEEE International Conference on Pervasive Computing and Communications Workshops*. IEEE, 697–702.

[23] Roddy Cowie, Ellen Douglas-Cowie, Susie Savvidou*, Edelle McMahon, Martin Sawey, and Marc Schröder. 2000. 'FEELTRACE': An instrument for recording perceived emotion in real time. In *ISCA tutorial and research workshop (ITRW) on speech and emotion*.

[24] Roddy Cowie, Gary McKeown, and Ellen Douglas-Cowie. 2012. Tracing emotion: an overview. *International Journal of Synthetic Emotions (IJSE)* 3, 1 (2012), 1–17.

[25] Erwan J David, Jesús Gutiérrez, Antoine Coutrot, Matthieu Perreira Da Silva, and Patrick Le Callet. 2018. A dataset of head and eye movements for 360 videos. In *Proceedings of the 9th ACM Multimedia Systems Conference*. 432–437.

[26] Yaling Deng, Meng Yang, and Renlai Zhou. 2017. A New Standardized Emotional Film Database for Asian Culture. *Frontiers in Psychology* 8 (2017), 1941. https://doi.org/10.3389/fpsyg.2017.01941

[27] Pieter MA Desmet, Martijn H Vastenburg, and Natalia Romero. 2016. Mood measurement with Pick-A-Mood: review of current methods and design of a pictorial self-report scale. *Journal of Design Research* 14, 3 (2016), 241–279.

[28] Julia Diemer, Georg W. Alpers, Henrik M. Peperkorn, Youssef Shiban, and Andreas Mühlberger. 2015. The impact of perception and presence on emotional reactions: a review of research in virtual reality. *Frontiers in Psychology* 6 (2015), 26. https://doi.org/10.3389/fpsyg.2015.00026

[29] Hui Ding, Goce Trajcevski, Peter Scheuermann, Xiaoyue Wang, and Eamonn Keogh. 2008. Querying and mining of time series data: experimental comparison of representations and distance measures. *Proceedings of the VLDB Endowment* 1, 2 (2008), 1542–1552.

[30] Kevin Doherty and Gavin Doherty. 2018. The construal of experience in HCI: Understanding self-reports. *International Journal of Human-Computer Studies* 110 (2018), 63 – 74. https://doi.org/10.1016/j.ijhcs.2017.10.006

[31] Anna Felnhofer, Oswald D. Kothgassner, Mareike Schmidt, Anna-Katharina Heinzle, Leon Beutl, Helmut Hlavacs, and Ilse Kryspin-Exner. 2015. Is virtual reality emotionally arousing? Investigating five emotion inducing virtual park scenarios. *International Journal of Human-Computer Studies* 82 (2015), 48 – 56. https://doi.org/10.1016/j.ijhcs.2015.05.004

[32] Julien Fleureau, Philippe Guillotel, and Izabela Orlac. 2013. Affective benchmarking of movies based on the physiological responses of a real audience. In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*. IEEE, 73–78.

[33] Jeffrey M Girard. 2014. CARMA: Software for continuous affect rating and media annotation. *Journal of Open Research Software* 2, 1 (2014).

[34] Jeffrey M Girard and Aidan GC Wright. 2018. DARMA: Software for dual axis rating and media annotation. *Behavior research methods* 50, 3 (2018), 902–909.

[35] James J Gross and Robert W Levenson. 1995. Emotion elicitation using films. *Cognition & emotion* 9, 1 (1995), 87–108.

[36] Uwe Gruenefeld, Abdallah El Ali, Susanne Boll, and Wilko Heuten. 2018. Beyond Halo and Wedge: visualizing out-of-view objects on head-mounted virtual and augmented reality devices. In *Proc. MobileHCI '18*. ACM, 40.

[37] Hatice Gunes and Björn Schuller. 2013. Categorical and dimensional affect analysis in continuous input: Current trends and future directions. *Image and Vision Computing* 31, 2 (2013), 120 – 136. https://doi.org/10.1016/j.imavis.2012.06.016 Affect Analysis In Continuous Input.

[38] Richard F Gunst and Robert L Mason. 2009. Fractional factorial design. *Wiley Interdisciplinary Reviews: Computational Statistics* 1, 2 (2009), 234–244.

[39] Jukka Häkkinen, Fumiya Ohta, and Takashi Kawai. 2019. Time Course of Sickness Symptoms with HMD Viewing of 360-degree Videos. *Electronic Imaging* 2019, 3 (2019), 60403–1.

[40] Kevin A Hallgren. 2012. Computing inter-rater reliability for observational data: an overview and tutorial. *Tutorials in quantitative methods for psychology* 8, 1 (2012), 23.

[41] Dini Handayani, Abdul Wahab, and Hamwira Yaacob. 2015. Recognition of emotions in video clips: the self-assessment Manikin validation. *Telkomnika* 13, 4 (2015), 1343.

[42] Robert M Haralick, Karthikeyan Shanmugam, and Its' Hak Dinstein. 1973. Textural features for image classification. *IEEE Transactions on systems, man, and cybernetics* 6 (1973), 610–621.

[43] Beverly L Harrison, Hiroshi Ishii, Kim J Vicente, and William Buxton. 1995. Transparent layered user interfaces: An evaluation of a display design to enhance focused and divided attention. In *CHI*, Vol. 95. 317–324.

[44] Sandra G Hart. 2006. NASA-task load index (NASA-TLX); 20 years later. In *Proceedings of the human factors and ergonomics society annual meeting*, Vol. 50. Sage Publications Sage CA: Los Angeles, CA, 904–908.

[45] Jennifer Healey, Pete Denman, Haroon Syed, Lama Nachman, and Susanna Raj. 2018. Circles vs. scales: an empirical evaluation of emotional assessment GUIs for mobile phones. In *Proceedings of the 20th International Conference on Human-Computer Interaction with Mobile Devices and Services*. 1–11.

[46] Andreas Henelius, Kati Hirvonen, Anu Holm, Jussi Korpela, and Kiti Muller. 2009. Mental workload classification using heart rate metrics. In *2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 1836–1839.

[47] Jason Jerald. 2015. *The VR book: Human-centered design for virtual reality*. Morgan & Claypool.

[48] Stephen C Johnson. 1967. Hierarchical clustering schemes. *Psychometrika* 32, 3 (1967), 241–254.

[49] H. Jun, M. R. Miller, F. Herrera, B. Reeves, and J. N. Bailenson. 2020. Stimulus Sampling with 360-Videos: Examining Head Movements, Arousal, Presence, Simulator Sickness, and Preference on a Large Sample of Participants and Videos. *IEEE Transactions on Affective Computing* (2020), 1–1.

[50] Robert S Kennedy, Norman E Lane, Kevin S Berbaum, and Michael G Lilienthal. 1993. Simulator sickness questionnaire: An enhanced method for quantifying simulator sickness. *The international journal of aviation psychology* 3, 3 (1993), 203–220.

[51] Robert S Kennedy, Kay M Stanney, and William P Dunlap. 2000. Duration and exposure to virtual environments: sickness curves during and across sessions. *Presence: Teleoperators & Virtual Environments* 9, 5 (2000), 463–472.

[52] Sander Koelstra, Christian Muhl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras. 2011. Deap: A database for emotion analysis; using physiological signals. *IEEE transactions on affective computing* 3, 1 (2011), 18–31.

[53] Vitaliy Kolodyazhniy, Sylvia D Kreibig, James J Gross, Walton T Roth, and Frank H Wilhelm. 2011. An affective computing approach to physiological emotion specificity: Toward subject-independent and stimulus-independent classification of film-induced emotions. *Psychophysiology* 48, 7 (2011), 908–922.

[54] Christian Krüger, Tanja Kojić, Luis Meier, Sebastian Möller, and Jan-Niklas Voigt-Antons. 2020. Development and Validation of Pictographic Scales for Rapid Assessment of Affective States in Virtual Reality. *arXiv preprint arXiv:2004.00034* (2020).

[55] Fanny Larradet, Radoslaw Niewiadomski, Giacinto Barresi, Darwin G Caldwell, and Leonardo S Mattos. 2020. Toward Emotion Recognition From Physiological Signals in the Wild: Approaching the Methodological Issues in Real-Life Data Collection. *Frontiers in Psychology* 11 (2020).

[56] Benjamin J Li, Jeremy N Bailenson, Adam Pines, Walter J Greenleaf, and Leanne M Williams. 2017. A public database of immersive VR videos with corresponding ratings of arousal, valence, and correlations between head movements and self report measures. *Frontiers in psychology* 8 (2017), 2116.

[57] Jie Li, Yiping Kong, Thomas Röggla, Francesca De Simone, Swamy Ananthanarayan, Huib de Ridder, Abdallah El Ali, and Pablo Cesar. 2019. Measuring and Understanding Photo Sharing Experiences in Social Virtual Reality. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1?14. https://doi.org/10.1145/3290605.3300897

[58] Yen-Chen Lin, Yung-Ju Chang, Hou-Ning Hu, Hsien-Tzu Cheng, Chi-Wen Huang, and Min Sun. 2017. Tell me where to look: Investigating ways for assisting focus in 360 video. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 2535–2545.

[59] David Lindlbauer, Klemen Lilija, Robert Walter, and Jörg Müller. 2016. Influence of display transparency on background awareness and task performance. In *Proc. CHI '16*. ACM, 1705–1716.

[60] Sean J. Liu, Maneesh Agrawala, Stephen DiVerdi, and Aaron Hertzmann. 2019. View-Dependent Video Textures for 360° Video. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology* (New Orleans, LA, USA) *(UIST '19)*. Association for Computing Machinery, New York, NY, USA, 249?262. https://doi.org/10.1145/3332165.3347887

[61] Wen-Chih Lo, Ching-Ling Fan, Jean Lee, Chun-Ying Huang, Kuan-Ta Chen, and Cheng-Hsin Hsu. 2017. 360 video viewing dataset in head-mounted virtual reality. In *Proceedings of the 8th ACM on Multimedia Systems Conference*. 211–216.

[62] Phil Lopes, Georgios N Yannakakis, and Antonios Liapis. 2017. RankTrace: Relative and unbounded affect annotation. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 158–163.

[63] Ian D Loram, Henrik Gollee, Martin Lakie, and Peter J Gawthrop. 2011. Human control of an inverted pendulum: is continuous control necessary? Is intermittent control effective? Is intermittent control physiological? *The Journal of physiology* 589, 2 (2011), 307–324.

[64] Ren C Luo, M-H Lin, and Ralph S Scherp. 1988. Dynamic multi-sensor data fusion system for intelligent robots. *IEEE Journal on Robotics and Automation* 4, 4 (1988), 386–396.

[65] Antoine Lutz, Julie Brefczynski-Lewis, Tom Johnstone, and Richard J Davidson. 2008. Regulation of the neural circuitry of emotion by compassion meditation: effects of meditative expertise. *PloS one* 3, 3 (2008).

[66] Siyuan Ma, Gangquan Si, Wenmeng Yue, and Zhiqiang Ding. 2016. An online monitoring measure consistency computing algorithm by sliding window in multi-sensor system. In *2016 IEEE International Conference on Mechatronics and Automation*. IEEE, 2185–2190.

[67] Andrew MacQuarrie and Anthony Steed. 2017. Cinematic virtual reality: Evaluating the effect of display type on the viewing experience for panoramic video. In *2017 IEEE Virtual Reality (VR)*. IEEE, 45–54.

[68] Javier Marín-Morales, Juan Luis Higuera-Trujillo, Alberto Greco, Jaime Guixeres, Carmen Llinares, Enzo Pasquale Scilingo, Mariano Alcañiz, and Gaetano Valenza. 2018. Affective computing in virtual reality: emotion recognition from brain and heartbeat dynamics using wearable sensors. *Scientific reports* 8, 1 (2018), 1–15.

[69] Soroosh Mariooryad and Carlos Busso. 2014. Correcting time-continuous emotional labels by modeling the reaction lag of evaluators. *IEEE Transactions on Affective Computing* 6, 2 (2014), 97–108.

[70] Gerd Marmitt and Andrew T Duchowski. [n.d.]. *Modeling visual attention in VR: Measuring the accuracy of predicted scanpaths*. Ph.D. Dissertation.

[71] Tara Matthews, Anind K. Dey, Jennifer Mankoff, Scott Carter, and Tye Rattenbury. 2004. A Toolkit for Managing User Attention in Peripheral Displays. In *Proc. UIST '04* (Santa Fe, NM, USA). ACM, New York, NY, USA, 247–256.

[72] Angeliki Metallinou and Shrikanth Narayanan. 2013. Annotation and processing of continuous emotional attributes: Challenges and opportunities. In *2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG)*. IEEE, 1–8.

[73] Frederik Nagel, Reinhard Kopiez, Oliver Grewe, and Eckart Altenmüller. 2007. EMuJoy: Software for continuous measurement of perceived emotions in music. *Behavior Research Methods* 39, 2 (2007), 283–290.

[74] Afshin Taghavi Nasrabadi, Aliehsan Samiei, and Ravi Prakash. 2020. Viewport prediction for 360° videos: a clustering approach. In *Proceedings of the 30th ACM Workshop on Network and Operating Systems Support for Digital Audio and Video*. 34–39.

[75] Donald A. Norman and Stephen W. Draper. 1986. *User Centered System Design; New Perspectives on Human-Computer Interaction*. L. Erlbaum Associates Inc., USA.

[76] Tiago Oliveira, Paulo Noriega, Francisco Rebelo, and Regina Heidrich. 2018. Evaluation of the Relationship Between Virtual Environments and Emotions. In *Advances in Ergonomics in Design*, Francisco Rebelo and Marcelo Soares (Eds.). Springer, Cham, 71–82.

[77] Timo Partala and Veikko Surakka. 2003. Pupil size variation as an indication of affective processing. *International journal of human-computer studies* 59, 1-2 (2003), 185–198.

[78] Xiaolan Peng, Jin Huang, Linghan Li, Chen Gao, Hui Chen, Feng Tian, and Hongan Wang. 2019. Beyond Horror and Fear: Exploring Player Experience Invoked by Emotional Challenge in VR Games. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI EA '19)*. Association for Computing Machinery, New York, NY, USA, 1?6. https://doi.org/10.1145/3290607.3312832

[79] Bastian Pfleging, Drea K Fekety, Albrecht Schmidt, and Andrew L Kun. 2016. A model relating pupil diameter to mental workload and lighting conditions. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. 5776–5788.

[80] Susanne Putze, Dmitry Alexandrovsky, Felix Putze, Sebastian Höffner, Jan David Smeddinck, and Rainer Malaka. 2020. Breaking The Experience: Effects of Questionnaires in VR User Studies. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–15.

[81] Silvia Rossi, Francesca De Simone, Pascal Frossard, and Laura Toni. 2019. Spherical clustering of users navigating 360 content. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 4020–4024.

[82] S. Rossi, F. De Simone, P. Frossard, and L. Toni. 2019. Spherical Clustering of Users Navigating 360° Content. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 4020–4024.

[83] Silvia Rossi, Cagri Ozcinar, Aljosa Smolic, and Laura Toni. 2020. Do Users Behave Similarly in VR? Investigation of the User Influence on the System Design. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 16, 2 (2020), 1–26.

[84] James A Russell. 1980. A circumplex model of affect. *Journal of personality and social psychology* 39, 6 (1980), 1161.

[85] Hiroaki Sakoe and Seibi Chiba. 1978. Dynamic programming algorithm optimization for spoken word recognition. *IEEE transactions on acoustics, speech, and signal processing* 26, 1 (1978), 43–49.

[86] Thomas Schubert, Frank Friedmann, and Holger Regenbrecht. 2001. The experience of presence: Factor analytic insights. *Presence: Teleoperators & Virtual Environments* 10, 3 (2001), 266–281.

[87] Valentin Schwind, Pascal Knierim, Nico Haas, and Niels Henze. 2019. Using Presence Questionnaires in Virtual Reality. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1?12. https://doi.org/10.1145/3290605.3300590

[88] Karan Sharma, Claudio Castellini, Freek Stulp, and Egon L Van den Broek. 2017. Continuous, real-time emotion annotation: A novel joystick-based analysis framework. *IEEE Transactions on Affective Computing* (2017).

[89] Karan Sharma, Claudio Castellini, Egon L van den Broek, Alin Albu-Schaeffer, and Friedhelm Schwenker. 2019. A dataset of continuous affect annotations and physiological signals for emotion analysis. *Scientific data* 6, 1 (2019), 1–13.

[90] Lin Shu, Jinyan Xie, Mingyue Yang, Ziyi Li, Zhenqi Li, Dan Liao, Xiangmin Xu, and Xinyi Yang. 2018. A review of emotion recognition using physiological signals. *Sensors* 18, 7 (2018), 2074.

[91] Jaana Simola, Kevin Le Fevre, Jari Torniainen, and Thierry Baccino. 2015. Affective processing in natural scene viewing: Valence and arousal interactions in eye-fixation-related potentials. *NeuroImage* 106 (2015), 21–33.

[92] Ashutosh Singla, Stephan Fremerey, Werner Robitza, and Alexander Raake. 2017. Measuring and comparing QoE and simulator sickness of omnidirectional videos in different head mounted displays. In *2017 Ninth international conference on quality of multimedia experience (QoMEX)*. IEEE, 1–6.

[93] V. Sitzmann, A. Serrano, A. Pavel, M. Agrawala, D. Gutierrez, B. Masia, and G. Wetzstein. 2018. Saliency in VR: How Do People Explore Virtual Environments? *IEEE Transactions on Visualization and Computer Graphics* 24, 4 (2018), 1633–1642.

[94] Mohammad Soleymani, Sadjad Asghari-Esfeden, Yun Fu, and Maja Pantic. 2015. Analysis of EEG signals and facial expressions for continuous emotion detection. *IEEE Transactions on Affective Computing* 7, 1 (2015), 17–28.

[95] Marco Speicher, Christoph Rosenberg, Donald Degraen, Florian Daiber, and Antonio Krúger. 2019. Exploring visual guidance in 360-degree videos. In *Proceedings of the 2019 ACM International Conference on Interactive Experiences for TV and Online Video*. 1–12.

[96] Anna Ståhl, Petra Sundström, and Kristina Höök. 2005. A foundation for emotional expressivity. In *Proceedings of the 2005 conference on Designing for User eXperience*. AIGA: American Institute of Graphic Arts, 33.

[97] Markus Andreas Stricker and Markus Orengo. 1995. Similarity of color images. In *Storage and retrieval for image and video databases III*, Vol. 2420. International Society for Optics and Photonics, 381–392.

[98] Mu-Chun Su and Yi-Chun Liu. 2002. A hierarchical approach to ART-like clustering algorithm. In *Proceedings of the 2002 International Joint Conference on Neural Networks. IJCNN'02 (Cat. No. 02CH37290)*, Vol. 1. IEEE, 788–793.

[99] S. Subramanyam, J. Li, I. Viola, and P. Cesar. 2020. Comparing the Quality of Highly Realistic Digital Humans in 3DoF and 6DoF: A Volumetric Video Case Study. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. 127–136.

[100] Wei Tang, Shiyi Wu, Toinon Vigier, and Matthieu Perreira Da Silva. 2020. Influence of Emotions on Eye Behavior in Omnidirectional Content. In *2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE, 1–6.

[101] Jordan Tewell, Jon Bird, and George R Buchanan. 2017. The heat is on: A temperature display for conveying affective feedback. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 1756–1767.

[102] Alexander Toet, Fabienne Heijn, Anne-Marie Brouwer, Tina Mioch, and Jan BF van Erp. 2019. The EmojiGrid as an Immersive Self-report Tool for the Affective Assessment of 360 VR Videos. In *International Conference on Virtual Reality and Augmented Reality*. Springer, 330–335.

[103] Alexander Toet, Fabienne Heijn, Anne-Marie Brouwer, Tina Mioch, and Jan B. F. van Erp. 2019. The EmojiGrid as an Immersive Self-report Tool for the Affective Assessment of 360 VR Videos. In *Virtual Reality and Augmented Reality*, Patrick Bourdot, Victoria Interrante, Luciana Nedel, Nadia Magnenat-Thalmann, and Gabriel Zachmann (Eds.). Springer International Publishing, Cham, 330–335.

[104] Paul Hendriks Vettehen, Daan Wiltink, Maite Huiskamp, Gabi Schaap, and Paul Ketelaar. 2019. Taking the full view: How viewers respond to 360-degree video news. *Computers in Human Behavior* 91 (2019), 24–32.

[105] Jan-Niklas Voigt-Antons, Eero Lehtonen, Andres Pinilla Palacios, Danish Ali, Tanja Kojic, and Sebastian Möller. 2020. Comparing Emotional States Induced by 360° Videos Via Head-Mounted Display and Computer Screen. In *2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE, 1–6.

[106] Guan Wang, Wenying Gu, and Ayoung Suh. 2018. The effects of 360-degree VR videos on audience engagement: Evidence from the New York Times. In *International Conference on HCI in Business, Government, and Organizations*. Springer, 217–235.

[107] David Watson. 1988. The vicissitudes of mood measurement: Effects of varying descriptors, time frames, and response formats on measures of positive and negative affect. *Journal of Personality and social Psychology* 55, 1 (1988), 128.

[108] Christopher D Wickens. 2008. Multiple resources and mental workload. *Human factors* 50, 3 (2008), 449–455.

[109] Chenglei Wu, Zhihao Tan, Zhi Wang, and Shiqiang Yang. 2017. A dataset for exploring user behaviors in VR spherical video streaming. In *Proceedings of the 8th ACM on Multimedia Systems Conference*. 193–198.

[110] Lan Xie, Zhimin Xu, Yixuan Ban, Xinggong Zhang, and Zongming Guo. 2017. 360probdash: Improving qoe of 360 video streaming using tile-based http adaptive streaming. In *Proceedings of the 25th ACM international conference on Multimedia*. 315–323.

[111] Mai Xu, Chen Li, Zhenzhong Chen, Zulin Wang, and Zhenyu Guan. 2018. Assessing visual quality of omnidirectional videos. *IEEE Transactions on Circuits and Systems for Video Technology* 29, 12 (2018), 3516–3530.

[112] Tong Xue, Surjya Ghosh, Gangyi Ding, Abdallah El Ali, and Pablo Cesar. 2020. Designing Real-time, Continuous Emotion Annotation Techniques for 360° VR Videos. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–9.

[113] Georgios N Yannakakis and Hector P Martinez. 2015. Grounding truth via ordinal annotation. In *2015 international conference on affective computing and intelligent interaction (ACII)*. IEEE, 574–580.

[114] Tianyi Zhang, Abdallah El Ali, Chen Wang, Alan Hanjalic, and Pablo Cesar. 2020. RCEA: Real-time, Continuous Emotion Annotation for Collecting Precise Mobile Video Ground Truth Labels. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–15.

[115] Sicheng Zhao, Yue Gao, Xiaolei Jiang, Hongxun Yao, Tat-Seng Chua, and Xiaoshuai Sun. 2014. Exploring principles-of-art features for image emotion recognition. In *Proceedings of the 22nd ACM international conference on Multimedia*. 47–56.

[116] Zhiwei Zhu, Kikuo Fujimura, and Qiang Ji. 2002. Real-time eye detection and tracking under various light conditions. In *Proceedings of the 2002 symposium on Eye tracking research & applications*. 139–144.