Head Movement Prediction using FCNN

Rabia Shafi[1], Wan Shuai[2*], Hao Gong[3] and Muhammad Usman Younus[4]
Northwestern Polytechnical University, Xian, China
E-mail: rabibabi@mail.nwpu.edu.cn Tel: +86-15529368601
[2*]Northwestern Polytechnical University, Xi'an, China
E-mail: swan@nwpu.edu.cn Tel: +86-29-88431253
[3]Northwestern Polytechnical University, Xi'an, China
E-mail: gong_h@mail.nwpu.edu.cn Tel: +86-29-88431253
[4]Ecole Mathématiques, Informatique, Télécommunications de Toulouse, Université de Toulouse, Toulouse, France.
E-mail: usman1644@gmail.com Tel: +86-33-753413108

Abstract— **Viewport adaptive streaming of 360-dgree videos relies on accurate prediction of the viewport, while the user generally suffers from significant quality degradation under long delay settings. To deal with this issue, advanced methods for long-term viewport prediction are highly desired to improve viewport prediction accuracy. To more accurately capture the non-linear relationship between the future and past viewpoints, this paper proposes a Fully Connected Neural Network (FCNN) model to make future predictions, which is light in computation. The input data such as yaw values, pitch values, Estimated Weighted Moving Average (EWMA) of yaw values, and EWMA of pitch values, are transformed into sine and cosine angles before feeding into the encoding layer of the FCNN model by considering the roll angle to zero. After transforming the data input into the proposed FCNN model, a long-term prediction length of up to 4 seconds has been explored, to capture the non-linear and long-term dependent relation between past and future viewport positions more accurately. Experimental results show that the proposed scheme performs well for the large size prediction window.**

Keywords: Viewport Prediction, FCNN, EWMA

I. INTRODUCTION

The head movement prediction is an intimidating task in 360-degree videos streaming among the various studies looking at the desired experience [1]. Usually, a 360-degree video is created using an omnidirectional camera and then by using projection, such as equirectangular projection, the spherical coordinates (longitude and latitude) are transformed to planar coordinates in a 2D space for achieving the immersive experience. The viewer is positioned at the center of rendered sphere where it is equipped with a Head-Mounted Display (HMD). In particular, the whole 360-degree video is downloaded and delivered to HMD when watching, but viewers only get to see a small viewable region, called viewport, which is a portion perceived by users. Wearing an HMD, a viewer's motion has three degree of freedom (pitch, yaw, and roll). Thus, streaming only the viewport area from the perspective of the user's head motion would make bandwidth usage more efficient [2]. Therefore, it is always being a challenge to provide a good immersive experience for 360-degree video streaming as they are vulnerable to unstable and insufficient bandwidth.

Different approaches have been proposed to accurately predict the head movement in 360-degree video streaming. One approach is to transmit only the corresponding frame in the user's viewport instead of the whole frame. The parts outside the viewport are not transmitted at all or are delivered with lower quality. Thus, this unique attribute of 360-degree videos saves the network bandwidth significantly. In addition, accurately predicting head movement will greatly reduce the motion-to-photon delay [3], [4] because the user viewport needs to be pre-fetched in advance by predicting the viewport. Therefore, in response to the user head motion, extracting and transmitting the viewport can add high latency that also need to be considered in 360-degree video streaming, otherwise it will adversely affect the user experience. If the latency value is too large, the delay will be noticeable for the end-user. To address the above-mentioned challenges, one needs to predict the user's viewport with high accuracy, otherwise the quality of the user declines.

Viewport prediction is a proto-typical complicated issue with dynamic changes and with unpredictable errors. Therefore, it is very important to embed an accurate head motion predictor to exploit the knowledge of past positions and to periodically predict the next position where the user will be likely looking at. Great efforts in [5]-[8] have been devoted to viewport prediction in 360-degree videos to tackle this issue with deep neural networks. In [5], two types of deep reinforcement learning models are proposed, in which the offline model estimates the heatmap of Field-of-View (FoV) of every frame, and online model predicts the head movement based on heatmaps and past head positions. In [7], a fixation prediction network predicts the FoV trajectory by concurrently leveraging the past FoV positions and video content characteristics. For long-term horizons, prediction error increases dramatically because the current user direction is not assumed to be a reliable predictor for directions in the next 3-4 seconds. Thus, it becomes very difficult to carry out the predictions in the distant future.

This paper evaluates the viewing behavior concerns in 360-degree videos by proposing a FCNN model that depends on the users' viewpoint information to predict and examine the future viewpoint. Experimental results show that the proposed scheme performs well for the large size of prediction window. The detailed information of the proposed work is described in the following sections.

The rest of the paper is organised as follows: Section 2 describes the related work in which the detail of head movement prediction-based techniques is given. Section 3 explains the proposed FCNN model, including its architecture for predicting yaw and pitch values. However, Section 4 describes the performance evaluation of the proposed model in comparison with alternative approaches. Finally, Section 5 concludes the paper.

II. RELATED WORK

Several research efforts have been proposed to focus on viewport prediction from the last couple of years, which aim to reduce the bandwidth consumption by predicting the user's area of interest and streaming the video portion that is likely to be watched with high priority. As, the head movement prediction is always being an indispensable part of 360-degree video streaming. Currently, many neural network-based approaches are proposed to predict the future viewport. This section elaborates and reviews the literature of prediction problems in 360-degree videos by defining their related issues.

Mostly, the existing systems introduced basic processing of head movement such as Linear Regression (LR) [9],[10], Simple Average [11], and Weighted Linear Regression (WLR) [12] by addressing a regression issue. Some researchers have recently investigated short-term FoV prediction. A logistic regression [13] is used in transmission improvement due to its simplicity to make the predictions by entirely streaming those tiles that will overlap with the estimated viewport. Another approach in [14] also estimates the user's future viewport by proposing a contextual bandit-based approach. Still, because of its lower accuracy than regression-based approaches, it does not use historical information. The authors in [15] conducted an experimental study of viewer motion by proposing ML mechanisms that will predict the viewer's behavior and prediction deviation itself. Their prediction results are then used to use network resources for a targeted streaming area efficiently.

Different prior studies have also analyzed the user's behavior instead of only targeting user's historical trajectories to boost the prediction performance, assuming that the users have similar Region-of-Interest (RoI) when watching the same video. The user's watching history is also exploited in [16] by using the K-Nearest-Neighbors (KNN) algorithm. Several existing approaches also predict the future viewing behavior by different prediction methods based on encoder-decoder architecture [17],[18]. These models parallelize the training phase that leads to better prediction accuracy compared to LR and Long-Short Term Memory (LSTM), as LSTM models have a time-consuming sequential training nature. However, in terms of how much FoV can be predicted for the future, the current prediction models are constrained. In [19], a machine translation model and sophisticated LSTM model are proposed that incorporate other viewers' history to predict the future user's orientation. While author in [20] first clustered the users according to their quaternion rotations to classify them into corresponding clusters and estimated the future fixation values as clustering centers. The last sample is used as a future viewport if no cluster will be available for the target user.

The viewport prediction is always being a vital enabler for 360-degree videos, which improves the prediction accuracy. In recent years, ML has developed rapidly, and its combination with image processing and big data has outstanding performance. The author in [5] proposed two DRL models to predict the head motion considering the motion trajectories and visual frames for better understanding. Their deep neural network only receives the user's view of interest and decides which direction and viewer's head will move. The online model predicts the viewer direction based on the saliency of each frame obtained by the offline model. The prediction horizon that predicts the next viewport position is about 30ms, i.e., one future frame. Thus, the positional information explicitly is not considered as input by predicting the FoV.

A saliency-driven model in [21] extracts the content-related features from the current frame, and predicts the next FoV based on the saliency algorithm. Moreover, this model does not work to consider the user's viewing behavior; and also fails to capture the properties, i.e., non-linearity and long-term dependency, resulting in undesirable performance regarding the prediction accuracy. This model can be considered a sub-study of [5]. Hence, these issues are addressed in [22], where a viewport prediction model has been developed using a Convolutional Neural Network (CNN), which reduces the pooling layers by introducing more convolutional layers to achieve a better non-linearity fitting ability. However, this work does not analyze the spherical CNN to process the spherical images directly. The work in [7] also proposed an LSTM network to predict the future viewport by adopting the visual saliency and user's head motion that achieves the prediction of 1 second later. [23] designed a hybrid architecture of CNN and LSTM model that predicts the gaze displacement based on gaze coordinates to perform the saliency computation associated with the gaze point, the viewport, and the whole image. Thus, the displacement prediction of a user for a future viewport is made by concatenating it with the viewer's historical head motion. The authors in [24] contributed a saliency model focused on user's fixation along with the head position of user because of central bias and multi-object confusion issues. Saliency mappings are used as input to LSTM by this model. Hence, they perform the viewport prediction for the next 2.5 seconds.

To more accurately capture the non-linear relationship between the future and past viewpoints, a Fully Connected

Neural Network (FCNN) model has been proposed, which predicts the future position of a specific user, where the input data (such as yaw values, pitch values, EWMA of yaw values, and EWMA of pitch values) are transformed into sine and cosine angles before inputting into the encoding layer. After transforming the data input into the proposed FCNN model, a long-term prediction length of up to 4 seconds has been explored.

III. PROPOSED FCNN MODEL

A 360-degree video display headset usually has three degree of freedom for rotational head movements (*yaw, pitch, and roll*) or six degree of freedom (such as transational movements in (*x, y, z)* in addition to three angles). Due to delivery platform and HMD technologies restrictions, this work only focusses on the rotational head movements and ignore the transational movements. Hence, the reference position (such as *O, i, j, k*) is set at boot time by HMD. *i* and *j* can change each time the HMD restarts but *k* is always taken as vertical. Regarding Euler' rotation theorem, any sequence of rotation of 3D coordinate system with fixed origin is equivalent to a single rotation around an axis, denoted by a unit vector $v(x, y, z)=xi + yj + zk$ in R^3. Thus, the viewport prediction can be modeled by the four-tuple of $f_t=(x, y, z, \theta)_t$, where *t* is time step index.

The viewport prediction problem can be represented as a multivariate time series prediction problem. For example, the prediction of a sequence of future viewport values f_{T+1}, $f_{T+2,\ldots,} f_{T+F}$ of length *F* is needed by giving a sequence of past predicted values f_1, $f_{2,\ldots,} f_T$ of length *T*. the prediction model takes the user's viewpoint position history as a sequence and then predicts the future viewpoint position as a sequence as well. Therefore, a sequence-to-sequence prediction model is developed by using an FCNN model. A three-layer FCNN model is used as a spatial combinatory for the prediction problem. The future head movement of a user in the 360-degree video streaming is related to multiple factors, such as current and past rotation status, treated as features in the prediction model. If the rotation angles for a fixed HMD are given, the viewpoint is determined. The architecture of FCNN model for viewport prediction has shown in Figure 1.

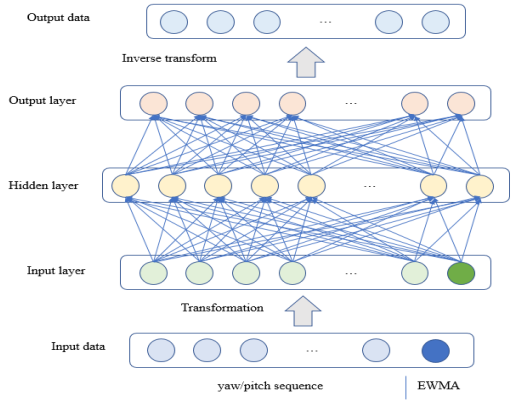


Figure 1: FCNN architecture for predicting yaw and pitch angles

Figure 1 shows the proposed three-layer FCNN model to implement the prediction process of head movement-based data, which is very light in computation. The proposed FCNN consists of one input layer, one hidden layer, and one output layer. The input layer of FCNN model takes the values of input data (yaw values, pitch values, EWMA of yaw values, and EWMA of pitch values) in this work as shown below, while the output layer provides the predicted output (yaw and pitch angles). In addition, the hidden layer of the proposed network encodes the additional non-linear information in the dataset. A hidden layer can have arbitrary multiple neurons where each neuron is connected with the nodes of input and output layer. The neural network with only one hidden layer is known as a single hidden layer neural network. The hidden layer chooses Rectified Linear unit (ReLu) [25] as its function that provides the non-linear transformation from input data to output data, whereas the sigmoid function is used by the output layer.

A. *Input Data Transform*

Each video of a participant is stored by considering the following factors as Timestamp *(T)*, playback time (*t*), unit quaternion *(x, y, z, w)* of HMD device, and the HMD position *(x, y, z)*. Quaternion is an algebraic structure that extends the familiar concept of complex numbers. While quaternions are much less intuitive than angles, rotations defined by quaternions can be computed more efficiently and with more stability. Therefore, it is reasonable to covert this dataset to be suitable for learning purposes. The unit Quaternion in a simple mathematical notation for the representation of orientations and object's rotation in a 3D space can be written as;

$$Q = w + x + y + z \ , \tag{1}$$

where *x, y,* and *z* are imaginary parts, while *w* is a real part. A rotation vector *(yaw, pitch, roll)* for each object is calculated from unit quaternion as follows:

$$\begin{Bmatrix} yaw \\ Pitch \\ roll \end{Bmatrix} = \begin{Bmatrix} a\tan 2(2(wz + xy), 1 - 2(w^2 + x^2)) \\ aSin(2(wy - xz)) \\ a\tan 2(2yz + wx), 1 - 2(x^2 + y^2) \end{Bmatrix} \tag{2}$$

In proposed work, the predictions based on viewpoint for yaw and pitch angles are made as stated in [15], where roll angle is mostly considered zero. It has been found that there is a very strong auto-correlation between yaw and pitch angles, as they are treated as independent variables for prediction. Therefore, sine and cosine of yaw and pitch angles are used for mapping on a unit circle. EWMA is used as input

of yaw and pitch values by using the following equation such as:

$$X_{ewma} = \frac{x_n(1-r)^{n-1} + x_{n-1}(1-r)^{n-2} +,...,x_2(1-r)^1 + x_1(1-r)^0}{(1-r)^{n-1} + (1-r)^{n-2} +,...,(1-r)^1 + (1-r)^0} \quad (3)$$

where $X = \{x_1, x_2, ..., x_n\}$ denotes the input sequence while $r = \frac{2}{N-1}$ gives the length of input sequence.

B. Output Data

Although After conversion using equations (2) and (3), the yaw and pitch angles are converted to radian by using equations (4) and (5), then encoded to sin (*yaw*), cos (*yaw*), sin (*pitch*), and cos (*pitch*) to reduce the angle periodicity that makes learning easier. Equation (6) is used for back conversion to the predicted values (yaw and pitch angles). The following equations show the relation between an angle θ_t (yaw angle) and its projected point (ω_1, ω_2) in 2D space.

$$\omega_1 = \sin(\theta_t) \quad (4)$$

$$\omega_2 = \cos(\theta_t) \quad (5)$$

$$\theta_t = \arctan(\omega_1 / \omega_2) \quad (6)$$

The same procedure is repeated for pitch values to predict the pitch angle (φ_t).

IV. Performance Evaluation

This section describes the experimental evaluation of the proposed model in comparison with alternative approaches.

A. EXPERIMENTAL SETUP AND METRICS

A public head movement dataset for 360-degree video is used for experimental purposes [26]. Compared to other 360-degree video datasets, the used dataset is comprehensive and has diverse dataset to explore the user behavior patterns during spherical video viewing that builds the new user identification mechanism. Table 1 shows the demographic profile for all the participants. A total of 48 users (24 males and 24 females) across 18 videos from 5 different categories participated in two separate experiments to carefully record how they watch the videos, how they move in each session, what type of content they can remember, and what direction they focus.

The model has implemented using the PyCharm environment. The proposed FCNN model is trained with the following hyperparameters settings such as batch size (32), Adam optimizer [27], and learning rate (0.002). In the training process, the network was trained for 50 normalized epochs with the ADAM optimizer that corrects the deviations and updates the weights to speed up the convergence during the model training. The proposed training model is a generalizable model that has been implemented for all the users. To perform the simulations, 80% processed log files and the remaining 20% user's log files are selected for all the videos as the training and testing datasets, respectively.

Table 1: Demographic Profile of the Participants.

Gender	Age	VR Experience	Academic Background
24 Female	≤ 20:10	Never: 1	Undergraduate: 10
24 Male	20=25:21	Sometimes:23	Master: 36
	≥ 26	Heard of before:22	PhD: 2
		Used frequency: 2	

For experimental work, three metrics, namely, Mean Square Error (MSE), Mean Absolute Error (MAE), and Average Euclidean Distance have been adopted between real and predicted viewpoints. For each metric, the performance of a model is better when the metric gets a lower value. The detail of each metric is given below.

- **Mean Square Error (MSE):** It is one of the most commonly used metrics to evaluate the proposed model over the prediction horizon. As shown below, *X* and *Y* are the model's predicted point and the real viewing point, respectively.

$$MSE(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i)^2 \quad (15)$$

where $X = (x_1, x_2, ..., x_n)$ and $Y = (y_1, y_2, ..., y_n)$.

- **Mean Absolute Error (MAE):** It is the commonly used regression error metric and can be employed to measure the average prediction accuracy by averaging the alleged error (the absolute value of each error) that is defined as:

$$MAE(X,Y) = \frac{1}{n}\sum_{i=1}^{n}|y_i - x_i|^2 \quad (16)$$

- **Average Euclidean Distance:** This metric gives the information to calculate the difference between the predicted and real viewpoints values, whose values are both in [-180,180). Mathematically, it is defined as:

$$d(X,Y) = \sum_{i=1}^{n} |((y_i - x_i + 180^0) mod 360^0) - 180^0| \quad (17)$$

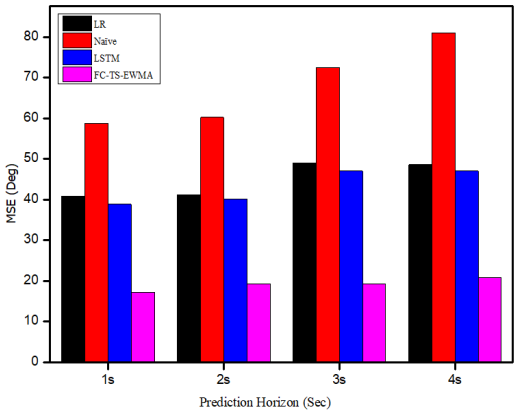
B. Performance Comparison

The proposed method is denoted as FCNN-TS-EWMA since it uses the transformation equations and EWMA function simultaneously. The results of the proposed FCNN-TS-EWMA model has been compared with techniques of LR [9], Naive [15], LSTM [7], FCNN model with EWMA (FCNN-EWMA), and only EWMA, respectively. In LR, a linear model predicts the future viewport by fitting all the data points in the sliding window. It uses the least square method to predict the user's orientation by using the history samples. In contrast, in the Naive method, the yaw and pitch values at the previous moment of the sequence are predicted for the next moment. However, the LSTM network learns the user's viewpoint motion patterns to predict the future viewport based on past movements. Whereas, the FCNN-EWMA means that the FCNN model is working without transformation of yaw and pitch values, while the EWMA model takes the input values of yaw, pitch, EWMA of yaw, and EWMA of pitch values. Here, the detail of each evaluating metrics is given below.

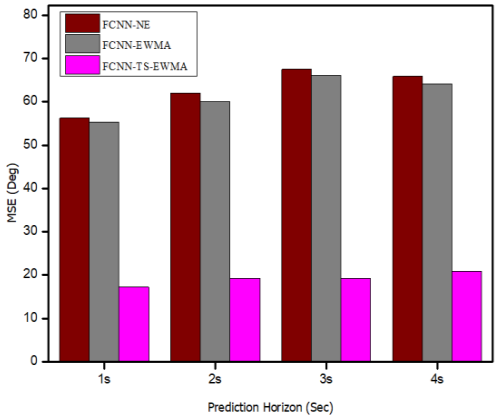
1. Evaluation of MSE

Figure 2 gives the performance of the proposed FCNN-TS-EWMA model for predicting the viewport and shows the prediction error based on MSE for the entire prediction length. A low MSE value means that the predicted value matches the real values. The MSE values for all methods tend to increase as the prediction window becomes larger. It is noted that, the proposed FCNN-TS-EWMA model has the lowest error rate at each position among all competitors.

Figure 2(a) clearly shows that the proposed FCNN-TS-EWMA approach outperforms the compared approaches from 1s to 4s. It represents that the proposed FCNN-TS-EWMA approach performs well than Naive method by a large margin. However, LSTM is performing well as compared to LR and Naïve approaches. In Figure 2(b), the prediction window is kept the same while the performance of FCNN-TS-EWMA is compared with other approaches, such as FCNN-NE and FCNN-EWMA. The proposed FCNN-TS-EWMA approach performs well and reduces the MSE value at each prediction length. The prediction made by FCNN-TS-EWMA model for each frame is based on viewpoint values that make it a low-complexity method. It also makes it possible to quickly adapt to the user's head movements since there is no need for expensive content processing computations. Figure 2 shows that the proposed FCNN-TS-EWMA model is effective in long-term prediction.



(a)



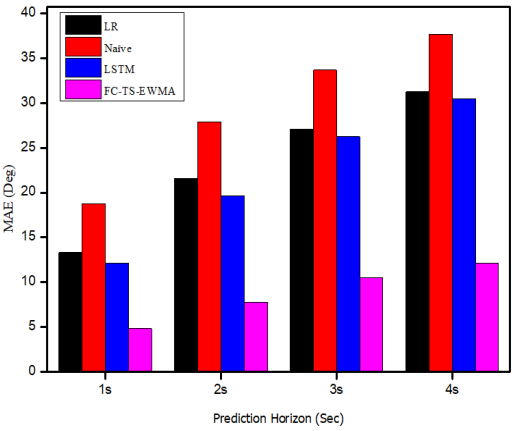
(b)

Figure 2: Evaluation of MSE

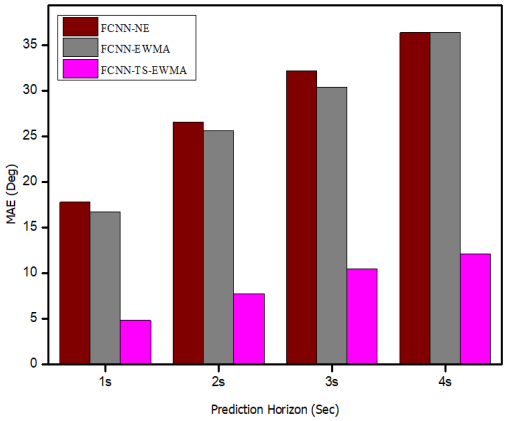
2. Evaluation of MAE

Figure 3 shows the MAE results of all benchmarkers with different prediction lengths. In terms of MAE, the proposed FCNN-TS-EWMA model outperforms when the prediction length increases from 1s to 4s, as shown in Figure 3. Compared to LSTM approach that performs best among the other competitors, the proposed FCNN-TS-EWMA model reduces MAE by 18% at the prediction length of 4s. It can be seen that FCNN-TS-EWMA gives better performance than other comparison approaches.

As the transformatted input has been used instead of using the direct input. Therefore, in Figure 3(b), It has found that the proposed FCNN-TS-EWMA approach again performs well by a large margin and has the lowest error rate at each position than its compititors. It represents that the proposed FCNN-TS-EWMA model significantly improves the prediction accuracy, especially when prediction length being large. This indicates that the proposed FCNN-TS-EWMA model has a stronger non-linearity ability and can perform better for a large prediction window.



(a)



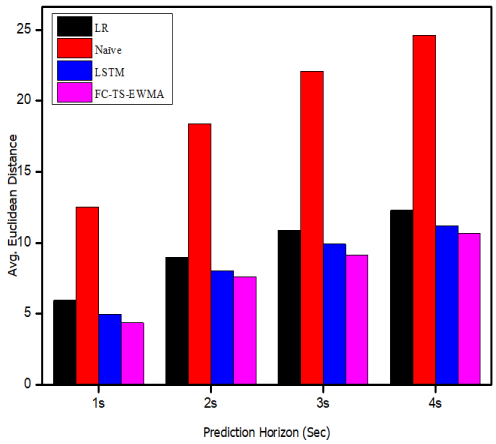
(b)

Figure 3: Evaluation of MAE

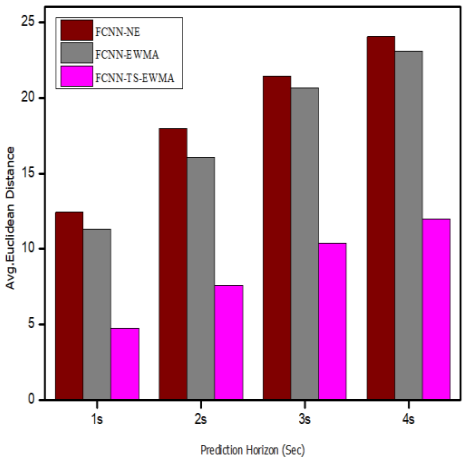
3. *Evaluation of Average Euclidean Distance*

To study how the prediction error at each position will increase/decrease in the prediction length, the Average Euclidean Distance at each position of the prediction window is also considered by the proposed FCNN-TS-EWMA model, which is shown in Figure 4. It can be seen that the LSTM performance is better than Naïve and LR, as shown in Figure 4 (a). While Naïve approach is increasing by a large margin than others and is not performing well in terms of Average Euclidean Distance to give the prediction accuracy. In Figure 4(b), it has found that the proposed FCNN-TS-EWMA model is very effective in long-term prediction and gives better performance than FCNN-EWMA.

Keeping all three-performance metrics in mind, it is noteworthy that the proposed FCNN-TS-EWMA model performs as the best model giving the highest prediction accuracy. It is concluded that the proposed FCNN-TS-EWMA model can accurately predicts future viewpoint positions of up to 4s.



(a)



(b)

Figure 4: Evaluation of Average Euclidean Distance.

V. CONCLUSIONS

This paper proposes a FCNN model that predicts the viewer's head movements in 360-degree video streaming to more accurately capture the non-linear relationship between the future and past viewpoints. The proposed FCNN model uses the transforming data instead of direct input to predict the future user movements. The prediction accuracy of the FCNN model has been compared with other comparison methods in terms of MAE, MSE, and Average Euclidean Distance. The experimental study shows that the proposed FCNN model outperforms the compared methods.

ACKNOWLEDGMENT

WE WOULD LIKE TO THANK HAO GONG FOR INSIGHTFUL COMMENTS TO IMPROVE THE QUALITY OF THE MANUSCRIPT

REFERENCES

[1] "Mtc360: A multi-tiles configuration for viewport-dependent360- degree video streaming," in *2020 IEEE 6th International Conference on Computer and Communications (ICCC)*. IEEE, 2020, pp. 1868–1873.

[2] M. Graf, C. Timmerer, and C. Mueller, "Towards bandwidth efficient adaptive streaming of omnidirectional video over http: Design, implementation, and evaluation," in *Proceedings of the 8th ACM on Multimedia Systems Conference*, 2017, pp. 261–271.

[3] Park, P. A. Chou, and J.-N. Hwang, "Rate-utility optimized streaming of volumetric media for augmented reality," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol.9,no.1,pp.149–162,2019.

[4] R. Yao, T. Heath, A. Davies, T. Forsyth, N. Mitchell, and P. Hoberman,"Oculus vr best practices guide," *Oculus VR*, vol. 4, pp. 27–35,2014.

[5] 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 transactions on pattern analysis and machine intelligence*, vol. 41, no. 11, pp. 2693–2708, 2018.

[6] Y. Li, Y. Xu, S. Xie, L. Ma, and J. Sun, "Two-layer fov prediction model for viewport dependent streaming of 360-degree videos," in *International Conference on Communications and Networking in China*. Springer,2018,pp.501–509.

[7] 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*, 2017, pp. 67–72.

[8] M. Assens Reina, X. Giro-i Nieto, K. McGuinness, and N. E. O'Connor, "Saltinet: Scan-path prediction on 360 degree images using saliency volumes," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2017, pp. 2331–2338.

[9] F. Duanmu, E. Kurdoglu, S. A. Hosseini, Y. Liu, and Y. Wang, "Prioritized buffer control in two-tier 360 video streaming," in *Proceedings of the Workshop on Virtual Reality and Augmented Reality Network*, 2017, pp.13–18.

[10] 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*,2017,pp.315–323.

[11] X. Corbillon, G. Simon, A. Devlic, and J. Chakareski, "Viewportadaptive navigable 360-degree video delivery," in *2017 IEEE international conference on communications (ICC)*. IEEE, 2017, pp. 1–7.

[12] A. T. Nasrabadi, A. Mahzari, J. D. Beshay, and R. Prakash, "Adaptive 360-degree video streaming using scalable video coding," in *Proceedings of the 25th ACM international conference on Multimedia*, 2017, pp. 1689–1697.

[13] 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*, 2016, pp. 1–6.

[14] J. Heyse, M. T. Vega, F. De Backere, and F. De Turck, "Contextual bandit learning-based viewport prediction for 360 video," in *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. IEEE, 2019,pp.972–973.

[15] 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)*. IEEE,2016,pp.1161–1170.

[16] Y. Ban, L. Xie, Z. Xu, X. Zhang, Z. Guo, and Y. Wang, "Cub360: Exploiting cross-users behaviors for viewport prediction in 360 video adaptive streaming," in *2018 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2018, pp. 1–6.

[17] J. Yu and Y. Liu, "Field-of-view prediction in 360-degree videos with attention-based neural encoder-decoder networks," in *Proceedings of the 11th ACM Workshop on Immersive Mixed and Virtual Environment Systems*,2019,pp.37–42.

[18] M. Jamali, S. Coulombe, A. Vakili, and C. Vazquez, "Lstm-based viewpoint prediction for multi-quality tiled video coding in virtual reality streaming," in *2020 IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE, 2020, pp. 1–5.

[19] C. Li, W. Zhang, Y. Liu, and Y. Wang, "Very long term field of view prediction for 360-degree video streaming," in *2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*. IEEE, 2019,pp.297–302.

[20] A. T. Nasrabadi, A. Samiei, and R. Prakash, "Viewport prediction for 360 videos: a clustering approach," in *Proceedings of the 30th ACM Workshop on Network and Operating Systems Support for Digital Audio andVideo*,2020,pp.34–39.

[21] A. D. Aladagli, E. Ekmekcioglu, D. Jarnikov, and A. Kondoz, "Predicting head trajectories in 360 virtual reality videos," in *2017 International Conference on 3D Immersion (IC3D)*. IEEE, 2017, pp. 1–6.

[22] Q. Yang, J. Zou, K. Tang, C. Li, and H. Xiong, "Single and sequential viewports prediction for 360-degree video streaming," in *2019 IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE, 2019, pp.1–5.

[23] Y. Xu, Y. Dong, J. Wu, Z. Sun, Z. Shi, J. Yu, and S. Gao, "Gaze prediction in dynamic 360 immersive videos," in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp.5333–5342.

[24] A. Nguyen, Z. Yan, and K. Nahrstedt, "Your attention is unique: Detecting 360-degree video saliency in head-mounted display for head movement prediction," in *Proceedings of the 26th ACM international conference on Multimedia*, 2018, pp. 1190–1198.

[25] J. He, L. Li, J. Xu, and C. Zheng, "Relu deep neural networks and linear finite elements," *arXiv preprint arXiv:1807.03973*, 2018.

[26] Chenglei Wu, Zhihao Tan, Zhi Wang, Shiqiang Yang. "A Dataset for Exploring User Behaviors in VR Spherical Video Streaming," In Proceedings of ACM Multimedia Systems (MMSys) 2017, Taipei, Taiwan, June 20-23, 2017.

[27] Y. Jiang and F. Han, "A hybrid algorithm of adaptive particle swarm
optimization based on adaptive moment estimation method," in *International Conference on Intelligent Computing*. Springer, 2017, pp. 658–667.