Two-Layer FoV Prediction Model for Viewport Dependent Streaming of 360-Degree Videos

Yunqiao Li, Yiling Xu(✉), Shaowei Xie, Liangji Ma, and Jun Sun

Shanghai Jiao Tong University, Shanghai, China
{liyunqiao,yl.xu,junsun}@sjtu.edu.cn, {xswjsnj,maliangji2012}@163.com
http://cmic.sjtu.edu.cn/CN/Default.aspx

Abstract. As the representative and most widely used content form of Virtual Reality (VR) application, omnidirectional videos provide immersive experience for users with 360-degree scenes rendered. Since only part of the omnidirectional video can be viewed at a time due to human's eye characteristics, field of view (FoV) based transmission has been proposed by ensuring high quality in the FoV while reducing the quality out of that to lower the amount of transmission data. In this case, transient content quality reduction will occur when the user's FoV changes, which can be improved by predicting the FoV beforehand. In this paper, we propose a two-layer model for FoV prediction. The first layer detects the heat maps of content in offline process, while the second layer predicts the FoV of a specific user online during his/her viewing period. We utilize a LSTM model to calculate the viewing probability of each region given the results from the first layer, the user's previous orientations and the navigation speed. In addition, we set up a correction model to check and correct the unreasonable results. The performance evaluation shows that our model obtains higher accuracy and less undulation compared with widely used approaches.

Keywords: Omnidirectional video · Field of view prediction
FoV-based transmission

1 Introduction

VR immerses the user into the virtual world by realizing free interaction, which reshapes the consumption experience for users. Intensive attention is being attracted from the industry and the academia [1]. Omnidirectional video is one of the most typical application format in VR. Obviously, to get the same visual perception quality as traditional videos, the whole omnidirectional videos covering 360-degree scenes will contain much more amount of data. However, due to the restriction of human visual system (HVS), users can only see the content

X. Liu et al. (Eds.): ChinaCom 2018, LNICST 262, pp. 501–509, 2019.
https://doi.org/10.1007/978-3-030-06161-6_49

within their FoV at a time, and the size of FoV equipped by the head mounted display (HMD) of HTC Vive system is about 110-degree in horizontal range.

To save the bandwidth while not sacrificing the quality of experience (QoE), FoV-based transmission [9] of omnidirectional videos is widely recognized as a effective scheme. The principle of this scheme is to transmit the content that covers user's FoV with high quality, while the other regions with lower quality to avoid the blank screen [3]. Studies about FoV adaptation model finds that the refinement duration affects the QoE of the user significantly [15]. After the user switching the viewport by turning head, the quality within the new FoV will decrease before the high quality content is received covering the new FoV. While, this can be improved by FoV prefetching which has to be motivated by accurate FoV prediction beforehand.

Recently, some researchers have investigated the very short-term FoV prediction. Feng Qian proposed to utilize weighted linear regression algorithm in the prediction based on previous viewing orientation information [2,12], and this approach has been widely adopted in transmission improvement thanks to its simplicity. Meanwhile, as a state-of-art technology, Convolutional Neural Network (CNN) is also applied to improve the prediction [16]. Furthermore, content features are being considered in prediction [5,13], while still limited to short-term.

The navigation trajectory of a user keeps close relevance to the personalized FoV prediction. We introduce the orientation and navigation speed which reflects the user's viewing status in our prediction model. In addition, content attractiveness detection helps to build reliable prediction as users are often stimulated to switch the viewport for certain objects. We propose a two-layer model to work on the FoV prediction. The first layer is designed as an offline process to detect the heat value of each region in the omnidirectional videos. Motion features of the content is paid more emphasis, which proves to better comply with this problem. In the second layer, FoV of the user is predicted in real time. The previous orientation and navigation speed as part of the inputs illustrates the short-term viewing trend and status. What's more, we input the results from the first layer related to short-term and long-term step into the second layer to provide distinguish guidance to the prediction. In this layer, a LSTM module is trained to learn the intricate connections in this problem, and a correction module is designed to check and correct the unreasonable results at last.

This paper is organized as follows. In Sect. 2, we describe the overall framework and concrete details of each step in our model. Experiment setup and results are presented in Sect. 3. Finally, in Sect. 4 we provide conclusions about our research and look into the future work related.

2 Prediction Model

To predict the FoV of a specific user, we investigate the viewing behavior of users. According to the research of Sitzmann, the viewing status can be divided into focusing and browsing status [13]. Under the focusing status, users are attracted by the specific content in the video. They switch the viewport to follow the

targeted content or pay attention to the extended part of it. In this status, the navigation speed appears to be lower than the average, and the heat region of the content close to current FoV has a strong correlation with the viewport switching. When the user is under browsing status, their movement shows great irregularity, for they do not pay real attention to the content in omnidirectional video, and the head rotation speed appears to be higher than the average. In this status, the prediction should mainly focus on heat region of the video, which may get the userś attention later and change one's viewing state into gazing status, with wider range than the previous circumstances. Therefore, we propose a two-layer FoV prediction model as shown in Fig. 1 considering above key factors to adapt to different viewing status.

In order to save the computing resources and to predict faster during the viewing process, the detection of the heat region of the video is completed offline before viewed, which refers to the first layer in our model. By detecting the saliency feature of each static frame and the moving object between frames, two main kinds of feature maps are obtained. As people tend to focus on and follow the moving objects when immerses in an omnidirectional video, we place extra emphasis on the motion features when combine two features into heat maps.

Besides the heat value already detected, user's orientation and navigation speed on previous frames are tracked and feeded back into the data-driven LSTM architecture during the user' consumption. Furthermore, to make up the oversight of the FoV characteristics, a correction module is added as the final part in second layer.

We adopt the tile-based transmission scheme in this paper [8], where a full omnidirectional video is partitioned into many coding independent rectangular tiles, so that each tile can be transmitted independently [6]. Our prediction model is established on this transmission scheme as well.

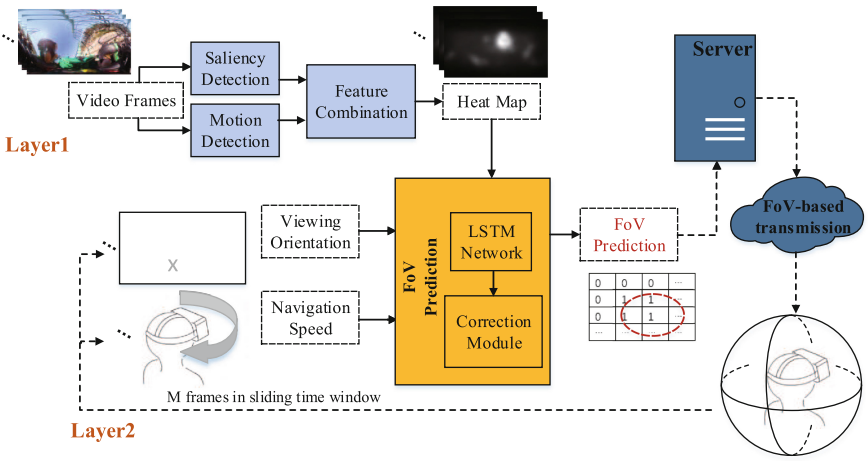

Fig. 1. Overview of the proposed two-layer FoV prediction model. The example of a tile-based streaming scheme is shown.

2.1 Layer1- User Attention Prediction

This layer is an offline process to detect the attractive regions in the content with higher probability to be viewed. The heat value $HT_n, n \in N$ of N tiles are obtained in this layer.

Saliency Detection. Image saliency map is obtained by detecting the objects that show distinct differences in features of colors, textures, etc. from the surroundings. The Fused Saliency Map (FSM) [4] is used for detection for it adapt current detection models to omnidirectional image characteristics.

Motion Detection. People tend to be attracted by moving objects such as moving animals or athletes in a sports game. This can be detected by analyzing the features between consecutive frames. We utilize the Lucas-Kanade optical flow approach to detect pixel-level motion features [14].

Content Feature Combination. We convolute the motion maps with a 2D Gaussian filter so that the motion features can present overall impact on the attractive region detection results. Pixel-wise heat maps are calculated as:

$$HP_k = \begin{cases} MP_k, & MP_k \geq SP_k \\ \frac{MP_k + SP_k}{2}, & MP_k < SP_k \end{cases} \tag{1}$$

HP_k is the heat value of k^{th} pixel,SP_k and MP_k are saliency value and filtered motion value. Then tile-wise heat value HT is obtained by calculating and normalizing the summation of the pixel-wise heat values HP within each tile.

2.2 Layer2- FoV Prediction

The second layer is an online process to predict the tiles to be viewed next.

LSTM Prediction Module. This module is to predict the probability of each tile to be viewed by a specific user, which is based on the recurrent neural network (RNN) architecture. This scheme presents high efficiency in temporal sequential problems as realizing weight sharing in time domain. We choose Long short-term memory(LSTM) [7,10] model replacing the ordinary node in network by the memory cell to avoid gradient vanishing and explosion by adding a forget gate towards the cell to rectify the long-term and short-term memory of the node.

As the impact of inputs towards the prediction gradually changes (the reliability of the head orientation in the sliding window gradually decreases), the output of the prediction should also reflect this trend. When omnidirectional videos are partitioned into N tiles, our model provides short-term prediction results $Pred_{short} = \{p\}_1^N$ and additional long-term prediction results $Pred_{long} = \{p\}_1^N$ to adapt to the variation of user's FoV in the prediction window.

The input to our LSTM model consists of the tile-wise heat maps of next $2 * M$ frames including $\{H\}_{t+1}^{t+M}$ related to short-term content and $\{H\}_{t+M+1}^{t+2*M}$ related to long-term content, along with user's orientation and navigation speed of previous M frames $\{O\}_{t-M}^{t}$, $\{S\}_{t-M}^{t}$. The head orientation information $O(x, y, z)$ is expressed in the form of the position on the x, y, z coordinates; and the navigation speed on sphere surface $S(yaw, pitch, raw)$ is calculated as the orientation change compared with the previous frame in rotation coordinates $yaw, pitch, raw$.

We adopt two layer LSTM model with 256 neurons each layer, which presents better performance compared with the other parameters we have tried. The prediction results are obtained at the last time step, while inner results obtained in each time step is retained in the form of state information and becomes the input into the same neuron at the next time step to establish the association in the time domain as shown in Fig. 2.

When training the LSTM module, learning rate is set to be 0.01, and a dropout layers with 50% drop rate is added to prevent the overfitting. We adopt the Adam Optimizer to minimize the cross-entropy loss. The sliding window is set to be 1 second, as well as the size of short-term and long-term prediction window. A down-sampling process is performed on frame rate of the input by 2 times to shorten the input length which accelerates the prediction as the content and viewing features of adjacent frames has minor changes. We divide the dataset into 80% and 20% for training and validation respectively. Different videos are trained and validated separately, so that a unique prediction network is trained for each video.

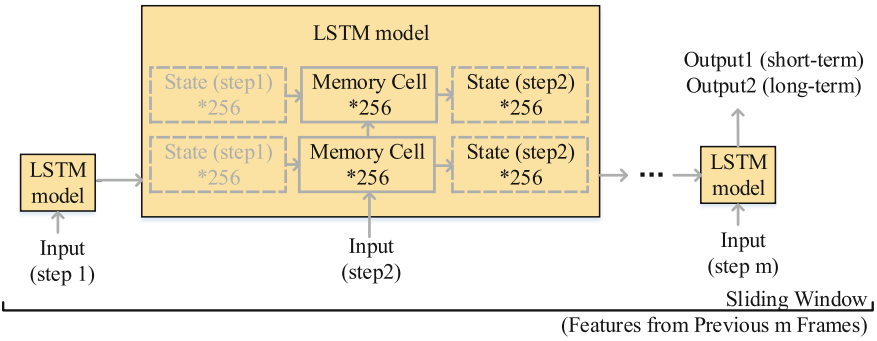

Fig. 2. LSTM prediction module in our approach.

Correction Module. By setting an appropriate threshold, tiles are decided whether to be covered by FoV. However, the separately judged tile-wise prediction results cannot always compose a reasonable FoV region. We propose a correction module with two steps to make sure the predicted FoV as an integrated region with reasonable size.

The first step is to make up the omitting error that a few tiles are misjudged to be negative, while their surroundings are all judged to be covered by FoV. This violate the integrity of FoV. When there is a wide range of misjudged tiles, we believe that this situation does not belong to omitting error, but the drawback of the prediction model which has been avoided in our model. We scan the prediction results for each row of all tiles. When the predicted FoV is partitioned to be at two side of the projected plane, we correct the negative predicted tiles among the range of the frame edge and the farmost positive predicted tiles away from the frame edge. Otherwise we find the edge tiles that predicted to be viewed and make sure each misjudged tile among them is reset as the positive tiles.

The second step is to determine the minimum range that the positive judged tiles should cover. Equirectangular projection is widely used to project the omnidirectional video from spherical 3-Dimention (3D) to 2D plane, sampling uniformly on latitude. However, this approach causes the distortion of the content near two poles. Equally partitioned tiles on projected 2D frames correspond to spherical tiles with larger size near the equator and smaller size when approaching poles. Therefore, during the playback of an omnidirectional video that has been rendered as a spherical three-dimension video, the range of tiles falling into the user's FoV is not always the same.So that the FoV shows a changing shape and size on the projected 2D plane video. We locate the center of the predicted FoV by averaging the position of all tiles. The shape and range of the FoV on the 2D projected video related to the above calculated center can be obtained according to the projection relationship. Then we complement the negative judged tiles around the center to fill in the range of FoV.

3 Performance Evaluation

The LSTM network of the prediction model has to be trained by the omnidirectional video dataset [11]. The dataset has ten omnidirectional videos all lasting one minutes. All videos are of 4K resolution with 30fps. The videos are projected with equirectangular projection. Each video is viewed by 30 viewers wearing HMD and the orientation of each frame has been collected by Open Track and presented in both Cartesian coordinates and Euler angles. The omnidirectional video is partitioned into 10?20 tiles, which are labeled whether is viewed by the user.

We evaluate the performance of our two-layer model of FoV prediction given the pre-detected heat map and continuously collected user's head movement information related to the previous frames in sliding window.

We compare the performance of our model with the typical and widely used prediction approach of weighted linear regression [12].

Table 1 provides the specific measurement of accuracy and F-score of the prediction on each video. Accuracy presents the primary performance of the prediction model by calculating the ratio of tiles correctly classified to all tiles. F-score is the weighted mean of the precision and recall (precision is the ratio of

Table 1. Performance of the model on each video.

Videos	Training set		Validation set	
	Accuracy	F-score	Accuracy	F-score
Mega coaster	89.53%	0.7312	88.90%	0.7011
Roller coaster	90.08%	0.7270	88.28%	0.6693
Driving with the 360	79.08%	0.3719	77.34%	0.2821
Shark shipwreck	83.46%	0.5097	83.26%	0.5259
Perils panel	90.16%	0.7408	89.12%	0.7246
Kangaroo Island	84.53%	0.5735	82.62%	0.5308
SFR spore	92.37%	0.8014	89.29%	0.7282
Hog rider	80.29%	0.3544	77.09%	0.2742
Pac-Man	88.74%	0.7151	87.45%	0.6826
Chariot race	88.26%	0.6818	87.79%	0.6040
Average(ours)	86.65%	0.6207	85.11%	0.5722
Average(w-reg)	73.00%	0.5352	72.54 %	0.5155

tiles classified to be positive correctly to all positive classified tiles, and recall is the ratio of tiles classified to be positive correctly to all viewed tiles).

The accuracy presents the primary performance of the prediction model. Our model outperforms over the weighted linear regression approach with a considerably increase of accuracy. Our prediction approach obtains balance between precision and recall over most video sequences. The performance improvement caused by our model is probably because (a) the newly introduced navigation speed leads the model to be more sensible to different viewing status, (b) correction module in the second layer improves the prediction to be more reasonable, and (c) our model predict both short-term FoV and long-term FoV to match the user's viewport switching in longer time (d) we synthetically utilize the content features and viewing movement information with our two-layer model.

Moreover, we evaluate the fluctuation of the prediction approach on each video sequence. As shown in the Fig. 3, our model obtains great stability compared with the weighted linear regression approach. Maintaining the prediction at a higher accuracy all the time during the user's viewing period still proves to be a great challenge. In the process of viewing most video sequences, the prediction accuracy of our model is stable near the average value in spite of the changeable viewing status of users. However, when our prediction model achieves higher accuracy, the stability prediction reduces. In contrast, the weighted linear regression approach not only obtains low average accuracy, but also shows strong instability. In other words, our approach capture the regularity of the FoV switching better, and is able to be adaptive to various users and videos.

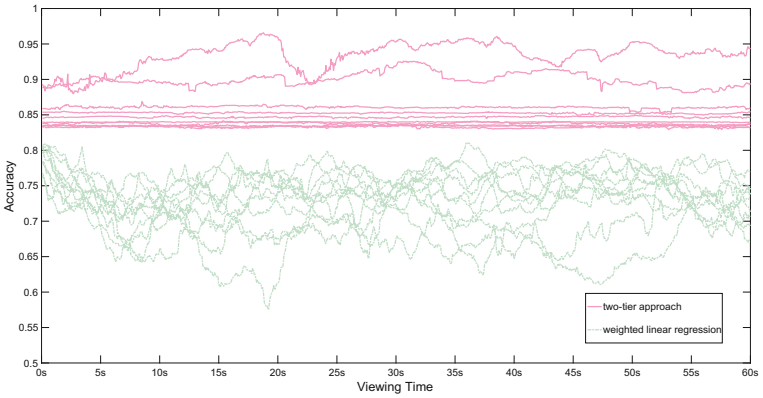

Fig. 3. The prediction performance comparison of our approach against weighted linear regression ap-proach on ten videos during the whole consumption period.

4 Conclusion

In this paper, we propose a two-layer model to predict user's FoV on omnidirectional video during one's viewing. The user's navigation speed is considered for the first time with user's viewing orientation and content features. LSTM model is trained and used to predict the probability of each region to be viewed, which is further optimized by a correction module. The experimental results show that our prediction approach is superior to other conventional approaches, obtaining great accuracy and stability improvement under different viewing status. In the future, we will utilize our prediction model in our realistic multi-network system to further evaluate the performance in real application.

Acknowledgments. This paper is supported in part by National Natural Science Foundation of China (61650101), Scientific Research Plan of the Science and Technology Commission of Shanghai Municipality (16511104203), in part by the 111 Program (B07022).

References

1. Virtual and augmented reality: Understanding the race for the next computing platform. Technical report. The Goldman Sachs Group Inc. (2016)
2. Bao, Y., Wu, H., Zhang, T., Ramli, A.A., Liu, X.: 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. IEEE (2016)
3. Corbillon, X., Simon, G., Devlic, A., Chakareski, J.: Viewport-adaptive navigable 360-degree video delivery. In: 2017 IEEE International Conference on Communications (ICC), pp. 1–7. IEEE (2017)

4. De Abreu, A., Ozcinar, C., Smolic, A.: Look around you: Saliency maps for omnidirectional images in vr applications. In: 2017 Ninth International Conference on Quality of Multimedia Experience (QoMEX), pp. 1–6. IEEE (2017)
5. Fan, C.L., Lee, J., Lo, W.C., Huang, C.Y., Chen, K.T., Hsu, C.H.: 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, pp. 67–72. ACM (2017)
6. Graf, M., Timmerer, C., Mueller, C.: Towards bandwidth efficient adaptive streaming of omnidirectional video over http: Design, implementation, and evaluation. In: Proceedings of the 8th ACM on Multimedia Systems Conference, pp. 261-271. ACM (2017)
7. Hochreiter, Sepp, Schmidhuber, Jürgen: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
8. Mohammad H., Swaminathan, V.: Adaptive 360 vr video streaming: Divide and conquer. In: 2016 IEEE International Symposium on Multimedia (ISM), pp. 107–110. IEEE (2016)
9. Hu, Y., Xie, S., Xu, Y., Sun, J.: Dynamic VR live streaming over MMT. In: 2017 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB), pp. 1–4. IEEE (2017)
10. Lipton, Z.C., Berkowitz, J., Elkan, C.: A critical review of recurrent neural networks for sequence learning (2015). arXiv:1506.00019
11. Lo, W.C., Fan, C.L., Lee, J., Huang, C.Y., Chen, K.T., Hsu, C.H.: 360 video viewing dataset in head-mounted virtual reality. In: Proceedings of the 8th ACM on Multimedia Systems Conference, pp. 211–216 (2017)
12. Qian, F., Ji, L., Han, B., Gopalakrishnan, V.: Optimizing 360 video delivery over cellular networks. In: Proceedings of the 5th Workshop on All Things Cellular: Operations, Applications and Challenges, pp. 1–6 (2016)
13. Sitzmann, Vincent, Serrano, Ana, Pavel, Amy, Agrawala, Maneesh, Gutierrez, Diego, Masia, Belen, Wetzstein, Gordon: Saliency in vr: How do people explore virtual environments? IEEE Trans. Vis. Comput. Graph. **24**(4), 1633–1642 (2018)
14. Wu, Z., Su, L., Huang, Q., Wu, B., Li, J., Li, G.: Video saliency prediction with optimized optical flow and gravity center bias. In: 2016 IEEE International Conference on Multimedia and Expo (ICME), pp. 1–6. IEEE (2016)
15. Xie, S., Xu, Y., Qian, Q., Shen, Q. Ma, Z., Zhang, W.: Modeling the perceptual impact of viewport adaptation for immersive video. In: 2018 IEEE International Symposium on Circuits and Systems (ISCAS), pp. 1–5. IEEE (2018)
16. Xu, M., Song, Y., Wang, J., Qiao, M., Huo, L., Wang, Z.: Modeling Attention in Panoramic Video: A Deep Reinforcement Learning Approach (2017)