PEM360: A dataset of 360° videos with continuous Physiological measurements, subjective Emotional ratings and Motion traces

Quentin Guimard*
quentin.guimard@univ-cotedazur.fr
Université Côte d'Azur, CNRS, I3S
Sophia-Antipolis, France

Florent Robert*
florent.robert@inria.fr
Université Côte d'Azur, CNRS, Inria, I3S
Sophia-Antipolis, France

Camille Bauce
Université Côte d'Azur, CNRS, I3S
Sophia Antipolis, France

Aldric Ducreux
Université Côte d'Azur, CNRS, I3S
Sophia Antipolis, France

Lucile Sassatelli
Université Côte d'Azur, CNRS, I3S
Institut Universitaire de France
Sophia-Antipolis, France

Hui-Yin Wu
Université Côte d'Azur, Inria
Sophia-Antipolis, France

Marco Winckler
Université Côte d'Azur, CNRS, Inria, I3S
Sophia-Antipolis, France

Auriane Gros
Université Côte d'Azur, CHU de Nice, CoBTeK
Nice, France

ABSTRACT

From a user perspective, immersive content can elicit more intense emotions than flat-screen presentations. From a system perspective, efficient storage and distribution remain challenging, and must consider user attention. Understanding the connection between user attention, user emotions and immersive content is therefore key. In this article, we present a new dataset, PEM360 of user head movements and gaze recordings in 360° videos, along with self-reported emotional ratings of valence and arousal, and continuous physiological measurement of electrodermal activity and heart rate. The stimuli are selected to enable the spatiotemporal analysis of the connection between content, user motion and emotion. We describe and provide a set of software tools to process the various data modalities, and introduce a joint instantaneous visualization of user attention and emotion we name *Emotional maps*. We exemplify new types of analyses the PEM360 dataset can enable. The entire data and code are made available in a reproducible framework.

CCS CONCEPTS

• **Human-centered computing** → **Virtual reality**; *User studies*.

KEYWORDS

360° videos, user experiment, emotions, physiological data

*Both authors contributed equally to this research.

MMSys '22, June 14–17, 2022, Athlone, Ireland

ACM ISBN 978-1-4503-9283-9/22/06…$15.00
https://doi.org/10.1145/3524273.3532895

ACM Reference Format:
Quentin Guimard, Florent Robert, Camille Bauce, Aldric Ducreux, Lucile Sassatelli, Hui-Yin Wu, Marco Winckler, and Auriane Gros. 2022. PEM360: A dataset of 360° videos with continuous Physiological measurements, subjective Emotional ratings and Motion traces. In *13th ACM Multimedia Systems Conference (MMSys '22), June 14–17, 2022, Athlone, Ireland.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3524273.3532895

1 INTRODUCTION

Immersive media is rising as new types of multimedia experiences are becoming more accessible over a wide spectrum of applications, e.g., virtual reality (VR) for games, training and rehabilitation, immersive 360° videos for journalism and new forms of storytelling, and advanced interaction in the *Metaverse*[1]. Increased accessibility is driven by affordability of VR equipment, and intense effort to design efficient distribution methods in constrained network conditions. Both the design and the efficient distribution of immersive experiences remain challenges, however, for two main reasons. First, it has been shown that immersive content can elicit more intense emotions than flat-screen presentation [2, 9, 19]. Second, compression and streaming decisions must be driven by quality of experience (QoE) models [7]. These models are dependent on user attention prediction. For example, Xu et al. [25] considered content saliency and field of view (FoV) preferences to extend the PSNR and SSIM metrics to 360° content. It is therefore key to understand the attentional and emotional processes of users in an immersive environment, and how these processes are associated with the content.

To enable the study of this connection between user attention, user emotions and immersive visual content, this article introduces:

- PEM360, a new dataset of user head movements and gaze recordings in 360° videos, along with self-reported emotional ratings of valence and arousal, and continuous physiological

[1] https://www.cnbc.com/2021/12/27/metas-oculus-virtual-reality-headsets-were-a-popular-holiday-gift.html

measurement of electrodermal activity and heart rate. The stimuli are selected based on high-level and low-level content saliency to enable the spatiotemporal analysis of the connection between content, user motion and emotion.
- a set of software tools to pre-process the data of gaze, electrodermal activity and content, and to visualize jointly instantaneous heat maps of gaze and arousal level superimposed on the frame, which we name *Emotional maps*.
- a preliminary analysis validating the data and verifying known results, and examples of new connections that can be investigated.

The entire collection of artifacts is presented as Python tools and notebooks to enable reproducibility of the data processing. The dataset and tools are now available in a public GitLab repository[2].

2 EXISTING DATASETS

While a number of works in the domain of human-computer interaction and cognitive sciences have studied emotions in immersive environments [2, 9, 19, 24], it has been only more recently that the multimedia community in particular has considered sensing and recording emotions along with head and gaze motion [16, 20, 22, 26, 27]. Human emotions are commonly decomposed along two main dimensions: valence, representing the negative or positive nature of an emotion (unpleasant-pleasant), and arousal, representing the intensity of the perceived emotion (calm-excited) [3].

Li et al. [16] introduced the first reference database obtained from 95 users freely watching 73 videos in 360° who provided their valence and arousal ratings after every clip using the self-assessment manikin (SAM) tool [5]. Their head positions were recorded. The material publicly available however only consists of the videos and the average valence and arousal value pair for each video (averaged over all users). Subsequently, Tang et al. [20] presented an experiment where 19 users watched 36 images in 360° while their self-reported emotions and eye motion were collected.

More recently, it has been shown that the capability of a single rating issued after experiencing the 360° content is limited and unable to fully represent the variations of user state [24]. This is why new tools have been introduced to enable the continuous collection of self-reports inside the immersive environment [22, 27]. The data collected by Toet et al. [22] and Xue et al. [27] also comprise physiological measurements of heart rate and electrodermal activity (EDA, as skin conductance), which has been shown to reliably represent user instantaneous arousal [4]. Toet et al. [22] presented a new emotions rating tool, named EmojiGrid, tested on 40 users viewing 62 videos from the reference database of Li et al. [16]. While they provide the per-user per-video valence and arousal ratings, only time averages are made available for EDA, and no gaze or head motion traces. Xue et al. [27] introduced a continuous grading tool of valence and arousal. They provide a dataset of 11 immersive videos from the same database [16] experienced by 32 users. Subjective emotional ratings, physiological measurements (including EDA) and head and gaze movements are continuously collected and made available.

This latter work [27] is closer to ours and has been made partly concurrently. Our dataset is however complementary and enables other types of studies. We provide EDA streams at a higher rate, acquired at 16Hz, compared to 4Hz in [27]. In the aforementioned objective of understanding the connection between attended regions and instantaneous emotions, like arousal, it is important to enable the detection of several peaks of the phasic component of EDA per second, requiring hence a higher acquisition rate. Also, we sample seven video stimuli from the same reference database [16] for our experiments, so that specific criteria on saliency are met, as detailed in Sec. 3.1. Out of the seven videos, six differ from the videos selected by Xue et al. [27]. Our dataset therefore enriches the existing datasets and enables extensive analysis to gain new insights on the connection between attention, emotion and content.

[2] https://gitlab.com/PEM360/PEM360/

3 USER EXPERIMENT AND DATASET DESCRIPTION

We conducted a controlled, indoor laboratory experiment where users watched 360° videos in a VR headset. We collected eye movement (EM), head movement (HM), heart rate (HR) and skin conductance (EDA) data as well as emotion annotations of valence and arousal. The user experiment has been approved by the university ethics committee.

3.1 Stimuli

The videos are selected to enable several levels of content analysis and description, to correlate with user motion and emotion. User attention in relation with the visual content is described with saliency maps, obtained either from gaze locations, or estimated from the content. Here we consider two levels of content description as two types of saliency maps, and select the videos so that for each, the overlap between both saliency maps is limited. Specifically, we consider low-level (LL) and high-level (HL) saliency. Low-level saliency maps are made up of a combination of colors, intensity and orientations as defined by Itti et al. [11]. Since we are dealing with videos and not images, we combine this definition with the one of optical flow [10], because we also consider motion in the video to be part of the low-level saliency. High-level saliency maps are composed of high-level semantic features, such as faces, cars, or animals. Inspiring from Chopra et al. [8], high-level saliency is obtained from YOLOv4 object detector, with object bounding boxes being used as binary saliency maps.

We selected 7 videos from the reference database of Li et al. [16]. The selected videos should have a range of valence and arousal as wide as possible of level of valence and arousal, and the LL saliency should be evenly distributed both within and outside object bounding boxes characterizing HL saliency. To select these videos we compare (i) the number of pixels inside and outside objects, and (ii) the per-pixel LL saliency (ranging between 0 and 255), computed as the total LL saliency inside and outside objects normalized with the corresponding number of pixels. Fig. 1 demonstrates this in videos 13 and 73. The number of pixels with such minimum LL saliency inside and outside objects is equivalent over time, as is the per-pixel LL saliency in both areas. Fig. 2 shows a frame where regions with high LL saliency can be seen outside of the detected objects. Table 1 lists the video details.

ID	Valence	Arousal	Start (s)	End (s)	Duration (s)	YouTubeID
12	7	4.6	5	103	98	T-aOVE22lEw
13	4.92	4.08	4	131	127	GJGfxfGEa9Y
17	5.22	5	5	69	64	g7btxyIbQQ0
23	7.2	3.2	8	143	135	CDfsFuDuHds
27	6	1.6	60	180	120	QxxXu_B–ZA
73	6.27	6.18	9	70	61	bUiP-iGN6oI
32	6.57	1.57	40	130	90	-bIrUYM-GjU

Table 1: Details of selected videos. Videos are accessible at youtube.com/watch?v=[YouTubeID].

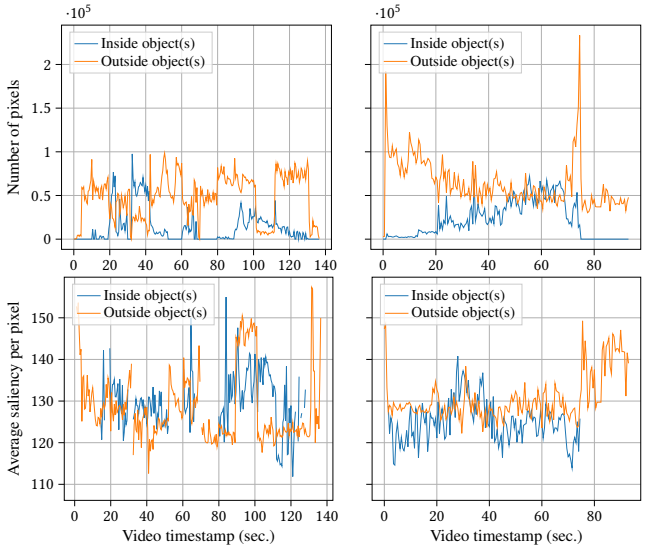

Figure 1: HL and LL saliency characterization of video 13 (left) and video 73 (right). Top: number of pixels inside and outside objects. Bottom: average LL saliency per pixel inside and outside objects.

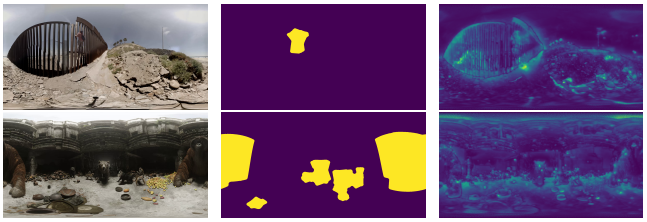

Figure 2: HL and LL saliency visualization for frame 2145 of video 13 (top) and frame 3630 of video 73 (bottom). Left: the frame. Center: HL saliency (detected objects, human on top, animals at bottom). Right: LL saliency.

3.2 Equipment

Recordings of head and eye movements have been made with a FOVE headset, equipped with an eye-tracker with a 120Hz acquisition rate, and tethered to a desktop computer. A Unity3D scene was used with a 360° sphere object to display the videos. We use the FOVE Unity plugin to record head and gaze positions.

Recordings of EDA and optical pulse have been made with a Shimmer3 GSR+ sensor with a frequency range of 15.9Hz and 51.2Hz, respectively. All of the measurements were resampled to 100Hz for analysis. The apparatus is depicted in Fig. 3.

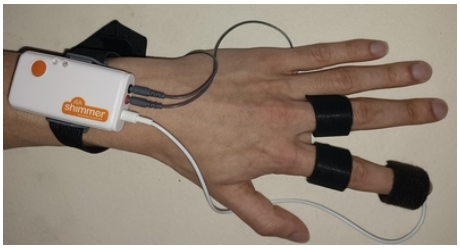

Figure 3: Shimmer3 GSR+ used to record EDA and optical pulse. Gray wires connect the EDA sensor, white wire connects the pulse sensor.

3.3 Participants

The experiment was carried out with a total of 34 users, in which 31 had complete data (10 women, 20 men, 1 non-binary; 18-29 years old, M=24, SD=3.26). 19 of them had a normal vision, 9 had corrected to normal vision and 3 did not have a normal vision. Most of them played games but rarely or never in VR, and the majority have seen only one or two 360° videos before the experiment. Participants received monetary compensation for their time. The seven videos were experienced by all 31 users for their entire duration (60 to 135 seconds, see Table 1).

3.4 Procedure

The lab experiment started with a pre-questionnaire assessing the user's background with VR and checking for visual deficiencies. Eye tracking calibration was done using the FOVE software for each user before beginning the experiment to make sure the eye tracking data is properly recorded. The VR experiment systematically started with a low-arousal (relaxing) video (ID 32) to bring EDA and HR levels to a user-relative baseline. The remaining six VR videos were then experienced in a random order by every user. Users were in standing position during the experience and could freely explore in 360° while holding the back of a chair to maintain balance and orientation. The videos were played without audio. After each viewing, the headset was removed and the SAM scale presented for arousal and valence rating. At least a 1-min break outside of the headset was observed between videos.

3.5 Dataset structure

The resulting dataset PEM360 is provided with the structure shown in Fig. 4. The `raw_data` folder contains 34 folders, one for each user. User folders contain a Shimmer CSV file containing the EDA and optical pulse data recorded over all the 360° videos experienced by the user, and seven CSV files, one per video, containing the gaze and head motion data recorded during the corresponding video.

Entries in the CSV files include system timestamps to synchronize the data modalities for analysis.

Valence and arousal ratings of each user for each video are stored in the root folder under `graded_valence_arousal.csv`. Finally, the root folder `PEM360` also contains the Python Jupyter notebook providing the software tools described in Sec. 4, and the entire data processing workflow to reproduce the analysis presented in Sec. 5.

```
PEM360/
├─ emotional_maps/                ┆
│  ├─ compute_emotional_map.py    ├─ saliency_maps/
│  ├─ requirements.txt            │  ├─ compute_gt_saliency.ipynb
│  └─ utils_shimmer.py            │  ├─ compute_hl_saliency.py
├─ raw_data/                      │  ├─ compute_ll_saliency.py
│  ├─ LICENSE                     │  ├─ nfov.py
│  ├─ user_01/                    │  ├─ pySaliencyMap.py
│  │  ├─ shimmer.csv              │  ├─ pySaliencyMapDefs.py
│  │  ├─ video_12.csv             │  └─ requirements.txt
│  │  ├─ video_13.csv             ├─ LICENSE
│  │  ├─ video_17.csv             ├─ README.md
│  │  ├─ video_23.csv             ├─ design_of_experiments.csv
│  │  ├─ video_27.csv             ├─ graded_valence_arousal.csv
│  │  ├─ video_32.csv             ├─ notebook.ipynb
│  │  └─ video_73.csv             ├─ requirements.txt
│  ├─ user_02/                    └─ resample_data.ipynb
```

Figure 4: Folder structure of the dataset with main files.

4 PRE-PROCESSING SOFTWARE

Along with the data, we provide a Jupyter notebook to reproduce the entire processing of head and gaze data, EDA, ratings of valence and arousal, and the code to produce saliency maps from the content.

4.1 Processing gaze data

For both HM and EM, 3D positions are logged in Cartesian coordinates $(x, y, z) \in \mathbb{R}^3$. We provide functions:

- to convert the positions from Cartesian to Eulerian (ϕ, θ, ψ) denoting respectively yaw, pitch and roll,
- to obtain speed and acceleration over yaw and pitch,
- to obtain global speed and acceleration by computing the derivatives of the orthodromic distance,
- to represent rotational motion with quaternions (hence enabling to compute non-linear motion on the sphere as changes in quaternion rotational axis).

4.2 Processing EDA data

The EDA signal is the raw measurement of skin electrical conductance in micro-Siemens (μS). Two main components can be distinguished in an EDA signal [4, 6]: the tonic level, also called skin conductance level (SCL), varies slowly and represents slow autonomic changes that may not be associated with stimulus presentation; and the phasic level, which represents faster changes in EDA, and can better reflect the impact of successive stimuli. Raw EDA, phasic and tonic components are shown in Fig. 5-top and 5-center. We use the Python toolbox Neurokit [17] to process EDA data, which uses the cvxEDA method to extract the phasic component. Finally, the physiological arousal to be analyzed in connection with experimental stimuli can be assessed from several metrics on the phasic level, such as peak frequency, duration and amplitude. This is called the skin conductance response (SCR), and can be defined in several ways. In our code, we choose to compute instantaneous SCR as the absolute value of the first-order time derivative of the phasic component, shown in 5-bottom. Note however that the code can easily be modified to implement other definitions of SCR from the phasic component. The obtained SCR is therefore a time series for every user-video pair. This enables analysis with SCR averaged over time for each such pair (as often done), or on a time-dependent basis.

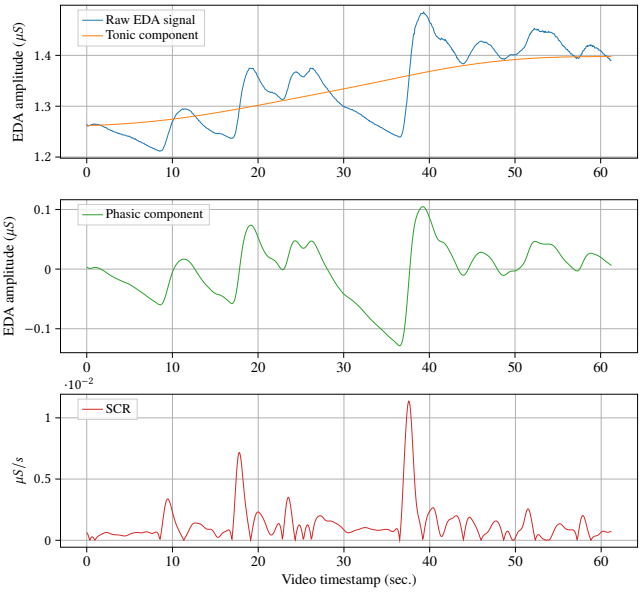

Figure 5: EDA signal recorded for user 03 while watching video 73. The three graphs from the top show the raw EDA data and the tonic component, the phasic component and the SCR (absolute value of phasic first derivative).

4.3 Processing video content

As introduced in Sec. 3.1, we use LL and HL saliency models designed for regular flat images. We therefore apply them on FoV projections of the entire frame. We first uniformly sample 100 points on the unit sphere and project them on the equirectangular frame using the equirectangular-toolbox [18]. Each "patch" is made of a projection centered on one of these points, it is a 512x512 image corresponding to a 108°x108° FoV. These patches can overlap each other and are separately given to the appropriate models for both LL and HL saliency. For LL saliency, we use a Python implementation of Itti's saliency map [13], which also allows the combination of Itti saliency with the optical flow between consecutive frames, which we do by using separate extractors for each patch. For HL saliency, we use the TensorFlow 2 implementation of YOLOv4 [21]. For each patch given to the YOLO model, we create a binary saliency map equal to 1 inside the bounding boxes of the detected objects. For both LL and HL saliency, the overlapping patches are back-projected by addition onto the equirectangular frame to obtain a single (LL or

HL) saliency map per frame. For LL saliency, the back-projection is normalized by dividing the value of each pixel by the number of patches it belongs to. The final value of a given pixel is the average over all existing projections for this pixel. For HL saliency, the back-projection is normalized by clipping the value of each pixel between 0 and 1. The final value of a given pixel is the maximum over all existing projections for this pixel. Finally, the saliency maps are downscaled by a factor of 5 both horizontally and vertically (from 1920x1080 to 384x216) for storage space reasons. The LL saliency is downscaled using average pooling over blocks of 5x5 pixels, whereas HL saliency is downscaled using max pooling over blocks of the same size. The files are stored in HDF5 format and can be accessed from a link given in the article's repository mentioned in Sec. 1, but can also be re-computed from the provided code.

4.4 Instantaneous visualization of gaze and emotions: *Emotional maps*

Figure 6: *Emotional map* visualizing instantaneous gaze locations (luminance) and user arousal (from blue to red for low to high SCR). Example with high arousal in a roller-coaster video.

As previously discussed, the stimuli choice and experimental procedure are designed to collect data enabling a time-dependent analysis of the connection between attention, emotion and content. That is why we provide a tool for the experimenter to play the 360° video and visualize the instantaneous gaze locations and arousal (SCR) of a given user from the recorded data. This tool implements a new way of visualizing arousal in connection with gaze, which we name *emotional maps*. An *emotional map* is a 4D-array represented as a frame where:

- pixel luminance reflects the time the user spent attending the area over a past window of T seconds. A Gaussian kernel of parameter σ is convolved with every gaze location, and accumulated over the sliding window of T seconds. A bright (resp. dark) area can therefore reflect a fixation (resp. a saccade).
- pixel color represents the user's SCR, from blue (low arousal) to red (high arousal).

Emotional maps generated from a record with our tool are accumulated into videos. Each point persists on the video for P seconds, creating a trail to easily visualize the gaze path and arousal changes. An example of such a video frame is shown in Fig. 6[3]. The script `compute_emotional_map.py` creates the emotional maps and blends them with the frames to produce the resulting video visualization from records of gaze and EDA data. We believe this tool can lead to important qualitative insights for diverse disciplines (including neuroscience) on the connection between visual attention and emotion.

[3]Demonstration of a resulting video is accessible at https://tinyurl.com/25vjwk2s.

5 PRELIMINARY ANALYSIS OF THE DATA

In this section we first verify the validity of our data and correspondence with the original dataset and between arousal and EDA. We then exemplify possible analyses of correlation between motion and emotion, and between attention, content saliency and emotion.

5.1 Data validation

Reliability of the collected ratings. We verify the reliability of the collected arousal and valence by assessing the similarity of the user ratings for each video. This is achieved with the intra-class correlation coefficient (ICC), with classes corresponding to the 360° videos. ICC estimates based on mean ratings with a two-way mixed effects model are 0.96 (95% CI 0.87-0.99) for arousal and 0.88 (95% CI 0.72-0.98) for valence. According to Koo and Li's guidelines [14], this is excellent and good inter-rater agreement, respectively.

Agreement between collected ratings and original dataset. Fig. 7 shows the valence and arousal ratings of our users as a boxplot for each video, along with a red dot representing the corresponding average values available in the original dataset [16]. We observe the good agreement between both sets, as the latter are all the times but one in the inter-quartile value range of our data. We also compute the median of the root square difference of averages of our valence and arousal ratings with the corresponding averages from [16]. This median is 1.17 (within a range of 1 to 9), showing the agreement between both.

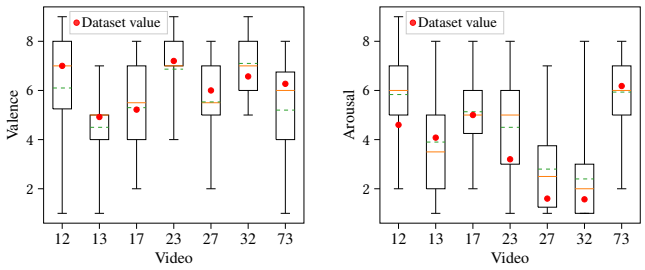

Figure 7: Arousal and valence ratings by users for each videos. The green dotted line corresponds to the mean and the orange solid line to the median.

5.2 Connecting EDA with graded arousal

We investigate the correspondence between SCR and arousal ratings. We gather the average SCR values $\mathrm{SCR}_{u,v}$ for every pair (u, v) of user u and video v, and corresponding graded arousal $\mathrm{GA}_{u,v}$. First, we average both variables over all users for every video, and obtain seven sample pairs $(\mathrm{GA}_v, \mathrm{SCR}_v)$, shown in Fig. 8-left. We verify as did Toet et al. [22] that the video ranking according to mean graded arousal is similar to the video ranking according to mean SCR. We also compute the Spearman correlation coefficient (CC) between GA_v and SCR_v for all seven videos. The Spearman CC between

mean graded arousal and mean SCR is $(0.92, p = 0.003)$. According to [12, appx. 6C, p. 79], such level of correlation is significant ($\alpha = 0.05$, $\beta = 0.2$) from 7 samples (see [23]).

We then consider the 217 sample pairs $(\text{GA}_{u,v}, \text{SCR}_{u,v})$. It is interesting to observe that the Pearson or Spearman CCs do not show any correlation between these pairs. Looking more closely at the data, we identify that the mean level of SCR per user, $\text{SCR}_u = \mathbb{E}_v[\text{SCR}_{u,v}]$ (averaged over all videos), varies significantly over the users ($M = 6.0e-4, S = 6.2e-4$). With the rationale that the *excitability* of a user is person-dependent and impacts the absolute SCR values, we verify whether the SCR variations relative to this individual's mean are better associated with graded arousal. To do so, we define centered SCR as $\text{cSCR}_{u,v} = \text{SCR}_{u,v} - \text{SCR}_u$, and do the same with graded arousal $\text{cGA}_{u,v} = \text{GA}_{u,v} - \text{GA}_u$. Fig. 8-right represents the scatter plot of $\text{cSCR}_{u,v}$ against $\text{cGA}_{u,v}$. The Spearman CC between both is $(0.25, p < 0.001)$.

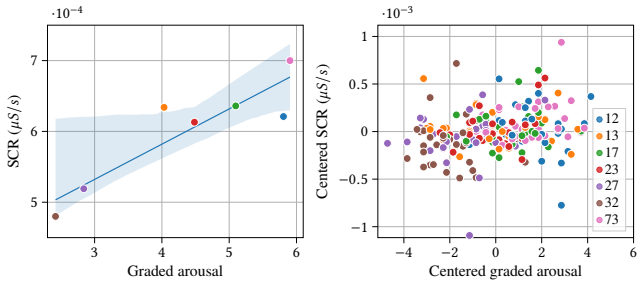

Figure 8: Dots colors code for video ID (legend on the right). Left: Scatter plot of SCR_v against GA_v. The shaded area represents the 95% CI of the linear regressor (solid blue line). Right: Scatter plot of $\text{cSCR}_{u,v}$ against $\text{cGA}_{u,v}$.

5.3 Analysis of correlations between attention, emotion and content

This section exemplifies exploratory analysis of correlations between attention (with head or gaze movements), emotion (with valence, arousal or SCR) and content (described with HL or LL saliency).

We first show examples of correlation between head motion and arousal, as already partly observed in other works [16, 26]. With the same centering process as above, we center the absolute pitch value. A mean absolute pitch at 0 means the user constantly keeps their head in the equatorial position. The centered absolute pitch value for pair (u, v) therefore indicates how much user u deviates in video v from their average slant. We can show that the Pearson CC between centered absolute head pitch and centered graded arousal is $(0.42, p < 0.001)$. Also, the Pearson CC between centered head speed (in rad/s) and centered graded arousal is $(0.26, p < 0.001)$. These results are examples of associations that can then be investigated more ahead (looking at the confounding, mediating or interacting factors for example).

Second, we exemplify how content description with HL and LL saliency can be leveraged to investigate the association between visual attention, content and emotion. We compare how well both types of saliency maps match the users' fixations over every frame of the 360° video. To do so, we compute the normalized scanpath saliency (NSS), that measures the amount of saliency around fixations [15]. We consider segments of 5 sec. to average the saliency maps of all frames in this interval, and aggregate the user's fixations in this interval, hence obtaining an NSS value for both saliency types $\text{NSS}^{HL}_{u,v,i}$ and $\text{NSS}^{LL}_{u,v,i}$ for every user u, video v, interval i. The averages over intervals are denoted $\text{NSS}_{u,v}$.

To study the relationship between both types of saliency and the user's arousal, we consider in Fig. 9 the difference $\text{NSS}^{Diff}_{u,v} = \text{NSS}^{HL}_{u,v} - \text{NSS}^{LL}_{u,v}$ plotted against mean-centered skin conductance response $\text{cSCR}_{u,v}$ (left) and graded arousal $\text{GA}_{u,v}$ (right) for all u, v, the points being colored per video. The major finding is the increasing trend of NSS^{Diff} with EDA and graded arousal. Specifically, the Pearson CC between NSS^{Diff} and EDA cSCR is $(0.25, p < 0.001)$, and the Pearson CC between NSS^{Diff} and graded arousal $\text{GA}_{u,v}$ is $(0.41, p < 0.001)$. There is therefore a moderate significant correlation between $\text{NSS}^{Diff}_{u,v}$ and $\text{cSCR}_{u,v}$ [1], meaning that HL saliency is more predictive of the attention in higher arousal states.

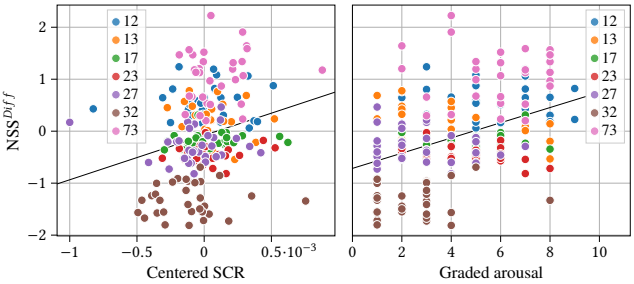

Figure 9: $\text{NSS}^{Diff}_{u,v}$ against $\text{cSCR}_{u,v}$ and $\text{GA}_{u,v}$ for all users u and videos , v. The black line shows a linear regression model fitted on the data.

6 CONCLUSION

In this article, we have presented the new PEM360 dataset of 360° videos with continuous physiological measurements, subjective emotional ratings and user motion traces. The stimuli are selected to enable investigating the spatiotemporal connection between user attention, user emotions and visual content. We have described the data collection process, the pre-processing workflow of the different data modalities, and exemplified some possible novel types of analyses to demonstrate the potential insights that can be drawn from PEM360. The artifacts are made available in a reproducible framework based on notebooks.

ACKNOWLEDGMENTS

This work has been partly supported by the French government, through the UCA JEDI and EUR DS4H Investments in the Future projects ANR-15-IDEX-0001 and ANR-17-EURE-0004. This work was partly supported by EU Horizon 2020 project AI4Media, under contract no. 951911 (https://ai4media.eu/).

REFERENCES

[1] 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

[2] Rosa María Baños, Cristina Botella, Isabel Rubió, Soledad Quero, Azucena García-Palacios, and Mariano Luis Alcañiz Raya. 2008. Presence and Emotions in Virtual Environments: The Influence of Stereoscopy. *Cyberpsychology & behavior : the impact of the Internet, multimedia and virtual reality on behavior and society* 11 1 (2008), 1–8.

[3] Lisa Feldman Barrett. 1998. Discrete emotions or dimensions? The role of valence focus and arousal focus. *Cognition and Emotion* 12, 4 (1998), 579–599. https://doi.org/10.1080/026999398379574 Place: United Kingdom Publisher: Taylor & Francis.

[4] Wolfram Boucsein. 2012. *Electrodermal activity, 2nd ed.* Springer Science + Business Media, New York, NY, US. https://doi.org/10.1007/978-1-4614-1126-0 Pages: xviii, 618.

[5] 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. https://doi.org/10.1016/0005-7916(94)90063-9

[6] Jason J. Braithwaite, Diana Patrícia Zethelius Watson, Roland S. G. Jones, and Michael A. Rowe. 2013. Guide for Analysing Electrodermal Activity & Skin Conductance Responses for Psychological Experiments. *CTIT technical reports series* (2013).

[7] Kjell Brunnström, Sergio Ariel Beker, Katrien De Moor, Ann Dooms, Sebastian Egger, Marie-Neige Garcia, Tobias Hossfeld, Satu Jumisko-Pyykkö, Christian Keimel, Mohamed-Chaker Larabi, Bob Lawlor, Patrick Le Callet, Sebastian Möller, Fernando Pereira, Manuela Pereira, Andrew Perkis, Jesenka Pibernik, Antonio Pinheiro, Alexander Raake, Peter Reichl, Ulrich Reiter, Raimund Schatz, Peter Schelkens, Lea Skorin-Kapov, Dominik Strohmeier, Christian Timmerer, Martin Varela, Ina Wechsung, Junyong You, and Andrej Zgank. 2013. Qualinet White Paper on Definitions of Quality of Experience. https://hal.archives-ouvertes.fr/hal-00977812 Qualinet White Paper on Definitions of Quality of Experience Output from the fifth Qualinet meeting, Novi Sad, March 12, 2013.

[8] Lovish Chopra, Sarthak Chakraborty, Abhijit Mondal, and Sandip Chakraborty. 2021. PARIMA: Viewport Adaptive 360-Degree Video Streaming. In *Proceedings of the Web Conference 2021*. ACM, 2379–2391. https://doi.org/10.1145/3442381.3450070

[9] 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. *Int. J. Hum.-Comput. Stud.* 82, C (oct 2015), 48–56. https://doi.org/10.1016/j.ijhcs.2015.05.004

[10] Berthold K.P. Horn and Brian G. Schunck. 1981. Determining optical flow. *Artificial Intelligence* 17, 1 (1981), 185–203. https://doi.org/10.1016/0004-3702(81)90024-2

[11] L. Itti, C. Koch, and E. Niebur. 1998. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20, 11 (1998), 1254–1259. https://doi.org/10.1109/34.730558

[12] Jeffrey J. Walline. 2001. Designing Clinical Research: an Epidemiologic Approach, 2nd Ed. *Optometry and Vision Science* 78, 8 (2001). https://journals.lww.com/optvissci/Fulltext/2001/08000/Designing_Clinical_Research__an_Epidemiologic.5.aspx

[13] Akisato Kimura. 2020. pySaliencyMap. https://github.com/akisatok/pySaliencyMap

[14] Terry K Koo and Mae Y Li. 2016. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of chiropractic medicine* 15, 2 (June 2016), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012 Edition: 2016/03/31 Publisher: Elsevier.

[15] Olivier Le Meur and Thierry Baccino. 2013. Methods for comparing scanpaths and saliency maps: strengths and weaknesses. *Behavior Research Methods* 45, 1 (March 2013), 251–266. https://doi.org/10.3758/s13428-012-0226-9

[16] 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 (Dec. 2017), 2116. https://doi.org/10.3389/fpsyg.2017.02116

[17] Dominique Makowski, Tam Pham, Zen J. Lau, Jan C. Brammer, François Lespinasse, Hung Pham, Christopher Schölzel, and S. H. Annabel Chen. 2021. NeuroKit2: A Python toolbox for neurophysiological signal processing. *Behavior Research Methods* 53, 4 (feb 2021), 1689–1696. https://doi.org/10.3758/s13428-020-01516-y

[18] Nitish Mutha. 2017. Equirectangular-toolbox. https://github.com/NitishMutha/equirectangular-toolbox

[19] Federica Pallavicini, Alessandro Pepe, and Maria Eleonora Minissi. 2019. Gaming in Virtual Reality: What Changes in Terms of Usability, Emotional Response and Sense of Presence Compared to Non-Immersive Video Games? *Simulation & Gaming* 50, 2 (2019), 136–159. https://doi.org/10.1177/1046878119831420 arXiv:https://doi.org/10.1177/1046878119831420

[20] 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, 1–6. https://doi.org/10.1109/QoMEX48832.2020.9123126

[21] TensorfFlow. 2021. TensorFlow 2 YOLOv4. https://wiki.loliot.net/docs/lang/python/libraries/yolov4/python-yolov4-about/

[22] Alexander Toet, Fabienne Heijn, Anne-Marie Brouwer, Tina Mioch, and Jan B. F. van Erp. 2020. An Immersive Self-Report Tool for the Affective Appraisal of 360° VR Videos. *Frontiers in Virtual Reality* 1 (Sept. 2020), 552587. https://doi.org/10.3389/frvir.2020.552587

[23] UCSF. 2021. Sample Size Calculators for designing clinical research. https://sample-size.net/correlation-sample-size/.

[24] 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)*. 1–6. https://doi.org/10.1109/QoMEX48832.2020.9123125

[25] Mai Xu, Chen Li, Zhenzhong Chen, Zulin Wang, and Zhenyu Guan. 2019. Assessing Visual Quality of Omnidirectional Videos. *IEEE Transactions on Circuits and Systems for Video Technology* 29, 12 (2019), 3516–3530. https://doi.org/10.1109/TCSVT.2018.2886277

[26] 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 *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–8. https://doi.org/10.1145/3411763.3451627

[27] Tong Xue, Abdallah El Ali, Tianyi Zhang, Gangyi Ding, and Pablo Cesar. 2021. CEAP-360VR: A Continuous Physiological and Behavioral Emotion Annotation Dataset for 360 VR Videos. *IEEE Transactions on Multimedia* (2021), 1–1. https://doi.org/10.1109/TMM.2021.3124080