HeMoG: A White-Box Model to Unveil the Connection Between Saliency Information and Human Head Motion in Virtual Reality

Miguel Fabian Romero Rondon*†, Dario Zanca‡, Stefano Melacci§, Marco Gori†§ and Lucile Sassatelli*¶
*I3S, CNRS, Université Côte d'Azur, 06900, Sophia Antipolis, France
Email: miguel-fabian.romero-rondon@univ-cotedazur.fr
†Inria, Maasai, 06902, Valbonne, France
‡Machine Learning and Data Analytics Lab, Friedrich-Alexander-Universität Erlangen-Nürnberg, 91052, Erlangen, Germany
§Department of Information Engineering and Mathematics, University of Siena, 53100, Siena, Italy
¶Institut Universitaire de France
Email: lucile.sassatelli@univ-cotedazur.fr

***Abstract*—Immersive environments such as Virtual Reality (VR) are now a main area of interactive digital entertainment. The challenge to design personalized interactive VR systems is specifically to guide and adapt to the user's attention. Understanding the connection between the visual content and the human attentional process is therefore key. In this article, we investigate this connection by first proposing a new head motion predictor named HeMoG. HeMoG is a white-box model built on physics of rotational motion and gravitation. Second, we compare HeMoG with existing reference Deep Learning models. We show that HeMoG can achieve similar or better performance and provides insights on the inner workings of these black-box models. Third, we study HeMoG parameters in terms of video categories and prediction horizons to gain knowledge on the connection between visual saliency and the head motion process.**

I. Introduction

Immersive environments for entertainment or training are gaining traction, in particular Virtual Reality (VR) for applications related to, e.g., gaming, museums, journalism, or rehabilitation. Designing VR experiences that are both interactive, comfortable and engaging is key to create immersive personalized environments. The challenge is to identify, adapt to and guide the attentional trajectory of the user. Visual attention is already considered in a number of such systems, be it for guidance in cinematic VR with 360° videos [1], [2], [3] or 3D-interactive environments [4], or to enable efficient VR streaming by predicting where the user is going to look at and send in high-quality only the attended Field of View (FoV) to save data rate [5], [6].

Understanding the connection between the audio-visual content and the human attentional process is therefore key for the design of immersive and personalized environments. Focusing only on the visual aspect, visual attention is a set of cognitive operations that allow us to filter the relevant locations in our visual field [7]. This mechanism also guides the movement of our head and eyes to center the selected location in our fovea, that is the area of the retina with the highest amount of photoreceptors and therefore allows sharp central vision [8].

Recently, VR in the form of 360° videos has been considered to study how people explore 360° environments with 3 degrees of freedom. The work of [9] collects user data to analyze and identify a few insights of human exploration in 360° videos (e.g., user congruence, the existence of an initial exploratory phase for ca. 18 sec. before a user focuses). Other works such as [10] aim at extracting the saliency maps, i.e., 2D-distributions of visual attention over a viewing period [11], from the content. To dynamically predict the head motion over a certain time *prediction horizon*, several Deep Learning (DL) models have been proposed, such as [12], [13] or [14].

These models, often referred to as "black-boxes", however do not provide any insight on the dependence of the head motion on the visual content. In this article, we address **2 research questions**:

Q1: To which extent can we investigate the inner workings of these DL models with a white-box model?

Q2: What knowledge can we obtain from a white-box model regarding the connection between saliency information and head motion?

We make **3 contributions**:

- [Sec. III] We design a new white-box model to predict head motion from the past motion and the 360° content. This model is built on the assumption that the head motion can be described by gravitational physics laws driven by virtual masses created by the content. This model is named HeMoG (Head Motion with Gravitational laws of attention).
- [Sec. IV] We evaluate the performance of HeMoG in comparison with reference DL models to predict head motion from the exact same inputs. When the prediction is made from past motion only (i.e., without content information), we show that HeMoG and the reference DL models achieve comparable performance. We interpret this as the DL model learning the curvature and friction dynamics of head motion that HeMoG is explicitly built on (1st answer to Q1). When HeMoG is fed

with saliency information, HeMoG can achieve comparable or better performance than the reference DL model TRACK (taken from [14]). We interpret this as the state-of-the-art DL models performing a similar type of fusion as HeMoG, which enables to benefit from both input modalities, past positions and visual content (2nd answer to Q1). We discuss in which case the representation learning of the DL models is key in Sec. VI.

- [Sec. V] In order to answer Q2, we take a closer look to the optimal hyper-parameters for HeMoG w.r.t. (i) the semantic category of the 360° video and (ii) the *prediction horizon*. On videos where the saliency maps render attractive areas (videos of categories *Static Focus*, *Moving Focus* and *Rides*), the optimal weight assigned in the motion equation to the content masses is higher than that when the video does not feature specific attractive areas (videos of category *Exploration*). Furthermore, analyzing the evolution of the saliency weight over the *prediction horizon* of 5 sec., we identify that the head motion momentum is most important first, and the content information starts being relevant after 3 sec. only.

The repository containing the code to use HeMoG and to reproduce the results in this paper is available at https://gitlab.com/miguelfromeror/hemog.

II. Related Work

Several DL models have been proposed to predict head or gaze motion in 360° videos:

Xu et al. in [13] designed a Deep Reinforcement Learning model to predict head motion. Their deep neural network receives the viewer's FoV as a 42×42 input image, and must decide to which direction and with which magnitude the viewer's head will move. Features obtained from convolutional layers processing each 360° frame cropped to the FoV are then fed into an LSTM to extract direction and magnitude. The prediction horizon is only one frame, i.e., 30ms.

In [12], Xu et al. predict the gaze positions over the next second in 360° videos based on the gaze coordinates in the past second and the video content. The time series of past head coordinates is processed by a doubly-stacked LSTMs. For the video information, spatial and temporal saliency maps are first concatenated with the RGB image, then fed to Inception-ResNet-V2 to obtain the "saliency features". The prediction horizon is 1 second long.

Nguyen et al. in [10], Li et al. in [15] and Fan et al. in [16] proposed a similar approach. They introduced LSTM-based networks fed with the concatenation of the head position encoded as a mask and the visual features extracted from pre-trained saliency extractor networks. The doubly-stacked LSTMs outputs the probability that each tile is viewed in the future trajectories.

Recently, [14] made a critical study of existing DL models, showing systematic weaknesses by comparing their performance with simple and stronger baselines. They also proposed a new DL model, named TRACK, that establishes state-of-the-art performance on several datasets. We therefore consider TRACK as the DL model our proposed HeMoG model must be compared with.

All these DL models are black-box models whose, to the best of our knowledge, explanability has not been studied. Explainability and interpretability of DL models decisions and predictions is a wide and highly active research area (see, e.g., [17]). In this work, our goal is not only to understand what type of inductive bias is exploited by existing DL models, but rather to design a white-box model that we can leverage both to gain insight on what the DL model learns, and unveil the connection between visual content and head motion.

In [18] Chen et al. proposed Sparkle, a model tailored to predict the exploration patterns of individual users in a 360° video. This model was evaluated against models based on Logistic Regression and the models from [10] and [16], which were found in [14] to be outperformed by baselines not modeling motion at all. Owing to the tiled equirectangular projection of the video frames considered in Sparkle, the prediction algorithm has to deal with the issues of the periodicity at the horizontal borders and the motion is limited at the poles in the vertical borders. In our model (HeMoG) this is solved by keeping the spherical nature of the data and using quaternions to represent the rotational velocity and acceleration. Another consideration in Sparkle is that the viewing information of other users is available for all the videos and there is a model learned per user. To avoid the systematical collection of user data [19], we considered in our problem modeling that the users' statistics for the specific video are *not* known at test time, furthermore, in HeMoG the parameters are not adjusted per user but the same parameters are shared across several users.

Finally, for regular 2D videos, [20] recently proposed a gravitational model to generate human-plausible visual scan-paths. We take inspiration from this model to design HeMoG, which, contrary to [20], is built on a 3D-rotational motion description with specific terms related to head/neck fatigue.

III. HeMoG: a model of head motion in 360° videos

In this section we present a new model of head motion in 360° videos named Head Motion with Gravitational laws of attention (HeMoG). We formulate the shift in human attention as the analogous mechanics of a ball rotating around a fixed origin. As shown in Fig. 1, the red ball represents the center of the FoV of a user exploring a virtual environment. All elements in a visual scene compete as attractors for the human attention process. This concept of attraction can be effectively described by means of gravitational models, where each location in the scene is associated with a virtual mass that is capable of attracting attention. The ball in this analogy rotates at a fixed length, the distance is normalized to a length of one for appropriate application of orthodromic distance through the arccos of vector dot product method.

The fundamental equation of rotational motion is:

$$\dot{\mathbf{L}} = \tau \; , \tag{1}$$

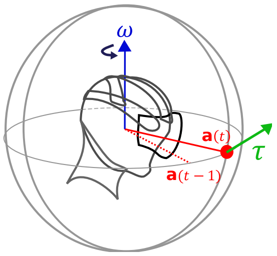

Fig. 1: Gravitational model of the head position of a person exploring a VR scene. The center of the FoV $\mathbf{a}(t)$ is modeled as a ball attached to a stick of fixed length that rotates with an angular velocity ω and with torque τ.

where:

- $\mathbf{L}$ is the angular momentum, expressed as $\mathbf{L} = I\omega$, with ω the angular velocity and I the moment of inertia. For a ball attached to a fixed point (red dot in Fig. 1), I can be expressed as the product of the ball's mass m with the norm of the ball's position vector $|\mathbf{a}(t)|$. There is not such a valid analogy in our modeling of the center of focus, and we shall keep I as a parameter in the mathematical formulation that follows.
- τ represents the torque applied to the system. This torque results from various forces, as described below.

We therefore have $\tau = \frac{d(I\omega)}{dt} = \dot{I}\omega + I\dot{\omega}$. Having constrained the attention on the unit sphere, the norm of $\mathbf{a}(t)$ does not change over time, resulting in $\dot{I} = 0$. Therefore we obtain

$$I\dot{\omega} = \tau \ . \tag{2}$$

Modeling of τ: The torque is the turning effectiveness of a force. To model head rotation, we assume that two types of forces are at play:

- forces that drive the head focus to salient areas of the 360° content. Every 360° frame therefore generates a field of force

$$\mathbf{E}(\mathbf{a}) = \int_{\mathbf{r}\in\Upsilon} \mathbf{F}(\mathbf{r},\mathbf{a})d\mathbf{r}, \tag{3}$$

where Υ is the set of points in the sphere. Given the virtual mass $\mu(\mathbf{r},t)$ of every point $\mathbf{r}$ at time t, the force exerted at the current focus point $\mathbf{a}(t)$ is assumed to decrease radially as:

$$\mathbf{F}(\mathbf{r},\mathbf{a}) = \gamma(t)\frac{1}{||\mathbf{r}-\mathbf{a}||^2}\mu(\mathbf{r},t)(\mathbf{r}-\mathbf{a}) \ . \tag{4}$$

The parameter $\gamma(t)$ weights the importance of the attraction force over time. We set

$$\gamma(t) = 1 - e^{(-\beta t)},$$

with parameter β to be a model parameter. This models the growing importance of the content over the *prediction horizon*: the motion continuity should be most important for short-term prediction, while the content diverts attention after a few seconds. The model input $\mu(\mathbf{r},t)$ for every pair $(\mathbf{r},t)$ can be set in different ways. Three cases are considered in this article. In Sec. IV-B, $\mu(\mathbf{r},t)$ is set to 0. In Sec. IV-C, $\mu(\mathbf{r},t)$ is set to the so called *ground-truth saliency map* $sal_{gt}(\mathbf{r},t)$. In Sec. IV-D, $\mu(\mathbf{r},t)$ is set to the element-wise product $so(\mathbf{r},t) \odot of(\mathbf{r},t)$, with $so(\mathbf{r},t)$ being a 0-1 pixel map of bounding boxes (1 inside, 0 outside) of detected objects, and $of(\mathbf{r},t)$ the optical flow at this pixel.

- a torque modeled as $-\lambda\omega$, corresponding to a force of friction modeling the energy dissipation when a user continues on their momentum, equivalently the fatigue or the principle of least effort in which humans tend to return static.

Computation of $\mathbf{a}(t)$: The final motion equation is therefore:

$$I\dot{\omega} = \left(\int_{\mathbf{r}\in\Upsilon} \mathbf{a} \times \mathbf{F}(\mathbf{r},\mathbf{a})d\mathbf{r}\right) - \lambda\omega \ , \tag{5}$$

where the first term in the right-hand-side is the torque associated with the field of force, $\times$ denoting the vector product. In the implementation, we drop I as parameters $\gamma(t)$ and λ in the right-hand-side can compensate for it. The evolution of $\mathbf{a}(t)$ is computed with quaternion rotations at each time instant:

$$\mathbf{a} = q \otimes \mathbf{a}_x \otimes q^{-1}, \tag{6}$$

where $\mathbf{a}_x$ is a constant unit vector, and $\otimes$ is the quaternion multiplication. As a consequence, considering the second order derivatives of the quaternion q:

$$\ddot{q} = \dot{q} \otimes q^{-1} \otimes \dot{q} + \frac{1}{2}\dot{\omega} \otimes q. \tag{7}$$

We can describe the dynamics of the system, by introducing the auxiliary variable $z(t) = \frac{d}{dt}(\mathbf{a}_x)$, with the system of first order differential equations:

$$\begin{cases} \mathbf{a}(t) = q(t) \otimes \mathbf{a}_x \otimes q^{-1}(t) \\ \dot{q}(t) = z(t) \\ \dot{z}(t) = \dot{q}(t) \otimes q^{-1}(t) \otimes \dot{q}(t) + \frac{1}{2}\dot{\omega} \otimes q(t), \end{cases} \tag{8}$$

subject to the boundary conditions $\mathbf{a}(t_0) = \mathbf{a}_0$, $\mathbf{a}_x = (1,0,0)$ and $z(t_0) = z_0$. If we pose $y = (q,z)$, then system (8) can be compactly re-written in the canonical form:

$$\dot{y} = \Phi(y,\mu,\gamma,\lambda), \tag{9}$$

that can be solved numerically by classic methods like Euler's and Runge-Kutta's. In this system of equations there are three parameters that are key in defining the head motion process:

- λ : models the fatigue of the user or the tendency to return to rest.
- $\gamma(t,\beta)$: models the strength of the forces from the visual input at each time-step.
- μ : the virtual masses generated from the visual input.

We vary these parameters throughout the paper to explain the usage of HeMoG to properly model the dynamics of head motion.

IV. Comparing Deep Models with HeMoG

In this section, we address Q1: To which extent can we investigate the inner workings of these DL models with a white-box model? To do so, we compare the performance of HeMoG with the reference DL models of [14].

A. Experimental Setup

1) Dataset: We selected the publicly available dataset of [12] to perform our experiments. This dataset consists of 208 omnidirectional videos. The duration of each video ranges between 15 and 80 seconds long (36s in average), each video is viewed by at least 31 participants. To perform the parameter estimation, we randomly selected a subportion of the traces of 166 videos (80%) and 15 users (50%) from the dataset, and exploited them to estimate the model parameters. Then the model with the parameters found is tested in the remaining traces (42 videos and 16 users), there is no overlap between videos or users in the train and test set. We subsampled all the videos in the dataset to 5 frames per second. The frames are resized to a resolution of 952×476.

Instead of using the equirectangular frame as visual input where the pixels in the poles are oversampled, the Vogel method [21] is employed to generate approximately uniformly distributed points on the sphere, as proposed in [22]. As illustration of the uniform sampling of the equirectangular frame using the Vogel method, the sampling of 200 points is shown in Fig. 2. In our experiments we used a sampling of 10000 points. Fig. 2 also shows the interaction between the field of forces $\mu(r, t)$, the position of the head $\mathbf{a}(t)$, the angular velocity ω and the torque τ.

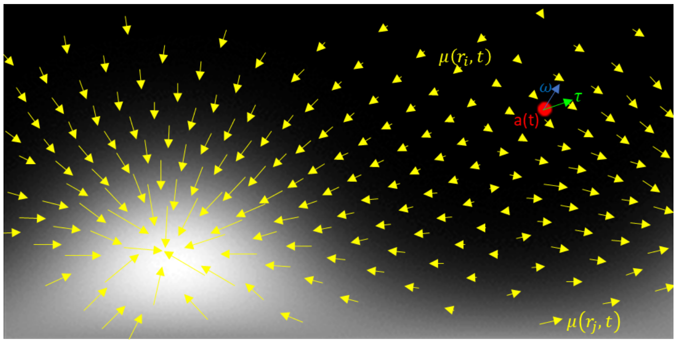

Fig. 2: Interaction between the Field of forces of a synthetic image and the position of the head $\mathbf{a}(t)$.

The integration of Eq. 9 that drives the focus of attention trajectory is based on the `odeint` function of Python SciPy library. The function is based on LSODA, which is a general purpose software that dynamically determines where the problem is stiff and chooses the appropriate solution method.

2) Problem definition and metric: We focus on the **dynamic prediction problem** that consists, at each video playback time t, in predicting the future user's head positions between t and $t + H$, with H being the *prediction horizon*. We set $H = 5$ sec. to match the settings of [14]. Let T be the video duration. We define the terms *prediction step* s, and video *time-stamp* t, such that: at every *time-stamp* $t \in [0, T]$, we run predictions $\hat{\mathbf{a}}(t+s)$, for all *prediction steps* $s \in [0, H]$. In what follows, t therefore identifies with t_0 and s with t in Sec. III, with initial conditions being position $\mathbf{a}(t)$ ($\mathbf{a}_0$) and current rotational velocity $\dot{q}(t)$ (z_0). We evaluate the predictions at every step s with the orthodromic distance between the ground-truth of the future position and the predicted positions. The orthodromic distance is the shortest distance of two points measured along the surface of the sphere, and is calculated as $D(\mathbf{a}(t + s), \hat{\mathbf{a}}(t + s)) = \arccos(\mathbf{a}(\mathbf{t} + \mathbf{s}) \cdot \hat{\mathbf{a}}(\mathbf{t} + \mathbf{s}))$, where $\cdot$ is the dot product operation.

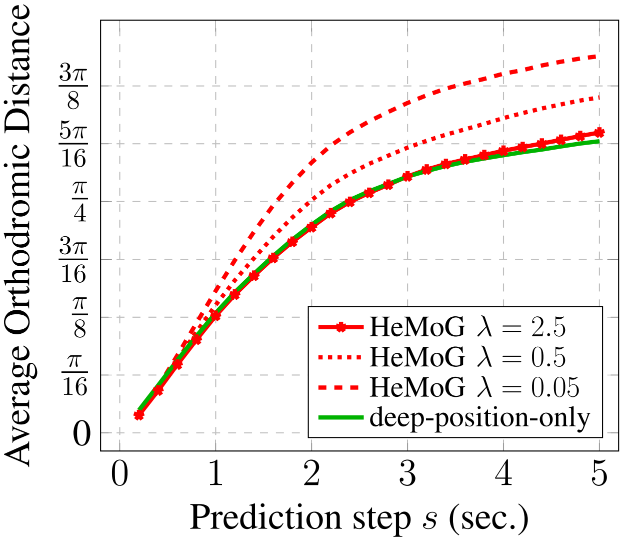

Fig. 3: Prediction error of HeMoG with $\lambda = 2.5$ (and $\beta = 0$) compared with the *deep-position-only* baseline. The performance of HeMoG with other values of $\lambda = 0.5$ and 0.05 are shown to illustrate the impact of the parameter.

B. HeMoG models well head motion continuity and attenuation

We first investigate the impact of parameter λ of HeMoG, which is meant to represent the attenuation of energy when the user continues on their momentum (modeled as a force of friction in Sec. III). To do so, we set the visual content weight $\gamma(s)$ to 0 by setting $\beta = 0$. We compare HeMoG against the DL model named *deep-position-only*, introduced in [14], because it uses only the history of past positions to make the predictions and it has been shown to outperform all previously existing DL models over all *prediction steps*. It is a Sequence-to-Sequence LSTM framework consisting in an encoder and a decoder that does not consider any visual input. The encoder receives the *historic window* input of past head positions and generates an internal representation that initializes the decoder producing the series of predictions.

1) Results: Fig. 3 depicts the results of HeMoG in the test set with the parameter $\lambda = 2.5$ tuned in the train set. We observe that $\lambda = 2.5$ yields performance of HeMoG close to that of the *deep-position-only* baseline. Fig. 3 also presents the results of HeMoG with other values of λ, a lower value of λ represents lower fatigue (more volatility), while a higher value of λ represents higher fatigue (motion reduced more quickly).

First, it is remarkable that such a white-box model predicts head motion as well as a DL model. Second, **we interpret this as the DL model *deep-position-only* learning the curvature**

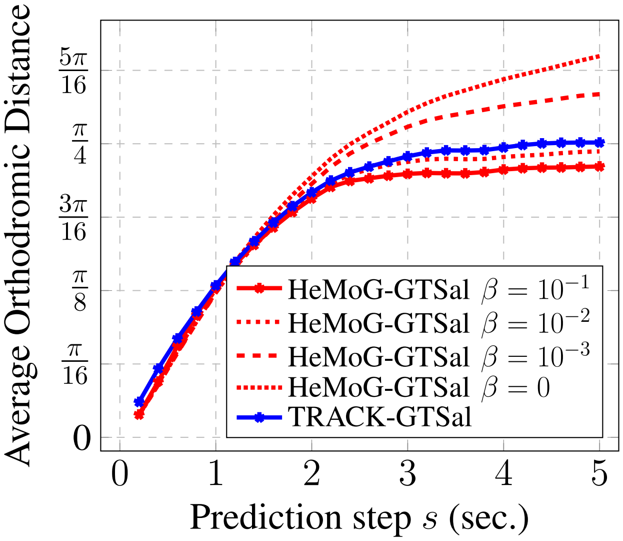

Fig. 4: Prediction error of HeMoG with $\lambda = 2.5$, $\beta = 10^{-1}$ and *ground-truth saliency* (GT-Sal) input, compared with TRACK. Other values of $\beta = 10^{-2}$ and 10^{-3} are shown to illustrate the impact of the parameter.

and friction dynamics of head motion that HeMoG is explicitly built on. This is the first element of answer to question **Q1**.

C. HeMoG combines well past motion and accurate content information

We study whether HeMoG correctly models the fusion between visual information and history of head positions. To do so, we keep λ set to 2.5 following the previous results. To be independent from the imperfection of any saliency predictor fed with the visual content, we consider here the *ground-truth saliency*: it is the heat map (2D-distribution) of the viewing patterns, obtained at each point in time from the users' traces. Here we compare HeMoG with the complete DL model TRACK from [14]. To compare HeMoG and TRACK fairly, we specify that TRACK is fed with the same type of visual content information as HeMoG.

1) Results: Fig. 4 presents the results of HeMoG fed with *ground-truth saliency* (named GTSal) with the value of $\beta = 0.1$ found in the train set. With $\beta = 0.1$, HeMoG performs similarly or slightly better than the DL model TRACK, which was shown to efficiently fuse the multi-modal inputs [14]. Fig. 4 also presents the results of HeMoG for lower values of β. The value of β affects the coefficient of the attraction force $\gamma(s)$ (the coefficient of the visual input) through $\gamma(s) = 1-e^{(-\beta s)}$. The higher the value of β the faster the growth of importance of the visual coefficient. Given that TRACK features a dedicated recurrent neural unit for each of both input modalities (past position and frame saliency) and a recurrent neural unit for the fusion of the so-obtained embeddings, TRACK has the flexibility to learn various ways of combining both modalities. The fact that **HeMoG, with its fixed fusion scheme shown in Eq. 5 performs as well or better can be interpreted as TRACK performing a similar type of fusion as HeMoG**, which enables to benefit from both types of information (the lowest curves in Fig. 4 are lower than those with the positional modality only in Fig. 3). This is the second element of answer to **Q1**.

D. HeMoG behaves as the DL model and lowers the impact of a noisy saliency estimate

In a non-ideal case where the saliency is not obtained from the viewing patterns but rather estimated from the content, we analyze the performance of HeMoG in comparison with TRACK. The extraction of visual saliency in 360° videos has been studied as an extension of image saliency [10]. However, additionally to the salient objects that can be found in images and videos, the motion of objects in the scene becomes an important cue specifically for videos [23]. For this reason we considered the *moving objects* as important cues to extract saliency from the 360° video content. The objects in each FoV of the scene are detected using YOLOv4 [24], the aggregation of all detected objects in all FoVs provides a binary map, shown in Fig. 5(c). To obtain the *moving objects* map shown in Fig. 5(d), we perform the element-wise product of the binary map and the norm of the pixel velocities computed in the 360° scene. This *moving objects* map obtained from the visual content is named the *content-based saliency* (CBSal).

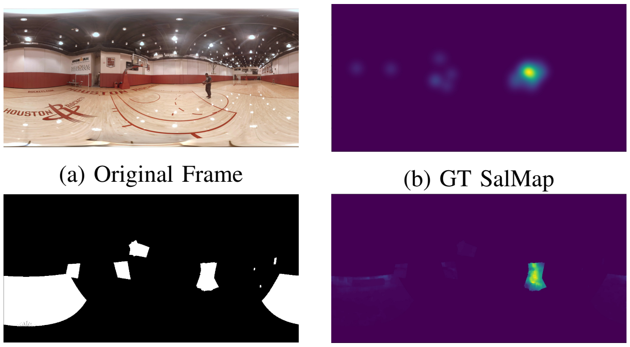
(a) Original Frame
(b) GT SalMap
(c) Detected Objects
(d) CB SalMap

Fig. 5: Saliency map extraction from a frame of video '072'. **(a)** Original frame. **(b)** *Ground-truth saliency* map. **(c)** Detected objects map. **(d)** *Content-based saliency* map: *moving objects* map.

1) Results: In Fig. 6, we present the results of our model HeMoG against the DL model TRACK using the same visual information CBSal as input. First, we observe that both models increase their error significantly when they use a noisy input for the visual saliency. Second, contrary to what occurred with *ground-truth saliency*, HeMoG performance improves by reducing the value of β, in other words, by minimizing the impact of the CBSal input. Using a value of $\beta = 10^{-5}$, HeMoG approaches the performance of TRACK. **This reinforces the hypothesis that TRACK and HeMoG perform the same type of fusion**.

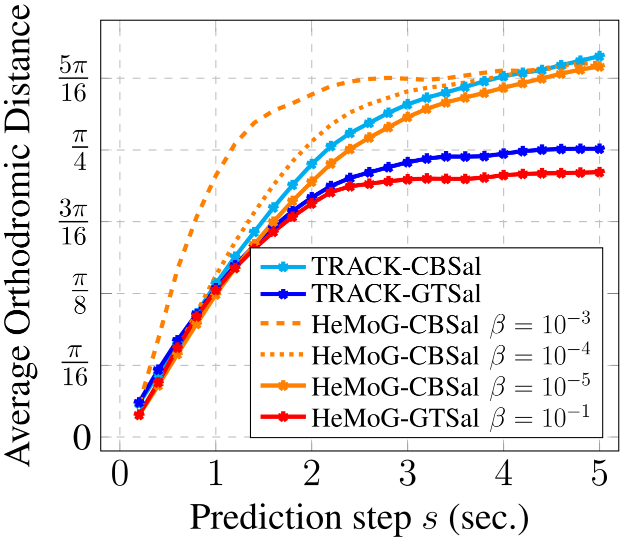

Fig. 6: Prediction error of HeMoG with $\lambda = 2.5$, $\beta = 10^{-5}$ and *content-based saliency* (CBSal), compared with TRACK using CBSal. The curves of HeMoG and TRACK with GTSal are shown for reference. Other values of $\beta = 10^{-3}$, 10^{-4} are shown to illustrate the impact of the parameter.

V. Impact of the visual saliency on head motion

In this section, we address Q2 by analyzing the impact of the visual saliency on head motion, in terms of the video category and the time-step in the *prediction horizon*.

A. *Visual saliency impacts head motion only for certain video categories*

The videos from the dataset of [12], contain heterogeneous scenes including music shows, documentaries, sports, movies, etc. More generally, [25] have identified the following main video categories for which they could discriminate significantly different users' behaviors: *Exploration, Static Focus, Moving Focus* and *Rides*. In *Exploration* videos, there is no specific attraction point and the spatial distribution tends to be more widespread and hence individual trajectories more difficult to predict. *Static Focus* videos are made of a single or few attraction areas (e.g., a standing person in an empty room). In *Moving Focus* videos, the attraction points move over the sphere. *Rides* videos are shot with the 360° camera moving. In this case, the attraction point for the user is usually the camera moving direction to minimize motion sickness.

We categorized each of the videos in the dataset of [12] into one of the four groups: *Exploration, Static Focus, Moving Focus* or *Rides*. The number of videos belonging to each of the classes is: 16 videos of *Rides*, 100 *Exploratory* videos, 74 *Moving Focus* and 18 *Static Focus* videos. In Fig. 7, we show some of the videos from the dataset with their respective category.

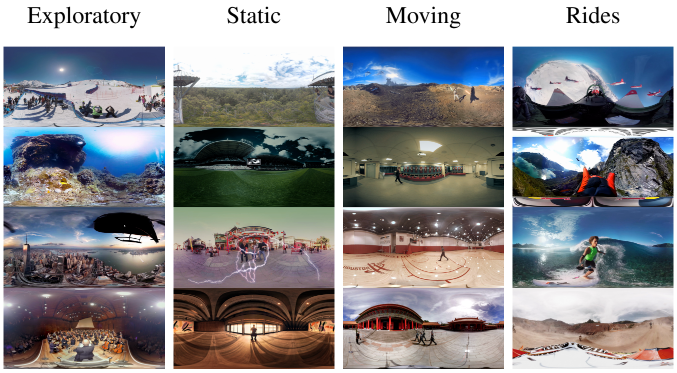

Fig. 7: Some videos from [12], categorized into *Exploratory, Static (focus), Moving (focus)* and *Rides*

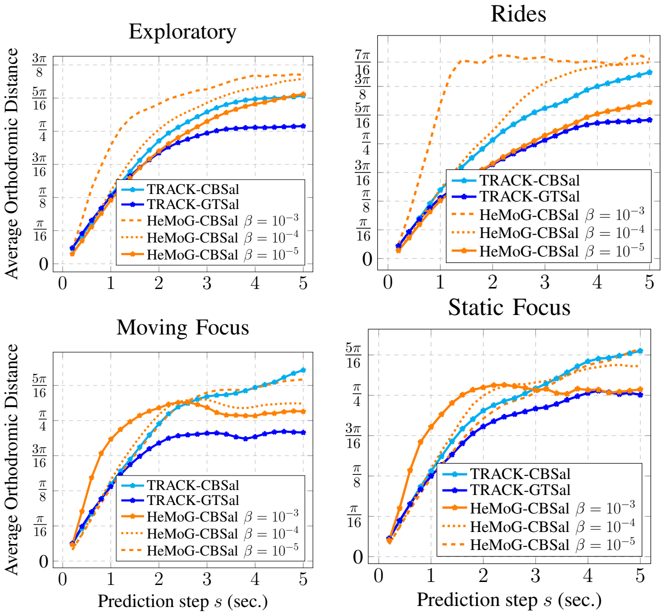

Fig. 8: Prediction error of HeMoG compared with TRACK grouped per category. **Top-left**: *Exploratory*. **Top-right**: *Rides*. **Bottom-left**: *Moving Focus*. **Bottom-right**: *Static Focus*.

In Fig. 8, we present the results of HeMoG per video category, with CBSal and for different values of β. For *Exploration* videos, as several objects around the sphere are equally attractive, we expect that there is no much information about saliency that could be captured in CBSal, and as in *Rides* videos the most salient point is around the direction of camera motion, CBSal cannot capture the relevant information from these videos. Indeed, CBSal is the product of the optical flow with the objects bounding boxes, and hence the camera direction where the optical flow is minimum cannot be highlighted as salient this way. Fig. 8 confirms that the lowest values of β are those providing best results for *Exploration* and *Rides* videos. HeMoG therefore reduces the weight of the saliency information in these cases (as it is also possibly the behavior of the DL model TRACK given its curve).

For the *Moving Focus* and *Static Focus* categories, we observe that when we increase the value of β, the error in the long-term decreases, showing the relevance of the saliency information for longer-term prediction. However, the error in the short-term increases, which we discuss in the next section. **These results with different optimal values of β per video category show that the impact of saliency on head motion is stronger for *Static Focus* and *Moving Focus* (and likely for *Rides* too) than for the *Exploration* category**.

B. Visual saliency impacts head motion only after 3 seconds

As discussed above, increasing β in *Static Focus* and *Moving Focus* videos lowers the error in the long-term *prediction steps* but degrades it in the short-term. This reveals a possibly not optimal choice of the $\gamma(s)$ function that controls the rapidity of importance growth of the saliency information over s. For now we have set, from Sec. III, $\gamma(s) = 1 - e^{(-\beta s)}$. We ran numerical searches and identified that the values of $\gamma(s)$ that give the best performance of the gravitational model per time-step are:

$$\gamma(s) = \begin{cases} 10^{-}5 & \text{if } 0 < s <= 3 \\ 10^{-}1 & \text{if } 3 < s <= 5, \end{cases} \quad (10)$$

from which we draw two conclusions. First, **the motion momentum is more important than the visual content in the first 2.5 seconds of the *prediction horizon*, and the visual content can inform the head motion prediction model only for horizons longer than 3 seconds**. Second, that a sigmoid-like function $\gamma(s) = \frac{C}{1+\exp(-\beta(s-S))}$ with additional parameters C for the scaling and S to center the transition from 0 to 1, would be a better fit. This is confirmed in Fig. 9 with the comparison of HeMoG when the parameters are set properly for the different categories and for each prediction step s. Let us note that this also shows that the DL model TRACK is capable to learn and adapt to the different video categories, while the white-box approach is limited by the right choices of parameters. However, we show here that the differential equation model of HeMoG captures the main dynamics and yields performance similar to the DL models in average.

VI. Discussion

Comparison of HeMoG with Deep Learning models: The main difference between both types of models is as follows. The latter are equipped with representation learning capability (learning how to extract relevant features from the saliency map they are fed) and able to modulate the weights assigned to momentum and saliency features in the fusion depending on the saliency and motion information (capabilities detailed in [26]). In comparison, HeMoG is able to properly fuse both types of information for any video, provided that the saliency information is the ground-truth. When it is not the ground-truth anymore (when fed with CB-sal), then the saliency weight β in the HeMoG model must be adapted to the video category. Also, we mention that when using other types of saliency extracted from the content, for example the saliency maps obtained from PanoSalNet [10], the performance of HeMoG shown in Fig. 10 is slightly worse than that of TRACK, which we explained by the lower level of information present in the estimated saliency map in relation with the user motion.

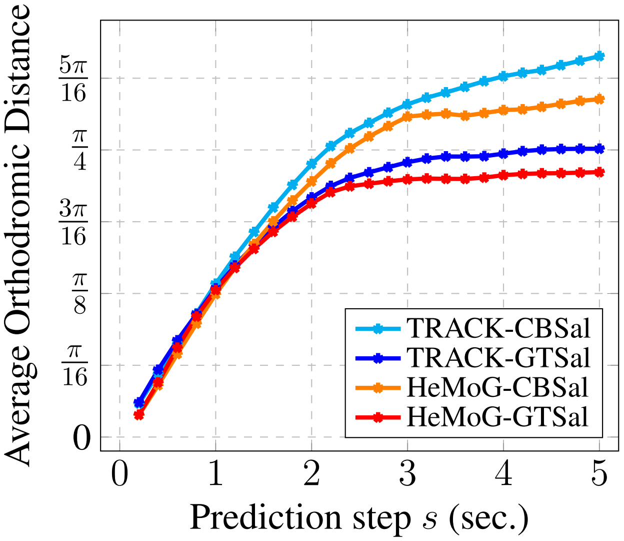

Fig. 9: Results averaged over all video categories. HeMoG is set with $\gamma(s) = 1 - e^{(-\beta s)}$ and $\beta = 10^{-}5$ for *Exploratory* and *Rides* videos, and $\gamma(s)$ from Eq. 10 for *Moving Focus* and *Static Focus* videos, compared with TRACK. The curves of HeMoG and TRACK with GTSal are shown for reference.

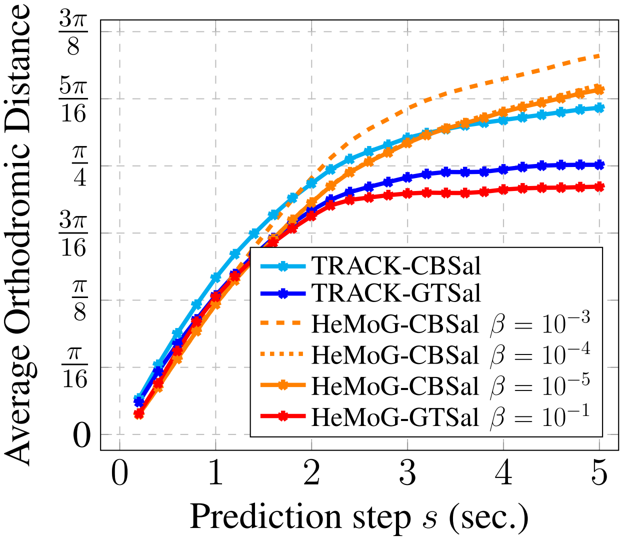

Fig. 10: Prediction error of HeMoG with $\lambda = 2.5$, $\beta = 10^{-5}$, and *content-based saliency* (CBSal) computed from PanoSalNet [10], compared with TRACK using the same CBSal-PanoSalNet. The curves of HeMoG and TRACK with GTSal are shown for reference. Other values of $\beta = 10^{-3}$, 10^{-4} are shown to illustrate the impact of the parameter.

On focus-type videos, TRACK is able to extract some useful information (improvement compared with *deep-position-only*), while HeMoG is not.

Indeed, the Content-Based saliency obtained from PanoSalNet can be noisy, as we show in Fig. 11 with an original frame of video '072' and its extracted saliency with PanoSalNet, and the Ground-truth saliency from user statistics on this frame. While the salient object in the frame is the human, low-level features like the lights reflected in the floor and high-level features like the text written in the floor are taken into account by the saliency extractor, making the resulting saliency map noisy and more difficult to get motion-relevant information from.

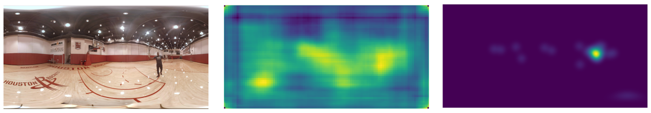

Fig. 11: Saliency map extraction from a frame of video '072'. **(left)** Original frame. **(center)** *PanoSalNet saliency* map. **(c)** *Ground-truth saliency* map

Generalizing to more comprehensive saliency maps: While the study of this paper has been restricted to saliency attractors based on moving objects, we can consider extending to static objects whose importance can be ranked, meaning that the trajectory of focus of attention is also subjected to a gravitational field created by static objects [27], [28]. We can also study how to improve for the case of *Rides* scenes characterized by camera motion. To have more solid estimates of pixel velocities, methods for camera motion estimation [29] are already present in the literature and can help in creating more suitable saliency estimates for the proposed model. The treatment of such complex scenes is left for future work.

Integration of HeMoG with Gaze-based models: The motion of eyes is important to determine the exploration behaviors of people watching 360° videos. While eye-tracking data is outside of the scope of this work, our model HeMoG could be extended to get both the head and eye movements from a 360° video source. Given the head orientation predicted by HeMoG, the FoV can be cropped from the 360° content. Then, a model for regular 2D videos as the one proposed in [20] can be used on this planar FoV section to predict plausible human gaze scanpaths.

VII. Conclusions

In this article we have investigated the human head motion process driven by attention when a user experiences an immersive 360° video. We have first introduced a new computational model named HeMoG, enabling to predict future head positions from the user's past positions and the visual content. HeMoG is built on differential equations obtained from the physics of rotational motion where the attractive salient areas in the 360° frames are represented as virtual masses. HeMoG is hence a white-box model and its (time-varying) parameters control the connection between visual content and head motion process. The performance of HeMoG can be comparable with those of DL predictors, which we interpret as the DL models learning the same type of fusion as HeMoG: curvature continuity and momentum attenuation from friction in the short-term, diversion of motion with saliency attraction in the longer-term. The evolution of best parameter values in terms of video categories and horizon reveals that, on videos that are not exploratory, the initial motion momentum is most important until ca. 3s, after which the saliency weights more in the motion equation. Future works include refining the saliency extractor to feed the model with, and incorporating these findings into an attention-driven system to produce personalized immersive environments.

Acknowledgment

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] S. Rothe, D. Buschek, and H. Hußmann, "Guidance in cinematic virtual reality-taxonomy, research status and challenges," *Multimodal Technologies and Interaction*, vol. 3, no. 1, p. 19, 2019.

[2] S. Dambra, G. Samela, L. Sassatelli, R. Pighetti, R. Aparicio-Pardo, and A.-M. Pinna-Déry, "Film editing: New levers to improve vr streaming," in *Proceedings of the 9th ACM Multimedia Systems Conference*, 2018, pp. 27–39.

[3] C. Marañes, D. Gutierrez, and A. Serrano, "Exploring the impact of 360 movie cuts in users' attention," in *2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. IEEE, 2020, pp. 73–82.

[4] H.-Y. Wu, T.-Y. Li, and M. Christie, "Logic control for story graphs in 3d game narratives," in *Smart Graphics*, Y. Chen, M. Christie, and W. Tan, Eds. Cham: Springer International Publishing, 2017, pp. 111–123.

[5] F. Qian, L. Ji, B. Han, and V. Gopalakrishnan, "Optimizing 360 video delivery over cellular networks," in *Workshop on All Things Cellular: Operations, Applications and Challenges*. ACM, 2016, pp. 1–6.

[6] J. Chakareski, R. Aksu, X. Corbillon, G. Simon, and V. Swaminathan, "Viewport-driven rate-distortion optimized 360° video streaming," in *2018 IEEE International conference on communications (ICC)*. IEEE, 2018, pp. 1–7.

[7] S. A. McMains and S. Kastner, *Visual Attention*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 4296–4302. [Online]. Available: https://doi.org/10.1007/978-3-540-29678-2_6344

[8] O. Le Meur, P. Le Callet, and D. Barba, "Predicting visual fixations on video based on low-level visual features," *Vision research*, vol. 47, no. 19, pp. 2483–2498, 2007.

[9] V. Sitzmann, A. Serrano, A. Pavel, M. Agrawala, D. Gutierrez, B. Masia, and G. Wetzstein, "Saliency in VR: How do people explore virtual environments?" *IEEE Trans. on Visualization and Computer Graphics*, vol. 24, no. 4, pp. 1633–1642, 2018.

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

[11] O. Le Meur and T. Baccino, "Methods for comparing scanpaths and saliency maps: strengths and weaknesses," *Behavior research methods*, vol. 45, no. 1, pp. 251–266, 2013.

[12] Y. Xu, Y. Dong, J. Wu, Z. Sun, Z. Shi, J. Yu, and S. Gao, "Gaze prediction in dynamic 360° immersive videos," in *IEEE CVPR*, 2018, pp. 5333–5342.

[13] M. Xu, Y. Song, J. Wang, M. Qiao, L. Huo, and Z. Wang, "Predicting head movement in panoramic video: A deep reinforcement learning approach," *IEEE Trans. on PAMI*, 2018.

[14] M. F. Romero Rondon, L. Sassatelli, R. Aparicio-Pardo, and F. Precioso, "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.

[15] Y. Li, Y. Xu, S. Xie, L. Ma, and J. Sun, "Two-layer FoV prediction model for viewport dependent streaming of 360° videos," in *EAI Int. Conf. on Communications and Networking (ChinaCom)*, Chengdu, China, Oct. 2018.

[16] C.-L. Fan, J. Lee, W.-C. Lo, C.-Y. Huang, K.-T. Chen, and C.-H. Hsu, "Fixation prediction for 360 video streaming in head-mounted virtual reality," in *ACM NOSSDAV*, 2017, pp. 67–72.

[17] N. Xie, G. Ras, M. van Gerven, and D. Doran, "Explainable deep learning: A field guide for the uninitiated," *CoRR*, vol. abs/2004.14545, 2020. [Online]. Available: https://arxiv.org/abs/2004.14545

[18] J. Chen, X. Luo, M. Hu, D. Wu, and Y. Zhou, "Sparkle: User-aware viewport prediction in 360-degree video streaming," *IEEE Transactions on Multimedia*, 2020.

[19] M. R. Miller, F. Herrera, H. Jun, J. A. Landay, and J. N. Bailenson, "Personal identifiability of user tracking data during observation of 360-degree vr video," *Scientific Reports*, vol. 10, no. 1, pp. 1–10, 2020.

[20] D. Zanca, M. Gori, S. Melacci, and A. Rufa, "Gravitational models explain shifts on human visual attention," *Scientific Reports*, vol. 10, no. 1, pp. 1–9, 2020.

[21] R. Swinbank and R. J. Purser, "Fibonacci grids," in *Conference on Numerical Weather Prediction*, vol. 13, 1999, p. 125.

[22] H. Fassold, "Adapting computer vision algorithms for omnidirectional video," in *Proceedings of the 27th ACM International Conference on Multimedia*, 2019, pp. 1026–1028.

[23] L. Maczyta, P. Bouthemy, and O. Le Meur, "Unsupervised motion saliency map estimation based on optical flow inpainting," in *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2019, pp. 4469–4473.

[24] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "Yolov4: Optimal speed and accuracy of object detection," *arXiv preprint arXiv:2004.10934*, 2020.

[25] M. Almquist, V. Almquist, V. Krishnamoorthi, N. Carlsson, and D. Eager, "The prefetch aggressiveness tradeof in 360 video streaming," in *ACM Int. Conf. on Multimedia Systems (MMSys)*, 2018.

[26] D. Alvarez-Melis and T. S. Jaakkola, "Towards robust interpretability with self-explaining neural networks," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, ser. NIPS'18. Red Hook, NY, USA: Curran Associates Inc., 2018, p. 7786–7795.

[27] G. Ma, S. Li, C. Chen, A. Hao, and H. Qin, "Stage-wise salient object detection in 360° omnidirectional image via object-level semantical saliency ranking," *IEEE Transactions on Visualization and Computer Graphics*, vol. 26, no. 12, pp. 3535–3545, 2020.

[28] J. Li, J. Su, C. Xia, and Y. Tian, "Distortion-adaptive salient object detection in 360° omnidirectional images," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 1, pp. 38–48, 2019.

[29] D. Kim, S. Pathak, A. Moro, A. Yamashita, and H. Asama, "Selfsphnet: Motion estimation of a spherical camera via self-supervised learning," *IEEE Access*, vol. 8, pp. 41 847–41 859, 2020.