A Comparative Analysis of Viewing Prediction Techniques for 360° Video Streaming Applications

Moatasim Mahmoud
R&D Department
Singular Logic
Aristotle University of Thessaloniki
mmoatasim@singularlogic.eu

Stamatia Rizou
Rizou Stamatia
R&D Department
Singular Logic
14564 Athens, Greece
srizou@singularlogic.eu

Andreas S. Panayides
VIDEOMICS Group
CYENS Centre of Excellence
Nicosia, Cyprus
a.panayides@cyens.org.cy

Pavlos I. Lazaridis
School of Computing and Engineering
University of Huddersfield
HD1 3DH Huddersfield, U.K
p.lazaridis@hud.ac.uk

Nikolaos V. Kantartzis
School of Electrical and Computer Eng.
Aristotle University of Thessaloniki
54124 Thessaloniki, Greece
kant@auth.gr

George K. Karagiannidis
School of Electrical and Computer Eng.
Aristotle University of Thessaloniki
54124 Thessaloniki, Greece
geokarag@auth.gr

Zaharias D. Zaharis
School of Electrical and Computer Eng.
Aristotle University of Thessaloniki
54124 Thessaloniki, Greece
zaharis@auth.gr

2024 Panhellenic Conference on Electronics & Telecommunications (PACET) | 979-8-3503-1884-5/24/$31.00 ©2024 IEEE | DOI: 10.1109/PACET60398.2024.10497019

***Abstract*—In this work, we implement multiple techniques for predicting users viewing directions while watching 360° videos. We utilize historical viewing traces to forecast future directions based on a real-life head tracking dataset. We compare the performance of linear regression (LR), artificial neural networks (ANN), long short-term memory (LSTM), and convolutional neural networks (CNN). We assess their efficiency in terms of viewing angles prediction errors. We also investigate tile viewing prediction in tile-based 360° video transmission scenarios. We built two classifiers based on ANN and LSTM to predict watched tiles and provide an evaluation of their performance in this article.**

Keywords—Virtual Reality (VR), 360° Videos, Viewport Prediction, Deep Learning

I. INTRODUCTION

360° videos and virtual reality (VR) applications are becoming increasingly popular paving the way toward an immersive era. Omnidirectional videos envelop users within an interactive form of content, as people equipped with head-mounted displays (HMDs) can move their heads to navigate the spherical scenes. Delivering VR videos at satisfying quality, however, requires high transmission rates and is subject to stringent latency levels [1]. This created a pressing challenge for network operators and service providers. Novel streaming paradigms have been developed to alleviate these challenges. Viewport-based 360° video streaming suggests the transmission of only viewed areas within the spherical scenes, rather than sending the entire bulky videos. Tile-based streaming has been an effective solution that attempts to divide the 360° video into rectangular tiles and only transmit tiles which are relevant to the user's field of view (FoV) [2].

Tiling of 360° videos has become a notable strategy in 360° video transmission. Panoramic videos are typically captured using a 360° camera or generated by stretching multiple videos captured from multiple cameras. The spherical video is then transformed into other 2D formats using projection methods. The equi-rectangular projection (ERP) scheme stretches the spherical shape and fits it into a rectangular shape. The resulting ERP video is then divided into independent tiles, where each tile is compressed and transmitted separately. Fig. 1 illustrates the projection and tiling of a 360° video.

Although adaptive and tiled 360° video streaming can assist in huge bandwidth savings, its practical implementation relies on having accurate viewport predictions. VR applications have low latency requirements. A motion-to-photon (MTP) latency of less than 20ms is usually required for a smooth watching experience. High delay can result in interrupting the video or in presenting empty parts within the user's FoV. In many cases, the low responsivity causes distress and headache to the users, as symptoms of cybersickness [3]. Typical video streaming frameworks rely on transmitting independent segments of the video (2-10 seconds long), in order to adapt to the network conditions. Proactive 360° video streaming, therefore, needs viewport prediction models to produce reliable estimations for future viewing directions [4], [5]. An extended literature review on the topic of optimized mobile 360 video delivery can be found in our survey paper [6].

The authors in [7] utilized spherical convolutional networks to assist the viewport prediction through visual feature extraction from 360° videos. In [8], an online learning framework has been proposed for predicting users' viewports in 360° video live streaming services. The suggested solution leverages saliency maps and implements a convolutional long short-term memory (ConvLSTM) network to produce accurate viewport predictions. Although such methods have good performance, they are complex and can be time consuming. Moreover, they require knowledge of the video content, which is not always practical. In [9], a trajectory-based clustering algorithm is developed to predict the viewing trajectories for new users.

Fig. 1: Equi-rectangular projection and tiling of 360° videos.

In our previous work, we performed an evaluation comparison of different codecs applied to tiled 360° videos

979-8-3503-1884-5/24/$31.00 ©2024 IEEE

[10]. In this article, we dedicate our work to investigating the viewing prediction problem. We focus on prediction schemes that only leverage the user's historical viewing path. We assess and compare the performance of multiple viewport prediction techniques. We also study the implication of prediction inaccuracies on tile-based 360° video transmission. In section II, we provide the necessary context and introduce the problem in hand. We present the used VR dataset and provide some analysis of users viewing behavior in Section III. Section IV discusses the evaluation settings, adopted solutions, and simulation results. Finally, we conclude our work in section V.

II. Problem Description

In this work, we focus on 360° viewing prediction with a given viewing historical path. At any given time, the user's viewport is defined by their head orientation, in addition to the HMD specifications. The user's head direction (i.e., the FoV center) at time τ can be represented by a yaw and a pitch angle pair $(\theta_\tau, \varphi_\tau)$. The yaw and pitch angles represent the horizontal and vertical orientation in a 360°x180° space.

The user's historical viewing path over H samples $\mathbf{V^H} = [(\theta_{-H}, \varphi_H), \ldots, (\theta_{-h}, \varphi_{-h}), \ldots, (\theta_0, \varphi_0)]$ is assumed to be given as a vector of pitch and yaw angle pairs. It is thus used for forecasting the future viewing path over F time stamps $\mathbf{V^*} = [(\theta^*_{+1}, \varphi^*_{+1}), \ldots, (\theta^*_f, \varphi^*_f), \ldots, (\theta^*_{+F}, \varphi^*_{+F})]$ in an attempt to match the true one $\mathbf{V^F} = [(\theta_{+1}, \varphi_{+1}), \ldots, (\theta_f, \varphi_f), \ldots, (\theta_{+F}, \varphi_{+F})]$. Fig. 2 demonstrates the viewport prediction problem in hand, showing the aforementioned viewing paths. In practice viewing direction information can be obtained from the HMD, as HMDs are typically equipped with head and eye gaze tracking mechanisms. The specifications of the HMD describe the sampling rate of the collected viewing information.

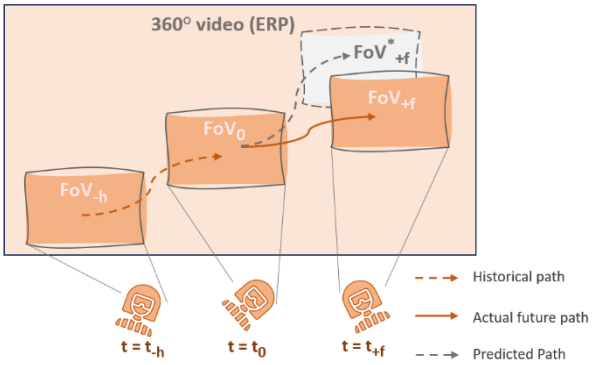

Fig. 2: Illustration of viewport prediction and viewing paths.

III. Dataset Analysis

In this section, we first introduce the VR viewing dataset which we utilize for carrying out our experiments. We also provide relevant analysis of the tracking records. We investigate the user behavior while watching 360° videos in terms of attention regions and head movement rates.

A. Dataset

For our analysis, we use the large-scale public dataset in [11]. The dataset in [11] contains the viewing information of 27 users watching 100 videos that belong to different categories. The data sequences, which include head and eye gaze tracking data, are captured at 120Hz sampling rate using the VIVE Pro Eye VR headset. We use the head viewing positions given as two-dimensional points (x,y) and convert them to the corresponding yaw and pitch angles (θ, φ).

B. Viewing behavior analysis

Before jumping to the viewing prediction part, it is important to study the viewing data and highlight the patterns within the users' viewing trajectories. We analyze the head movement information found in the dataset. We convert the given two-dimensional points into yaw and pitch pairs. Based on this data, we examine the areas where users focus their attention the most, and the speed at which the users move their heads.

Engaging parts within 360° videos are usually found along the equator, i.e., the center of the vertical axis. The upper and lower parts usually contain still and uninteresting content such as an empty sky or an idle ground. Thus, it is common for VR users to focus their attention toward central areas and move their heads around the horizontal axis, following or looking for relevant objects. We investigate this behavior by analyzing the available tracking records. Fig. 3 shows a heat map that features the center of attention for users. This figure entails the watching data from all 27 videos in the dataset across their entire lengths. It is clear that most of the time users in this dataset focused on central regions within the videos, with some navigation across the horizontal axis. This observation can be regarded as a general case for different video types, regardless of the genre they belong to.

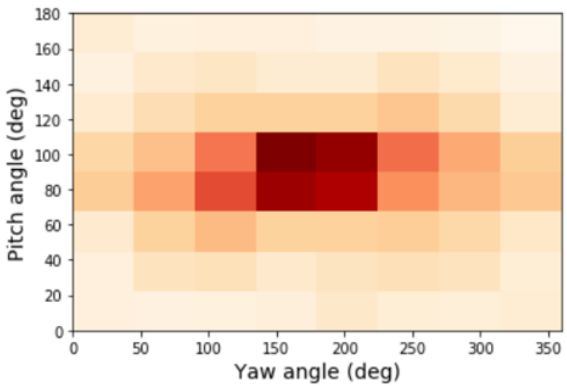

Fig. 3: Users viewing heatmap.

Another important aspect of users' viewing behavior is the pace at which they move their viewing directions. The prediction horizon of future viewing paths is limited by the change rate in yaw and pitch angles. In general, VR users move their heads at steady paces, where sudden and fast changes in their FoV positions do not frequently occur. As explorations in the horizontal and vertical directions are independent of each other, we analyze these changes in yaw and pitch angles separately. We find the maximum angle changes within 1, 2, and 3 second time windows for both yaw and pitch angles.

Fig. 4 depicts the distribution of the maximum changes within these time windows. Short viewing windows exhibit small changes in both directions, as the majority of one-second windows have changes in the yaw and pitch angles of only a few degrees. This enables the development of accurate viewing prediction models for short timeframes. However, when the length of the observation window increases (up to 3 seconds, in Fig. 4), high angle changes become more frequent. Intuitively, this makes the prediction task more challenging as the prediction window increases.

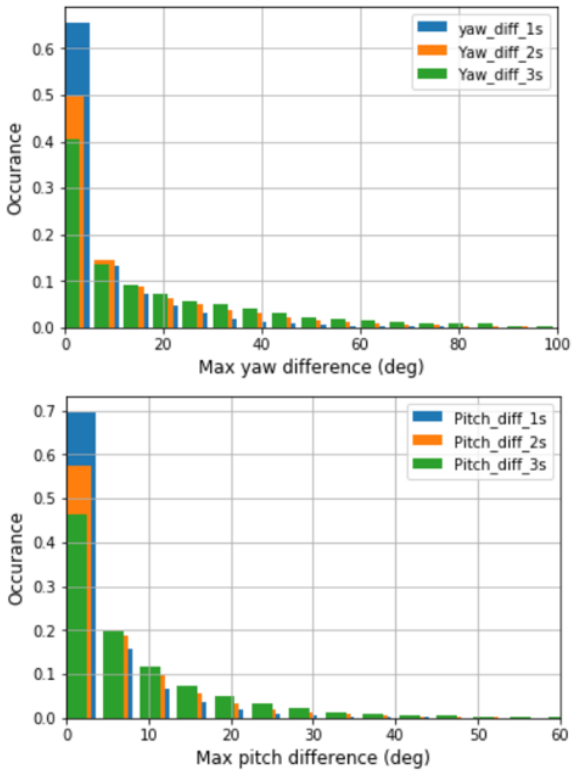

Fig. 4: Changes in yaw and pitch angles within 1, 2, and 3 second time windows.

IV. Evaluation

In this section, we discuss the simulated prediction methods and provide the evaluation results. We carried out our experiments in two parts. First, we implemented four regression models to predict future yaw and pitch angles based on historical paths. In the second part, we study tile viewing prediction. We show how prediction inaccuracies in viewing angles translate as suboptimal tile coverage in tile-based 360° transmission. Therefore, we developed two binary classification models that directly predict which tiles are watched by the users and compare their performance to the regression-based solutions.

We compare the performance of the different techniques in terms of the resulting prediction errors. Then we study how prediction inaccuracies in viewing angles translate as suboptimal tile coverage in tile-based 360° transmission.

A. Evaluation of viewing direction prediction

In this part, we introduce the four regression models that we use to predict future yaw and pitch angles $\mathbf{V}^*$. These solutions rely only on historical viewing paths $\mathbf{V}^{\mathbf{H}}$. To account for the spherical nature of 360° videos, we use the sine and cosine values of the angles. This avoids prediction errors across the video edges. For instance, a yaw angle of 0° and another of 359° should be regarded as adjacent to one another. After the prediction of the trigonometric values, the predicted angles are obtained (in degrees) using the inverse tangent function.

The four regression models used in this part are summarized as follows:

- ANN: An artificial neural network (ANN) containing two layers that output the trigonometric values (i.e., sine and cosine) of the predicted angles across all future samples.
- Conv1D: A 1D convolutional neural network (CNN) that takes as input two sequences, containing the historical sine and cosine values to be processed through two different convolution paths. The convolution outputs of the two sequences are then merged using an output dense layer.
- LSTM: A long short-term memory (LSTM) model. The implemented LSTM network follows a similar hierarchy to the Conv1D network. Each of the two sequences go through two LSTM layers to be merged using an output dense layer.
- LR: a simple linear regression (LR) model that fits the data to form a linear relationship between the historical and future angles. The trigonometric values of the angles are also used in this model, where the actual angle values are retrieved afterwards.

To assess the viewing direction prediction, we calculate the absolute error in angle values and average all the observation samples (S) found at each timestamp. The following formula shows the error calculation for yaw angle prediction at future timestamp f, where the same formula is used for calculating prediction errors in yaw angles.

$$Error(\theta_f) = \frac{1}{S}\sum_{S} |\theta_{f,s} - \theta^*_{f,s}|$$

The yaw and pitch prediction errors resulting from the four regression models are depicted in Fig. 5. The figure shows the prediction errors over time within a three-second window. The average estimation error increases over time to reach 30° for yaw angles and around 10° for pitch angles at 3 seconds estimation period. All four prediction methods achieve comparable performance, while LSTM slightly outperforms the other models when considering both yaw and pitch prediction results. However, LSTM needs a long training time compared to the other models.

B. Evaluation of viewing tile prediction

As tile-based transmission is becoming the norm in 360° videos streaming, we develop two classifiers for tile viewing prediction. We also use the regression models from the precious subsection for finding their performance in tiled based streaming settings. At any given time, the user's viewport corresponds to a set of tiles that need to be presented, entirely or partially, to the user. However, unreliable viewport predictions can result in losing some of these tiles due to inaccurate mapping. We evaluate the described regression models in terms of tile coverage and compare their performance to two tile classification-based models.

With tile classification, we can directly predict which tiles will be watched by the user. We implemented two tile classification models to perform this task, which are described as follows:

- ANN-tile: A two-layer ANN model that uses historical trigonometric values of yaw and pitch angles to perform binary classification on future tiles. The binary cross-entropy model is used as the loss function in the output layer. Watched tiles are

predicted at each timestamp within the prediction window.

- LSTM-tile: A LSTM model that takes four sequences comprised of the sine and cosine angles of yaw and pitch angles. An output dense layer merges the output of the four paths and uses the sigmoid activation function.

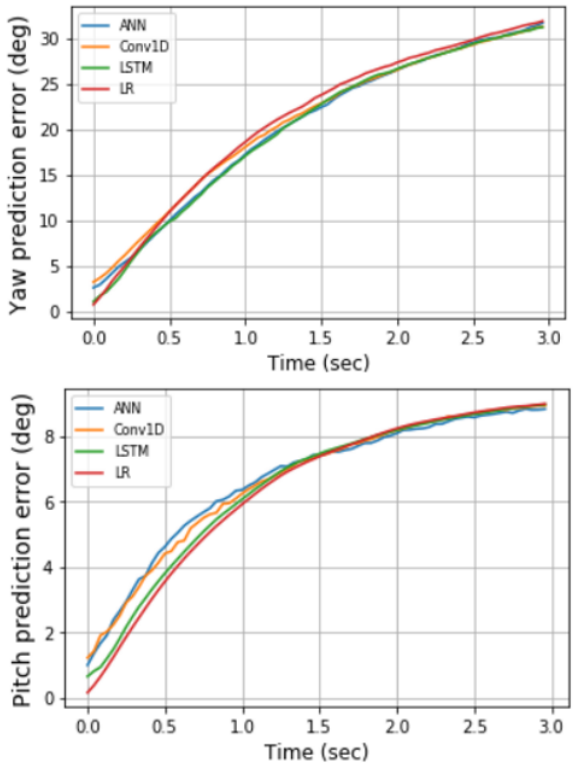

Fig. 5: Prediction error for yaw and pitch angles as a function of prediction time.

Table 1: Evaluation metrics of tile viewing prediction models.

Model	Accuracy	Precision	Recall	F1
ANN	0.853	0.604	0.621	0.613
CNN	0.850	0.596	0.612	0.604
LSTM	0.858	0.617	0.626	0.622
LR	0.855	0.613	0.609	0.611
ANN-tile	**0.931**	**0.793**	**0.832**	**0.812**
LSTM-tile	**0.928**	**0.782**	**0.826**	**0.803**

We investigated the performance of all six techniques when applied for tile coverage prediction. We assume an 8x8 tiling scheme and classify the tiles at each timestamp into watched and unwatched tiles. To assess the performance of the simulated schemes, we calculate their accuracy, precision, recall, and the F1 score. Table 1 shows the evaluation of the different tile viewing prediction models. ANN-tile and LSTM-tile produce the best scores over all evaluation metrics. These results demonstrate the advantages of direct tile classification over mapping them from regression-based angle viewing predictions.

V. Conclusion

In this paper, we motivated the need to accurately predict user viewing directions to improve performance of 360° video streaming applications. We formally introduced the problem of viewing direction prediction, and we analyzed the viewing behavior based on a rich large-scale public dataset. To tackle the viewing prediction problem, we compared four different regression models. In addition, we developed two binary classification models that directly predict which tiles are watched by the users and compare their performance to the regression-based solutions. Our evaluation results show that all four regression models achieve comparable performance. However, the results also show that classification models are more appropriate for tile-based streaming settings.

Acknowledgment

This research was supported by the European Union, through the Horizon 2020 Marie Skłodowska-Curie Innovative Training Networks Programme “Mobility and Training for beyond 5G Ecosystems (MOTOR5G)” under grant agreement no. 861219.

References

[1] Westphal, Cedric. "Challenges in networking to support augmented reality and virtual reality." IEEE ICNC (2017).

[2] Yaqoob, Abid, Ting Bi, and Gabriel-Miro Muntean. "A survey on adaptive 360 video streaming: Solutions, challenges and opportunities." IEEE Communications Surveys & Tutorials 22.4 (2020): 2801-2838.

[3] Stauffert, Jan-Philipp, Florian Niebling, and Marc Erich Latoschik. "Latency and cybersickness: Impact, causes, and measures. A review." Frontiers in Virtual Reality 1 (2020): 582204.

[4] Hou, Xueshi, et al. "Predictive adaptive streaming to enable mobile 360-degree and VR experiences." IEEE Transactions on Multimedia 23 (2020): 716-731.

[5] Nguyen, Anh, and Zhisheng Yan. "Enhancing 360 Video Streaming through Salient Content in Head-Mounted Displays." Sensors 23.8 (2023): 4016.

[6] Mahmoud, Moatasim, et al. "A Survey on Optimizing Mobile Delivery of 360° Videos: Edge Caching and Multicasting." IEEE Access (2023).

[7] Wu, Chenglei, et al. "A spherical convolution approach for learning long term viewport prediction in 360 immersive video." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 34. No. 01. 2020.

[8] Wang, Mu, et al. "CoLive: An Edge-Assisted Online Learning Framework for Viewport Prediction in 360° Live Streaming." 2022 IEEE International Conference on Multimedia and Expo (ICME). IEEE, 2022.

[9] Petrangeli, Stefano, Gwendal Simon, and Viswanathan Swaminathan. "Trajectory-based viewport prediction for 360-degree virtual reality videos." 2018 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR). IEEE, 2018.

[10] Mahmoud, Moatasim, et al “Versatile Video Coding Performance Evaluation for Tiled 360° Videos.” EUROPEAN WIRELESS 2023 (2023) .

[11] Jin, Yili, et al. "Where Are You Looking? A Large-Scale Dataset of Head and Gaze Behavior for 360-Degree Videos and a Pilot Study." Proceedings of the 30th ACM International Conference on Multimedia. 2022.