An Optimal Spatial-temporal Smoothness Approach for Tile-based 360-degree Video Streaming

Yixuan Ban[1,3], Lan Xie[1], Zhimin Xu[1], Xinggong Zhang[1,2], Zongming Guo[1,2,*], Yueyu Hu[1]

[1] *Institute of Computer Science & Technology, Peking University, Beijing, China*
[2] *Cooperative Medianet Innovation Center, Shanghai, China*
[3] *Beijing University of Posts and Telecommunications, Beijing, China*
banyixuan@bupt.edu.cn, {xielan, xuzhimin, zhangxg, guozongming, huyy}@pku.edu.cn

***Abstract*—The world is becoming more and more virtual than we ever thought it would be. Many video service providers have rolled out 360-degree videos which provide immersive experience to users. However, huge bandwidth occupation of 360-degree video hinders its wide spreading over the Internet. Besides, only part of the video is displayed on the screen, transmitting whole video results in waste of bandwidth and computational resources. Tile-based adaptive streaming is regarded as a bandwidth-friendly approach which only delivers specific portion of the whole video. It requires the clients to decide which portion and at which bitrates to deliver. However, due to both space and time partition of 360-degree videos in tile-based adaptive streaming, there still exists a challenge on the quality inconsistence on spatial and temporal domains. In this paper, we propose a optimal spatial-temporal smoothness approach under restricted network for tile-based adaptive streaming. The bitrates of tiles are determined optimally, aiming at maximizing the overall quality while minimizing the spatial and temporal quality variation. By conducting extensive experiments over real bandwidth dataset and user's head movement traces, our approach can get a significant improvement. Specifically, the Viewport-PSNR can be raised by 24.1% compared with traditional delivery of whole 360 video; while the spatial and temporal stability can be improved by 40.5% and 24.6% respectively compared with tile-based streaming.**

***Index Terms*—360-degree video, tile-based adaptive streaming, spatial and temporal quality smoothness, QoE-driven optimization**

I. Introduction

The past few years have witnessed the increasing interest on 360-degree video (360 video). Most video websites, such as YouTube, have already provided this service. However, compared with ordinary video, 360 video's high resolution and bitrates demand hinders the wide spread over the Internet. Besides, due to the field of view (FoV) limitation of head mounted displays (HMDs), only part of the video is rendered in the screen one time, which results in inevitable waste of bandwidth and computational resources. Implemented over Dynamic Adaptive Streaming over HTTP (DASH) [1], tile-based viewport adaptive streaming [2] is regarded as a promising way to deliver 360 video through the Internet. As shown in Fig. 1, it spatially crops an original 360 video into several tiles. Then, each tile is cropped into continuous chunks, and encoded at multiple bitrates storing at the server side separately. Once requiring playbacks, the client—running on a HMD or smart phone, conducts *FoV adaptation*, i.e. viewport adaptation, and *rate adaptation* to request the appropriate set of tiles to deliver. Specifically, rate adaptation is designed to adapt to the time-varying bandwidth, choose the proper bitrates of each chunk aiming at providing high quality and continuous video playback; while FoV adaptation is expected to cope with user's head movement and determine which tiles are necessary to be delivered by predicting user's orientation. Ideally, in tile-based adaptive streaming, the bandwidth consumption would be largely reduced as only content within FoV is requested, rather than the whole video.

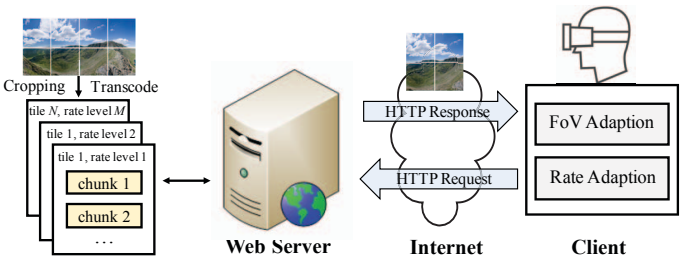


Fig. 1. Overview of Tile-based Streaming of 360-degree Video

According to the existing literature, rate adaptation has been investigated [3] [4]. And for tile-based adaptive 360 video streaming, user's orientation prediction and weight assignment for tiles have been studied [2] [5]. However, despite all this, providing high quality of experience (QoE) of 360 video for users is still challenging. While the benefit of high quality is clear, due to both space and time partition of video content

*Corresponding author. E-mail:guozongming@pku.edu.cn

This work was supported by National Natural Science Foundation of China under contract No. 61471009 and Culture Development Funding under Grant No.2016-288.

978-1-5386-0462-5/17/$31.00 ©2017 IEEE.

in tile-based streaming, the detrimental impact of spatial variability and temporal variability on QoE should also be considered. Indeed, [6] suggests that spatial variability in quality could lead to QoE reduction, while [7] indicates that temporal variability in quality can also result in QoE drop. Three prominent sources for these variability are: 1) time-space varying nature of video content 2) time varying network capacity 3) user's head movement.

In this paper, we propose an optimal temporal-spatial smoothness approach for tile-based adaptive 360 video streaming. Given a total bitrate for chunks (determined by rate adaptation logic) and weights for tiles (according to viewport prediction), the problem is to select tiles to request aiming at maximizing the overall quality while minimizing the spatial and temporal quality variation. While details are presented in the paper, some highlights of our contributions include the followings:

- We identify the QoE model in tile-based adaptive 360 video streaming which includes video quality and its temporal and spacial variability.
- We formulate the FoV adaptation into a QoE-driven optimization problem. By solving this problem, the client can select which tiles to request that maximize the overall quality while minimizing the spatial and temporal quality variation.
- Extensive experiments are conducted over real bandwidth datasets and user's head movement traces. The results demonstrate that our proposed approach can provide high video quality while decreasing the spatial and temporal quality variation.

This paper is organized as follows. Section II explains our approach specifically. The experiment evaluations are detailed in Section III. Finally, we concluded this paper in Section IV.

II. Model System and Problem Formulation

A. Overview of the System

As shown in Fig. 2, in tile-based adaptive streaming, the 360 video sources in equirectangular projection (ERP) format are temporally partitioned into L *chunks* with same duration. For each chunk, it is spatially cropped into N tiles which are indexed in raster-scan order. Then, each tile is encoded at M kinds of bitrates, each of which we denote it as a *fragment*. The client conducts rate adaptation to decide the total bitrate of request fragments; and FoV adaptation to decide which fragments should be requested by predicting the user's orientation and assigning different weights to fragments.

Noticing that even tiles are encoded at same bitrate, the difference of content will still cause spatial and temporal quality variance. More specifically, the mixed-quality among fragments will lead to spatial and temporal variability in quality. Visible blocking artifacts could be severe and decrease user's QoE. To solve this problem, in this paper, we propose an optimal temporal-spatial smoothness approach. By implementing this, the clients can select the best set of fragments to request.

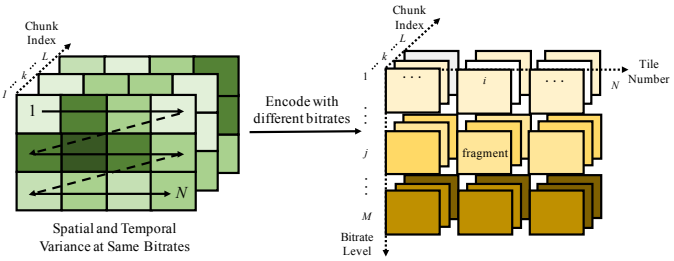


Fig. 2. Partition and encode in tile-based adaptive streaming

B. Proposed Optimization Framework

The notations we used in this paper are introduced as below. Let $k \in \{1...L\}$ denote the index of chunk, $i \in \{1...N\}$ denote the index of tile and $j \in \{1...M\}$ denote bitrate level. We use $r_{i,j}^k$ to denote the bitrate of i-th tile in k-th chunk that encode at j-th bitrate level. To represent video quality, we use $d_{i,j}^k$ representing the distortion of fragment compared with the source 360 video. Due to the fact that only a portion of the video is displayed to user at one time, a weight, i.e. $w_{i,j}^k$, is assigned to each fragment to represent its viewing likelihood [5]. Further, we denote $\boldsymbol{X} = \{x_{i,j}^k\}$ as the results of selection where $x_{i,j}^k = 1$ represents the fragment is selected and $x_{i,j}^k = 0$ otherwise. Noticing that the weight variable has the same value where current viewport covers to prevent it from introducing spatial variability. As for other tiles, the value is diminishing from the viewport margin to the other side of the video sphere to reduce the prediction error's influence.

We define three factors in our QoE optimization problem: 1) expected quality distortion $\Phi(\boldsymbol{X})$, which represents requesting chunk's quality. 2) spatial quality variance $\Psi(\boldsymbol{X})$, which reflects the quality variance among chosen fragments within one chunk. 3) temporal quality variance $\Theta(\boldsymbol{X})$, which refers to the quality variance among continuous chunks in the time sliding window.

To maximize content quality while getting tradeoffs with both spatial and temporal quality variation, we denote η as the weight of spatial quality variance $\Psi(\boldsymbol{X})$ while ζ as the weight of temporal quality variance $\Theta(\boldsymbol{X})$. Therefore, our optimization problem can be formulated as below :

$$\begin{aligned} \min_{\boldsymbol{X}} \quad & \Phi(\boldsymbol{X}) + \eta \cdot \Psi(\boldsymbol{X}) + \zeta \cdot \Theta(\boldsymbol{X}) \\ \text{s.t.} \quad & \sum_{i=1}^{N} \sum_{j=1}^{M} x_{i,j}^k \cdot r_{i,j}^k \le \Re_k \\ & \sum_{j=1}^{M} x_{i,j} \le 1, \quad x_{i,j} \in \{0,1\}, \quad \forall i. \end{aligned} \tag{1}$$

In this formulation, $\boldsymbol{X}$ is the only variable, which represents the selection results for each chunk including the tile order and bitrates. The restrictions above represent the total rates limitation for each chunk and the forbiddance of tile's repeated selection.

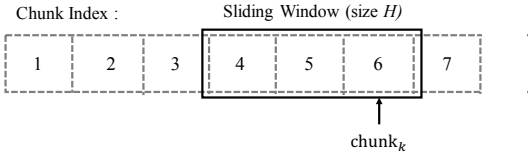


Fig. 3. Temporal Variance in Sliding Window

C. Expected Quality Distortion

In our optimization, we use Mean Squared Error (MSE) on points indicating the distortion $d_{i,j}^k$ compared with source 360 video. Considering tiles content and user's movement's influence, each chunk's expected quality distortion can be formulated as :

$$\Phi_k(\boldsymbol{X}) = \frac{\sum_{i=1}^{N}\sum_{j=1}^{M} d_{i,j}^k \cdot x_{i,j}^k \cdot w_{i,j}^k}{\sum_{i=1}^{N}\sum_{j=1}^{M} x_{i,j}^k}. \tag{2}$$

D. Spatial Quality Variance

Based on expected quality distortion discussed above, the spatial quality variance of each chunk should be :

$$\Psi_k(\boldsymbol{X}) = \frac{\sum_{i=1}^{N}\sum_{j=1}^{M} (d_{i,j}^k \cdot x_{i,j}^k \cdot w_{i,j}^k - \Phi_k(\boldsymbol{X}))^2}{\sum_{i=1}^{N}\sum_{j=1}^{M} x_{i,j}^k}. \tag{3}$$

E. Temporal Quality Variance

To get the temporal quality variance, we denote the sliding window as H chunks' long. As shown in Fig. 3, for chunk$_k$, the temporal variance is as below:

$$\begin{aligned} \overline{\Phi_k}(\boldsymbol{X}) &= \frac{\sum_{n=k-H+1}^{k} \Phi_n(\boldsymbol{X})}{H}, \\ \Theta_k(\boldsymbol{X}) &= \frac{\sum_{n=k-H+1}^{k} (\Phi_n(\boldsymbol{X}) - \overline{\Phi_k}(\boldsymbol{X}))^2}{H}. \end{aligned} \tag{4}$$

In practice, our optimization framework can be designed as a script storing at the client side, which is asynchronous and activated each time the current chunk is downloaded. It can release the computation pressure on servers, which is beneficial to download time, and thus to a continuous playback.

III. Experiment Results

In our experiments, we evaluate various approaches under realistic network bandwidth and users head movement trajectories. We use HSDPA bandwidth dataset [8] which consists of 81 throughput traces and covers a wide range of network conditions. The head movement trajectories are generously provided by AT&T [5] which include 20 traces of 5 users. The same as [5], we use the video Roller Coaster as test video sequence in ERP format with resolution 2880×1440. We spatially partitioned the video stream into 6×12 tiles ($N = 72$) with 240×240. The video sequences are about 180s. We crop it into multiple chunks every one second, which consists of 180 chunks ($L = 180$). To meet practical requirements, each tile is encoded at 5 kinds of rates ($M = 5$), including {20kbps, 50kbps, 100kbps, 200kbps, 300kbps} by open source encoder x264. Further, we calculate the distortion $d_{i,j}^k$ by comparing encoded tiles with raw video. To get the best temporal stability, the sliding window is set to 3 seconds ($H = 3$) in this experiment. Finally, at the client, we apply the buffer-based rate adaptation [4] aiming at providing continuous playback.

Considering the existing works, we evaluate the following algorithms:

- ERP: It regards the 360 video as ordinary video, which is widely used in some video websites such as YouTube.
- Tile-W: The bitrates of different tiles are purely decided according to their weights [9], which means the weighted tiles can always get more rates.
- Tile-OPT1: To evaluate the effect of different terms in our optimization framework, this method removes the objective function of temporal variability in our proposed method.
- Tile-OPT2: This method is our proposed approach, it considers content quality, spatial variability and temporal variability together.

In performance comparison, we take the following measurement metrics into consideration:

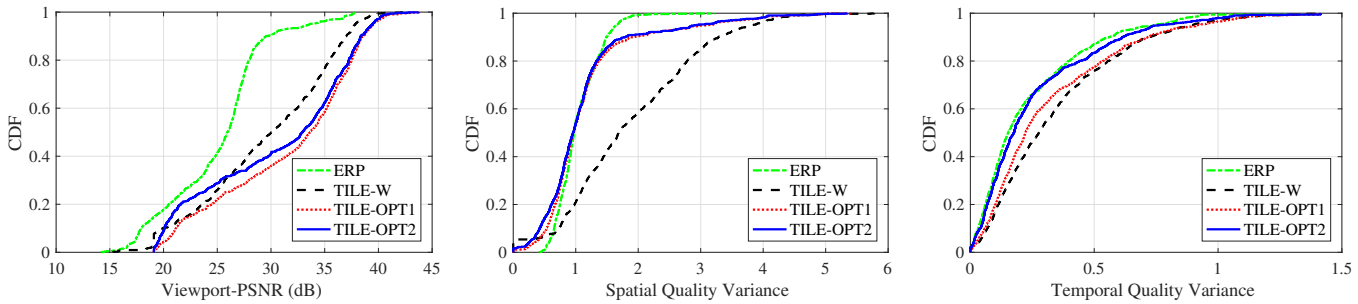
- Viewport-PSNR: The Peak Signal to Noise Ratio (PSNR) in viewport is regarded as an effective measurement of video quality in research fields [10]. It directly evaluates user's experience.
- Bandwidth Utilization: This metric reflects different approaches' ability of utilizing data.
- Spatial Quality Variance: The content quality variance among different tiles in one chunk will lead to decrease of user's QoE notably [6]. We evaluate this performance according to the coefficient of variation(CV) of quality.
- Temporal Quality Variance: The quality variance among different chunks in sliding window will result in QoE drop, which is more serious than the spatial quality variance on fluctuating network conditions. CV is still used as above to evaluate.

A. Bandwidth Utilization and Viewport-PSNR

TABLE I
Performance of PSNR and Bandwidth Utilization

Method	Viewport-PSNR	Bandwidth Utilization
ERP	25.08	74.21%
Tile-W	29.53	78.87%
Tile-OPT1	31.62	97.80%
Tile-OPT2	31.13	97.51%

Fig.4(a) depicts the CDF distribution of viewport-PSNR over different methods. First, because the ERP method has to deliver the whole 360-degree video, the viewport-PSNR is the lowest compared with other tile-based methods. As for the Tile-W method, it has significant improvement on viewport-PSNR over ERP format. However, when compared with Tile-



(a) CDF of Viewport-PSNR (b) CDF of Spatial Quality Variance (c) CDF of Temporal Quality Variance

Fig. 4. Internet Experiment results

OPT methods, the waste caused by weight-driven quantization is obvious. In Tile-OPT1 and Tile-OPT2 methods, they both optimized the video delivery to make the best use of available network resources, improve the Viewport-PSNR. The tradeoff between content quality and temporal stability is reflected by the slightly decrease of Viewport-PSNR in Tile-OPT2.

Table I summarizes average Viewport-PSNR and bandwidth utilization over four methods. The ERP format has the lowest viewport-PSNR and bandwidth utilization. The Tile-W method provides improvement on bandwidth utilization because of the weight-driven allocation approach. However, since the quantization's limitation, the improvement on bandwidth utilization is basically unnoticed. Our proposed optimization approaches Tile-OPT1 and Tile-OPT2 get significant improvement on Viewport-PSNR and bandwidth utilization because of the purpose of maximizing user's QoE. According to the experiment results, in our proposal, the Viewport-PSNR can be raised by 24.1% compared with ERP format, 5.4% compared with Tile-W. As for the bandwidth utilization, our approach can be raised by 31.4% compared with ERP, 23.6% compared with Tile-W. Due to the temporal stability factor of Tile-OPT2, the PSNR and utilization of it is slightly poor than Tile-OPT1 reasonably.

B. Temporal and Spatial Quality Variance

To provide an immersive watching experience, the quality smoothness in space and time domain is of importance. To evaluate further, we compared spatial and temporal quality variance over different methods. According to Fig.4(b), the ERP format is the most spatial stable method because the video content is never partitioned and it is always transmitted as a whole. As for the Tile-W method, with weight-driven allocation algorithm, is fluctuant obviously. The Tile-OPT1 and Tile-OPT2 methods including optimization process are much stable than Tile-W. Especially, the Tile-OPT1 can reduce the spatial variance by 40.5%. Tile-OPTs are all introduced in spatial stability factor, tend to pick fragments more stable on space domain. The spatial stability can also be reflected on the curves' closeness with ERP format. The Fig.4(c) indicates that the ERP method is temporal stable because the server has to deliver the whole content once. The method we proposed has significant improvement compared to Tile-W and Tile-OPT1, which is short of temporal factor. The closeness of curves of ERP and Tile-OPT2 also proves that. Numerically, the promotion can reach up by 24.6% compared with Tile-W.

IV. Conclusions

In this paper, we propose a QoE-driven optimization framework for tile-based adaptive 360 video streaming. Our purpose is to get highest video quality while minimizing the spatial and temporal variability. By conducting extensive experiments over real datasets, we observe that our proposed approach can effectively improves the video quality by 24.1% compared with ERP, and the spatial and temporal stability can be raised by 40.5% and 24.6% compared with weight-driven approach. The results over these metrics demonstrate our contributions and our algorithm's robustness on 360 video delivery.

References

[1] I. J. W13533, "Mpeg dash: The standard for multimedia streaming over the internet," 2012.

[2] J. Le Feuvre and C. Concolato, "Tiled-based adaptive streaming using mpeg-dash," in *Proceedings of the 7th International Conference on Multimedia Systems*. ACM, 2016, p. 41.

[3] S. Akhshabi, A. C. Begen, and C. Dovrolis, "An experimental evaluation of rate-adaptation algorithms in adaptive streaming over http," in *Proceedings of the second annual ACM conference on Multimedia systems*. ACM, 2011, pp. 157–168.

[4] T.-Y. Huang, R. Johari, N. McKeown, M. Trunnell, and M. Watson, "A buffer-based approach to rate adaptation: Evidence from a large video streaming service," *ACM SIGCOMM Computer Communication Review*, vol. 44, no. 4, pp. 187–198, 2015.

[5] 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*. ACM, 2016, pp. 1–6.

[6] H. Wang, V.-T. Nguyen, W. T. Ooi, and M. C. Chan, "Mixing tile resolutions in tiled video: A perceptual quality assessment," in *Proceedings of Network and Operating System Support on Digital Audio and Video Workshop*. ACM, 2014, p. 25.

[7] C. Yim and A. C. Bovik, "Evaluation of temporal variation of video quality in packet loss networks," *Signal Processing: Image Communication*, vol. 26, no. 1, pp. 24–38, 2011.

[8] H. Riiser, P. Vigmostad, C. Griwodz, and P. Halvorsen, "Dataset: Hsdpa-bandwidth logs for mobile http streaming scenarios," 2012.

[9] J. Le Feuvre and C. Concolato, "Tiled-based adaptive streaming using mpeg-dash," in *Proceedings of the 7th International Conference on Multimedia Systems*. ACM, 2016, p. 41.

[10] M. Yu, H. Lakshman, and B. Girod, "A framework to evaluate omnidirectional video coding schemes," in *Mixed and Augmented Reality (ISMAR), 2015 IEEE International Symposium on*. IEEE, 2015, pp. 31–36.