2018 IEEE International Symposium on Multimedia (ISM)

Tile-based Rate Assignment for 360-Degree Video based on Spatio-Temporal Activity Metrics

Robert Skupin, Yago Sanchez, Lei Jiao, Cornelius Hellge, Thomas Schierl
Fraunhofer Heinrich-Hertz-Institute
Berlin, Germany
{*forename.surname*@hhi.fraunhofer.de}

***Abstract*—Tile-based video systems have recently emerged as a viable solution to overcome the challenges of 360-degree video. For instance, the HEVC based viewport-dependent profile of MPEG OMAF allows serving clients independently coded tiles of the 360-degree video at varying resolution to enhance fidelity within the actual user viewport. During streaming, the client constantly adapts its tile selection and feeds a single merged bitstream to the video decoder. This paper addresses the open issue of rate assignment in a distributed encoding system in such a multi-resolution tiled streaming scenario. A model for tile rate assignment based on the spatio-temporal activity of the video is presented to reduce variance of the quality distribution and experimental results are reported.**

***Keywords*—Rate assignment, spatio-temporal activity, tile-based streaming, HEVC, 360-degree video**

I. Introduction

Tile-based video systems have recently emerged as a viable solution to overcome the challenges of 360-degree video, especially the limited resolution of end device decoder. Codec levels limits are typically targeted towards traditional broadcast scenarios, e.g. 4K resolution, which is insufficient in 360-degree video applications where only a small subset of the video plane is actually presented to the user. Representing the 360-degree video in an adequate fidelity, e.g. a resolution of 4K in the user viewport or roughly 20K for the whole sphere, goes well beyond the limits of todays deployed hardware. Therefore, means to efficiently and dynamically distribute the available pixel budget over the video plane are highly disable for 360-degree video streaming. Tile-based video systems allow to quickly adapt the delivered video to the user viewing orientation, e.g. for offering higher fidelity in terms of resolution in the user viewport relative to the remaining areas of the video plane [1].

Although this can also be achieved by encoding multiple variants of a 360-degree video, each of which being adjusted to the viewport in the pixel domain (e.g. using a viewport-dependent projection such as truncated pyramid), such solutions come at a considerable overhead for generation, encoding and caching of each variant. Typically, multiple dozens such variants are necessary to offer sufficient adaptivity to changes of the user viewport [2].

Tile-based video systems allow to reduce the costs of viewport adaptivity by preparing data in a viewport independent way and allowing the client to handle transport of the 360-degree video in a viewport-dependent manner. For instance, the HEVC based viewport-dependent profile of MPEG OMAF allows serving clients independently coded tiles of the 360-degree video at varying resolution to enhance fidelity within the actual user viewport. Multiple combinations of the tiles, each with a different viewport emphasis as through the included resolution variants of each tile, are offered to the client to select between. However, all combinations are built upon the same set of multi-resolution tile encodings, thereby mitigating the overhead of a myriad of viewport-dependent encodings at full video resolution. During streaming, the client constantly adapts its tile selection to ensure the best possible match to the current (or expected) user viewport. After streaming, the tiles are merged using the ISO base media file format to form a conforming HEVC bitstream and fed into a single video decoder instance [3].

However, an open issue in such systems is the selection of suitable bitrates for the encoding of individual video tiles. As a client mixes tiles of, for instance, two resolution variants of a video, a distributed encoding system is typically used. Furthermore, tiles in such a tile-based 360-degree streaming system, are more constrained than general HEVC tiles which are independent only in terms of entropy coding and intra-prediction. A standard conform extension of HEVC tiles is a technique referred to as Motion-Constrained Tile Sets (MCTS) in which the encoding of tiles (or pictures) is also constrained at tile (or picture) boundaries on encoder side with respect to various inter-prediction aspects. Such MCTS allow to recombine tiles without inducing encoder/decoder mismatches and are a prerequisite for tile-based streaming using a single decoder instance. Furthermore, using MCTS allows for straightforward parallelization of the encoding process using completely separate encoder instances per tile. As none of the resulting bitstreams which are ultimately fed to client decoders is the result of a single encoding process, particular care has to be taken to ensure the conformity of merged bitstreams with respect to codec level limits. Also, as typical in bitrate adaptive HTTP streaming, it is desirable to offer multiple different bitrate variants to a client to choose from in order for the client to adapt to its current network conditions such as throughput. Each of these variants is in turn one of multiple tile combinations.

The remainder of the paper is structured as follows. Section II gives an overview on tile-based encoding, while Sect. III introduces the proposed spatio-temporal activity-based rate assignment model. Section IV provides a description of the conducted experiments and results with a conclusion being given in Sect. V.

II. Tile-based Encoding

As evident from the above, rate distribution among tiles in the encoding is a key element for a tile-based 360-degree video service and, at foremost, the rate assignment for each

978-1-5386-6857-3/18/$31.00 ©2018 IEEE
DOI 10.1109/ISM.2018.00019

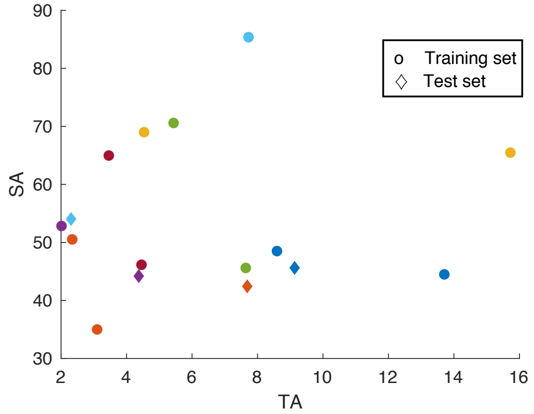

Figure 1: SA and TA content complexity measurements for an exemplary set of sixteen 360-degree video sequences.

tile has to ensure that no tile combination overshoots a given target bitrate B_{Target}. Ideally, a joint rate control algorithm should be used to determine the bitrate of each of the tiles so that $B_{\text{Tile(i)}}$ fulfills:

$$\sum_{i\in s} \mathrm{B_{Tile(i)}} \leq \mathrm{B_{Target}}\ \forall s \in S \qquad (1)$$

where s is one of the possible combinations of tiles at high and low resolution corresponding to one of the potential viewing orientations of the users in the set of all the combinations S.

However, such a rate control would be very complex as a single encoder would need to consider all potential combinations of tiles and make sure that the sum of bitrate of the tiles in each of combinations fulfills the constraint described in Eq. (1). In addition, doing so would prevent from one of the benefits of tile-based streaming, which is the parallelization at content preparation.

A more suitable and straightforward approach, referred to as uniform rate model, is to assign rates using a uniform bitrate distribution among tiles, in which each tile at a given resolution is assigned the same portion of available rate as, for instance, in [4]. As the tiling setup envisioned in this work builds upon tiles of varying resolution, i.e. tile of varying spatial dimensions, the bitrate limit B_r assigned to each tile resolution would follow Eq. (2).

$$\sum_{r} \mathrm{N_r} * \mathrm{B_r} = \mathrm{B_{Target}} \mid \frac{\mathrm{B_i}}{\mathrm{B_j}} = \mathrm{R_{ij}} \qquad (2)$$

where r is an index identifying one of the tile resolutions available and N_r is the number of tiles at resolution r that are contained within each of the potential combinations. R_{ij} is a ratio that represents the bitrate ratio of tiles when encoded with similar quality but at different resolutions. The notion of equal quality of encoded video within this paper is that the same QP is used for all sample blocks or coding units (CUs) in case of HEVC. The drawback of the uniform rate model following Eq. (2) is that it is assumed that all tiles at a given resolution are equally complex and require the same bitrate to achieve similar fidelity. However, this is not a valid assumption as content complexity and, hence, coding performance considerably varies locally in reality.

In order to determine the complexity of each of the tiles, spatio-temporal activity metrics have been used. In [5], the metrics spatial perceptual information or Spatial Activity (SA) and temporal perceptual information or Temporal Activity (TA) are introduced to help selecting sets of test sequences in order for them to span a largest possible portion of the spatio-temporal information plane and ensure that sequences of different complexity are selected. The spatio-temporal complexity or activity metrics SA and TA used within this paper follow the work done in [6], where instead of using max values as in [5], mean values for each sequence are used as shown in Eq. (3) and Eq. (4).

$$SA = mean\{std(Sobel[F_n])\} \qquad (3)$$

$$TA = mean\{std([F_n - F_{n-1}])\} \qquad (4)$$

where F_n is the frame number n within a given video sequence.

Figure 1 gives measurements of SA and TA of the 360-degree video sequence set, which the experiments in this paper are built upon. Circles represent training sequences while diamonds represent test sequences. It can be seen that the sequences cover a wide range of the SA and TA plane. Given that sequences are intended to be encoded tile-wise, Figure 2 illustrates that SA and TA of individual tiles of a given video sequence may themselves span across considerable value ranges, i.e. content complexity varies immensely on a local scale, entirely depending on the content characteristics. This contributes to the fact that a rate assignment following Eq. (2) may yield varying quality across the coded video picture plane. The applied QP and its distribution over CUs allows to measure this quality variance across the picture plane.

Figure 3 provides a histogram of QP per coding unit of all frames of one of the tile combinations of one sequence (Seq1) of the test set, when encoded with a uniform rate distribution following Eq. (2). It is apparent from the figure, that the quality is distributed very unevenly across the resulting picture plane. Minimizing the variance of the QP distribution through an improved rate distribution is the optimization target in this work and will serve as metric to evaluate the proposed model.

III. Spatio-Temporal Activity Based Rate Assignment

The basic idea presented in this paper is to facilitate complexity metrics for a rate assignment model, referred to as the activity-based model, that on one hand ensures that all tiles adhere to the constraint formulated in Eq. (1) and, on the other

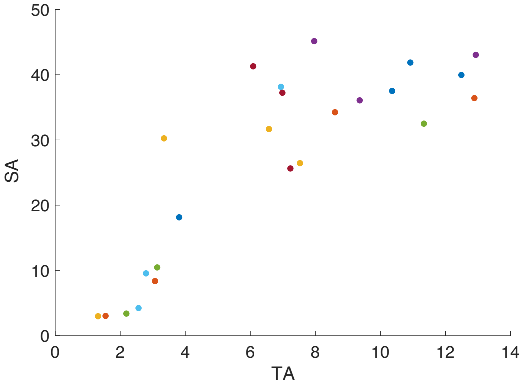

Figure 2: SA and TA measurements of 24 individual tiles of test video sequence Seq1.

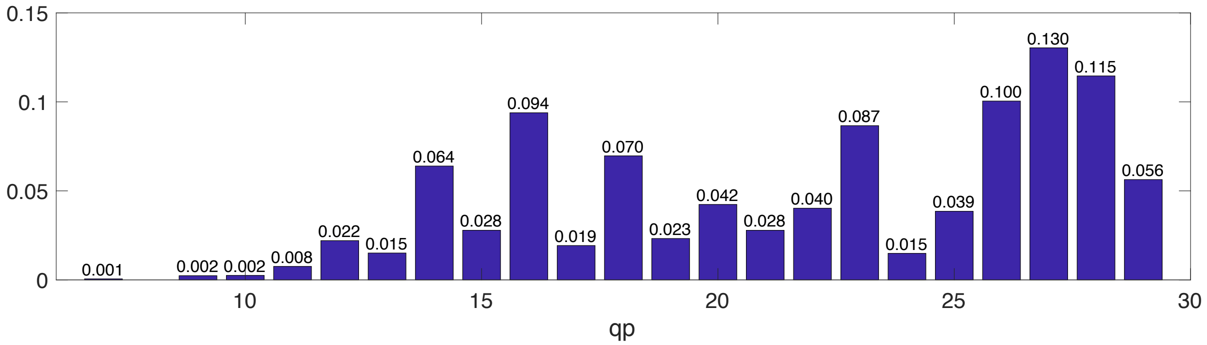

Figure 3: Histograms of QP over the picture plane of a combination of tiles from video test sequence Seq1 when encoded with a uniform rate model.

hand, result in a more even quality distribution among tiles of a video sequence, i.e. low variance in the QP distribution.

Spatio-temporal activity metrics such as SA and TA have been used for different purposes related to video compression in the past. For instance, the authors in [6] developed a model based on spatio-temporal activity metrics to derive the bitrate of an encoded video depending on the QP and frame rate, so that it can be used in context of rate-control.

Therefore, in our work, we also use SA and TA as defined in [6], since they have been proved to be suitable for bitrate estimation. The idea is to use a non-linear regression model to approximate a rate assignment function for tiles of a video sequence, when encoded with the same QP.

Since the model needs to be used for different values of B_{Target}, a model is used to derive BP_i, which is the bitrate percentage of a tile i over the overall bitrate of all tiles, also referred to as normalized bitrate within this paper and shown in Eq. (5).

$$\mathrm{BP_i} = \frac{\mathrm{Bitrate_{tile}(i,QP_{\mathit{init}})}}{\sum_{t=0}^{maxTiles-1}\mathrm{Bitrate_{tile}(t,QP_{\mathit{init}})}} \qquad (5)$$

where $\mathrm{Bitrate_{tile}(i,QP_{init})}$ is the bitrate of tile i when encoded with a QP equal to $\mathrm{QP_{init}}$ and maxTiles being the number of tiles into which the whole 360-degree video is split.

The data used for fitting the activity-based model stems from constant quality tile-based encodings of a large video sequence set, referred to as training set, and the respective complexity metric measurements of SA and TA per tile.

For encoding of the training set, a constant QP encoder operation mode was chosen, using a $\mathrm{QP_{init}}$ equal to 22 and the high-resolution variant of all tiles. The non-linear regression was then performed using the following exponential mode to model $\mathrm{Bitrate_{tile}(i,QP_{init})}$ as a function of SA and TA as shown in Eq. (6). The model for $\mathrm{Bitrate_{tile}(i,QP_{init})}$ has been derived using the high-resolution version of the training set sequence encodings following:

$$\mathrm{Bitrate'_{tile}(SA,TA)} = \alpha * SA^{\beta} + \gamma * TA^{\delta} + \varepsilon \qquad (6)$$

Varying $\mathrm{QP_{init}}$ for performing the regression did not show significant impact. Once the model $\mathrm{Bitrate_{tile}(i,QP_{init})}$ is derived for the training set, it can be used for the test set with Eq. (5). First, the tile set combination that leads to the maximum bitrate is identified as follows:

$$\mathrm{s_{max}} = \underset{\mathrm{s}}{\arg\max} \sum_{j\in s} \frac{\mathrm{BP_j(SA,TA)}}{R_j} \qquad (7)$$

where $\mathrm{R_j}$ corresponds to:

$$\mathrm{R_j} = \begin{cases} 1 & if \;\; res(j) = N_{res} - 1 \\ \mathrm{R}_{N_{res}-1,\mathrm{res(j)}} \; \big(Eq.(2)\big) & otherwise \end{cases} \qquad (8)$$

with res(j) → [0, …, $\mathrm{N_{res}}$-1] being a function that determines the resolution of tile j within the combination s and $\mathrm{N_{res}}$-1 corresponding to the highest resolution available.

Note that the Eq. (7) leads to a value below 1 since some of the tiles in the set $\mathrm{s_{max}}$ correspond to low-resolution tiles. Thus, the bitrate of each individual tile at high resolution that fulfills the requirement in Eq. (1) is computed as:

$$\mathrm{B_{Tile(i)}(SA,TA)} = \frac{\mathrm{BP_i(SA,TA)}}{\sum_{j\in s_{max}} \frac{\mathrm{BP_j(SA,TA)}}{R_j}} * \mathrm{B_{Target}} \qquad (9)$$

with $\mathrm{s_{max}}$ being derived from Eq. (7). In order to compute the bitrate limits for other resolutions different to the highest one $\mathrm{B_{Tile(i)}(SA,TA)}$ is divided by $\mathrm{R}_{\mathrm{N}_{res}-1,\mathrm{res}}$ where res is the target resolution.

IV. Experimental Evaluation

For the experimental evaluation, the training set used to estimate the parameters of Eq. (6) as well as $\mathrm{R_{ij}}$ of Eq. (2) encompassed 12 video sequences with various frame rates between 25 and 60 and durations between 10 to 30 seconds.

Table 1: Parameter values

Parameter	Value	Description
$\mathrm{N_{res}}$	2	Number of resolutions
$\mathrm{R_{10}}$	2.28	Ratio from bitrates of high resolution to low resolution
$\mathrm{N_0}$	16	Number of low resolution tiles in one tile combination
$\mathrm{N_1}$	8	Number of high resolution tiles in one tile combination
maxTiles	24	Number of tiles in which the videos are divided
$\mathrm{B_{Target}}$	16 Mbps	Target bitrate limit for any tile combination

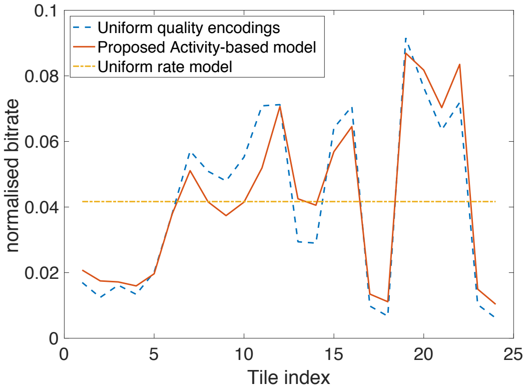

Figure 4: Normalized bitrates over tiles of Seq1.

As apparent from Figure 1, the spatio-temporal complexity of the training set sequences, marked with circles, is diverse. The activity-based model is evaluated using four randomly selected video sequences, referred to as the test set and which are not part of the training set. The spatio-temporal activity of the test set is also visualized in Figure 1 and marked with diamonds. In particular, for the sequence Seq1 of the test set, a tile-wise analysis of spatio-temporal activity is given in Figure 2.

The tile-based streaming approach considered here is described in more depth in [1]. The sequences in the training set and test set use the Cubemap Projection (CMP) to represent the 360-degree sphere. All sequences are sampled to two resolutions and tiled at a 2x2 tile granularity per cube face. Therefore, the total number of tiles (maxTiles) is equal to 24. The parameter values considered in the paper are summarized in the Table 1.

An important benchmark are the actual rate distributions that stem from actual constant QP encodings of the test set sequences. Figure 4 reports the resulting normalized bitrate BP_i of the 24 tiles for encoding using the uniform rate model in Eq. (2), the proposed activity-based model in Eq. (6) and encodings based on a single QP value, i.e. the target of uniform quality. It can be seen that although there are small differences between the activity-based model and the uniform quality encodings, the error is much smaller than for the uniform rate model encodings.

In addition to evaluating the performance of the proposed model, it is also important to assess whether the quality variability is reduced in terms of chosen QPs compared to the uniform rate model. Figure 6 shows the histogram of QP per coding unit for the proposed method. The same tile combination of sequence Seq1 as in Figure 3 for the uniform rate model has been used. It can be seen that for that exemplary combination of tiles, the bitrate distribution has a smaller variance with QPs between 18 and 27 in comparison to QPs between 7 and 29.

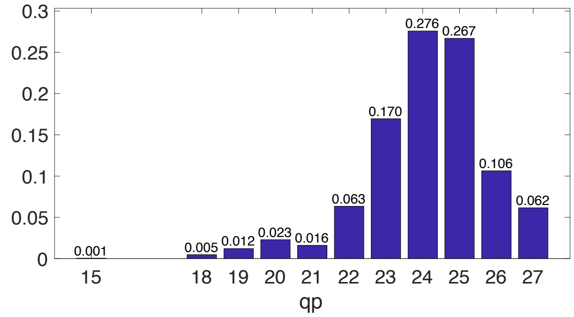

Figure 6: QP distribution of a combination of tiles from Seq1 when using the activity-based model.

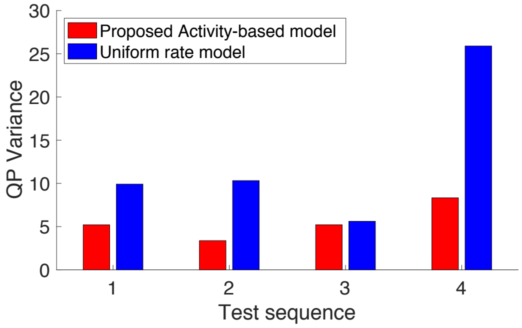

Figure 5: QP distribution variance.

Figure 5 shows the variance for the four test sequences in the test set for both the uniform rate approach and the activity-based approach. It can be seen that the activity-based approach outperforms the uniform rate approach halving the variance of the QP distributions of the tiles or even reducing it to a third. For sequence number 3 the performance of both methods is similar but for the other three sequences there is a great improvement when the proposed model based on spatio-temporal activity metrics is used.

V. Conclusion

The proposed activity-based model for rate assignments in context of tiled encoding of 360-degree video results in a reduced variance of the QP distribution when compared to a uniform rate approach. The bitrates derived from such a model can, for instance, be helpful in 360-degree streaming services that employs a capped variable bitrate encoding strategy. Future work targets to extend the model towards rate-controlled encoding strategies.

References

[1] R. Skupin, Y. Sánchez, C. Hellge, T. Schierl. "Tile based HEVC video for head mounted displays." Multimedia (ISM), 2016 IEEE International Symposium on. IEEE, 2016.

[2] E. Kuzyakov, D. Pio , “Next-generation video encoding techniques for 360 video and VR”, online:https://code.fb.com/virtual-reality/next-generation-video-encoding-techniques-for-360-video-and-vr/

[3] Skupin, R., Sanchez, Y., Wang, Y. K., Hannuksela, M. M., Boyce, J., & Wien, M. (2017, December). Standardization status of 360 degree video coding and delivery. In *Visual Communications and Image Processing (VCIP), 2017 IEEE*(pp. 1-4). IEEE.

[4] Inoue, Masayuki, et al. "Interactive panoramic video streaming system over restricted bandwidth network." Proceedings of the 18th ACM international conference on Multimedia. ACM, 2010.

[5] ITU-T Rec. P.910 (04/2008) Subjective video quality assessment methods for multimedia applications. pages 1–42, 2009.

[6] Lottermann, Christian, et al. "Bit rate estimation for H. 264/AVC video encoding based on temporal and spatial activities." Image Processing (ICIP), 2014 IEEE International Conference on. IEEE, 2014.