A Review of Deep Learning Solutions in 360° Video Streaming

Moatasim Mahmoud
European Projects Department
Singular Logic
14564 Athens, Greece
mmoatasim@singularlogic.eu

Stamatia Rizou
European Projects 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 Engineering
Aristotle University of Thessaloniki
54124 Thessaloniki, Greece
kant@auth.gr

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

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

2023 12th International Conference on Modern Circuits and Systems Technologies (MOCAST) | 979-8-3503-2107-4/23/$31.00 ©2023 IEEE | DOI: 10.1109/MOCAST57943.2023.10176729

***Abstract*—The spread of virtual reality and 360° video applications has raised research interest in developing new streaming techniques. On one hand, 360° videos rely on strict network requirements compared to conventional 2D videos. Realizing an adequate user experience is subject to ultra-low latency and huge bitrate requirements. On the other hand, 360° videos have distinct characteristics that allow for innovative streaming solutions. These solutions have benefited from the advancements in deep learning for optimizing the transmission under restricted network resources. In this paper, we review existing works employing deep learning in 360° video transmission and we highlight the challenges associated with 360° video streaming.**

Keywords—360° videos, deep learning, video streaming, augmented reality (AR), virtual reality (VR).

I. INTRODUCTION

The latest developments in multimedia technologies have led to the popularization of virtual reality (VR) and 360° videos. Users equipped with head-mounted displays (HMDs) can explore the spherical view of an omnidirectional video by moving their heads toward the desired directions. Current and future communication systems aim to accommodate the emerging multimedia applications which raise critical network challenges. 360° videos for instance rely on very high bitrates and tight latency requirements [1]. To overcome these challenges, viewport information has been leveraged for the development of resource-efficient streaming algorithms. Rather than transmitting whole bulky video files, viewport-dependent streaming transmits only the parts that are within or adjacent to users' viewports [2]. This in turn ensures substantial bandwidth reductions and more efficient use of resources. Projection methods such as equirectangular projection (ERP) and cube-map projection are usually utilized to transform 360° videos into 2D formats [3]. Consequently, the resulting 2D shapes are divided into rectangular tiles that can be encoded independently. Hence allowing for a more flexible and optimized transmission by streaming only tiles related to users' viewports. HMDs normally collect users' head movements and eye gaze traces and translate them into their corresponding viewports. These traces can also be utilized for predicting future users' viewports and thus allowing for proactive network decisions.

In the case of conventional 2D videos, many network solutions have been investigated and applied for smooth video streaming. Adaptive video streaming readjusts transmission parameters according to the varying network conditions [4], [5]. Edge caching can bring popular videos closer to end users for reduced latency and optimized bandwidth [6], [7]. In situations where video content is streamed to multiple clients, multicast transmission suggests a simultaneous transmission to reduce the required resources [8]. Several optimization techniques have been adopted to attain the best possible performance. Nonetheless, 360° videos differ from regular videos as they entertain stricter quality of service (QoS) requirements. Moreover, 360° videos have unique characteristics that can be integrated for more innovative solutions.

In the past decades, machine learning (ML) and deep learning (DL) techniques have evolved into powerful tools that demonstrated significant breakthroughs in domains such as natural language processing and computer vision [9]. Advancements in communication systems made it difficult for conventional modeling and optimization techniques to capture the full complexity of such systems. Thus, the power of ML and DL has been also investigated and utilized in communication and networking problems [10], [11]. Such techniques prove promising in the context of 360° video transmission and recent work towards this direction exist.

In this article, we review the current approaches on deep learning methods applied for the optimization of 360° video transmission. In section II, we introduce important aspects of viewing 360° videos and in section III we review DL-based viewport prediction methods. Section IV summarizes the use of DL in 360° video transmission. We then provide a short discussion and conclude our work in section V.

979-8-3503-2107-4/23/$31.00 ©2023 IEEE

II. 360° Videos Characteristics

360° videos are typically captured by 360° camera or created by combining multiple scenes of distributed cameras. A user watching a 360° video using an HMD can freely navigate the spherical view by looking in the desired direction. In a three degrees of freedom (3DoF) setting, a user can move his/her head along the pitch, yaw, and roll axes to cover the whole spherical scene [12]. A user's viewport or field of view (FoV) is constructed based on the user's viewing direction and the HMD specifications. Fig. 1 shows a user while watching a 360° and his defined FoV.

The spherical scene is usually mapped into a rectangular format and then partitioned into nonoverlapping tiles that are encoded and transmitted independently. Among the varied existing mapping schemes, equirectangular projection is the most common approach. In ERP, the spherical video is projected into a rectangular grid resulting in non-uniform distribution of pixels [13]. Cubemap projection stretches the spherical shape and projects it into the six faces of a cube. Although cubemap improves over ERP, the pixel density still varies within the sides of the cube. In view of this, equi-angular cubemap (EAC) adjusts the sampling steps to be at uniform distances and assigns equal pixel densities across the whole scene [14]. Many other shapes and formats have been investigated to overcome the nonuniform pixel distribution including the dyadic projection [15] and the barrel projection [16]. Pyramid projection [17] provides a viewport-dependent encoding method that relies on the watched viewport. The pyramid encoding scheme expands the spherical video to cover a pyramid shape where only the user's FoV is rendered at full quality. Transforming omnidirectional videos into 2D formats allows for leveraging powerful 2D video encoding methods including Advanced Video Coding (AVC) / H.264 [18], High-Efficiency Video Coding (HEVC) / H.265 [19], and AV1 [20]. In 2020, Versatile Video Coding (VVC) / H.266 [21] was introduced to outperform previously existing codecs and to serve a range of applications including 360° video streaming.

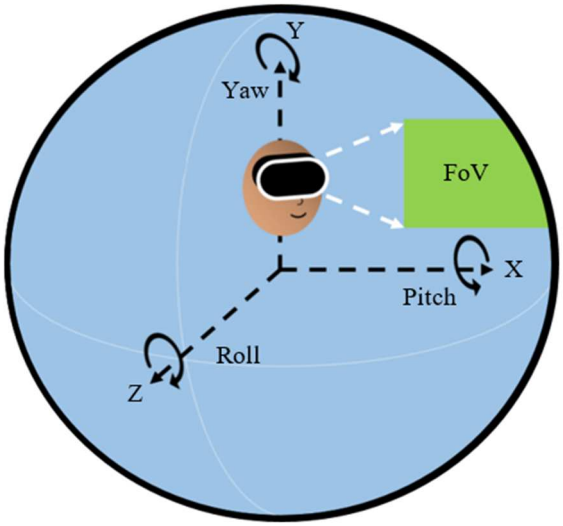

Figure 1: Viewing 360° videos.

Efficient transmission schemes attempt to send tiles that comprise the user's FoV leading to significant bandwidth reductions. This utilizes knowledge about the user's viewing direction which can be obtained from the HMD recorded traces. Due to the crucial latency requirements of omnidirectional videos, many viewport prediction methods have been proposed to anticipate users' requests and act accordingly in advance. Humans tend to focus on engaging parts within spherical videos. Analyzing viewing trajectories demonstrate that users' attention is directed toward specific objects or scenes with high correlation among different users [22]. Moreover, horizontal explorations of videos (yaw) prevail over vertical explorations (pitch) as the vertical direction is usually concentrated in central regions. These observations can be utilized for making intelligent viewport predictions and for optimizing proactive 360° video streaming strategies.

III. DL-Based Viewport Prediction

Accurate viewport predictions are needed for a reliable proactive service. Consequently, many solutions have employed the recent advancements in DL methods. [23] combines content-related information including image saliency maps and motion maps with viewing directions recorded by the HMDs to obtain tile viewing probabilities. In this work, image saliency maps are produced offline using a convolutional neural network (CNN). Saliency information, along with motion maps and previous viewing orientations, are then fed to a fixation prediction network to predict future viewing directions. The fixation prediction network utilizes long short-term memory (LSTM) networks to estimate the upcoming viewing probabilities of video tiles. In addition to relying on the user's own past viewing information, [24] leverages other users' future FoVs. Considering trajectory-based and heatmap-based input formats, the authors propose multiple models to tackle the different scenarios including LSTM seq2seq models, attentive mixture of experts (AME), and fully convolutional networks (FCNs). [25] employs LSTM for encoding users' recorded gaze trajectories and CNN for extracting saliency features of the video content. The outputs of the two modules are combined using fully connected layers to predict gaze displacements between current and future times. A transformer-based architecture that uses the user's past viewport scanpath is proposed in [26] to estimate the upcoming scanpath. Transformer networks employ self-attention layers and allow for capturing time dependencies in the input scanpaths. The proposed model results in a good long-term prediction performance with low computational complexity and without relying on any content-based information. LiveROI [27] provides online viewport prediction to be used in live VR streaming settings. A 3D-CNN model is adopted to identify important actions within the video content and interpret them through natural language. The extracted information is then combined with the user's viewing trajectory to obtain the final predictions.

Proactive transmission of 360° videos requires having reliable viewport predictions. In some transmission scenarios, however, it is wise to anticipate inaccuracies in viewport predictions by adding a margin to the predicted viewports to be also transmitted. In other cases, the whole panoramic video is rendered and sent at low resolution and only the predicted viewport is transmitted at the full resolution. Therefore, the user experience will be less affected by inaccurate predictions.

IV. DL-BASED 360° VIDEO TRANSMISSION

Transmitting only the tiles relevant to users' FoVs results in a considerable bandwidth reduction. With tile-based streaming, tiles can be encoded and transmitted independently at different rates. Choosing appropriate tile rates according to

channel conditions and users' viewports to optimize the transmission quality can be a challenging problem [28], [29]. [30] first calculates future viewport predictions using LSTM to be utilized in optimizing the tile bitrate selection. Specifically, a reinforcement learning (RL) algorithm is used for choosing proper viewport and non-viewport tile bitrates. Another RL-based solution for adaptive 360° video streaming is proposed in [31]. A 3D-CNN model is used for obtaining spatio-temporal features within the video content and utilizes the acquired information to provide the viewport predictions. The selection of tiles and tile qualities is performed using RL according to changes in the available bandwidth and user viewport. Another 360° video delivery system is presented in [32]. This framework aims at maximizing users' quality of experience (QoE) under network capacity conditions. The proposed tiled adaptive transmission relies on viewport predictions which are acquired based on the saliency and motion maps of the videos and head-tracking historical information. Two different deep learning approaches are investigated in this work for obtaining the tile probability maps. The first one is based on a CNN+LSTM architecture, whereas the second approach leverages 3D-CNNs. PARSEC [33] makes use of the underutilized computational power on the client side to reduce the bandwidth requirements for streaming 360° videos. It employs deep neural networks (DNNs) based super-resolution at the client side to retrieve the high-resolution content. Depending on the available computational power, available bandwidth, and viewport predictions, a rate adaptation algorithm is proposed to select which tiles should be fetched at full resolutions from the server and which tiles should be reconstructed from lower resolutions at the client side to maximize the overall QoE.

In the presence of multiple VR users who are expected to consume common content, correlations among users' viewports can be leveraged by multicasting the common viewed areas. Optimizing the multicast transmission includes grouping and scheduling the users according to their network conditions and similarities between their FoVs. In [34], a multicast transmission scheme is proposed in a millimeter wave (mmWave) communication system where groups of users are served by nonorthogonal beams. A deep recurrent neural network (DRNN) model is trained to predict users' FoVs. Users are then clustered according to spatial and content correlations. In this work, the admission and scheduling subproblems are optimized to maximize the average frame quality under latency constraints. [35] employs online reinforcement learning for smart transmission mode selection for VR broadcasting services. The considered network consists of a macro cell, mmWave small cells, and device-to-device (D2D) clusters. Mobile VR users dynamically select their transmission modes based on the online RL scheme that maximizes the system throughput.

Caching popular tiles closer to users reduces the experienced latency and avoids sending the same data multiple times, hence alleviating the burden on core networks. Optimizing caching policies involves answering the questions of what tiles to cache and where. [36] proposes a multi-neural network solution to maximize the cache hit ratio (CHR) of tiled 360° video transmission. Specifically, an LSTM network is used for video popularity predictions and a CNN network is adopted to perform content-based tile classification. A 360° video caching and delivery framework is introduced in [37]. Collaborative transcoding-enabled caching is optimized using multi-agent deep reinforcement learning with the aim of reducing service latency. The assumed network consists of a macro base station (MBS) and multiple small base stations (SBSs) and employs nonorthogonal multiple access (NOMA) for the multicast transmission. In [38], a federated deep reinforcement learning (FDRL) method is adopted to optimize caching and rate adaptation for VR video transmission in hierarchical clustered mobile edge computing networks. An agent is trained to optimize a reward function that incorporates CHR, video quality, quality changes, rebuffering time, as well as bandwidth and transcoding costs where the performance analysis shows that FDRL results in improving the CHR and the user QoE.

V. Conclusion

The need for transmitting VR and 360° videos introduces new network challenges for researchers and network operators. Omnidirectional videos have higher bitrates and tolerate lower latencies than conventional 2D videos. In this context, innovative viewport-based and tile-based streaming methods aim at optimizing the transmission of 360° videos. In this article, we first introduced the features of 360° videos and the accompanying network challenges. We then discussed the DL techniques used for content-based and trajectory-based viewport prediction. Next, a review of DL solutions applied in optimizing 360° transmission has been provided. A range of DL models can be used in viewport prediction and tile-based 360° video streaming. Applying the advanced DL techniques in optimizing the transmission of omnidirectional videos is a promising track for achieving low-cost seamless 360° video delivery.

ACKNOWLEDGEMENT

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] Bastug, E., Bennis, M., Médard, M., & Debbah, M. (2017). Toward interconnected virtual reality: Opportunities, challenges, and enablers. IEEE Communications Magazine, 55(6), 110-117.

[2] He, D., Westphal, C., & Garcia-Luna-Aceves, J. J. (2018, August). Joint rate and fov adaptation in immersive video streaming. In Proceedings of the 2018 Morning Workshop on Virtual Reality and Augmented Reality Network (pp. 27-32).

[3] Chen, Zhenzhong, Yiming Li, and Yingxue Zhang. "Recent advances in omnidirectional video coding for virtual reality: Projection and evaluation." Signal Processing 146 (2018): 66-78.

[4] Yu, L., Tillo, T., & Xiao, J. (2017). QoE-driven dynamic adaptive video streaming strategy with future information. IEEE Transactions on Broadcasting, 63(3), 523-534.

[5] Esakki, G., Panayides, A. S., Jalta, V., & Pattichis, M. S. (2021). Adaptive video encoding for different video codecs. IEEE Access, 9, 68720-68736.

[6] Pang, H., Liu, J., Fan, X., & Sun, L. (2018, June). Toward smart and cooperative edge caching for 5G networks: A deep learning based approach. In 2018 IEEE/ACM 26th International Symposium on Quality of Service (IWQoS) (pp. 1-6). IEEE.

[7] Fu, Sibao, Fan Yang, and Ye Xiao. "AI inspired intelligent resource management in future wireless network." IEEE Access 8 (2020): 22425-22433.

[8] Lecompte, David, and Frédéric Gabin. "Evolved multimedia broadcast/multicast service (eMBMS) in LTE-advanced: Overview and Rel-11 enhancements." IEEE Communications Magazine 50.11 (2012): 68-74.

[9] Shinde, Pramila P., and Seema Shah. "A review of machine learning and deep learning applications." 2018 Fourth international conference

on computing communication control and automation (ICCUBEA). IEEE, 2018.

[10] Zhang, C., Patras, P., & Haddadi, H. (2019). Deep learning in mobile and wireless networking: A survey. IEEE Communications surveys & tutorials, 21(3), 2224-2287.

[11] Wang, Fangxin, et al. "Deep learning for edge computing applications: A state-of-the-art survey." IEEE Access 8 (2020): 58322-58336.

[12] 5G; Extended Reality (XR) in 5G, Standard ETSI TR 126 928 (V16.0.0), ETSI, Tech. Rep., Nov. 2020.

[13] Yu, M., Lakshman, H., & Girod, B. (2015, September). A framework to evaluate omnidirectional video coding schemes. In 2015 IEEE international symposium on mixed and augmented reality (pp. 31-36). IEEE.

[14] Brown, Chip. "Bringing Pixels Front and Center in VR Video." Google, Google, 14 Mar. 2017, https://blog.google/products/google-ar-vr/bringing-pixels-front-and-center-vr-video/.

[15] Benko, H., Wilson, A. D., & Zannier, F. (2014, October). Dyadic projected spatial augmented reality. In Proceedings of the 27th annual ACM symposium on User interface software and technology (pp. 645-655).

[16] Chen, S., Kuzyakov, E., & Peng, R. (2017, April 19). Enhancing high-resolution 360 streaming with view prediction. Engineering at Meta. Retrieved April 2, 2023, from https://engineering.fb.com/2017/04/19/virtual-reality/enhancing-high-resolution-360-streaming-with-view-prediction/

[17] Kuzyakov, Evgeny, and David Pio. "Next-Generation Video Encoding Techniques for 360 Video and VR." Engineering at Meta, 21 Jan. 2016, https://engineering.fb.com/2016/01/21/virtual-reality/next-generation-video-encoding-techniques-for-360-video-and-vr/.

[18] Wiegand, Thomas, et al. "Overview of the H. 264/AVC video coding standard." IEEE Transactions on circuits and systems for video technology 13.7 (2003): 560-576.

[19] Sullivan, Gary J., et al. "Overview of the high efficiency video coding (HEVC) standard." IEEE Transactions on circuits and systems for video technology 22.12 (2012): 1649-1668.

[20] P. de Rivaz and J. Haughton, "Av1 bitstream & decoding process specification." [Online]. Available: https://aomediacodec.github. io/av1-spec/av1-spec.pdf

[21] Bross, Benjamin, et al. "Overview of the versatile video coding (VVC) standard and its applications." IEEE Transactions on Circuits and Systems for Video Technology 31.10 (2021): 3736-3764.

[22] Duanmu, F., Mao, Y., Liu, S., Srinivasan, S., & Wang, Y. (2018, July). A subjective study of viewer navigation behaviors when watching 360-degree videos on computers. In 2018 IEEE International Conference on Multimedia and Expo (ICME) (pp. 1-6). IEEE.

[23] Fan, C. L., Lee, J., Lo, W. C., Huang, C. Y., Chen, K. T., & Hsu, C. H. (2017, June). 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).

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

[25] Xu, Y., Dong, Y., Wu, J., Sun, Z., Shi, Z., Yu, J., & Gao, S. (2018). Gaze prediction in dynamic 360 immersive videos. In proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 5333-5342).

[26] Chao, F. Y., Ozcinar, C., & Smolic, A. (2021, October). Transformer-based Long-Term Viewport Prediction in 360° Video: Scanpath is All You Need. In MMSP (pp. 1-6).

[27] Feng, X., Li, W., & Wei, S. (2021, July). LiveROI: region of interest analysis for viewport prediction in live mobile virtual reality streaming. In Proceedings of the 12th ACM Multimedia Systems Conference (pp. 132-145).

[28] Nguyen, D., Ngan, L., Thuong, L. H., & Huong, T. T. (2022, June). LL-VAS: Adaptation Method for Low-Latency 360-degree Video Streaming over Mobile Networks. In 2022 IEEE Symposium on Computers and Communications (ISCC) (pp. 1-6). IEEE.

[29] Zhao, Lingzhi, et al. "Adaptive streaming of 360 videos with perfect, imperfect, and unknown fov viewing probabilities in wireless networks." IEEE Transactions on Image Processing 30 (2021): 7744-7759.

[30] Jiang, X., Chiang, Y. H., Zhao, Y., & Ji, Y. (2018, October). Plato: Learning-based adaptive streaming of 360-degree videos. In 2018 IEEE 43rd Conference on Local Computer Networks (LCN) (pp. 393-400). IEEE.

[31] Park, S., Hoai, M., Bhattacharya, A., & Das, S. R. (2021). Adaptive streaming of 360-degree videos with reinforcement learning. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (pp. 1839-1848).

[32] Park, S., Bhattacharya, A., Yang, Z., Das, S. R., & Samaras, D. (2021). Mosaic: Advancing user quality of experience in 360-degree video streaming with machine learning. IEEE Transactions on Network and Service Management, 18(1), 1000-1015.

[33] Dasari, M., Bhattacharya, A., Vargas, S., Sahu, P., Balasubramanian, A., & Das, S. R. (2020, July). Streaming 360-degree videos using super-resolution. In IEEE INFOCOM 2020-IEEE Conference on Computer Communications (pp. 1977-1986). IEEE.

[34] Perfecto, C., Elbamby, M. S., Del Ser, J., & Bennis, M. (2020). Taming the latency in multi-user VR 360°: A QoE-aware deep learning-aided multicast framework. IEEE Transactions on Communications, 68(4), 2491-2508.

[35] Feng, L., Yang, Z., Yang, Y., Que, X., & Zhang, K. (2020). Smart mode selection using online reinforcement learning for VR broadband broadcasting in D2D assisted 5G HetNets. IEEE Transactions on Broadcasting, 66(2), 600-611.

[36] Kumar, S., Bhagat, L., & Jin, J. (2022). Multi-neural network based tiled 360° video caching with Mobile Edge Computing. Journal of Network and Computer Applications, 201, 103342.

[37] Xiao, Han, et al. "A transcoding-enabled 360 VR video caching and delivery framework for edge-enhanced next-generation wireless networks." IEEE Journal on Selected Areas in Communications 40.5 (2022): 1615-1631.

[38] Li, Y. (2022). Federated Deep Reinforcement Learning-Based Caching and Bitrate Adaptation for VR Panoramic Video in Clustered MEC Networks. Electronics, 11(23), 3968.