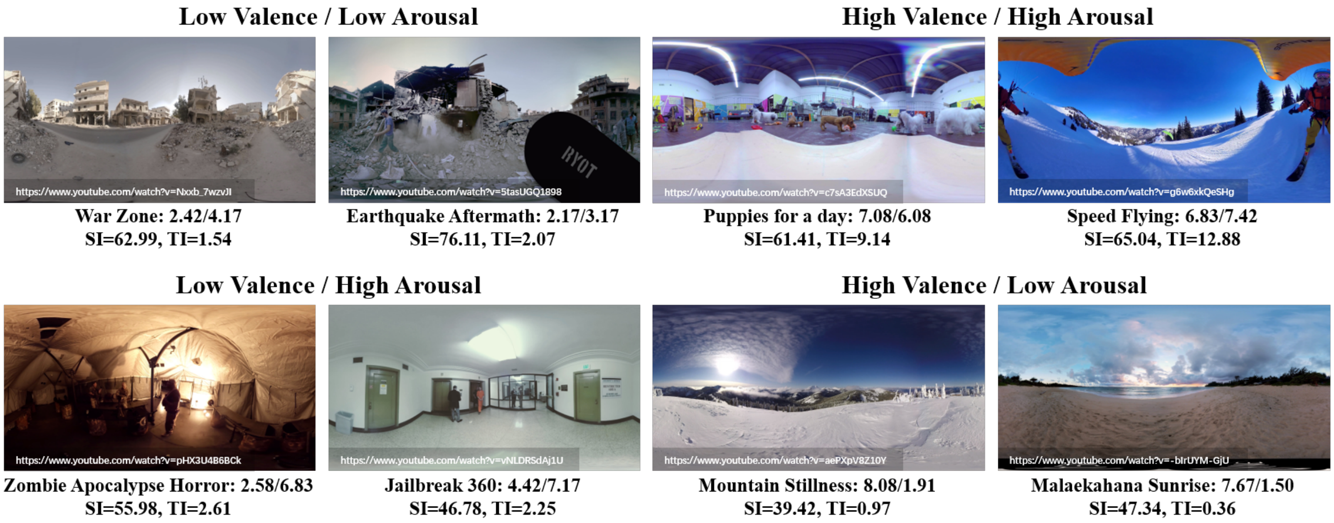
Investigating the Relationship between Momentary Emotion Self-reports and Head and Eye Movements in HMD-based 360° VR Video Watching

Tong Xue
Beijing Institute of Technology
Beijing, China
xuetong@bit.edu.cn

Abdallah El Ali
Centrum Wiskunde & Informatica
Amsterdam, The Netherlands
aea@cwi.nl

Gangyi Ding
Beijing Institute of Technology
Beijing, China
dgy@bit.edu.cn

Pablo Cesar
Centrum Wiskunde & Informatica
Delft University of Technology
Amsterdam, The Netherlands
p.s.cesar@cwi.nl


https://www.youtube.com/watch?v=Nxxb_7wzvJI
https://www.youtube.com/watch?v=5tasUGQ1898
https://www.youtube.com/watch?v=c7sA3EdXSUQ
https://www.youtube.com/watch?v=g6w6xkQeSHg
https://www.youtube.com/watch?v=pHX3U4B6BCk
https://www.youtube.com/watch?v=vNLDRSdAj1U
https://www.youtube.com/watch?v=aePXpV8Z10Y
https://www.youtube.com/watch?v=-bIrUYM-GjU

Low Valence / Low Arousal

War Zone: 2.42/4.17
SI=62.99, TI=1.54

Earthquake Aftermath: 2.17/3.17
SI=76.11, TI=2.07

High Valence / High Arousal

Puppies for a day: 7.08/6.08
SI=61.41, TI=9.14

Speed Flying: 6.83/7.42
SI=65.04, TI=12.88

Low Valence / High Arousal

Zombie Apocalypse Horror: 2.58/6.83
SI=55.98, TI=2.61

Jailbreak 360: 4.42/7.17
SI=46.78, TI=2.25

High Valence / Low Arousal

Mountain Stillness: 8.08/1.91
SI=39.42, TI=0.97

Malaekahana Sunrise: 7.67/1.50
SI=47.34, TI=0.36

Figure 1: Screenshots of the eight 1-minute 360° videos tested [45]. Under each video: "Name: Mean Valence/Arousal. Spatial Perceptual Information (SI), Temporal Perceptual Information (TI)". Mean V-A scores shown are based on 1-minute clipped videos.

ABSTRACT

Inferring emotions from Head Movement (HM) and Eye Movement (EM) data in 360° Virtual Reality (VR) can enable a low-cost means of improving users' Quality of Experience. Correlations have been shown between retrospective emotions and HM, as well as EM when tested with static 360° images. In this early work, we investigate the relationship between momentary emotion self-reports and HM/EM in HMD-based 360° VR video watching. We draw on HM/EM data from a controlled study (N=32) where participants watched eight 1-minute 360° emotion-inducing video clips, and annotated their valence and arousal levels continuously in real-time. We analyzed HM/EM features across fine-grained emotion labels from video segments with varying lengths (5-60s), and found significant correlations between HM rotation data, as well as some EM features, with valence and arousal ratings. We show that fine-grained emotion labels provide greater insight into how HM/EM relate to emotions during HMD-based 360° VR video watching.

CHI '21 Extended Abstracts, May 8–13, 2021, Yokohama, Japan

ACM ISBN 978-1-4503-8095-9/21/05…$15.00
https://doi.org/10.1145/3411763.3451627

CCS CONCEPTS

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

KEYWORDS

Emotion; head movement; eye movement; 360° video; virtual reality

ACM Reference Format:
Tong Xue, Abdallah El Ali, Gangyi Ding, and Pablo Cesar. 2021. Investigating the Relationship between Momentary Emotion Self-reports and Head and Eye Movements in HMD-based 360° VR Video Watching. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts (CHI '21 Extended Abstracts), May 8–13, 2021, Yokohama, Japan.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3411763.3451627

1 INTRODUCTION

Unlike with desktop environments, wearing a Virtual Reality (VR) Head-Mounted Display (HMD) and being in a virtual environment where users watch 360° videos content has the capacity to further stimulate audio-visual perception of users. This has been shown to result in a stronger sense of immersion and presence [29]. Furthermore, given that users can rotate their heads freely to interact with the displayed content, a growing research topic is the development of reliable visual attention models for improving processing, coding, delivering and rendering techniques for immersive media [20]. These can enable a low-cost means of improving users' Quality of Experience (QoE), where affective state plays a strong role in where users look. Given this, it is becoming increasingly important to explore the relationship between Head Movements (HM) and Eye Movements (EM) on the one hand, and the moment by moment experienced emotions on the other, while users are immersed in VR experiences [25, 39].

Previous studies have indicated that head posture and movement could reflect emotion states [16, 22, 33, 40]. For example, people tend to raise their heads when they are happy, but lower their heads when they are sad. However, few works explored the possible links between movement behavior and emotions in virtual environments. Recently Tang et al. [39] discussed the influence of emotions on eye behavior while viewing omnidirectional 360° image content. Furthermore, Li et al. [25] examined the relationship between rotational HM and emotions while users watched 360° videos, where emotion labels were obtained via post-stimuli self-reports.

Typically emotion data collection takes place via post-interaction or post-stimuli self-reports of valence and arousal (V-A) (cf., [31]), which are retrospective and discrete in nature (e.g., Self-Assessment Manikin (SAM) [7]). However, such self-reports are temporally imprecise, especially for video content, since one can experience multiple emotions throughout [38, 48] (e.g., experiencing >1 emotion when entire video is labeled 'happy'). Moreover, retrospective evaluations rely on episodic memory (cf., self-report construal in HCI [15]), which can introduce episodic memory biases (e.g., peak-and-end effects) [11]. In this work, since our task involves simultaneously watching 360° videos using HMDs and annotating in real-time continuously, we follow prior work on real-time and continuous emotion annotation [12, 18, 35, 45, 48]. Here, we draw on Russell's Circumplex model [31] using the two dimensions of valence and arousal to capture the finer granularity of emotion annotations throughout the user's immersive experience. These fine-grained emotion self-reports enable us to study the relationship between momentary emotion states and HM/EM, under varying interaction durations (or video segments). Given this, in this work we ask: is there a (statistical) relationship between emotions and HM and EM in HMD-based 360° VR video watching, and how is this affected by video segment duration?

In this exploratory work, we build on our prior (upcoming) work [45] where we collected HM/EM data from a controlled study (N=32) where participants watched eight 1-minute 360° emotion-inducing video clips, and annotated their V-A levels continuously and in real-time[1] [44]. We analyzed HM and EM features across fine-grained emotion labels from video segments with varying lengths (5-60s (seconds)), and found that: (1) Standard deviation of HM yaw (for 5, 10, and 20s segments) negatively correlated with valence, while HM pitch positively correlated with arousal. (2) Standard deviation of EM yaw (for 5 and 10s segments) negatively correlated with valence, while EM pitch negatively correlated with arousal. (3) Eye fixation amount was significantly higher for exciting videos, with lower saccade duration. Our early work contributes a novel means to assess the relationship between objective HM and EM measures, and the moment-by-moment affective states (through fine-grained annotations) during immersive 360° VR video watching experiences. It should be noted that we only look at correlations, so findings should be interpreted cautiously since we cannot make statements about the direction of the causal arrow. Below, we start with a survey of related work.

2 RELATED WORK

Two research strands influenced our approach (relationship between HM/EM and emotions, and datasets for understanding 360° media), which we describe below.

2.1 Relationship between HM/EM and Emotions

Compared with facial expressions, studies [1, 4, 22, 32] have shown that head movements can convey additional important information. Lhommet et al. [24] and Gross et al. [19] showed that there exists a significant relationship between particular head movements and certain emotions. Livingstone et al. [26] tracked vocalists' head movements while speaking and singing passages of varying emotions and findings showed that head pitch is effectively associated with emotions. Lemos et al. [14] analyzed gaze features in eye movements (including blinks and pupil changes), and showed it is possible to infer valence and arousal. Wiebe et al. [42] showed that users spend more time on watching pictures with positive or negative emotions than neutral pictures. Kusano et al. [23] focused on stress prediction, where they proposed a machine learning method to extract heart rate features from head motion to predict stress. More recently, Tang et al. [39] explored the influence of emotions on eye movement behavior while users watched 360° images, and found significant effects of negative emotions on fixation and EM saccade features. Li et al. [25] investigated the relationship between HM and valence and arousal, where they found a significant positive

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

relationship between head pitch and arousal, while the standard deviation of yaw positively correlated with valence. Together, these foregoing studies underscore the relationship between tracking HM and EM, and inferring emotional states from such measures. Importantly, in none of these works do they have fine-grained emotion self-report labels, which allows a temporal analysis of affective state and its association with HM and EM.

2.2 Datasets for Understanding 360° Media

Rai et al. [30] created a dataset of 360° images with HM and EM data captured during a user study with 63 users. They calculated and provided processed head saliency maps, head-eye saliency maps, and scanpaths. Later, it was extended to 360° video content. David et al. [13] captured HM and EM from 57 participants freely viewing 19 video sequences each with a duration of 20 seconds, which can be used to support research on visual attention and behavior exploration of 360° content. Slater et al. [36] conducted an experiment with 20 users who were required to walk through a virtual field and count the number of trees with diseased leaves, where results showed a positive association between head yaw and reported presence. Won et al. [43] found a relationship between lateral head rotations and anxiety in a virtual learning experience. Li et al. [25] provided a public dataset of 360° videos together with results of HM and SAM ratings. As mentioned earlier (Sec. 2.1), they found significant positive relationships between HM and emotion states, however a limitation of their work is that the duration of video clips were long (which may result in cybersickness and thus lower presence [41]) and the ratings were retrospective, which may reduce the accuracy of HM feature analysis on emotion. While these previous works have taken steps to investigate user behavior and affect in VR environments, they are focused on 360° images, or looked at 360° videos but for only HM, not EM.

3 DATA COLLECTION SETUP

We draw on our upcoming work [45], where we conducted a controlled, indoor laboratory experiment (N=32; 16f,16m; 18-33 years old, M=25, SD=4.0) to collect HM and EM data as well as continuous emotion annotations while users watched 360° videos. We used and analyzed the same data we already collected, albeit from an HM and EM perspective. In that data collection study, we draw on the Circumplex model [31] of emotion, where four types of videos were shown depending on V-A video ratings, as shown in Figure 2a. These are: high valence / high arousal (HVHA), high valence / low arousal (HVLA), low valence / low arousal (LVLA), low valence / high arousal (LVHA). Eight 360° videos with emotion labels (see Figure 1) were selected from Li et al.'s [25] public database (https://vhil.stanford.edu/360-video-database/), two videos per emotion type. The videos are of different lengths where most are longer than 2 minutes, and this can result in motion sickness and fatigue [8, 25]. Therefore, each video was clipped to 1-minute in length (cf., [27]), where a pre-study showed that the emotion labels of clipped videos were consistent with the original database labels [45]. We also provide Spatial Perceptual Information (SI) and Temporal Perceptual Information (TI) for eight selected videos in equirectangular format [62] to depict spatial and temporal complexity, as shown in Figure 1. Whereas SI indicates the amount of spatial detail and is higher for more spatially complex scenes, TI indicates the amount of temporal changes and is higher for high motion sequences.

Participants viewed the 360° video clips (see Figure 2b) through an HTC Vive Pro Eye[2] 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. Participants annotated the videos using the HaloLight and DotSize peripheral visualization techniques [44, 46]. For annotation input, a wireless digital gaming joystick, called Joy-Con[3] was used. With a return spring, the proprioceptive feedback could aid in realigning to center position under no force, which makes it suitable for continuous annotation (cf., [34]) while wearing an HMD. While watching a 360° video, participants rated their emotional states (as V-A) continuously using the joystick. Following prior work [25, 28], carry over effects (so-called Halo effects) of one emotion to another were avoided, as well as to reduce fatigue of viewing 360° video. Therefore, a delay of 15s between videos was enforced, with an additional time gap of 5 minutes between each experimental block. At the end of each video, participants were asked to report their emotional experience using a within-VR Self-Assessment Manikin (SAM) [7] rating scale. A custom scene in Unity Engine[4] was used 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[5] to collect HM and EM data from the HMD, along with the SteamVR SDK[6] which provides virtual reality support. The project ran on a 2.2 GHz Intel i7 Alienware laptop with an Nvidia RTX 2070 graphics card.

4 RESULTS

4.1 Preprocessing

We recorded participants' head rotation and eye movement raw data through the HMD, and then extracted pitch and yaw values of HM and EM based on these. Pitch represents the movement around the X-axis, where pitch values are between $(-90, 90)$ with 0 indicating the vertical center. Yaw refers to the movement around the Y-axis, where yaw values are between $(-180, 180)$ with 0 indicating the horizontal center of the original equirectangular video. These are shown in Figure 2c. We first divided each video into varying length segments (in seconds), which were: 5s, 10s, 20s, 30s and 60s. The sample size (segments x videos x participants) of 5s-segment is (12 x 8 x 32), 10s-segment is (6 x 8 x 32), 20s-segment is (3 x 8 x 32), 30s-segment is (2 x 8 x 32), 60s-segment is (1 x 8 x 32). The sample size per emotion quadrant within each segment durations are as follows: 5s segment (HVHA: 1073, HVLA: 754, LVLA: 419, LVHA: 826); 10s segment (HVHA: 487, HVLA: 381, LVLA: 236, LVHA: 432);

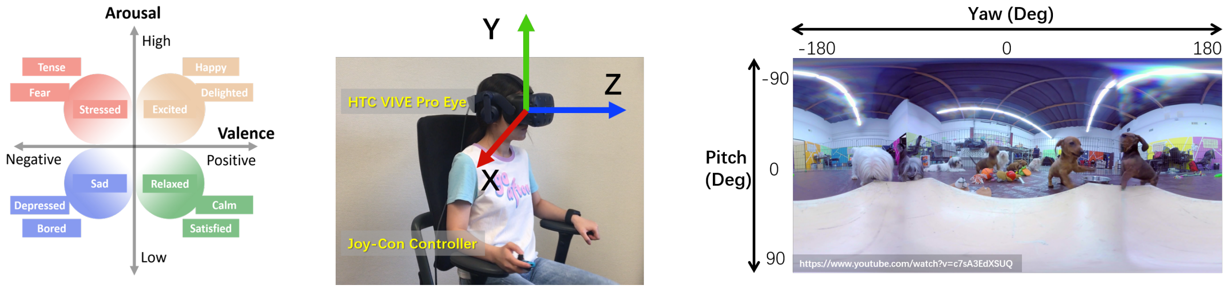
[2] https://enterprise.vive.com/us/product/vive-pro-eye/; last retrieved: 21.02.2021

[3] https://www.nintendo.com/switch/choose-your-joy-con-color/; last retrieved: 21.02.2021

[4] https://unity.com/; last retrieved: 21.02.2021

[5] http://developer.tobiipro.com/unity/unity-getting-started.html; last retrieved: 21.02.2021

[6] https://store.steampowered.com/app/250820/SteamVR/; last retrieved: 21.02.2021



(a) Valence-Arousal model space based on the Circumplex model of emotion [31].

(b) One user in our data collection setup, wearing HMD and annotating with Joy-con controller.

(c) One frame in Equirectangular format with pitch and yaw.

Figure 2: Data collection setup.

Seg Length (s)	HM Data	Valence (Mean)		Valence (Median)		Arousal (Mean)		Arousal (Median)	
		Corr	p	Corr	p	Corr	p	Corr	p
5	pitch_mean	-0.147	0.154	**0.442**	**0.000**	-0.207	0.043	**0.458**	**0.000**
	yaw_mean	**0.306**	**0.002**	**0.317**	**0.002**	**0.264**	**0.009**	**0.264**	**0.009**
	pitch_std	**0.304**	**0.003**	-0.249	0.015	0.243	0.017	-0.190	0.064
	yaw_std	0.051	0.620	**-0.315**	**0.002**	0.033	0.750	**-0.265**	**0.009**
10	pitch_mean	-0.157	0.288	**0.483**	**0.001**	-0.229	0.117	**0.516**	**0.000**
	yaw_mean	0.318	0.027	0.327	0.023	0.299	0.039	0.242	0.098
	pitch_std	0.335	0.020	-0.237	0.105	0.272	0.062	-0.112	0.448
	yaw_std	0.038	0.799	**-0.441**	**0.002**	0.002	0.989	-0.325	0.024
20	pitch_mean	-0.167	0.436	**0.522**	**0.009**	-0.266	0.209	**0.585**	**0.003**
	yaw_mean	0.343	0.100	0.374	0.071	0.321	0.126	0.316	0.132
	pitch_std	0.272	0.199	-0.263	0.214	0.135	0.530	-0.145	0.500
	yaw_std	0.085	0.692	**-0.532**	**0.007**	0.020	0.927	-0.417	0.042
30	pitch_mean	-0.188	0.486	0.514	0.050	-0.287	0.282	0.560	0.024
	yaw_mean	0.381	0.145	0.363	0.167	0.350	0.184	0.333	0.208
	pitch_std	0.370	0.158	-0.237	0.377	0.283	0.288	-0.136	0.615
	yaw_std	0.173	0.522	-0.547	0.028	0.149	0.581	-0.380	0.146
60	pitch_mean	-0.194	0.645	0.517	0.189	-0.161	0.703	0.509	0.198
	yaw_mean	0.532	0.174	0.363	0.377	0.635	0.091	0.203	0.630
	pitch_std	0.304	0.464	-0.402	0.323	0.320	0.440	-0.416	0.306
	yaw_std	0.227	0.588	-0.673	0.067	0.183	0.664	-0.508	0.199

Table 1: Pearson's product-moment correlations for head movement data and continuous valence and arousal ratings. Significant values are shown in bold ($p < 0.01$).

20s segment (HVHA: 224, HVLA: 195, LVLA: 124, LVHA: 225); 30s segment (HVHA: 145, HVLA: 129, LVLA: 92, LVHA: 146); 60s segment (HVHA: 70, HVLA: 65, LVLA: 47, LVHA: 74). We observe that users annotated LVLA the least, across all segment sizes.

The mean and median of continuous V-A ratings are calculated for each segment. These continuous annotations were validated by our upcoming work [45], where we found that continuous V-A annotations are consistent with discrete within-VR and original stimuli ratings from Li et al. [25]. For HM and EM data, we calculated the mean and standard deviation (std) value of pitch and yaw. Since a Shapiro-Wilk test showed that these segment sequences are normally distributed ($p > 0.05$), we calculated Pearson's product-moment correlations between participants' HM/EM data and their continuous V-A ratings. Following prior work [25], this was done to explore statistical relationships between HM/EM and emotion labels. Since we conduct multiple correlation comparisons, results may be prone to a higher number of false positives (Type I errors) [37]. Using Bonferroni adjustment however is too conservative: while it lowers Type I errors, it can also increase Type II errors [6]. Given our exploratory work, we therefore lowered our alpha level from 0.05 to 0.01. As a cautionary measure, we also tested correction using the False Discovery Rate (FDR) [5] method, which has been shown to be more balanced. However results with FDR

Seg Length (s)	EM Data	Valence (Mean)		Valence (Median)		Arousal (Mean)		Arousal (Median)	
		Corr	p	Corr	p	Corr	p	Corr	p
5	pitch_mean	0.068	0.512	**-0.321**	**0.001**	0.136	0.185	**-0.354**	**0.000**
	yaw_mean	**0.299**	**0.003**	**0.275**	**0.007**	0.257	0.011	0.223	0.029
	pitch_std	**0.286**	**0.005**	-0.248	0.015	0.223	0.029	-0.166	0.106
	yaw_std	0.144	0.161	**-0.284**	**0.005**	0.123	0.234	-0.200	0.051
10	pitch_mean	0.073	0.624	-0.360	0.012	0.162	0.273	**-0.421**	**0.003**
	yaw_mean	0.314	0.030	0.288	0.047	0.293	0.043	0.200	0.174
	pitch_std	**0.395**	**0.005**	-0.307	0.034	0.339	0.018	-0.178	0.226
	yaw_std	0.124	0.400	**-0.418**	**0.003**	0.083	0.575	-0.276	0.058
20	pitch_mean	0.075	0.728	-0.390	0.060	0.191	0.370	-0.472	0.020
	yaw_mean	0.339	0.105	0.332	0.113	0.316	0.133	0.271	0.201
	pitch_std	0.305	0.147	-0.350	0.094	0.170	0.426	-0.194	0.363
	yaw_std	0.175	0.414	-0.429	0.036	0.108	0.615	-0.297	0.159
30	pitch_mean	0.091	0.738	-0.373	0.154	0.211	0.432	-0.443	0.085
	yaw_mean	0.383	0.144	0.318	0.230	0.351	0.182	0.282	0.290
	pitch_std	0.374	0.154	-0.354	0.178	0.268	0.315	-0.218	0.417
	yaw_std	0.239	0.373	-0.436	0.091	0.216	0.423	-0.248	0.354
60	pitch_mean	0.089	0.833	-0.382	0.350	0.093	0.826	-0.394	0.335
	yaw_mean	0.555	0.153	0.315	0.448	0.656	0.077	0.146	0.730
	pitch_std	0.381	0.351	-0.537	0.170	0.345	0.403	-0.525	0.181
	yaw_std	0.345	0.403	-0.586	0.127	0.282	0.499	-0.399	0.327

Table 2: Pearson's product-moment correlations for eye movement data and continuous valence and arousal ratings. Significant values are shown in bold ($p < 0.01$).

correction were less conservative than setting alpha to 0.01, so we report on results only considering a lower alpha level.

4.2 Key Findings

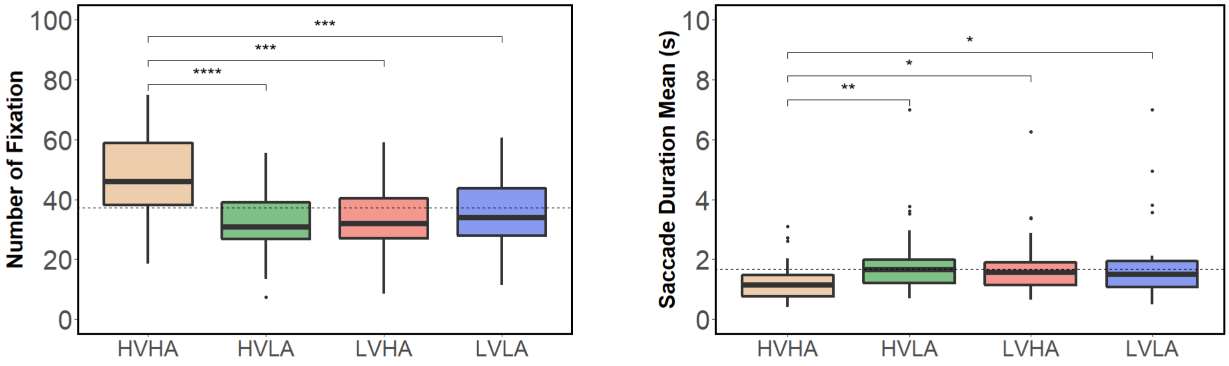
We follow recommendations (cf., psychology) [3] for determining correlation strength: (low: $0.1 < |corr| < 0.3$; moderate; $0.3 < |corr| < 0.6$; high: $0.6 < |corr| < 1.0$). For HM data, we found a moderate significant positive correlation between pitch mean and median of arousal ratings for 5, 10, and 20s segments. We found a moderate significant negative correlation between yaw std and valence median for 5, 10, and 20s segments. Correlations and corresponding p-values are shown in Table 1. For EM data, the results showed that there are moderate significant negative correlations between pitch mean and median of arousal ratings for 5 and 10s segments. There are low significant negative correlations between yaw std and valence median for 5s segments, and moderate significant negative correlations for 10s segments. Correlations and corresponding p-values are presented in Table 2. We also explored the effects of labelled video emotion on two EM features, fixation and saccade, two widely used eye features in affective computing and cognitive research [21, 39]. The number of fixations, fixation duration sum/std/mean, saccade duration mean/std are computed for each participant viewing each video. The results showed that there are significant differences on number of fixations and mean of saccade duration among different types of videos. The number of fixations is larger for HVHA videos than other video types, while the saccade duration is smaller for HVHA than others, as shown in Figure 3.

5 DISCUSSION

5.1 Time-segmented relationships between HM/EM and Valence and Arousal

First, for HM data, we found positive correlations between pitch mean and median of arousal ratings for 5, 10, and 20s segments, which suggests that participants usually raised their heads while reporting high arousal, and lowered their heads while reporting low arousal for the videos. This is consistent with Li et al. [25] who observed a similar effect. Also, Lhommet et al.'s [10] work indicated that people tend to move their head backwards during expressions of fear or surprise, which supports this finding. However, for EM data, the results showed that there are negative correlations between pitch mean and median of arousal ratings for 5 and 10s segments. Consistent with our earlier analysis, users seemed to often raise their heads when reporting high arousal. Given that primary content of 360° videos are displayed near the equator [2, 17], users usually look down when they raise their head, and up when they lower it.

Second, we found negative correlations between yaw std and valence medians for 5, 10, and 20s segments, which suggests that those who displayed greater side-to-side head movements reported lower ratings for valence. This negative relationship parallels research conducted by Won et al. [43] who showed a significant relationship between the amount of head yaw and reported anxiety, but contrasts with Li et al.'s [25] findings. One possible reason is that the video stimuli used by Li et al. are longer (> two minutes), where according to Li et al., participants simply viewed the content presented to them without the need for navigation. Furthermore,



(a) Fixation Number Mean

(b) Saccade Duration Mean

Figure 3: Boxplots for eye movement measures across emotion types. (**, p<0.0001; ***, 0.0001<p<0.001; **, 0.001<p<0.01; *, 0.01<p<0.05)**

their ratings were retrospective (i.e., post-stimuli), which may have had an effect. In our case, the videos were clipped to one minute segments, and users rated their emotions in real-time while exploring the content. Thus for each short segment (< 30 seconds), participants gave lower ratings of valence when they moved their heads for navigation. Similarly, there are negative correlations between yaw std and valence median for 5 and 10s segments for EM data. Finally, previous studies [39] have shown a significant impact of negative emotions on fixation and saccade features, with more visual agitation and avoidance behavior from larger, longer, and faster saccades. This is in line with our LVHA and LVLA findings, which lead to more fixation points and less saccade durations compared to HVHA videos. However for our HVLA videos, these were seaside (TI: 0.36) and snow mountain scenes (TI: 0.97) with relatively low temporal complexity, which may have prompted users to explore the scenes more freely. Thus there are smaller fixation points and longer saccadic durations for HVLA than HVHA.

5.2 Limitations and Future Work

Our work has some limitations and leaves open questions for future work: first, instead of dividing a video into fixed segments, we plan to select unmodified clips of differing lengths with emotion labels for direct testing. This would allow us to further test the effect of video length on emotions and HM/EM behavior. Second, we found that the shorter the segment duration, the more significant the correlation between HM/EM and reported emotions were. As we saw in Sec. 4.1, the shorter segment durations result in larger datasets to calculate correlations, which can result in more significant effects. Such a potential artifact warrants further scrutiny in future work.

Third, we do not consider the effects of video content and characteristics on 360° video watching (cf., spatial and temporal saliency maps [47]). While we have shown SI and PI per video, currently the segmentation sizes are fixed (e.g., 5s, 10s, ...) and not based on video content to warrant further analysis. For future work, we aim to investigate more closely the link between content analysis (what does the clip depict?), momentary emotion (how do users feel at that given moment?), and HM/EM data. Fourth, it is worthwhile to test the relationship between post-stimuli SAM ratings and HM/EM data, and then compare with continuous annotations. Fifth, one can consider more eye features for analysis, such as pupil size, blink, gaze location and direction, as these have been additionally shown to link with emotions [9, 14]. Finally, while here we looked at correlations, findings should be interpreted cautiously since we cannot make statements about the direction of the causal arrow: does experiencing and reporting emotion states result in certain HM/EM movements, or does performing certain HM/EM movements lead to observed differences in reported emotion states?

6 CONCLUSION

This early work provides the basis for further investigating the relationship between real-time and continuous emotion annotations, and time-segmented HM/EM data while users watch 360° videos. Our early findings contribute to a better understanding of the relationship between objective HM and EM measures tracked during VR-based HMD usage, and the momentary reported affective states during immersive 360° VR video watching experiences. Carrying out further research in this direction can help enable a low-cost means for improving users? Quality of Experience (QoE) during immersive VR interactions at a temporally fine-grained level, which can be used not only for improved processing, coding, delivering and rendering techniques, but also as a means to dynamically adapt displayed emotion content based on implicit user behavior during the viewing experience.

ACKNOWLEDGMENTS

This work is supported by the National Key Research and Development Program of China (No. 2020YFF0305200).

REFERENCES

[1] Andra Adams, Marwa Mahmoud, Tadas Baltrušaitis, and Peter Robinson. 2015. Decoupling facial expressions and head motions in complex emotions. In *2015*

International Conference on Affective Computing and Intelligent Interaction (ACII). IEEE, Xi'an, China, 274–280.

[2] Shahryar Afzal, Jiasi Chen, and KK Ramakrishnan. 2017. Characterization of 360-degree videos. In *Proceedings of the Workshop on Virtual Reality and Augmented Reality Network.* Association for Computing Machinery, New York, NY, USA, 1–6.

[3] Haldun Akoglu. 2018. User's guide to correlation coefficients. *Turkish Journal of Emergency Medicine* 18, 3 (2018), 91–93. https://doi.org/10.1016/j.tjem.2018.08.001

[4] Hillel Aviezer, Y. Trope, and A. Todorov. 2012. Body Cues, Not Facial Expressions, Discriminate Between Intense Positive and Negative Emotions. *Science* 338 (2012), 1225 – 1229.

[5] Yoav Benjamini and Yosef Hochberg. 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B: Methodological* 57, 1 (1995), 289–300.

[6] Yoav Benjamini and Daniel Yekutieli. 2001. The Control of the False Discovery Rate in Multiple Testing Under Dependency. *Ann. Stat.* 29 (08 2001). https://doi.org/10.1214/aos/1013699998

[7] 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.

[8] 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.* Association for Computing Machinery, New York, NY, USA, 762–768.

[9] Manuel G Calvo and Peter J Lang. 2004. Gaze patterns when looking at emotional pictures: Motivationally biased attention. *Motivation and Emotion* 28, 3 (2004), 221–243.

[10] R Calvo, S D'Mello, J Gratch, A Kappas, M Lhommet, and SC Marsella. 2015. Expressing emotion through posture and gesture. , 13 pages.

[11] 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.

[12] 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*, Vol. None. ISCA-speech, Newcastle, Northern Ireland, UK, None.

[13] 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.* Association for Computing Machinery, New York, NY, USA, 432–437.

[14] Jakob De Lemos, Golam Reza Sadeghnia, Íris Ólafsdóttir, and Ole Jensen. 2008. Measuring emotions using eye tracking. In *Proceedings of measuring behavior*, Vol. 226. Proceedings of measuring behavior, (Maastricht, The Netherlands, 225–226.

[15] 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

[16] P. Ekman and W. V. Friesen. 1967. Head and body cues in the judgment of emotion: A reformulation.. In *Perceptual and Motor Skills*, Vol. 246. Sage Publications, Los Angeles, CA, 711–724.

[17] Stephan Fremerey, Ashutosh Singla, Kay Meseberg, and Alexander Raake. 2018. AVtrack360: An open dataset and software recording people's head rotations watching 360° videos on an HMD. In *Proceedings of the 9th ACM Multimedia Systems Conference.* Association for Computing Machinery, New York, NY, USA, 403–408.

[18] 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.

[19] M Melissa Gross, Elizabeth A Crane, and Barbara L Fredrickson. 2010. Methodology for assessing bodily expression of emotion. *Journal of Nonverbal Behavior* 34, 4 (2010), 223–248.

[20] Jesús Gutiérrez, Erwan J David, Antoine Coutrot, Matthieu Perreira Da Silva, and Patrick Le Callet. 2018. Introducing un salient360! benchmark: A platform for evaluating visual attention models for 360 contents. In *2018 Tenth International Conference on Quality of Multimedia Experience (QoMEX).* IEEE, Cagliari, 1–3.

[21] Omid Kardan, Marc G Berman, Grigori Yourganov, Joseph Schmidt, and John M Henderson. 2015. Classifying mental states from eye movements during scene viewing. *Journal of Experimental Psychology: Human Perception and Performance* 41, 6 (2015), 1502.

[22] M. Karg, A. Samadani, R. Gorbet, K. Kühnlenz, J. Hoey, and D. Kulić. 2013. Body Movements for Affective Expression: A Survey of Automatic Recognition and Generation. *IEEE Transactions on Affective Computing* 4, 4 (2013), 341–359. https://doi.org/10.1109/T-AFFC.2013.29

[23] Hitoshi Kusano, Yuji Horiguchi, Yukino Baba, and Hisashi Kashima. 2020. Stress Prediction from Head Motion. In *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA).* IEEE, Sydney, NSW, Australia, 488–495.

[24] Margaux Lhommet and Stacy C Marsella. 2014. Expressing emotion through posture. *The Oxford handbook of affective computing* 273 (2014), 1085–1101.

[25] 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.

[26] Steven R Livingstone and Caroline Palmer. 2016. Head movements encode emotions during speech and song. *Emotion* 16, 3 (2016), 365.

[27] 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.* Association for Computing Machinery, New York, NY, USA, 211–216.

[28] 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 (2008), e1897.

[29] Tiago Oliveira, Paulo Noriega, Francisco Rebelo, and Regina Heidrich. 2017. Evaluation of the relationship between virtual environments and emotions. In *International Conference on Applied Human Factors and Ergonomics.* Springer International Publishing, Cham, 71–82.

[30] Yashas Rai, Jesús Gutiérrez, and Patrick Le Callet. 2017. A dataset of head and eye movements for 360 degree images. In *Proceedings of the 8th ACM on Multimedia Systems Conference.* Association for Computing Machinery, New York, NY, USA, 205–210.

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

[32] Atanu Samanta and Tanaya Guha. 2017. On the role of head motion in affective expression. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, New Orleans, LA, 2886–2890.

[33] Sanneke J Schouwstra and Johan Hoogstraten. 1995. Head position and spinal position as determinants of perceived emotional state. *Perceptual and motor skills* 81, 2 (1995), 673–674.

[34] 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* 11, 1 (2017), 78–84.

[35] 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.

[36] Mel Slater, John McCarthy, and Francesco Maringelli. 1998. The influence of body movement on subjective presence in virtual environments. *Human factors* 40, 3 (1998), 469–477.

[37] George Davey Smith and Shah Ebrahim. 2002. Data dredging, bias, or confounding: They can all get you into the BMJ and the Friday papers.

[38] 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.

[39] 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, Athlone, Ireland, Ireland, 1–6.

[40] Harald G Wallbott. 1998. Bodily expression of emotion. *European journal of social psychology* 28, 6 (1998), 879–896.

[41] Séamas Weech, Sophie Kenny, and Michael Barnett-Cowan. 2019. Presence and Cybersickness in Virtual Reality Are Negatively Related: A Review. *Frontiers in Psychology* 10 (2019), 158. https://doi.org/10.3389/fpsyg.2019.00158

[42] Alex Wiebe, Anette Kersting, and Thomas Suslow. 2017. Deployment of attention to emotional pictures varies as a function of externally-oriented thinking: An eye tracking investigation. *Journal of behavior therapy and experimental psychiatry* 55 (2017), 1–5.

[43] Andrea Stevenson Won, Brian Perone, Michelle Friend, and Jeremy N Bailenson. 2016. Identifying anxiety through tracked head movements in a virtual classroom. *Cyberpsychology, Behavior, and Social Networking* 19, 6 (2016), 380–387.

[44] Tong Xue, Abdallah El Ali, Gangyi Ding, and Pablo Cesar. 2020. Annotation Tool for Precise Emotion Ground Truth Label Acquisition while Watching 360° VR Videos. In *2020 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR).* IEEE, Utrecht, Netherlands, 371–372.

[45] 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 *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems.* Association for Computing Machinery, Yokohama, Japan, 1–15.

[46] 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.* Association for Computing Machinery, New York, NY, USA, 1–9.

[47] Yun Zhai and Mubarak Shah. 2006. Visual Attention Detection in Video Sequences Using Spatiotemporal Cues. In *Proceedings of the 14th ACM International Conference on Multimedia* (Santa Barbara, CA, USA) *(MM '06).* Association for Computing Machinery, New York, NY, USA, 815?824. https://doi.org/10.1145/1180639.1180824

[48] 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*. Association for Computing Machinery, New York, NY, USA, 1–15.