Effects of Emotions on Head Motion Predictability in 360° Videos

Quentin Guimard, Lucile Sassatelli

Effects of Emotions on Head Motion Predictability in 360° Videos

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

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

ABSTRACT

While 360° videos watched in a VR headset are gaining in popularity, it is necessary to lower the required bandwidth to stream these immersive videos and obtain a satisfying quality of experience. Doing so requires predicting the user's head motion in advance, which has been tackled by a number of recent prediction methods considering the video content and the user's past motion. However, human motion is a complex process that can depend on many more parameters, including the type of attentional phase the user is currently in, and their emotions, which can be difficult to capture. This is the first article to investigate the effects of user emotions on the predictability of head motion, in connection with video-centric parameters. We formulate and verify hypotheses, and construct a structural equation model of emotion, motion and predictability. We show that the prediction error is higher for higher valence ratings, and that this relationship is mediated by head speed. We also show that the prediction error is lower for higher arousal, but that spatial information moderates the effect of arousal on predictability. This work opens the path to better capture important factors in human motion, to help improve the training process of head motion predictors.

CCS CONCEPTS

• **Human-centered computing** → **Virtual reality**; *User models*; • **Mathematics of computing** → Equational models.

KEYWORDS

360° videos, emotions, head motion, predictability

ACM Reference Format:
Quentin Guimard and Lucile Sassatelli. 2022. Effects of Emotions on Head Motion Predictability in 360° Videos. In *International Workshop on Immersive Mixed and Virtual Environment Systems (MMVE '22), June 14, 2022, Athlone, Ireland.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3534086.3534335

1 INTRODUCTION

Virtual Reality (VR) headsets allow users to experience 360° videos while being entirely immersed in a virtual environment, be it shot from the real-world or computer-generated. With the increasing affordability of VR headsets and popularity of 360° videos, the demand for streaming 360° videos is also growing. However, the necessary bandwidth to stream these videos with a quality high enough to provide real world-like experience can be up to two orders of magnitude that of a regular video [17]. To decrease the required bandwidth, a well-investigated approach is to send every 360° video segment with spatially heterogeneous quality, so as to focus the bandwidth budget to maximize the visual quality in the field of view (FoV), while decreasing it outside (e.g., with tile-based approaches). Doing so in a networking scenario requires predicting in advance what the user motion is going to be, so that the server can decide which sectors to send in higher or lower quality.

That is why a number of works have developed head motion prediction methods in the last few years, often based on deep neural networks [5, 6, 19]. For stored video content, major methods [5, 19, 24] consider predicting the user's head motion for any new user and any new video, therefore relying only on the knowledge of the past user head position and the past and future video content. However, human motion is a complex process which can depend on many more parameters, including the type of attentional phase the user is currently in [1, 21] and their emotions [14, 22, 26]. These parameters may be difficult or impossible to fully infer from the sole video content and past motion, therefore introducing uncertainty in the prediction task impacting the prediction error.

To the best of our knowledge, this is the first article to investigate the following research question: *What are user-centric parameters (emotion and motion) and video-centric parameters impacting the head motion predictability in immersive 360° videos, and what are the relationships?* This is an important question to understand how well can the human motion be captured, and how to improve prediction approaches, by augmenting the videos to be labelled/experienced wisely, or by changing the architectures or training losses of the deep models.

We make the following contributions:

• We propose a subset of user-centric and video-centric measures to investigate the connection between these and head motion predictability. We consider two datasets where the user movements and subjective emotions are made available, one of which we have collected. The considered measures are valence and arousal graded by every user on every video, head motion speed, spatial information (SI) and temporal information (TI), shown to provide important insights into this type of emotion-video feature-predictability relationships.

• We formulate three hypotheses that we verify, and model the data with a directed graph of causal relationships formalized in a structural equation model (SEM). We show that the prediction error is generally lower (higher predictability) for users having provided higher arousal ratings. We also show that the prediction error is higher for higher valence ratings, and that this relationship is mediated by head speed. Finally, we exhibit an interaction effect between SI and arousal, SI moderating the effect of arousal on the prediction error.

MMVE '22, June 14, 2022, Athlone, Ireland
2022. ACM ISBN 978-1-4503-9382-9/22/06…$15.00
https://doi.org/10.1145/3534086.3534335

Section 2 positions our approach with respect to existing works. Sec. 3 defines the head motion prediction problem and presents the prediction method we consider. Sec. 4 describes both datasets considered, made of 360° videos with user emotions and motion traces. Sec. 5 states the hypotheses based on previous works, the validity of which we analyze from the data, and Sec. 6 expresses how the effects of user emotions and motion on predictability are modeled in a SEM, allowing us to quantify the effect sizes of mediation and interactions. Sec. 8 discusses limitations and perspectives.

2 RELATED WORK

Several approaches have recently investigated head motion prediction, most being based on deep recurrent neural networks fed with both the series of past coordinates and the video content [5, 19, 23, 24]. In this article, we consider the prediction methods presented by Romero et al. [19] to benchmark our approach. Considering the analysis of how the video content or user emotions impact the head motion predictability, to the best of our knowledge only Romero et al. [19] analyzed the prediction performance disaggregated over video categories. However, none of the above works has looked at motion predictability based on felt emotions, nor did they formalize the relationship between predictability and video features.

The spectrum of human emotions is generally described with two main components of the circumplex model of emotions [20]: valence, denoting the pleasantness or unpleasantness of emotions (from positive or happiness to negative or sadness/fear), and arousal, representing the intensity of the felt emotion. In the two-dimensional space where valence can be represented on the x-axis and arousal on the y-axis, any point symbolizes an emotion that is a combination of a certain amount of valence and arousal. Circumplex models have been used most commonly to test stimuli of emotion words, emotional facial expressions, and affective states. Immersive environments experienced in a VR headset have been shown to provide more intense emotions than planar presentations of omnidirectional content [2, 8, 16]. Understanding the relationships between the components of emotion and the components of the immersive experience (such as presence and immersion), has attracted interest [3, 13]. For example, Jicol et al. [13] have recently investigated how emotions and afforded agency combine and interact to create the feeling of presence. In particular, they resort to a SEM approach to quantify the size of direct and indirect effects. To investigate the emotional process elicited by 360° videos, a first database [14] made of 73 omnidirectional videos with collected valence and arousal was made publicly available, but without head motion. A more recent dataset was presented by Xue et al. [27], making both subjective ratings (both after-viewing and continuous inside the video) and head and gaze movements available. The connections between user motion and emotion in virtual environments have been investigated by many [14, 22, 26]. For example, Xue et al. [26] show that the yaw standard deviation (connected to mean speed) negatively correlates with arousal, while Li et al. [14] show that it positively correlates with valence.

No work has so far investigated and formalized the effect of emotions and video features on head motion predictability, that is on the performance of prediction methods. To do so, we consider the most recent predictors introduced by Romero et al. [19], as motivated below, and formulate working hypotheses from initial results obtained in the works mentioned above [14, 26].

3 HEAD MOTION PREDICTION

We first define the problem of head motion prediction in Sec. 3.1, then describe the chosen method and the motivation behind this choice in Sec. 3.2.

3.1 Problem definition

The problem we consider is formally described as follows. We consider that a given 360° video v of duration T seconds is being watched by a user u. The head trajectory of the user is denoted $\mathbf{P}^{u,v}_{0:T}$, with $\mathbf{P}$ storing the head coordinates on the unit sphere (as, e.g., Euler angles, Cartesian coordinates or quaternions).

At any time t in $[0, T]$, we want to predict the future trajectory $\mathbf{P}^{u,v}_{t:t+H}$ over a prediction horizon H, assuming only $\mathbf{P}^{u,v}_{0:t}$ and the video content of v are known. That is, we do not assume any knowledge of traces other than u on this video v.

3.2 Prediction method

Amongst the existing methods tackling the above prediction problem [5, 19, 23, 24], we choose two main methods presented by Romero et al. [19], named Deep-position-only and TRACK. We make this choice because (i) these approaches are representative of other prior approaches relying on sequence-to-sequence architectures, (ii) they are recent, and (iii) the models and entire framework are made publicly available [18].

To conduct our study, we consider both models trained on two different head motion datasets from David et al. [7] and Xu et al. [25], and selected the trained models that obtained the best results when testing (without re-training or fine-tuning) on our data described in Sec. 4. All the trained models were similar in performance on our data, but the models trained on the dataset by Xu et al. [25], the largest dataset, performed slightly better. Then, we inspected the mutual effects, such as those shown in Fig. 4, when the prediction error is obtained with Deep-position-only and TRACK. As results were qualitatively similar, for the rest of the paper, we have chosen to only present results obtained with TRACK.

TRACK is a sequence-to-sequence deep model using separate long short-term memory (LSTM) units to encode both the past positions and the visual saliency. The same kind of LSTM units, combined with fully connected layers are then used to decode the future positions based on the embeddings given by the encoder. The visual saliency is made up of 384x216 saliency maps extracted from the video frames by PanoSalNet. TRACK was fully re-implemented using PyTorch and trained on multiple head motion datasets as provided in the repository [18].

4 DATASETS AND MEASURES

In this section, we present the datasets considered for our data analysis, and our choice of user-centric and video-centric measures. The effect of these measures on motion predictability in investigated next in Sec. 5 and 6.

4.1 Datasets

We consider the only two datasets available where both user movements and emotions have been collected from immersive viewing of 360° videos.

The first dataset we consider is CEAP360-VR [27]. This dataset is made from user experiments with 32 participants each watching 8 videos in a VR headset equipped with an eye-tracker, recording head and eye movements. After each video, the users grade their emotions valence and arousal. Additionally, emotional ratings are continuously annotated by the users thanks to a controller in their hand, along with physiological measurements with a wristband. We do not use this latter data in this article.

Our second source of data comes from PEM360 [9], our own dataset collected from user experiments. In conditions similar as above, 7 videos were experienced in an eye-tracker equipped VR headset by 31 participants, who rated their valence and arousal perceptions after every video clip. Physiological measurements were also collected but not used in this article. This dataset is now publicly available on public GitLab repository[1].

In both datasets, users were asked to rate each video using the self-assessment manikin (SAM) [4], giving individual ratings of valence and arousal.

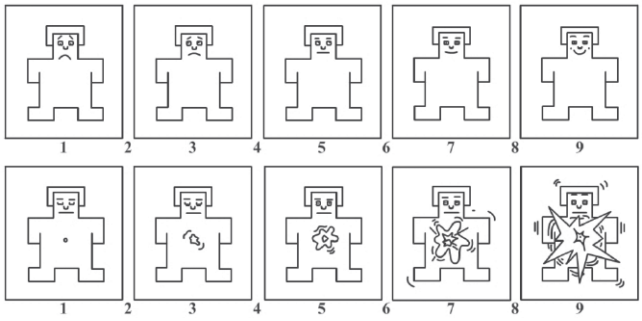

Figure 1: Self-Assessment Manikin (SAM) scale for rating of valence (top row) and arousal (bottom row). Taken from [4].

The videos shown to users in both datasets come from the same 360° video database [14], with average ratings of valence and arousal from 95 participants. We report in Table 1 the details of every video from both datasets. The left-most column "ID" refers to the video ID the author used in their dataset. The right-most column "Database ID" indicates the original ID in the 360° video database [14]. Ratings of valence and arousal given in this table are the original average ratings of the database. In each of the datasets, the videos were trimmed, so the videos are not exactly the same as in the database. The start offset of each video as well as the duration of the trimmed clip are specified in the table.

We should note that videos 32 and V6 are different versions of the same video, which makes a total of 14 distinct videos experienced in VR by 63 different users.

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

ID	Valence	Arousal	Start (s)	Duration (s)	Database ID
12	7.00	4.60	5	98	12
13	4.92	4.08	4	127	13
17	5.22	5.00	5	64	17
23	7.20	3.20	8	135	23
27	6.00	1.60	60	120	27
32	6.57	1.57	40	90	32
73	6.27	6.18	9	61	73
V1	7.47	5.35	0	60	50
V2	6.13	1.80	10	60	38
V3	3.20	5.60	65	59	21
V4	2.53	3.82	3	60	14
V5	6.75	7.42	0	60	52
V6	6.57	1.57	0	60	32
V7	4.40	6.70	127	60	68
V8	2.73	3.80	41	60	19

Table 1: Details of selected videos. The ID refers to the original database [14].

4.2 Measures

4.2.1 Outcome measure. The objective of this analysis is to evaluate the influence of various factors on the prediction error of head movements. We define the prediction error as the average displacement error (ADE) between the predicted head positions and the actual future head positions over a prediction horizon H. We set $H = 5$ seconds, the standard prediction horizon in recent deep prediction methods [5, 19], which covers both user inertia and content saliency [19].

We define the displacement error between two head positions (x_1, y_1, z_1) and (x_2, y_2, z_2) as the great circle distance between these two points. Since (x, y, z) are the Cartesian coordinates of a point on the unit sphere, we can easily compute the great circle distance $\Delta\sigma$ from the Euclidean distance d between these two positions as $\Delta\sigma = 2 \cdot \arcsin \frac{d}{2}$.

4.2.2 User-centric measures. We consider two types of user-centric measures: those related to emotions, and those related to motion. The measures of emotions are considered as the subjective ratings of valence and arousal made by each user after experiencing each 360° video, as detailed above.

The user motion can be characterized by various metrics, such as mean values of the head or gaze yaw and pitch angles, or the standard deviations of these positional components [14, 27]. Here, we choose to combine these elements and consider the angular speed of the head movements. Specifically, to compute head speed, we first convert the head coordinates collected from the VR headset into Cartesian coordinates, where each recorded head position at time t is a point on the unit sphere of coordinates (x_t, y_t, z_t).

The head motion data in CEAP360-VR is originally in the format $(\psi_t, \theta_t, \phi_t)$, where ψ is the yaw, θ is the pitch, and ϕ is the roll. These coordinates were first transformed to have $\psi \in [0, 2\pi[$ where 0 is the left edge of the equirectangular frame, and $\theta \in [0, \pi[$ where 0 is the top edge of the equirectangular frame. Cartesian (x_t, y_t, z_t) coordinates are then obtained as projections of these angles using

this set of equations:

$$\begin{cases} x_t = \cos\psi_t \cdot \sin\theta_t \\ y_t = \sin\psi_t \cdot \sin\theta_t \\ z_t = \cos\theta_t \end{cases}$$

We define the instantaneous head speed at time t as the total angular speed, noted ω_t, computed from the great-circle distance between two consecutive positions divided by the sampling rate of the recordings. The average head speed is then taken as the mean of all instantaneous head speeds for a given user on a given video.

4.2.3 Video-centric measures. As for user motion, several characterizations of 360° video content is possible. For example, Almquist et al. [1] propose a taxonomy in four categories depending on the location of the regions of interest. Romero et al. [19] inspire on this taxonomy and categorize videos based on the entropy of the head location heat maps. David et al. [7] and Xue et al. [27] consider spatial information and temporal information to characterize the 360° videos. In the preliminary study presented in this article, we consider legacy spatial and temporal information, and show their relevance characterizing the effects of emotions on motion predictability.

Spatial information and temporal information are scene-specific metrics defined in ITU-T Recommendation P.910 [12]. According to the ITU-T recommendation, SI and TI are "critical parameters" playing "a crucial role in determining the amount of video compression that is possible".

Spatial information (SI) or spatial perceptual information is "a measure that generally indicates the amount of spatial detail in a picture. (...) It is usually higher for more spatially complex scenes." "The SI is based on the Sobel filter. Each video frame (luminance plane) is first filtered with the Sobel filter. The standard deviation over the pixels in each Sobel-filtered frame is then computed", resulting in the SI for a single frame. We consider SI_v, the average SI for all the frames of video v.

Temporal information (TI) or temporal perceptual information is "a measure that generally indicates the amount of temporal changes of a video sequence. (...) It is usually higher for high motion sequences." "TI is based upon the motion difference feature, that is the difference between the pixel values (of the luminance plane) at the same location in space but at successive frames." The standard deviation over the pixels of all the differences between successive frames is then computed to give the TI for two consecutive frames. We consider TI_v, the average TI for all the frames of video v. "More motion in adjacent frames will result in higher values of TI."

5 HYPOTHESIS TESTING

Based on previous works [14, 26], we make the following *a priori* hypotheses:

H1 Prediction error is lower for higher user arousal.
H2 Prediction error is higher for higher user valence.
H3 Head speed mediates the effect of valence on error.

To analyze the validity of the above hypotheses, we first binarize some variables and perform analysis of variance (ANOVA) testing, shown in Table 2. The analysis of linear correlations on continuous data is incorporated into the structural equation modeling in Sec. 6.

The binarization is performed on *SI*, *TI*, *Arousal* and *Valence* (denoting continuous variables) to obtain SI_{bin}, TI_{bin}, $Arousal_{bin}$ and $Valence_{bin}$. For *SI* and *TI* of every video v, binarization thresholds are chosen so that approximately half of the videos are in each partition: SI_{bin} = -1 (resp. 1) for $SI_v \leq 45$ (resp. > 45), and TI_{bin} = 0 (resp. 1) for $TI_v \leq 3$ (resp. > 3). In Fig. 4, SI_{bin} is denoted "Low SI" or "High SI" with the same threshold. For *Arousal* and *Valence* of every user-video pair (u, v), $Arousal_{bin}$ = 0 (resp. = 1) for $Arousal_{u,v} \leq 5$ (resp. > 5), and the same to obtain $Valence_{bin}$. In Fig. 2, $Arousal_{bin}$ (resp. $Valence_{bin}$) is also referred to as "LA" for low *Arousal* (resp. "LV" for low *Valence*) and "HA" for high *Arousal* (resp. "HV" for high *Valence*), with the same thresholds as defined above.

	SI_{bin}	TI_{bin}	$Arousal_{bin}$	$Valence_{bin}$
Prediction error	70.89**	79.09**	15.15**	7.67*
Head speed	17.77**	19.94**	2.76	15.78**
Arousal	51.90**	55.50**	(1253**)	0.37
Valence	30.60**	2.69	0.42	(1266**)

Table 2: F-scores of one-way ANOVA. The significance of group difference is denoted with * for $p < 10^{-2}$ and ** for $p < 10^{-3}$.

We first observe from Table 2 that SI_{bin} significantly impacts all variables (*Prediction error*, *Head speed*, *Valence* and *Arousal*), while TI_{bin} does not significantly impact *Valence*.

The relations mentioned in H1 and H2 are significant. Fig. 2-left and 2-center show the direction of the association with 95% confidence intervals. We can therefore accept H1 and H2.

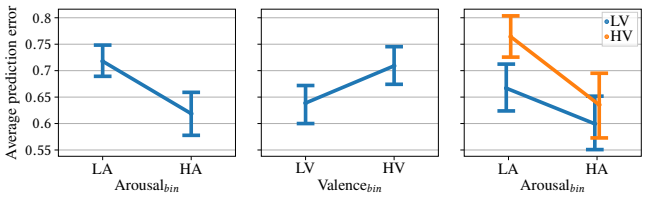

Figure 2: Prediction error against binarized $Arousal_{bin}$ (left) and $Valence_{bin}$ (center). Right: Difference in variation of *Prediction error* against $Arousal_{bin}$ depending on $Valence_{bin}$.

To investigate the sizes of the significant effects, Fig. 3 shows scatter plots of error as a function of *Arousal* and *Head speed*, as well as how *Head speed* varies with graded *Arousal* and *Valence*. We first observe that there is a strong correlation between *Prediction error* and *Head speed*. We also observe that, as hinted in preliminary results from Li et al. [14] and Xue et al. [26], *Prediction error* tends to decrease with *Arousal*. While *Head speed* does not seem to significantly vary with *Arousal*, as confirmed by the ANOVA result in Table 2, the scatter plot of *Head Speed* versus *Valence* shows that the significant association between both, shown by the corresponding ANOVA result in Table 2, is an increasing function. This is in line with H3, which will be validated in the next section.

As Fig. 3-top-right shows the strong association of *Prediction error* with *Head speed* and Table 2 shows significant associations of

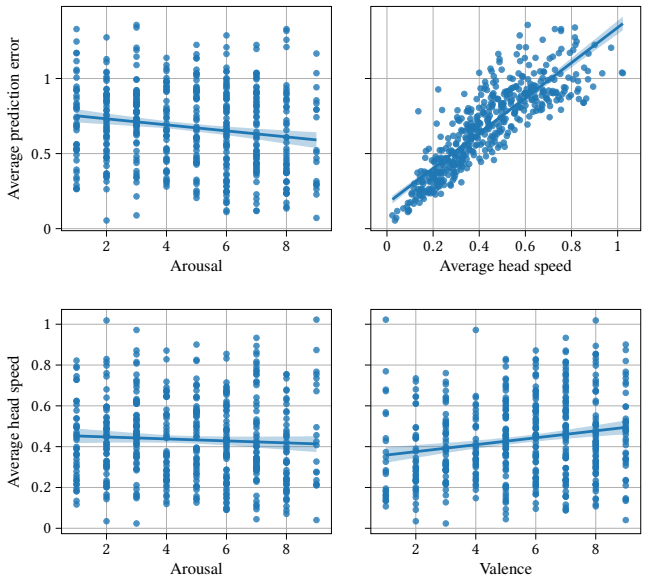

Figure 3: Scatter plots of *Prediction error* against *Arousal* and *Head speed* (top row), and *Head speed* against *Arousal* and *Valence* (bottom row). Straight lines are linear regression models fitted on the data. Shaded areas represent 95% confidence intervals.

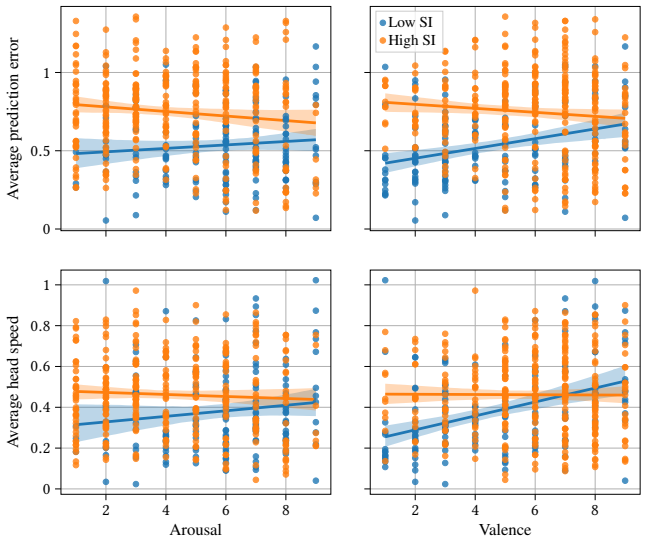

Figure 4: Scatter plots of *Prediction error* and *Head speed* against *Arousal* and *Valence*, disaggregated over SI_{bin}. Straight lines are linear regression models fitted on the data. Shaded areas represent 95% confidence intervals.

Prediction error, *Head speed*, *Arousal* and *Valence* with SI_{bin}, we analyze whether some variables interact with *Arousal* and *Valence* in their effect on *Head speed* and *Prediction error*. Fig. 4 shows that there is a possible interaction between the video feature SI_{bin} and *Arousal*, and SI_{bin} and *Valence*, in their effect on *Prediction error* and *Head speed*. This can be seen in the different slopes of linear model fitting the cloud of points, for each set of (u, v) points, for all users $u \in \mathcal{U}$ and videos v such that $SI_{bin}(v)$ = -1, or 1.

Also, it is interesting to observe in Fig. 2-right that *Prediction error* does not decrease in the same way with increased *Arousal*, depending on whether *Valence* is graded high or low. Indeed, *Prediction error* decreases more when *Arousal* increases when *Valence* is high. We may assume that the user tends to move more when they enjoy the video, and higher *Arousal* means more involvement/attentional capture, and hence synchronization between motion and the content's salient regions, facilitating the prediction. This corresponds partly to H3 and is investigated in the next section.

6 MODELLING THE EFFECT OF EMOTIONS AND VIDEO CHARACTERISTICS ON MOTION PREDICTABILITY

We now construct a structural equation model of the data. SEM is established as a methodological approach to represent how different variables affect each other [10]. It allows to build a network of causal relations, and to investigate direct and indirect effects with mediating variables and external moderators interacting on the effect. A SEM therefore gathers significant linear relations, enabling to both incorporate the correlation coefficient and measure the size of the effect. We construct a SEM based on accepted H1 and H2, and incorporating the possible interaction of SI_{bin} with *Arousal* and *Valence*. An interaction effect is modeled as the product of two variables, one of which is binary. Owing to the above analysis of Fig. 4, we define interaction variables *Arousal* $\times$ SI_{bin} and *Valence* $\times$ SI_{bin}. We then consider possible causal relationships from *Arousal*, *Valence*, *Arousal* $\times$ SI_{bin} and *Valence* $\times$ SI_{bin} to both *Head speed* and *Prediction error*, as well as relationship from *Head speed* to *Prediction error*.

We use the Python toolkit Semopy [11, 15], using the Wishart log-likelihood objective function. The resulting SEM is shown in Fig. 5, where only edges with regression coefficients significantly different from 0 have been kept (with $p \leq 0.01$). Every edge is tagged with the unstandardized coefficient of the linear relationship between both participating variables, and with the corresponding standardized coefficient in parenthesis. The unstandardized coefficient is impacted by the difference in the relative scale of the variables, while the standardized coefficient is independent of the scale and represents by how many standard deviations the end variable varies when the regressor variable increases by one standard deviation.

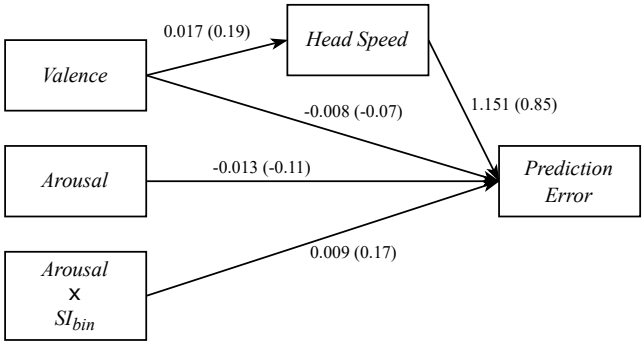

Figure 5: Structural equation model (SEM) describing the direct and indirect effects of user *Valence* and *Arousal* onto *Prediction error*, mediated by *Head speed* and moderated by video measure SI_{bin}.

First, the model shows that the major impact of *Valence* on *Prediction error* is mediated by *Head speed*. Indeed, the standardized indirect effect of *Valence* on *Prediction error* is $0.19 \times 0.85 = 0.16$, while the direct effect is only -0.07. H3 is therefore validated. This suggests that users rating the video with higher *Valence* tend to move more. This connects with the results obtained by Li et al. [14]. Second, the model confirms that the prediction error varies inversely with *Arousal*, with a standardized linear coefficient of -0.11. Third, the interaction effect is significant. Indeed, for a high *SI* video with SI_{bin} = 1, the total effect of *Arousal* on *Prediction error* is $-0.11 + 0.17 \times 1 = 0.06$. In this case, the effect of *Arousal* on *Prediction error* is low and not significantly negative. However, for a low *SI* video with SI_{bin} = -1, the total effect of *Arousal* on *Prediction error* is $-0.11 + 0.17 \times -1 = -0.28$. SI_{bin} is therefore a strong moderator of the effect of *Arousal* on *Prediction error*. To interpret this result, one may investigate how SI_{bin} connects with video categories, such as those proposed by Almquist et al. [1]. We may think that a high SI_{bin} describes videos with numerous salient areas in frames, hence yielding more exploratory head movements difficult to predict, even though the person rates their arousal/involvement in the video as high. Verifying such an interpretation is left for future works, as described in Sec. 8. Finally, this last result means that the video feature SI is a strong confounding factor which must be taken into consideration when one chooses 360° videos to investigate the impact of user emotion on motion prediction.

7 DISCUSSION

The results presented above open the path to promising directions to understanding the human motion process in immersive environments, as detailed in Sec. 8. Let us mention here some limits and perspectives of the presented data analysis.

First, obtaining results on more than 14 videos will be important for generalization, and to avoid possible spurious correlation between arousal and valence ratings that may impact our findings. Second, the main outcome variable considered here being the prediction error, the results may depend on the type of predictor considered. While, for the reasons described in Sec. 3.2, we verified that the results were similar between both methods taken from [19], other families of approaches might lead to different effects of emotion on predictability. Third, we have considered head motion in this work, but it would be important to identify how the effects of emotions differ when predicting eye motion. More generally, while we have focused on only three types of user-centric measures (arousal, valence and head motion speed) and two types of video-centric measures (SI and TI), it will be most interesting to generalize this approach to more user-centric measures such as electrodermal activity and gaze, and video-centric measures such as video categories (focus or exploration [1], fear or happiness [13], possibly relating SI and TI to these).

8 CONCLUSION AND FUTURE WORK

In this article, we have presented a first investigation into the effect of emotion on head motion predictability. We considered two datasets totalling 14 videos and 63 users, providing head motion traces and arousal and valence subjective ratings. Through hypothesis testing and structural equation modelling, we have shown that the predictability of head motion increases with arousal but decreases with valence, that the effect of valence on predictability is mediated by head speed, and that video SI interacts in the effect of arousal on predictability, a high SI moderating the effect.

This work opens the way to better understand factors impacting the human motion and their effect on the performance of head motion predictors, and how such knowledge can be leveraged to improve prediction. This can be done by augmenting the datasets with videos with specific emotional and visual features where head motion prediction is harder, or by designing ancillary training losses where a deep neural model would have to learn how to predict the user emotional state from the video content and the user's past motion. An important outcome of this work is also to estimate the motion predictability from user emotional state. Such an estimation of the confidence of head motion prediction can readily be leveraged in the optimization of a 360° streaming system, even more so if the user emotional state is estimated with lightweight non-invasive device such as finger straps to measure electrodermal activity. This is the subject of our very next work.

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] Mathias Almquist, Viktor Almquist, Vengatanathan Krishnamoorthi, Niklas Carlsson, and Derek Eager. 2018. The prefetch aggressiveness tradeoff in 360° video streaming. In *Proceedings of the 9th ACM Multimedia Systems Conference (MMSys '18)*. ACM, New York, NY, USA, 258–269. https://doi.org/10.1145/3204949.3204970

[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. https://doi.org/10.1089/cpb.2007.9936

[3] Miguel Barreda-Ángeles, Sara Aleix-Guillaume, and Alexandre Pereda-Baños. 2020. An "Empathy Machine" or a "Just-for-the-Fun-of-It" Machine? Effects of Immersion in Nonfiction 360-Video Stories on Empathy and Enjoyment. *Cyberpsychology, Behavior, and Social Networking* 23, 10 (Oct. 2020), 683–688. https://doi.org/10.1089/cyber.2019.0665

[4] 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

[5] Fang-Yi Chao, Cagri Ozcinar, and Aljosa Smolic. 2021. Transformer-based Long-Term Viewport Prediction in 360° Video: Scanpath is All You Need. In *IEEE 23nd International Workshop on Multimedia Signal Processing (MMSP)*. IEEE. https://doi.org/10.1109/MMSP53017.2021.9733647

[6] Jinyu Chen, Xianzhuo Luo, Miao Hu, Di Wu, and Yipeng Zhou. 2021. Sparkle: User-Aware Viewport Prediction in 360-Degree Video Streaming. *IEEE Transactions on Multimedia* 23 (2021), 3853–3866. https://doi.org/10.1109/TMM.2020.3033127

[7] 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 (MMSys '18)*. ACM, New York, NY, USA, 432–437. https://doi.org/10.1145/3204949.3208139

[8] 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

[9] 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 *Proceedings of the 13th ACM Multimedia Systems Conference (MMSys '22)*. ACM. https://doi.org/10.1145/3524273.3532895

[10] Rick H. Hoyle (Ed.). 1995. *Structural equation modeling: Concepts, issues, and applications.* Sage Publications, Inc, Thousand Oaks, CA, US. Pages: xxii, 289.

[11] Anna A. Igolkina and Georgy Meshcheryakov. 2020. semopy: A Python Package for Structural Equation Modeling. *Structural Equation Modeling: A Multidisciplinary Journal* 0, 0 (2020), 1–12. https://doi.org/10.1080/10705511.2019.1704289

[12] ITU-T P.910 2021. *Subjective video quality assessment methods for multimedia applications*. Recommendations. International Telecommunication Union: Telecommunication Standardization Sector.

[13] Crescent Jicol, Chun Hin Wan, Benjamin Doling, Caitlin H Illingworth, Jinha Yoon, Charlotte Headey, Christof Lutteroth, Michael J Proulx, Karin Petrini, and Eamonn O'Neill. 2021. Effects of Emotion and Agency on Presence in Virtual Reality. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–13. https://doi.org/10.1145/3411764.3445588

[14] 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

[15] Georgy Meshcheryakov, Anna A. Igolkina, and Maria G. Samsonova. 2021. semopy 2: A Structural Equation Modeling Package with Random Effects in Python. arXiv:stat.AP/2106.01140

[16] 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

[17] Jounsup Park, Philip A Chou, and Jenq-Neng Hwang. 2019. Rate-utility optimized streaming of volumetric media for augmented reality. *IEEE Journal on Emerging and Selected Topics in Circuits and Systems* 9, 1 (2019), 149–162. https://doi.org/10.1109/JETCAS.2019.2898622

[18] Miguel Fabián Romero-Rondón, Lucile Sassatelli, Ramón Aparicio-Pardo, and Frédéric Precioso. 2020. A unified evaluation framework for head motion prediction methods in 360° videos. In *Proceedings of the 11th ACM Multimedia Systems Conference (MMSys '20)*. ACM, 279–284. https://doi.org/10.1145/3339825.3394934

[19] Miguel Fabián Romero-Rondón, Lucile Sassatelli, Ramón Aparicio-Pardo, and Frédéric Precioso. 2021. TRACK: A New Method from a Re-examination of Deep Architectures for Head Motion Prediction in 360-degree Videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021). https://doi.org/10.1109/TPAMI.2021.3070520

[20] James A. Russell. 1980. A circumplex model of affect. *Journal of Personality and Social Psychology* 39, 6 (12 1980), 1161–1178. https://doi.org/10.1037/h0077714

[21] Vincent Sitzmann, Ana Serrano, Amy Pavel, Maneesh Agrawala, Diego Gutierrez, Belen Masia, and Gordon Wetzstein. 2018. Saliency in VR: How Do People Explore Virtual Environments? *IEEE Transactions on Visualization and Computer Graphics* 24, 4 (2018), 1633–1642. https://doi.org/10.1109/TVCG.2018.2793599

[22] 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

[23] Chenglei Wu, Zhi Wang, and Lifeng Sun. 2021. PAAS: a preference-aware deep reinforcement learning approach for 360° video streaming. In *Proceedings of the 31st ACM Workshop on Network and Operating Systems Support for Digital Audio and Video (NOSSDAV'21)*. ACM, Istanbul Turkey, 34–41. https://doi.org/10.1145/3458306.3460995

[24] Yanyu Xu, Yanbing Dong, Junru Wu, Zhengzhong Sun, Zhiru Shi, Jingyi Yu, and Shenghua Gao. 2018. Gaze Prediction in Dynamic 360° Immersive Videos. In *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 5333–5342. https://doi.org/10.1109/CVPR.2018.00559

[25] Yanyu Xu, Yanbing Dong, Junru Wu, Zhengzhong Sun, Zhiru Shi, Jingyi Yu, and Shenghua Gao. 2018. Gaze Prediction in Dynamic 360° Immersive Videos. In *IEEE CVPR*. 5333–5342. https://doi.org/10.1109/CVPR.2018.00559

[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