Viewport Forecasting in 360° Virtual Reality Videos with Machine Learning

1st Johanna Vielhaben
Machine Learning Group
Fraunhofer Heinrich-Hertz-Institute
Berlin, Germany
johanna.vielhaben@hhi.fraunhofer.de

2nd Hüseyin Camalan
Machine Learning Group
Fraunhofer Heinrich-Hertz-Institute
Berlin, Germany
camalanhuseyin@gmail.com

3nd Wojciech Samek
Machine Learning Group
Fraunhofer Heinrich-Hertz-Institute
Berlin, Germany
wojciech.samek@hhi.fraunhofer.de

4nd Markus Wenzel
Machine Learning Group
Fraunhofer Heinrich-Hertz-Institute
Berlin, Germany
markus.wenzel@hhi.fraunhofer.de

***Abstract*—*Objective*. Virtual reality (VR) cloud gaming and 360° video streaming are on the rise. With a VR headset, viewers can individually choose the perspective they see on the head-mounted display by turning their head, which creates the illusion of being in a virtual room. In this experimental study, we applied machine learning methods to anticipate future head rotations (a) from preceding head and eye motions, and (b) from the statistics of other spherical video viewers. *Approach*. Ten study participants watched each 3⅓ hours of spherical video clips, while head and eye gaze motions were tracked, using a VR headset with a built-in eye tracker. Machine learning models were trained on the recorded head and gaze trajectories to predict (a) changes of head orientation and (b) the viewport from population statistics. *Results*. We assembled a dataset of head and gaze trajectories of spherical video viewers with great stimulus variability. We extracted statistical features from these time series and showed that a Support Vector Machine can classify the range of future head movements with a time horizon of up to one second with good accuracy. Even population statistics among only ten subjects show prediction success above chance level. *Significance*. Viewport forecasting opens up various avenues to optimize VR rendering and transmission. While the viewer can see only a section of the surrounding 360° sphere, the entire panorama has typically to be rendered and/or broadcast. The reason is rooted in the transmission delay, which has to be taken into account in order to avoid simulator sickness due to motion-to-photon latencies. Knowing in advance, where the viewer is going to look at may help to make cloud rendering and video streaming of VR content more efficient and, ultimately, the VR experience more appealing.**

***Index Terms*—machine learning, virtual reality, cloud gaming, 360°video, body motion prediction, eye tracking, head-mounted display.**

I. Introduction

Virtual reality (VR) headsets let us immerse in virtual rooms, because our perspective changes when we move the head. Wearing the VR glasses, we can play games generated with computer graphics and watch spherical (360°) videos, filmed with omnidirectional camera rigs. The headset creates the sensory illusion of being in a virtual room by tracking our head pose and by displaying only the corresponding section of the VR content, which matches our current direction of view (cf. Fig. 1). In the present work, we let statistical algorithms learn to forecast the viewport from head and gaze motion trajectories of viewers of spherical videos wearing a VR headset. The question of *"Where do people look at in spherical videos?"* has practical implications, and is also of interest for basic research. Practical implications are manifold, ranging from insights for rendering, editing and presenting virtual reality content [1], [2] to efficient solutions for streaming [3]–[10]. Obviously, we turn not only our head, but also our eye gaze towards points of interest, because the fovea centralis of the retina provides the best visual acuity. The advent of affordable tiny eye trackers that can be built into VR goggles, as in the present study, has led to the idea of allocating more computational resources or bandwidth to the image area captured by the fovea [11], [12]. In a recent study, head and eye motions in spherical video were predicted based on the saliency and optical flow of the video content in combination with the head and eye motion trajectory [13]. In another approach, long-term viewport prediction was attempted by matching a users viewport evolution to the representative head trajectory of clustered trajectories from past viewers [14].

Adding to previous investigations, we forecast the head pose in the present study from user-generated data only, and let machines learn to predict head motion trajectories (a) from individual head pose and eye tracking signals, and (b) by exploiting patterns in the user population statistics of these signals. Our study started with the considerations that eye motions are likely to precede and thereby indicate future head motions and that people are most likely to look where other people look. In order to test these hypotheses, we have conducted an experimental study where every participant watched 3⅓ hours of spherical videos, wearing a virtual-reality headset with a

This work was supported by the Bundesministerium für Bildung und Forschung (BMBF) through the Berlin Big Data Center under Grant 01IS14013A and the Berlin Center for Machine Learning under Grant 01IS180371.

built-in eye tracker (cf. Fig. 1). So far, experimental studies measuring head and eye motions of viewers of spherical videos are still rare [13]. Crucially, our study is characterized by a great stimulus variability, because each participant viewed various, realistic videos, which resulted in an extensive data set. The data-driven methods presented here are expected to further virtual reality applications, both on the technical side (e.g., for efficient streaming of 360° video, or for remote rendering of virtual reality content in cloud-based gaming), and on the artistic side (e.g., for interactive games, which anticipate and play with the user actions, or for improved 360° movie editing).

II. VR Experiment

A. Experimental Design

Each participant of the study watched 50 spherical video clips in random order, wearing a virtual reality headset, while standing, and while head and eye motions were measured. The content of the videos is diverse and comprises, for instance, filmed documentations, concerts, short films and movie trailers, as well as rendered animations. The clips have an average duration of about four minutes, ranging from 1½ minutes to 10½ minutes, and sum up to a total duration of about 3⅓ hours. The spherical videos were downloaded from the content platforms "YouTube" and "Vimeo" in the best possible resolution (mostly in 4K) using the software "4K Video Downloader" (Open Media LLC) The video clips sum to 29.4 Gigabyte, where 20 clips were projected in the equirectangular layout and 30 were projected in the cube map layout. An overview of the video clips is given in Table III.

B. Experimental Apparatus

The virtual reality headset "HTC Vive" (HTC Corporation, Taiwan, and Valve Corporation, USA) displayed the videos with a refresh rate of $90\,\mathrm{Hz}$ and with a resolution of 1080×1200 pixels per eye. The field of view is approximately circular and has a diameter of about $110°$. The software "Whirligig Player" served as spherical video player (http://www.whirligig.xyz), running on the software platfom "SteamVR" (Valve). The following signals streams were measured and recorded in synchrony with the "LabRecorder" of the "Lab Streaming Layer" (LSL; Swartz Center for Computational Neuroscience, University of California, San Diego, USA): The eye gaze was measured by the "HTC Vive Binocular Add-on" integrated in the virtual reality headset, and the corresponding software "Pupil" (Pupil Labs, Germany) [15]. "Pupil" measures the gaze as x and y coordinates $\mathbf{r}$ within the viewport at a frame rate of $119\,\mathrm{Hz}$. The head pose was tracked by the headset that is equipped with a gyroscope, an inertial measurement unit, infrared light receivers and external infrared emitters mounted on tripods in an overhead position. The head orientation (represented by rotation matrices $\mathbf{R}$) was accessed with the software "OpenVR" (Valve) and recorded at $119\,\mathrm{Hz}$ with LSL. The current running time of the videos was obtained from the "Time Code Server" of the Whirligig Player at $10\,\mathrm{Hz}$ and forwarded to LSL.

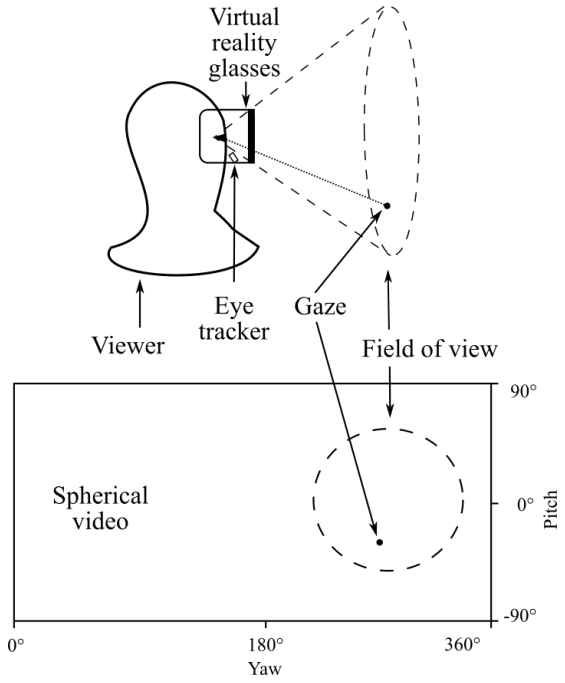

Fig. 1. The participants of the study viewed spherical videos through virtual reality glasses with a built-in eye tracker. The head pose determines the viewport on the spherical video. The retina of each eye provides the best visual acuity in the fovea with a range of about $5°$. Gaze movements shift the location of the fovea across the viewport.

C. Data Acquisition

Ten people participated in the experimental study (age: 27.2 ± 1.9 years; mean $\pm$ standard deviation; 7 female, 3 male; 5 with normal vision, 4 with contact lenses, 1 wearing spectacles). Before starting the experiment, the test subjects gave their informed written consent to participate in the study and to the data acquisition and processing. After the experiment, the test subjects received an allowance for the participation. A fast-track self-assessment of the study had resulted in a positive evaluation from the ethics commission of the Technische Universität Berlin (N°: FT_2017_36). The participants were instructed to watch the videos at their pleasure. They could take breaks after every video if desired. While the test subject rested, the experimenter had the opportunity to cool the virtual reality headset and built-in eye tracker with cool packs (about every half hour), which would have otherwise resulted in display glitches due to rapid heating-up. One study participant skipped few videos, which were a bit eerie. Therefore, four videos were removed from the analysis in Section III. In order to map the output of the eye tracker to orientation angles in the spherical video, we calibrated and validated the eye tracker before the experiment and after each break. For this purpose, we had prepared a calibration and a validation video where dots appeared at known locations on the sphere within the viewport, while the test subject kept the head still.

III. Forecasting the Viewport

We follow two approaches to forecast the viewport: based on the recognition of patterns in individual head and gaze trajectories (Section III-A) and by exploiting user population statistics of these signals (Section III-B).

A. Forecasting changes in head orientation

We anticipated how the head orientation changed within a time window $[t_0, t_0 + \Delta t]$. Past head and gaze information from an epoch $[t_0 - w, t_0]$ are exploited to infer whether the future angular head translation $\Delta\Phi$ during $[t_0, t_0 + \Delta t]$ will surpass a threshold $\Delta\Phi_c$ (cf. Fig. 2).

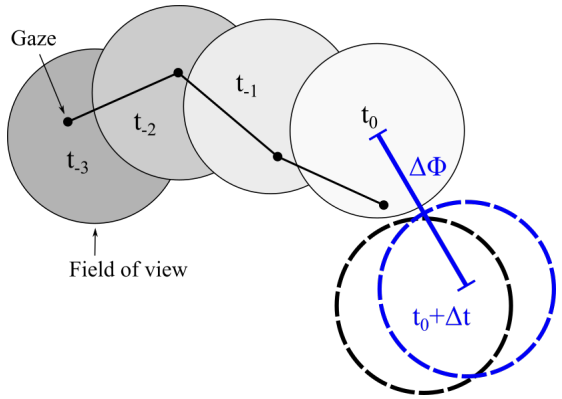

Fig. 2. Classification of head movements during $[t_0, t_0 + \Delta t]$, given past and present head and eye gaze trajectories (represented by discs and dots). We predicted if the non-directional angular translation $\Delta\Phi$ of the viewport was likely to surpass a threshold $\Delta\Phi_c$ within Δt.

For this purpose, we *a)* extracted statistical features from the head and gaze trajectories, *b)* applied a Support Vector Machine to classify if the head translation $\Delta\Phi$ surpassed a threshold $\Delta\Phi_c$ or not, and *c)* evaluated the predictive performance.

a) Feature extraction: First, we converted the head rotation matrices $\mathbf{R}_t$ to quaternions $\mathbf{q}_t$ for a lower dimensional representation and to be able to compute meaningful averages. Furthermore, the original multimodal data were measured in synchrony but not necessarily at the same time. To obtain measurements sampled regularly at equidistant time points, we resampled the recorded head and eye gaze trajectories to a frequency of 100 Hz. The data were binned into intervals of 0.01 s for which the head orientation quaternion average and the gaze point average were computed. The quaternions were averaged using the Spherical Linear Interpolation (SLERP) equation for two quaternions $\mathbf{q}_1, \mathbf{q}_2$ [16]:

$$SLERP(\mathbf{q}_1, \mathbf{q}_2, \alpha) = \mathbf{q}_1 \left(\mathbf{q}_1^{-1}\mathbf{q}_2\right)^{\alpha} \tag{1}$$

with a weight $\alpha \in [0, 1]$. To average over a set of more than two quaternions $\{\mathbf{q}_1, \mathbf{q}_2, \dots \mathbf{q}_N\}$ the *SLERP* equation is iterated:

Initialize $i := 0$
while $i < N - 1$ **do**
  $i < N - 1$
  $\alpha := \frac{1}{i+1}$
  $q_{i+1} := SLERP(\mathbf{q}_i, \mathbf{q}_{i+1}, \alpha)$
end while

We then segmented the head and gaze trajectories into non-overlapping epochs of window length w, represented by quaternions $\{\mathbf{q}_1, \mathbf{q}_2, \dots, \mathbf{q}_w\}$ and two dimensional gaze coordinates $\{\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_w\}$, respectively. We labeled each epoch with the actual translation $\Delta\Phi$ of the viewport during the time interval $[t_0, t_0 + \Delta t]$ and binarized the labels depending on whether the translation surpassed a threshold $\Delta\Phi_c$. Each epoch contains points on a trajectory recorded for one subject and one video, only. The following statistical features were computed and normalized to represent each epoch:

- Standard deviation of the gaze coordinates r_x and r_y

$$\text{std}_{\text{eye}x} = \sqrt{\frac{1}{w}\sum_{k=1}^{w}\left(r_{x,k} - \overline{r_x}\right)^2} \tag{2}$$

$$\text{std}_{\text{eye}y} = \sqrt{\frac{1}{w}\sum_{k=1}^{w}\left(r_{y,k} - \overline{r_y}\right)^2} \tag{3}$$

- Autocorrelation of the gaze coordinates r_x and r_y

$$\text{aucorr}_{\text{eye}x}(\tau) = \frac{\sum_{i=1}^{w-\tau}(r_{x,k} - \overline{r_x})(r_{x,k+\tau} - \overline{r_x})}{\sum_{i=1}^{w}(r_{x,k} - \overline{r_x})^2} \tag{4}$$

$$\text{aucorr}_{\text{eye}y}(\tau) = \frac{\sum_{i=1}^{w-\tau}(r_{y,k} - \overline{r_y})(r_{y,k+\tau} - \overline{r_y})}{\sum_{i=1}^{w}(r_{y,k} - \overline{r_y})^2} \tag{5}$$

- Root-mean-square of Euclidean distances dr_k of successive gaze points $\mathbf{r}_t, \mathbf{r}_{t+1}$

$$\text{rmssd}_{\text{eye}} = \sqrt{\frac{1}{w-1}\sum_{k=1}^{w-1} dr_k^2} \tag{6}$$

$$\text{d}r_k = \|\mathbf{r}_{k+1} - \mathbf{r}_k\| \tag{7}$$

- Standard deviation of Euclidean distances between successive gaze points $\text{d}r_t$

$$\text{stdd}_{\text{eye}} = \sqrt{\frac{1}{w-1}\sum_{k=1}^{w-1}\left(\text{d}r_k - \overline{\text{d}r_k}\right)} \tag{8}$$

- Root-mean-square of angular displacement dq_k between successive head orientations $\mathbf{q}_t, \mathbf{q}_{t+1}$

$$\text{rmssd}_{\text{head}} = \sqrt{\frac{1}{w-1}\sum_{k=1}^{w-1} dq_k^2} \tag{9}$$

$$\text{d}q_k = 2\arccos\left(\mathbf{q}_k \cdot \mathbf{q}_{k+1}\right) \tag{10}$$

- Standard deviation of angular distances between successive head orientations $\text{d}q_t$

$$\text{stdd}_{\text{head}} = \sqrt{\frac{1}{w-1}\sum_{k=1}^{w-1}\left(\text{d}q_k - \overline{\text{d}q_k}\right)} \tag{11}$$

The time lag τ is chosen as $50\,\mathrm{ms}$.

We aimed for a straightforward collection of features. Regarding the gaze coordinates, we hypothesized that the standard deviation and autocorrelation for each direction capture the relevant information for future head movements. As the representation of the head orientation is more complex, we chose two simple features based on angluar displacement between successive points and also added these for the Euclidean distances between gaze points. These measures, among others, were also employed to classify heartbeat time series using Support Vector Machines [17].

b) Classification with a SVM: We trained a linar SVM to classify the binary labeled epochs represented by the eight statistical features described above.

c) Evaluation: We evaluated the predictive performance of the SVM by the Area Under the Curve of Receiver Operating Characteristics (AUC ROC) and by precision (fraction of the number of truly positive classified instances and the number of positive classified instances) in two grouped 5-fold cross validations (CV). For the first CV, we partitioned the epochs into 5 non-overlapping folds grouped by subjects (2 subjects per fold). For the second CV, we partitioned the epochs into 5 folds grouped by videos (10 videos per fold).

B. Forecasting the viewport from population statistics

In a second line of research, we applied several regression models to predict one subject's head orientation at every time point of each video based on the head of other viewers (cf. Fig 3). The recorded data went through a processing pipe line of *a)* feature extraction, *b)* regression of the head orientation and *c)* and evaluation of the predictive performance in a leave-one-out cross validation.

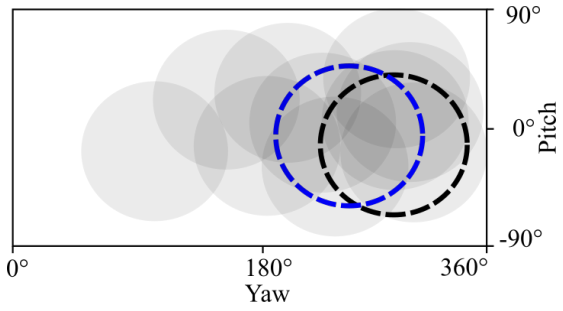

Fig. 3. Predicting the head pose of one viewer (blue circle) at a given time point from the head poses of all other viewers (illustrated by shaded discs). In this illustration, the black dashed circle represents the actual head pose of the viewer at this time point of the video.

a) Feature extraction: We re-sampled the head orientation trajectories and the video time series with a much larger bin size of $1\,\mathrm{s}$ and an inter-bin interval of $0.5\,\mathrm{s}$. We assigned the head orientation $\mathbf{q}_{t,i,j}$ as a label to a time point t for each video i and each subject j. The features of this time point t are the head orientations $\{\mathbf{q}_{t,i,k}\}$ of other subjects $k \neq j$ watching the same video i.

b) Regression: We used linear regression [18], linear support vector regression [19] and multilayer perceptron regression [20] to map the head orientation of other viewers to the head orientation of one subject for a time point in a video. As a baseline for comparison, we simply averaged the head orientations of other viewers to estimate the head orientation of the subject under consideration using the SLERP equation 1 (AVRG). Parameters for the models were set as follows:

- Linear regression (LR): ordinary least squares with intercept term
- Linear support vector regression (LSVR): epsilon-insensitive loss function ($\epsilon = 0$), crossing penalty parameter C set to 1, intercept calculated, each quaternion element predicted separately (SVR does not allow for multidimensional targets)
- Multilayer perceptron regression (MLPR): one hidden layer with 100 units, ReLu activation function for hidden layer units, no activation function for the output

c) Evaluation: The trajectories corresponding to one video were held out and served for testing the predictive performance. The data of all other videos were used for training the regressor. In a leave-one-out cross-validation scheme, feature extraction, regression and evaluation were repeated such that the data of each participant and video were assigned to the test set once (see Fig. 4 for a graphical explanation). The predicted head orientations obtained from the regressors were divided by their norms as quaternions are of unit length. These predictions were then evaluated by their geodesic distance to the measured head orientations. The geodesic distance d (GD) between two quaternions $\mathbf{q}_1$ and $\mathbf{q}_2$ was calculated as [21]

$$d = \arccos\left(2(\mathbf{q}_1 \cdot \mathbf{q}_2) - 1\right)$$

To give an intuition, the geodesic distance is the distance between two head orientations as they are mapped onto a unit sphere. It ranges from 0 to π, since π is the distance between two points on the sphere that are exactly opposite to each other (e.g. north and south poles). When comparing the distance of a head orientation and its prediction, the chance level is thus considered to be $\pi/2$, as this is the expected difference between two points randomly picked on a sphere.

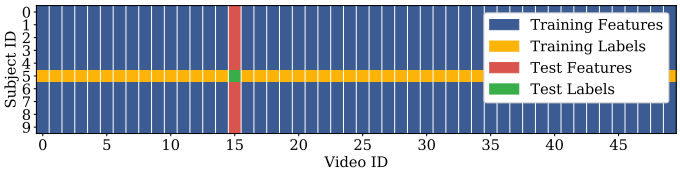

Fig. 4. Example partitioning of the dataset for one step of leave-one-out CV. This procedure is repeated (#subjects × #videos) times, with each subject-video pair i, j as test labels once. Each tile i, j refers to a time series $\{\mathbf{q}_{t,i,j}\}$.

IV. Results

A. Forecasting changes in head orientation

a) Statistical features: The trajectories were split into 113 630 epochs of length $w = 1000\,\text{ms}$. We labeled the epochs in nine different ways by chosing Δt in $[250\,\text{ms}, 500\,\text{ms}, 1000\,\text{ms}]$ and $\Delta\Phi_c$ in $[5°, 10°, 20°]$. The class ratios, i.e. the fraction of epochs, for which the translation $\Delta\Phi$ during the time interval $[t_0, t_0 + \Delta t]$ exceeds the threshold $\Delta\Phi_c$, are depicted in Table I for the nine scenarios.

TABLE I
Class ratios (fraction of epochs, for which the head orientation $\Delta\Phi$ surpassed $\Delta\Phi_c$ during $[t_0, t_0 + \Delta t]$) for nine different sets of epochs

$\Delta\Phi_c$	Δt		
	250 ms	500 ms	1000 ms
5°	0.41	0.58	0.74
10°	0.22	0.418	0.59
20°	0.06	0.74	0.40

To give an overview of how much the single features of the 8 dimensional feature vector differ between the classes, Table II lists the results T_1 of a two-sample Kolmogorov-Smirnov (K-S) test [22] for each feature for all scenarios with n positive and m negative instances. If $T_1 > 1.63\sqrt{\frac{n+m}{nm}}$ the null-hypothesis that the two samples are drawn from the same distribution is neglected at the significance level of $\alpha = 0.005$ [22]. This is the case for all features for all scenarios $(\Delta t, \Delta\Phi_c)$. The head trajectory based features $\text{rmssd}_{\text{head}}$ and $\text{stdd}_{\text{head}}$ show the highest K-S statistic for all scenarios. Among the gaze trajectory based features, the standard deviation of the gaze x coordinate has the highest K-S statistic. Thus, past head movements seem to be more relevant for future head movements but past gaze trajectories are still informative.

b) Performance of a linear SVM: The performance of a linear SVM with penalty parameter $C = 1$ is tested in two 5-fold cross validations (CV) grouped by subjects and videos, respectively (cf. Section III-A).

We report the mean CV test scores with standard deviation in Fig. 5. The predictions of the model show a higher AUC ROC the smaller the time horizon Δt and the higher the head translation threshold $\Delta\Phi_c$. With smaller Δt, the classification task becomes easier, since the model has to make inferences about a smaller amount of time in the future. Larger head translations are apparently preceded by head and gaze movements whose statistical properties are more distinct from a stationary head state.

The highest precisions are achieved for the scenarios $(250\,\text{ms},\ 5°)$, $(500\,\text{ms},\ 10°)$ and $(1000\,\text{ms},\ 20°)$. For $(1000\,\text{ms},\ 5°)$ the precision equals the class ratio (cf. Table I), implying that the SVM classifies all instances as positive. The precision is very low for $(250\,\text{ms},\ 20°)$. The reason is most likely rooted in the low class ratio of only 6% for this scenario (cf. Table I).

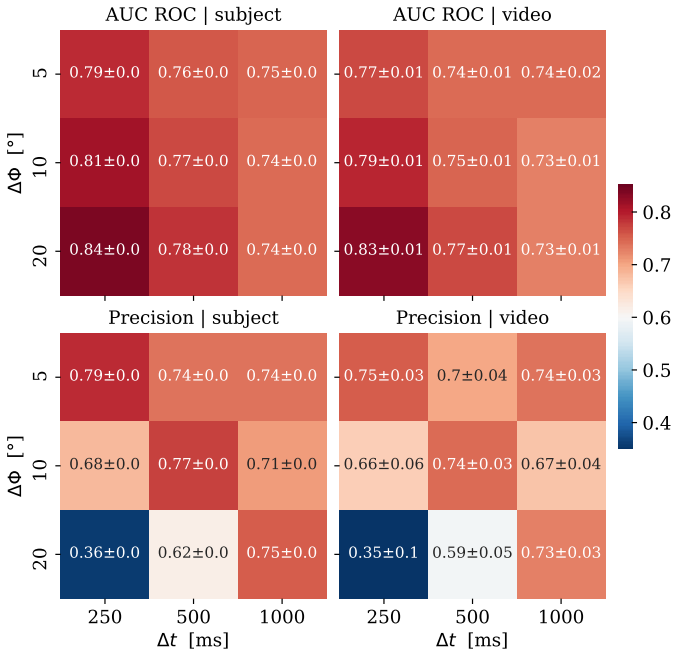

Fig. 5. Performance of a linear SVM with penalty parameter $C = 1$ measured by the mean test AUC ROC (upper row) and precision (lower row) across a grouped 5-fold cross validation with standard deviation. Left column: CV grouped by subjects, e.g. the SVM was trained on the data of 8 subjects and tested on the remaining data of 2 subjects. Right column: grouping by videos, e.g. the training set involved epochs from 40 videos and the test set epochs from the remaining 10 videos.

Overall, AUC ROC and precision, are higher in the subject grouped CV than in the video grouped one. This implies that generalization works better across subjects than across videos and that the reactions to the videos are distinct for each video.

B. Forecasting the viewport from population statistics

We report the average geodesic distance between predicted and actual head orientations for all models per video and in a grand average over all videos (Fig. 6). All predictions are above the chance level of $0.5\,\pi$. Simple averaging (AVGR) achieved reasonable geodesic distances that are not outperformed by the regression models MLPR, LSVR and LR.
Across all videos, MLPR performs slightly worse than the other models, which perform similarly. Regarding the grand model averages, only the LSVR yields slightly better results than the toy model AVRG. For the videos 12, 28 and 29 our approach worked significantly better than for the other videos. These videos were produced in a way that most of the relevant information is at the center orientation.

V. Discussion

We compiled a dataset of head and gaze trajectories of VR users watching a broad variety of 360° videos. Using this data, we followed two approaches to investigate the ability of machine learning models to forecast the viewport to help optimize VR rendering and transmission.

TABLE II

TEST STATISTIC T_1 OF A TWO-SAMPLE KOLMOGOROV-SMIRNOV TEST FOR EACH FEATURE EXTRACTED FROM THE LABELED EPOCHS WITH PARAMETERS Δt IN $[250\,\text{ms}, 500\,\text{ms}, 1000\,\text{ms}]$ AND $\Delta\Phi_c$ IN $[5°, 10°, 20°]$ WITH n POSITIVE AND m NEGATIVE INSTANCES. IF $T_1 > 1.63\sqrt{\frac{n+m}{nm}}$, THE NULL-HYPOTHESIS THAT THE SAMPLES FROM THE TWO CLASSES ARE DRAWN FROM THE SAME DISTRIBUTION CAN BE NEGLECTED AT THE SIGNIFICANCE LEVEL OF $\alpha = 0.005$. IT IS $1.63\sqrt{\frac{n+m}{nm}} = 0.01$ FOR ALL SCENARIOS, EXCEPT FOR $(250\,\text{ms},\ 20°)$ WHERE $1.63\sqrt{\frac{n+m}{nm}} = 0.02$. THE KOLMOGOROV-SMIRNOV TEST IMPLIES THAT THE (UNIVARIATE) FEATURE SAMPLES FROM THE TWO CLASSES ARE DRAWN FROM DISTINCT DISTRIBUTIONS FOR ALL SCENARIOS (AS INDICATED BY BOLD FONT).

Δt [ms]	250			500			1000		
$\Delta\Phi_c$ [°]	5	10	20	5	10	20	5	10	20
$\text{aucorr}_{\text{eye}x}$	**0.04**	**0.04**	**0.04**	**0.05**	**0.04**	**0.04**	**0.07**	**0.05**	**0.04**
$\text{aucorr}_{\text{eye}y}$	**0.04**	**0.03**	**0.06**	**0.05**	**0.04**	**0.03**	**0.06**	**0.05**	**0.04**
$\text{rmssd}_{\text{eye}}$	**0.16**	**0.16**	**0.18**	**0.15**	**0.14**	**0.14**	**0.15**	**0.13**	**0.12**
$\text{rmssd}_{\text{head}}$	**0.48**	**0.45**	**0.43**	**0.50**	**0.45**	**0.41**	**0.53**	**0.47**	**0.41**
$\text{std}_{\text{eye}x}$	**0.24**	**0.24**	**0.26**	**0.24**	**0.22**	**0.21**	**0.23**	**0.21**	**0.18**
$\text{std}_{\text{eye}y}$	**0.17**	**0.15**	**0.14**	**0.17**	**0.15**	**0.13**	**0.17**	**0.15**	**0.12**
stdd_{eye}	**0.18**	**0.17**	**0.19**	**0.17**	**0.15**	**0.15**	**0.16**	**0.14**	**0.13**
$\text{stdd}_{\text{head}}$	**0.48**	**0.44**	**0.42**	**0.50**	**0.45**	**0.41**	**0.53**	**0.46**	**0.41**
$1.63\sqrt{\frac{n+m}{nm}}$	0.01	0.01	0.02	0.01	0.01	0.01	0.01	0.01	0.01

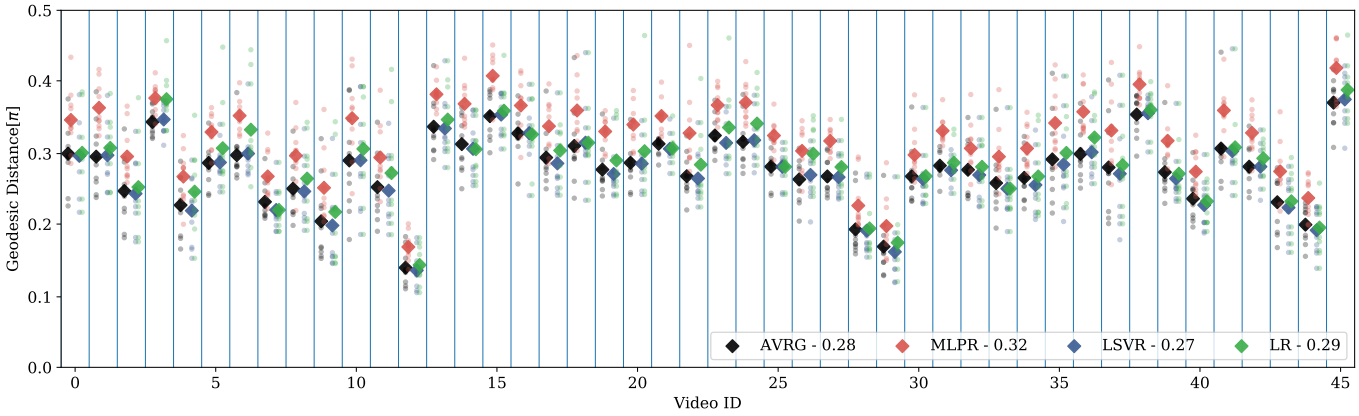

Fig. 6. Comparison of model predictions with head orientations of other subjects as input. Dots indicate individual subject-video pairings, while diamonds indicate subject group averages for a given model. Numbers in the legend indicate the grand average performance of each model. AVRG: Average using the SLERP method. MLPR: Multilayer Perceptron Regression. LSVR: Linear Support Vector Regression. LR: Linear Regression.

First, we applied SVMs to forecast changes in head orientation for different scenarios. We measured their performance by AUC ROC and precision in different cross validation schemes. AUC ROC scores show how well the classifier can distinguish between the two classes of epochs. The scores (cf. Fig. 5) reflect the difficulty of the task and decrease with a longer prediction horizon Δt and a smaller translation threshold $\Delta\Phi_c$. With AUC ROC ranging from 0.73 to 0.84 for the different scenarios, the classifier reached good results in distinguishing thresholded head translations. Precision is a meaningful measure with regards to an application in viewport adaptive rendering as it states the probability that the head translation will indeed surpass the threshold given that the classifier predicts so. Even for the longest prediction horizon of $\Delta t = 1000\,\text{ms}$ in our test, we observed a good precision of 0.67 to 0.75. Further preprocessing of the gaze data, i.e. the extraction of saccades with a duration of 100-150 ms (cf. [23]) and tracking motions might improve our scheme. With the results of the current approach as a proof of concept, we aim to train a recurrent neural network (RNN) like a LSTM RNN directly on the head and gaze signal to classify or regress future head movements.

Second, we attempted to predict the viewport of subjects given head orientations from other subjects, using a variety of simple regression methods. The fact that we achieved predictive performance above chance level having a very small sample size of ten subjects, promises that predicting salience patterns of a viewer using information of other viewers is possible. In the video streaming practice, the viewer size is expected to be much larger (than ten), which is likely to enhance our approach and allow for for sophisticated regression algorithms.

In conclusion, with our approach, we have developed a simple model to forecast the viewport in a spherical video, independent of any content information like video frames or sound, that shows a high predictive performance and might be applied for more efficient 360° video streaming.

REFERENCES

[1] A. Serrano, V. Sitzmann, J. Ruiz-Borau, G. Wetzstein, D. Gutierrez, and B. Masia, "Movie editing and cognitive event segmentation in virtual reality video," *ACM Trans. Graph.*, vol. 36, pp. 47:1–47:12, July 2017. doi: 10.1145/3072959.3073668.

[2] H. Hu, Y. Lin, M. Liu, H. Cheng, Y. Chang, and M. Sun, "Deep 360 pilot: Learning a deep agent for piloting through 360° sports video," *arXiv preprint*, vol. abs/1705.01759, 2017. http://arxiv.org/abs/1705.01759.

[3] Y. Bao, H. Wu, T. Zhang, A. A. Ramli, and X. Liu, "Shooting a moving target: Motion-prediction-based transmission for 360-degree videos," in *2016 IEEE International Conference on Big Data (Big Data)*, pp. 1161–1170, Dec 2016. doi: 10.1109/BigData.2016.7840720.

[4] Yanan Bao, Huasen Wu, A. A. Ramli, Bradley Wang, and Xin Liu, "Viewing 360 degree videos: Motion prediction and bandwidth optimization," in *2016 IEEE 24th International Conference on Network Protocols (ICNP)*, pp. 1–2, Nov 2016. doi: 10.1109/ICNP.2016.7784458.

[5] Y. Bao, T. Zhang, A. Pande, H. Wu, and X. Liu, "Motion-prediction-based multicast for 360-degree video transmissions," in *2017 14th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON)*, pp. 1–9, June 2017. doi: 10.1109/SAHCN.2017.7964928.

[6] L. Xie, Z. Xu, Y. Ban, X. Zhang, and Z. Guo, "360ProbDASH: Improving QoE of 360 Video Streaming Using Tile-based HTTP Adaptive Streaming," in *Proceedings of the 25th ACM International Conference on Multimedia*, MM '17, (New York, NY, USA), pp. 315–323, ACM, 2017. doi: 10.1145/3123266.3123291.

[7] S. Xie, Q. Shen, Y. Xu, Q. Qian, S. Wang, Z. Ma, and W. Zhang, "Viewport adaptation-based immersive video streaming: Perceptual modeling and applications," *arXiv preprint*, vol. abs/1802.06057, 2018. http://arxiv.org/abs/1802.06057.

[8] F. Qian, L. Ji, B. Han, and V. Gopalakrishnan, "Optimizing 360 video delivery over cellular networks," in *Proceedings of the 5th Workshop on All Things Cellular: Operations, Applications and Challenges*, ATC '16, (New York, NY, USA), pp. 1–6, ACM, 2016. doi: 10.1145/2980055.2980056.

[9] A. Ghosh, V. Aggarwal, and F. Qian, "A rate adaptation algorithm for tile-based 360-degree video streaming," *arXiv*, vol. abs/1704.08215, 2017. http://arxiv.org/abs/1704.08215.

[10] 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 *Proceedings of the 27th Workshop on Network and Operating Systems Support for Digital Audio and Video*, NOSSDAV'17, (New York, NY, USA), pp. 67–72, ACM, 2017. doi: 10.1145/3083165.3083180.

[11] P. Lungaro, R. Sjberg, A. J. F. Valero, A. Mittal, and K. Tollmar, "Gaze-aware streaming solutions for the next generation of mobile VR experiences," *IEEE Transactions on Visualization and Computer Graphics*, vol. 24, pp. 1535–1544, April 2018. doi: 10.1109/TVCG.2018.2794119.

[12] A. Patney, J. Kim, M. Salvi, A. Kaplanyan, C. Wyman, N. Benty, A. Lefohn, and D. Luebke, "Perceptually-based foveated virtual reality," in *ACM SIGGRAPH 2016 Emerging Technologies*, SIGGRAPH '16, (New York, NY, USA), pp. 17:1–17:2, ACM, 2016. doi: 10.1145/2929464.2929472.

[13] Y. Xu, Y. Dong, J. Wu, Z. Sun, Z. Shi, J. Yu, and S. Gao, "Gaze prediction in dynamic 360 immersive videos," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5333–5342, June 2018. doi: 10.1109/CVPR.2018.00559.

[14] S. Petrangeli, G. Simon, and V. Swaminathan, "Trajectory-based viewport prediction for 360-degree virtual reality videos," in *2018 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR)*, pp. 157–160, Dec 2018. doi: 10.1109/AIVR.2018.00033.

[15] M. Kassner, W. Patera, and A. Bulling, "Pupil: An open source platform for pervasive eye tracking and mobile gaze-based interaction," in *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*, UbiComp '14 Adjunct, (New York, NY, USA), pp. 1151–1160, ACM, 2014. doi: 10.1145/2638728.2641695.

[16] K. Shoemake, "Animating rotation with quaternion curves," *SIGGRAPH Comput. Graph.*, vol. 19, pp. 245–254, July 1985. doi: 10.1145/325165.325242.

[17] A. Kampouraki, G. Manis, and C. Nikou, "Heartbeat time series classification with support vector machines," *IEEE Transactions on Information Technology in Biomedicine*, vol. 13, pp. 512–518, July 2009. doi: 10.1145/325165.325242.

[18] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Berlin, Heidelberg: Springer-Verlag, 2006.

[19] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statistics and Computing*, vol. 14, pp. 199–222, Aug. 2004. doi: 10.1023/B:STCO.0000035301.49549.88.

[20] S. Haykin, *Neural Networks: A Comprehensive Foundation*. Upper Saddle River, NJ, USA: Prentice Hall PTR, 1st ed., 1994.

[21] F. C. Park and B. Ravani, "Smooth invariant interpolation of rotations," *ACM Trans. Graph.*, vol. 16, pp. 277–295, July 1997. doi: 10.1145/256157.256160.

[22] W. Conover, *Practical Nonparametric Statistics*. New York, NY, USA: John Wiley & Sons, 1971.

[23] B. Fischer and E. Ramsperger, "Human express saccades: extremely short reaction times of goal directed eye movements," *Experimental Brain Research*, vol. 57, pp. 191–195, Jan 1984. doi: 10.1007/BF00231145.

TABLE III
LIST OF THE VIDEOS PLAYED IN THE EXPERIMENT

Cluster	ID	Name	Duration [s]
Animated	0	Aether	240
	1	PlantRoom	195
	4	TheFutureofMusic	178
	13	Zurich	187
	14	DeathVanAndTheZeitshowFolks	102
	15	WakingUpInTheUniverse	588
	16	TheGiant	210
	23	Pleasure	322
	33	RachelPlatten	199
	40	CrystalHabit	362
	48	Ouch	300
Documentary	2	SwimmingWithBearsNationalGeographic	111
	5	AntarcticaJourneyThroughTheIceNationalGeographic	119
	6	BattleofWaterlooNationalGeographic	159
	7	BabyPandasNatGeoWild	166
	8	TigerSharkEncounterInTheBahamas	146
	9	TotalSolarEclipseExperienceIndonesia	388
	10	UnderLaRefikAnadolAndPeggyWeil	126
	11	TheSource	543
	17	WomenOnTheMove	203
	18	HelvetiaByNight	199
	19	VikingBattleNationalGeographic	299
	20	GetBarreledInTahiti	138
	24	HawaiiThePaceOfFormation	146
	25	FossilHuntersoftheGobi	224
	26	VictoriaFallsTheDevilsPoolNationalGeographic	108
	29	KletternSieMitDaniArnoldAufDenEiger	232
	30	OrangutanSchoolNationalGeographic	539
	31	DangerousHoneyHuntingExplorerNationalGeographic	218
	32	UnderwaterNationalParkNationalGeographic	349
	34	LionsNationalGeographic	270
	36	InsideTheBiennale	148
	37	TakeEveryWave	255
	39	DiveThroughanOilRigEcosystemNationalGeographic	192
	41	GiantSequoiasPartTwoNationalGeographic	155
	42	DavidHockneyMuseeGeorgesPompidou	425
	43	HowToProduceAppenzellCheeseSwitzerland	286
	44	VictoriaFallsTheCanyonsBelowNationalGeographic	101
	45	MitDemDampfschiffUeberDenVierwaldstaettersee	336
	47	ClimbingGiantsNationalGeographic	205
Movie Trailers	3	DunkirkGerman	276
	21	StarWarsHuntingOfTheFallen	630
	27	AssassinsCreed	151
	28	TheConjuringTwoExperienceEnfield	188
	35	Annabelle	210
	49	RingsScaryExperience	93
Short Movie	12	GoogleSpotlightStoryHelp	293
	22	ThroughMowglisEyesPartTwoTheJungleBook	93
	38	Interrupture	162
	46	ThroughMowglisEyesPartOneTheJungleBook	99