ABSTRACT

Delivery of 360-degree videos poses a challenge to the existing video distribution chains due to the high volume of data to be transmitted. Viewport-aware delivery schemes, e.g. tile-based streaming, constitute a promising approach to reduce the transmitted data volume but require knowledge of the future user viewport to reach its full potential. We discuss a predictive tile-based streaming scenario using head orientation prediction and study the upper limits in terms of throughput savings for different anticipation times and system end-to-end delays. Our results show that using head motion prediction for tile-based streaming systems can bring throughput gains up to 46% compared to a benchmark tile-based streaming system without prediction.

Index Terms— tile-based streaming, head motion prediction, omnidirectional video, 360 video, virtual reality

1. INTRODUCTION

A major fraction of today’s global Internet traffic is caused by video data. Although traditional 2D videos still constitute a large part of this traffic, the emerging immersive media services have already started to deliver omnidirectional (360-degrees) videos to various consumer devices. It is expected that Virtual Reality (VR) and Augmented Reality (AR) traffic will increase 12-fold between 2017 and 2022 globally [1], and a major part of this traffic will be caused by the delivery of high-quality omnidirectional content.

Omnidirectional videos are typically captured by multiple cameras that cover 360 degrees of a scene and rendered commonly through head-mounted displays (HMD) allowing the viewers to look around and freely explore the scene, providing a more interactive and immersive visual experience compared to 2D video. One major limiting factor of the current 360-degree video services is the limited Quality of Experience (QoE) provided to the users as the resolution and thus the visual quality in user’s viewport are not on par with the traditional 2D video services. Delivery of the full 360-video with a sufficiently high resolution puts a major burden on the existing video streaming systems, both in terms of the required network throughput and the decoder capabilities of the available end devices [2]. Therefore, it is crucial to limit the number of pixels transmitted to the user. Considering that only a small portion of the 360-video is presented to the user at once, due to the limited field-of-views (FoV) of HMDs and the human eye\(^1\), a promising idea is to tailor the delivered content according to the user’s viewport and emphasize regions inside the user viewport. Such techniques are commonly known as viewport-dependent approaches. One approach is to encode a distinct version for different user viewports such that the viewport is encoded with a high sample density (i.e. high resolution) whereas the remaining, non-viewport areas are encoded with a lower sample density [3]. This approach requires little change to the existing video streaming chain but may cause a massive storage and encoding overhead for large-scale deployments considering the number of different possible viewports at different bitrates [2]. An alternative approach is to divide the picture into independent, rectangular regions by using HEVC tiles [4] and encode multiple tiles of different resolution (or quality) for each region. In this scheme, the current user viewport is delivered using high-resolution tiles whereas the non-viewport regions are delivered using low-resolution tiles such that the amount of samples to be transmitted is reduced, yet the full 360-content is always available on the end device.

Recently, tile-based streaming has received a lot of attention both in scientific literature and from the industry. Driven by the need and requirements from the industry, various standard development organizations have been working on representation, delivery and other system aspects of immersive media. MPEG developed an application format called Omnidirectional MediA Format (OMAF) [5] which includes a viewport-dependent streaming profile enabling the use of HEVC tiles in HTTP-based adaptive streaming frameworks such as DASH [6]. 3GPP has also specified an operation point and a media profile for enabling tile-based VR streaming in 3GPP ecosystems [7]. An overview of the standardization activities related to the coding and delivery of immersive media can be found in [8] and [9].

Many works in the literature present end-to-end tile-based streaming systems [10–12]. Different approaches to improve the performance of tile-based streaming in terms of bandwidth utilization and user QoE are also considered. Some

\(^1\)Actually the binocular FoV of human eye is about 190\(^\circ\) (horizontal) and 135\(^\circ\) (vertical), considerably larger than the FoVs of the existing HMDs.
works investigate different tiling schemes [13, 14] whereas others propose improvements based on edge computing [15, 16], application layer [17, 18] and cross-layer [19, 20] optimizations. Another crucial optimization domain is the prediction of the future user viewport based on available sensor or historical data and the information extracted from the content. In a tile-based streaming scenario, user movement prediction may reduce the bandwidth requirements by allowing the client to request only the tiles that will actually be visible inside the user’s viewport as well as by prefetching and caching the tiles corresponding to the future viewports when the network conditions are favourable [21]. Fetching only the visible pixels inside the user viewport is a very efficient solution for saving bandwidth; however, it requires very low end-to-end (E2E) latency for VR applications, in which user interactions dynamically determine which part of the omnidirectional content is shown to the user. Despite the advances of the upcoming 5G technology towards ultra low latency in the order of a few milliseconds, such low latency is not expected to be readily available in existing networks in the near future [22]. Therefore, effective prediction techniques are essential in order to accurately request tiles inside the future user viewport and, in adaptive streaming scenarios, determine the required bitrate level for each tile, together with the system parameters like network throughput and buffer level [23].

In this paper, we investigate a predictive tile-based streaming scenario in which the future user viewport can be predicted with absolute certainty a pre-determined time before the onset of the actual movement. We perform our analysis for a range of anticipation times and also analyze the effect of system E2E delay on the performance of our theoretical predictor. Thus, we experimentally obtain the upper limits of the gains obtained by a perfect prediction scheme, as a function of the anticipation time and the E2E delay, in a tile-based adaptive streaming scenario.

This paper is structured as follows: Section 2 briefly summarizes several prediction techniques in the literature. Section 3 describes our theoretical predictor and adaptive streaming system model. Section 4 provides experimental results comparing our method to a state-of-the-art tile-based streaming system. Section 5 concludes the paper.

2. USER MOVEMENT PREDICTION

One important optimization to reduce the bandwidth requirements of the system and improve the user’s QoE is the prediction of the user’s future viewport. Several methods have been proposed in the literature for viewport prediction which rely on different kind of information obtained from i) various sensors, ii) content analysis, iii) statistical analysis (data-driven).

Sensor-based methods can be divided into two categories. Works such as [24–28] are specifically designed for VR applications and use the sensor information from HMD whereas the works in [29–32] focus on inferring user motion based on some physiological data such as EEG and EMG signals. Bao et al. [24] collected motion data in the three dimensions (pitch, yaw, roll) and exploited the auto-correlations in all three dimensions to predict viewer motions using regression techniques. Their findings indicate that a prediction time frame of 100-500 ms is feasible, and they obtain 45% bandwidth reduction compared to streaming the whole 360 content, with a 0.1% failure ratio (defined as the percentage of frames in which the viewers FoV is not completely transmitted). Sanchez et al. [26] analyze the effect of E2E latency on a tile-based streaming system and propose an angular acceleration based prediction method to mitigate its impact on the observed fidelity. Barniv et al. [29] exploit the myoelectric signals obtained from EMG devices to predict the impending movement. They train a neural network to map such signals to trajectory outputs and experiment combining EMG output with inertial (sensor) data. Their findings indicate that an anticipation time of 30-70 ms is achievable with low error rates. Bai et al. [31] attempt to utilize EEG signals to anticipate human movement before it occurs. They report an average prediction time of 0.62±0.25 s; however, they also note that some predictions occur without subsequent actual movements (false positives). Kirchner et al. [32] report that a combination of EEG and EMG can be used to reliably predict movements before a physical movement onset. Their results show that EEG and EMG devices show very different characteristics in terms of the achievable anticipation time. The invention described in [33] is more generic and describes embodiments in which the prediction may occur based on both inertial sensors of an HMD as well as bio-signals from EMG or EEG devices.

Content-based methods such as [23, 34, 35] employ saliency maps which are estimated using viewport trajectories and visual attention models. De Abreu et al. [34] collect the center trajectories of viewports from several users and propose a method to transform the gathered data into saliency maps. Ozcinar et al. [35] create a new visual attention user dataset for omnidirectional videos and analyze the prediction performance of state-of-the-art visual attention models. In their follow-up work [23], they introduce an adaptive omnidirectional video streaming pipeline that optimizes DASH representations of omnidirectional content considering their visual attention maps. Fan et al. [36] employ both sensor and content-related features to train two neural networks which accurately predict user fixations.

Data-driven methods collect user statistics and try to predict the future user orientations by developing statistical models based on historical data. Petrangeli et al. [37] model the viewport evolution of a given user over time as a trajectory and cluster trajectories with similar viewing behaviors to calculate functions for each cluster which are used to predict future viewport at run-time. Liu et al. [16] combine multiple data sources such as the viewing statistics of a video across users, viewing behaviors of multiple videos by a single user.
and other contextual cues to apply prediction-based prefetching of the future chunks. Xu et al. [38] leverage the probability distribution of the user’s orientation, and design an algorithm that prefetches segments by maximizing the expected quality.

3. PROPOSED METHOD

Our method is based on the tile-based streaming system proposed by Skupin et al. [10]. They dynamically adapt the resolution of the 360-video content based on the current user viewport by encoding the video at two resolutions using motion-constrained HEVC tiles at the desired tiling granularity, and render the user viewport using high resolution tiles. Also, non-viewport regions are served in low resolution tiles in order to compensate for latency of the tile updates. Thus, the whole 360-content is always transmitted to the user while emphasizing the viewport using high resolution tiles.

The system in [10] updates the tile selection solely based on the orientation feedback obtained from the sensors in HMD. As an enhancement, we assume in our model that a prediction mechanism is available which allows predicting the future orientation of a user with 100% accuracy, a predetermined amount of time before the actual head movement occurs. The main motivation of our assumption is to analyze the upper bound of the performance that can be attained via such a prediction mechanism (an algorithm or a sensor, e.g. an EMG device). We call the amount of time, for which the future user orientation can be predicted, the Look-ahead Time or LAT. Depending on the capabilities of the employed prediction device, LAT may have different values. For example, according to [29], head motion may be anticipated with an LAT up to 70 ms using an EMG device. Our analysis is motivated by the prediction capabilities of such devices but assumes that a theoretical perfect prediction device is available which provides the desired fine-granular viewport predictions.

An important parameter that effects the performance of a predictor is the E2E delay of the system. The E2E delay affects the latency of the decisions, which may greatly degrade its performance, and thus the reliability of a prediction algorithm. Relying on the orientation predictions, our system aims to request only the tiles that correspond to the future user orientation. However, it may not be possible to determine the user orientation for the next segment of the video to be requested with high reliability, if the E2E delay of the system is high. This may especially be problematic, if LAT of the employed prediction mechanism is not sufficiently high as to deal with such high E2E delay. In order to alleviate such uncertainty, we define a guard band (fallback region) around the predicted user viewport, based on LAT of the predictor, E2E delay of the system and a pre-defined maximum angular velocity. The tiles corresponding to the fallback region are requested in low resolution and kept as fallback content in case the predictor fails to deliver accurate results in a timely manner. Thanks to the prediction of the future viewport, our scheme manages to reduce the extent of the guard bands significantly compared to [10] where the whole non-viewport area is requested as fallback, due to the complete lack of knowledge on the future user viewport.

4. EXPERIMENTS AND RESULTS

The omnidirectional videos in our dataset are first projected onto the faces of a cube using Cube Map Projection (CMP). Each cube face is equally divided into four rectangular areas and a total of 24 HEVC tiles are encoded in two different resolutions, 512×512 and 256×256. The effective resolution of the complete 360-degree video is 3072×2048. In each tile stream, a random access point is inserted with a frequency of 32 frames and each tile stream is encoded in different quality representations for each QP value in [22, 27, 32, 37]. We recorded head orientation traces (per-frame) from 17 subjects for three different video sequences using an Oculus Rift CV1 HMD. The recorded user traces are provided as input to our simulation system.

We simulate a HTTP based adaptive streaming (HAS) scenario. Figure 1 shows an overview of our system. Depending on the viewport predictions, the HAS client determines which tiles will be downloaded in the original resolution, and also utilizes the orientation feedback from HMD (together with the predictions) to determine the extent of the guard bands. The selected tiles are concatenated to a single HEVC bitstream (as described in [10]), decoded, rendered and presented to the user. In our HAS system, we set the segment...
duration to 8 frames (ca. 264 ms at 30 fps) and the buffer size to one segment duration. The viewport size is defined as 90 × 90 degrees. We assume that the user can move with a maximum angular velocity of 200 deg/sec and 100 deg/sec in yaw and pitch directions, respectively. We also assume that the network throughput is sufficient such that even a non-optimized, full delivery scheme as in [10] is able to run smoothly without any interruptions in playback due to buffer underruns.

In our first analysis, we set the E2E delay to a reasonable value (for example 16 frames, ca. 528 ms) and analyze the effect of varying LAT on the rate-distortion (RD) performance of our method. As a benchmark, we compare the performance of our method to the tile-based streaming system proposed in [10] which does not use viewport prediction. Figure 2 shows the average Viewport PSNR (V-PSNR) [39] values attained by our method for different LATs and those from [10] over different bitrates (corresponding to the QP levels [22, 27, 32, 37]). The results show that our hypothetical perfect predictor outperforms the system presented in [10] and its performance improves with increasing LAT. Higher LAT values enable the client to make better long-term decisions and thus provide better quality. It is expected that in cases where a predictor can provide a high LAT, more performance gain can be obtained by employing larger buffers since the client would then be able to prefetch more segments.

Secondly, we investigate the effect of system E2E delay on the performance of our method. Here, we differentiate between an ideal case where the E2E delay is zero, and a more realistic case with varying non-zero E2E delays. This setup may be interpreted as a scenario in which sensors with variable prediction capabilities (in terms of LAT) are compared with each other. Table 1 shows the BD-rate savings relative to [10] obtained for different LATs under the assumption of different E2E delays. Since we assume a buffer length of one segment duration (264 ms), the achieved BD-rate savings saturate if LAT is greater than the E2E delay by a margin larger than one segment duration. The simulation results show that it is possible to obtain up to 46% BD-rate savings compared to the no-prediction method in [10]. We observe that for a given E2E delay, predictors with a higher LAT bring more BD-rate savings. It is also observed that for a given LAT, the BD-rate savings decrease with increasing E2E delay, confirming our expectations.

5. CONCLUSION

In this paper, we propose a hypothetical perfect predictor that accurately predicts a user’s future viewport for a given look-ahead time. We analyzed our predictor for varying look-ahead times by employing user traces obtained through viewing sessions, and compared its performance to a state-of-the-art tile-based streaming system that does not use viewport prediction. We also investigated the effect of end-to-end delay on the performance of our predictor. Simulation results show that i) prediction-based method outperforms the benchmark no-prediction method, ii) increasing look-ahead times bring more gains in terms of viewport PSNR, iii) increasing end-to-end delay reduces the gains obtained through a predictor. In future work, we will jointly investigate the look-ahead times and accuracy of a predictor to deepen our analysis.

**REFERENCES**


