Video Quality Evaluation for Tile-Based Spatial Adaptation

Hiba Yousef, Jean Le Feuvre
LTCI
Telecom ParisTech
Paris, France
{hiba.yousef, jean.lefeuvre}
@telecom-paristech.fr

Giuseppe Valenzise
L2S,CNRS
CentraleSupèlec
Université Paris-Sud
Paris, France
giuseppe.valenzise@l2s.centralesupelec.fr

Vedad Hulusic
Dep of Creative Technology
Fac of Science and Technology
Bournemouth University
Poole, UK
vhulusic@bournemouth.ac.uk

***Abstract*—The demand for very high-resolution video content in entertainment services (4K, 8K, panoramic, 360 VR) puts an increasing load on the distribution network. In order to reduce the network usage in existing delivery infrastructure for such services while keeping a good quality of experience, dynamic spatial video adaptation at the client side is seen as a key feature, and is actively investigated by academics and industrials. However, the impact of spatial adaptation on quality perception is not clear. In this paper, we propose a methodology for the evaluation of such adapted content, conduct a series of perceived quality measurements and discuss results showing potential benefits and drawbacks of the technique. Based on our results, we also propose a signaling mechanism in MPEG-DASH to assist the client in its spatial adaptation logic.**

***Index Terms*—Spatial-Based adaptation, HTTP adaptive streaming, Video quality evaluation, HEVC tiling.**

I. Introduction

Super high resolution and immersive video are expected to significantly increase the bandwidth consumption in video streaming services. In some cases, such as for 360 and panoramic video, only portions (viewports) of the video are displayed to the user at a given time. In those applications, spatially adapting the quality of the video, e.g., reducing the quality of those parts of the content which are outside the region of interest (ROI), is a key tool to achieve a more efficient use of network resources.

In video coding, spatial quality adaptation is typically implemented by bit allocation algorithms. They can operate at different levels of granularity, from group-of-pictures to the coding unit or macroblock level. While this flexibility enables to achieve good coding gains, it might be too complex and even impractical in the context of adaptive video streaming targeting a large number of viewers. Indeed, in order to address all the users and their viewing conditions, the provider should encode a very large number of versions of the content, one per possible active combination of viewing criteria, resulting in overloading the server and/or CDN capacity. Another approach for spatial quality adaptation, discussed in this work and, currently, in MPEG-DASH, is to divide the video into a grid of N×M videos (either independent or using schemes such as HEVC tiles [1]), encoded at different qualities, and let the client decide which video to pick based on its current viewing conditions.

When mixing tiles at different qualities, an important question is *how the quality difference between tiles may affect the overall visual quality*. In fact, boosting the quality of the tile containing the ROI might introduce artificial horizontal or vertical edges, due to the quality gap with neighboring tiles. This might in turn distract the visual attention and affect quality judgment [2]. In this paper, we study how overall video quality is affected by spatial quality adaptation in tiled video, by means of a subjective study on five 4K video contents. Specifically, for each content we encode tiles at two different bitrates: a higher bitrate for the ROI (identified in a preliminary visual attention test), and a lower one for the background. We then compare the visual quality of the video, measured through Mean Opinion Score (MOS), to a uniform tiling scheme (i.e., where all tiles have the same bitrate), for the same total bitrate. Our results show that non-uniform allocation might lead to important quality gains, mostly at low-to-medium bitrates. However, the gains highly depend on the semantic and perceptual masking properties of the content. Based on these results, we propose a new MPEG-DASH descriptor to signal which quality differences between ROI/non-ROI tiles are acceptable for a given content.

The rest of this paper is organized as follows. Section II presents related work in the domain. Section III presents the methodology used in video quality evaluation and describes the conducted experiment. Section IV presents the results of the experiments and discusses their impact on the adaptive streaming client and related options proposed for MPEG-DASH. Finally, Section V concludes this paper and proposes future work.

II. Related Work

Coding video with spatially varying quality has been largely studied in the past decades. Examples include

sprite coding [3], MPEG-4 Visual Objects [4], [5] and other more recent ROI-based bit allocation schemes for H.264/AVC and H.265/HEVC [6], [7], [8], [9], [10], [11]. In most cases, ROI-based allocation targets specific applications scenarios, e.g., assigning higher quality to faces in teleconferencing [8], [10]. Previous work demonstrates that a proper rate allocation strategy might lead to substantial (albeit content-dependent [12]) gains in terms of perceived visual quality. In [9] these gains were shown to be more important at low bitrates; however, their results were found on mobile video, and it is not clear how this conclusion extends to high-resolution content.

Our work relies on this previous art, but considers a different perspective: we consider the more general case of content with ROI bringing any semantics, and in the particular case of streamed tiled video [13]. Instead of defining at the encoding time the ROI and adjust its quality, we focus on using uniform quality at the encoder side, unaware of any ROI/spatial quality adaptation, and mix the qualities at the receiver side, using the HEVC tiles mechanism which enables parallel processing of different parts (Tiles) within the same picture [14]. We chose the HEVC Tiling approach for simplicity and interoperability reasons, since all the existing HEVC profiles include Tiles. For the same reason, tile-based streaming has been recently studied for 360° video [15] and panoramic streaming [16]. Tiling approach, on the other hand, imposes a number of constraints, e.g., the tiling scheme is much simpler and has coarser-granularity compared to previously proposed methods; in addition, quality adaptation is not a part of the rate-distortion optimization such as previously proposed ROI-based coding schemes, which implies that only a small set of different representations of the video can be stored and streamed to the user. Finally, the large spatial extent of tiles introduces stronger, more detectable visual artifacts (horizontal and vertical tile edges) compared to a fine-granularity bit allocation approach. The purpose of this work is to evaluate the impact of such artifacts and to understand to which extent and under which conditions adaptive bit allocation in tiling is advantageous.

III. Video Quality Evaluation Experiment

In this section, we describe the subjective experiment we conducted in order to assess the impact of spatial adaptation in tiled video streaming.

1) Material Selection: For the study, six 4K sequences from the Joint Video Exploration Team (JVET) dataset[1] were used. Example frames for each sequence are reported in Figure 1. The sequences were selected based on their temporal and spatial information [17], see Figure 2. To span the largest possible range of the spatio-temporal plane, *ToddleFountain*, *Tango*, *Drums* and *CatRobot* were chosen. *DaylightRoad* and *TrafficFlow* were close on SI-TI plane, so *TrafficFlow* was chosen as the fifth test sequence since it is the only sequence that does not have a specific ROI according to the visual attention test. The *DaylightRoad* was kept as a training sequence.

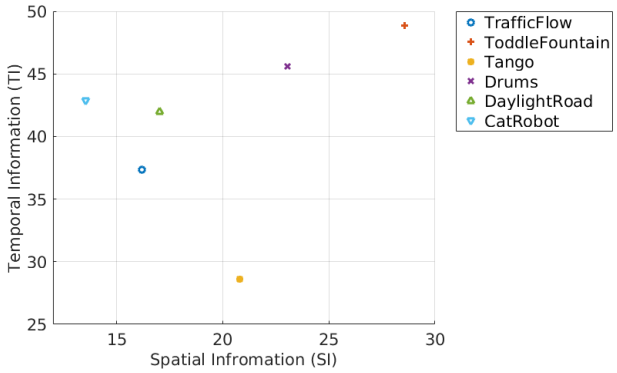

Fig. 2. Spatial and Temporal Information for the test sequences

A. Identification of Regions of Interest

In order to spatially adapt the bitrate across tiles, we first need to detect the region(s) of interest for each scene. To this end, we employed an eye tracker and recorded eye fixations using the *GazeRecorder*[2] software. The visual attention (VA) experiment was conducted with 10 participants (5F/5M) with normal or corrected-to-normal vision. They were asked to watch the sequences naturally, i.e., without a specific task. Based on the the VA results, the test sequences can be classified into three types: a) content with a nearly static ROI: *Drums*, *CatRobot*, *DaylightRoad*, b) content with a dynamic ROI: *Tango*, *ToddlerFountain* and c) content without a clearly identifiable ROI: *TrafficFlow*.

B. Design

We used the double stimulus impairment scale (DSIS) methodology for the test [17], [18], [19]. The subjects were presented once with a pair of stimuli, and had to evaluate

[1]https://mpeg.chiariglione.org/standardsexploration/future-video-coding/n16512-work-plan-assessment-test-material

[2]https://sourceforge.net/projects/gazerecorder/

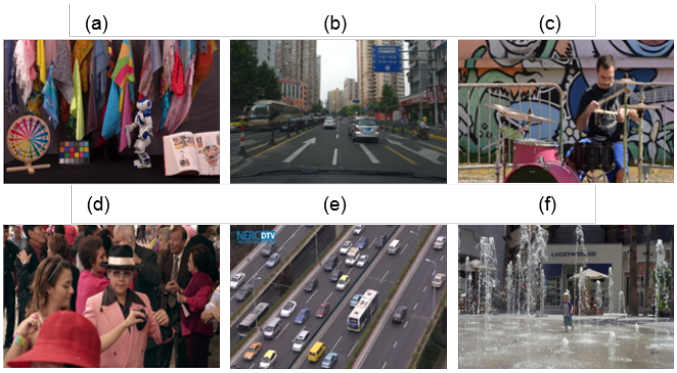

Fig. 1. Example frames of the selected sequences; (a) CatRobot, (b) DaylightRoad, (c) Drums, (d) Tango, (e) TrafficFlow, (f) ToddlerFountain

the impairment of the test stimulus with respect to the reference, using a 5-grades continuous degradation scale. The associated terms "Imperceptible", "Perceptible but not annoying", "Slightly annoying", "Annoying" and "Very annoying" are included for general guidance. Hereafter we will refer to these qualities on the impairment scale as $Q_m = \{Q_1, Q_2, Q_3, Q_4, Q_5\}$, Q_1 being "Imperceptible". We performed a selection between more than 1500 encoded sequences including uniform and non-uniform tiling grid, mixing tiles with the same and different qualities for the 6 input sequences. Overall, 130 video scenes in a random sequence order, all lasting for 8 seconds, were evaluated in the study by each participant.

1) Participants: 15 participants (7F/8M), naive to the purpose of the experiment, aged between 20 and 50, volunteered for the experiment. All of them reported normal or corrected-to-normal visual acuity.

2) Apparatus: The subjective test was performed in a dark and quiet test room dedicated to subjective visual tests [20]. The display used in the subjective test was a commercial 64.5" 4K TV screen (TV SONY KD-65XD7505) at 3840×2160 pixels resolution, with the peak luminance of 350 cd/m^2. The distance from the screen was fixed to 1.5 height of the display, with the eyes located at the center of the display.

3) Stimuli: In the main study, the same video sequences as for the VA test were used. In total, for each video content, 26 test stimuli were used for the test (5 without tiling, 5 with uniform tiling, and 16 tiled videos with spatial adaptation), as explained below.

For each original video, we used the *Kvazaar*[3] open source HEVC codec to encode 200 frames into 40 different qualities. The encoder was set to produce a uniform tile grid of $T = 3 \times 3$ tiles. Motion vectors constraints are taken into account as suggested in [13]. Five different qualities Q_m, $m = 1, \ldots, 5$, with Q_5 being the highest quality level, corresponding to a uniform bitrate allocation across tiles $BR_{m,c}$ (where c denotes video content), were chosen visually by 5 expert viewers to match the 5-grades of the continuous impairment scale. Then, we adapted each quality $\{Q_1, Q_2, Q_3, Q_4\}$ with 4 different adaptations (Q_5 was not considered since the quality of all tiles was already high enough and the artifacts were imperceptible).

To this end, we first identify the N_{ROI} tiles that contain the ROI. For each video, these tiles are coded at the same bitrate, BR_k^{ROI}, where $k = 1, \ldots, 4$, denotes the adaptation level, in order to avoid quality discontinuities across the ROI. Conversely, the $N_{\overline{\text{ROI}}} = T - N_{\text{ROI}}$ tiles that do *not* contain the ROI are coded with a bitrate $BR_k^{\overline{\text{ROI}}}$. The bitrate for the ROI tiles is found by increasing the bitrate of uniform allocation at a given quality level, i.e., $BR_k^{\text{ROI}} = \delta_k \cdot BR_{m,c}$, with $\delta_k > 1$. The bitrate for non-ROI tiles, $BR_k^{\overline{\text{ROI}}}$, is then found by ensuring that the total bitrate for each adaptation setting equals the bitrate of uniform allocation, by solving the following equation:

$$N_{\text{ROI}} \cdot BR_k^{\text{ROI}} + N_{\overline{\text{ROI}}} \cdot BR_k^{\overline{\text{ROI}}} = T \cdot BR_{m,c}. \quad (1)$$

The division between ROI and non-ROI tiles might create artificial edges, that we name *Quality-Switch Edges (QSE)*. The number of ROI tiles, ROI tile indices (in raster-scan order, starting at index 1 for top-left tile) and QSEs for each content are given in Table I.

TABLE I
THE NUMBER OF ROI TILES, ROI TILE INDICES AND QSES FOR THE TEST SEQUENCES. THE MAXIMUM NUMBER OF QSE IS 12, SINCE WE USE A 3×3 TILING CONFIGURATION.

Scene	N_{ROI}	**ROI tile indices**	**QSEs**
CatRobot	4	4,5,7,9	7/12
Drums	4	2,3,5,6	4/12
Tango	6	1,2,4,5,7,8	3/12
Toddler	2	4,5	5/12
Traffic[4]	3	4,5,6	6/12

It remains to find the values of δ_k to determine the four adaptations for each content. We adopted the following strategy. A panel of 4 expert viewers inspected videos coded with different tiles mixed at different bitrates, in order to determine the maximum bitrate increment factor, $\delta_4 = \delta_{\max}$, for each content and quality level. $\delta_{\max}$ was chosen in such a way to avoid having different tiles in a frame clearly corresponding to different attributes on the impairement scale, as this might hinder the use of the scale in the test. In practice, we found that $\delta_{\max}$ is the largest value past which QSE visibility becomes the dominant artifact in the quality assessment procedure. Once δ_4 has been found, the remaining values of δ_k, $k = 1, \ldots, 3$, are set to $\delta_k = \dfrac{\delta_{\max} - 1}{4}$.

4) Procedure: All videos were evaluated by each participant. The test was divided into three sessions to keep the duration of each session below 20 minutes. The test started with a training phase using the *DaylightRoad* sequence. The reference stimulus was the sequence encoded in a high bitrate without tiling; the test stimuli were the tiled and the non-tiled version with 5 different qualities of the sequence. Each subject observed the reference and then the test stimuli once for 8 seconds each and had 5 seconds to evaluate the overall quality through quality-rating forms.

IV. Results and Discussion

To analyze the results, we first computed the mean opinion scores (MOS) with the corresponding 95% confidence intervals. The values of MOS for the different spatial adaptation bitrate mixing (indicated by δ_k), as well as the MOS of uniform tiling and no-tiling configurations, are

[3]http://ultravideo.cs.tut.fi/# encoder

[4]For the Traffic scene, no particular ROI was found in the VA test, so an horizontal tile line was selected to reduce intersections of tile borders with the content.

TABLE II
TOTAL BITRATE PER SCENE [*Mbps*] AND ROI BITRATE INCREMENT FACTORS δ_k.

Scene	**BR1**	**BR2**	**BR3**	**BR4**
CatRobot (S1)	1.5	2.3	3	4.5
δ_1	1.05	1.1	1.125	1.1875
δ_2	1.1	1.2	1.25	1.375
δ_3	1.15	1.3	1.375	1.5625
δ_4	1.2	1.4	1.5	1.75
Drums (S2)	2.7	3.2	4.2	6
δ_1	1.125	1.15	1.15	1.1875
δ_2	1.25	1.3	1.3	1.375
δ_3	1.375	1.45	1.45	1.5625
δ_4	1.5	1.6	1.6	1.75
Tango (S3)	2	2.5	3	4
δ_1	1.1	1.075	1.1	1.15
δ_2	1.2	1.15	1.2	1.3
δ_3	1.3	1.225	1.3	1.45
δ_4	1.4	1.3	1.4	1.6
Toddler (S4)	10	20	30	35
δ_1	1.5	1.65	1.6	1.6
δ_2	2	2.3	2.2	2.2
δ_3	2.5	2.95	2.8	2.8
δ_4	3	3.6	3.4	3.4
Traffic (S5)	1.4	1.5	1.7	3
δ_1	1.0625	1.1	1.1	1.25
δ_2	1.125	1.2	1.2	1.5
δ_3	1.1875	1.3	1.3	1.75
δ_4	1.25	1.4	1.4	2

reported in Figure 3, for different bitrates corresponding to quality levels Q_1 to Q_4. Figure 4 aggregates these results by picking the best bitrate mix (i.e., the δ_k yielding the highest MOS) across bitrates.

A. Effect of bitrate

At lower bitrates, we can see that the spatial adaptation gives better result than the uniform tiling grid for content with ROI, confirming that our approach can be used in adaptive streaming to provide better quality at the client side. Furthermore, it seems that configurations with higher QSE (*CatRobot*, *Traffic*) have a lesser perceived quality than configurations with lower QSE (*Drums*, *Tango*, *Toddler*). We plan on further looking at that possible correlation in future work. From the results, we can see that higher potential gains for spatial adaptations are found at low-to-medium bitrates, the artifact introduced by QSE being more visible than other at such rates. It is important to note that no post-processing other than HEVC standard one was applied on the content after decoding, in order to remain compatible with existing HEVC deployments. This implies that no deblocking filter was applied on QSE, making these more visible at higher rates. This shows the major complexity of HEVC tiling spatial adaptation: encoders have to ensure that tiles borders are not more visible than other coding artifacts, to avoid degrading the perceived quality of the (uniform tiled) content.

B. Presence of ROI

As expected, we also found out that the sequence without ROI (*Traffic*) did not benefit much from the spatial adaptation. We can derive the recommendation that spatial adaptation in sequences without a clear ROI should be used carefully, as it is in general not advantageous and can actually be detrimental to quality due to the presence of QSE's.

C. Effect of δ

We can observe that the choice of δ, i.e., quality difference between tiles, impacts the average quality gains, due to the trade-off between higher ROI quality vs. more QSE. It also makes the evaluation more difficult as distortion is spatially localized and overall quality is more difficult to judge. The choice of δ is very important and affects significantly the overall quality: by running factorial ANOVA we found a significant effect of the tiling scheme for all contents. Most importantly, the best δ is both content and rate dependent.

D. Applicability to HTTP Adaptive Streaming

We can see from this study that the benefit of spatial adaptation can be important at low bitrates, but most importantly that the quality variation δ across tiles has different impact based on the content. This implies that an adaptive streaming client cannot make the decision of selecting the quality using only bitrates/quality ranking of the media, but requires additional signaling derived from the content type. Based on this work, we have proposed in MPEG-DASH a mechanism to signal at the manifest level which tile quality combinations are acceptable for the given content. This is achieved by a DASH descriptor called *QualityEquivalence*, which allows specifying a target maximum quality difference between representations from different adaptation sets played together. This can apply to spatial HEVC adaptation, where each tile corresponds to one adaptation set. By carefully choosing the qualities of each tile, it is possible to instruct the client to perform tile adaptation on lower rates but not on higher rates for which the benefit is not obvious. This work is currently under standardization at MPEG[21]. In order for the community to reproduce and extend this experiment, the test sequences and tools used are made available at http://gpac.io.

V. CONCLUSION

In this paper, we presented a subjective study to evaluate the impact on perceived video quality when doing spatial-based adaptation on HEVC content using tiles. We have shown that especially at lower bitrates one can benefit from spatial adaptation, in some cases achieving better quality than non-tiled version at the same overall bitrate. At the same time, we have shown that too high

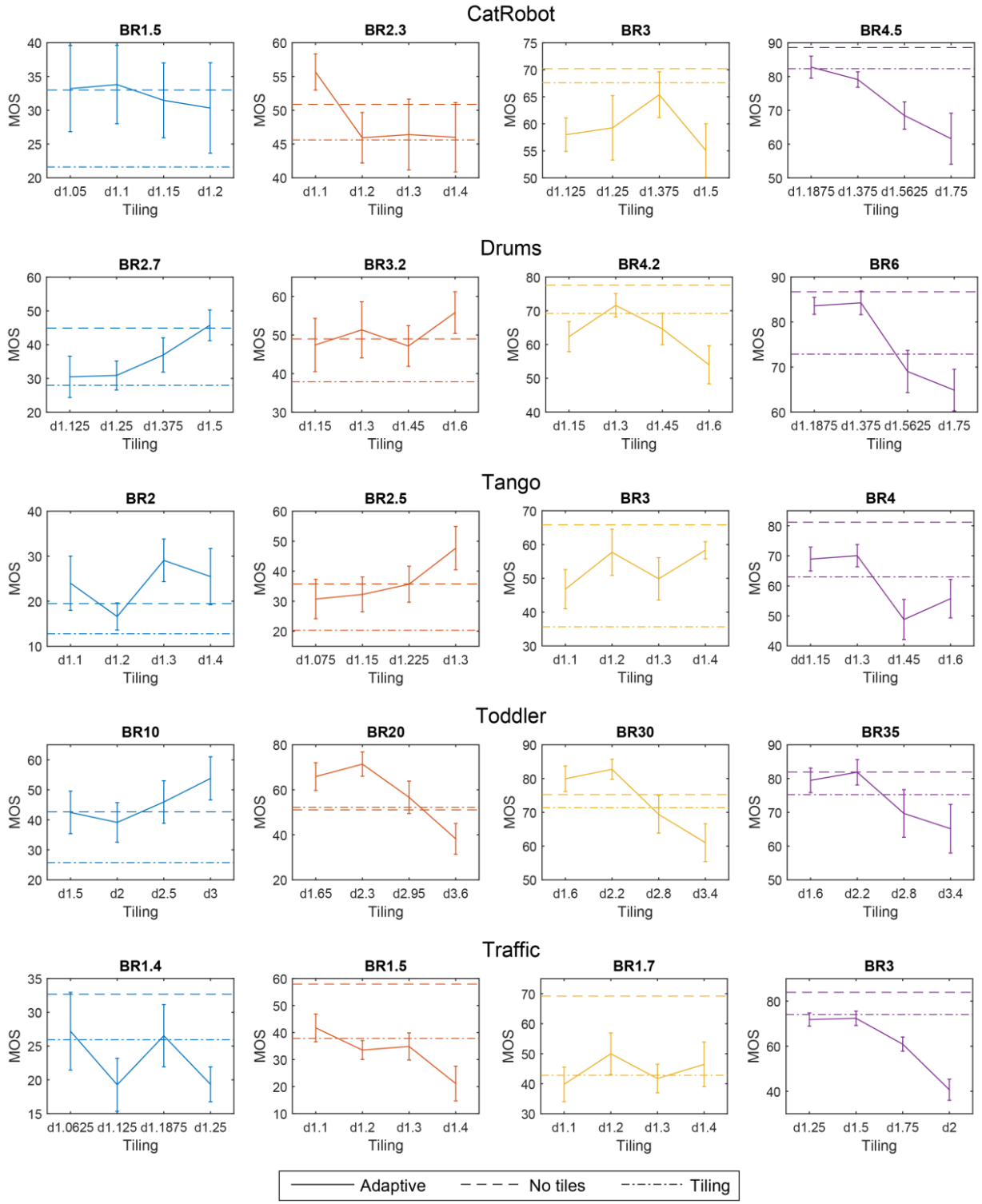

Fig. 3. Comparison of mean opinion scores (MOS) across different scenes, bitrates and tiling conditions with the 95% confidence intervals. The MOS axis is adjusted on each graph for better precision

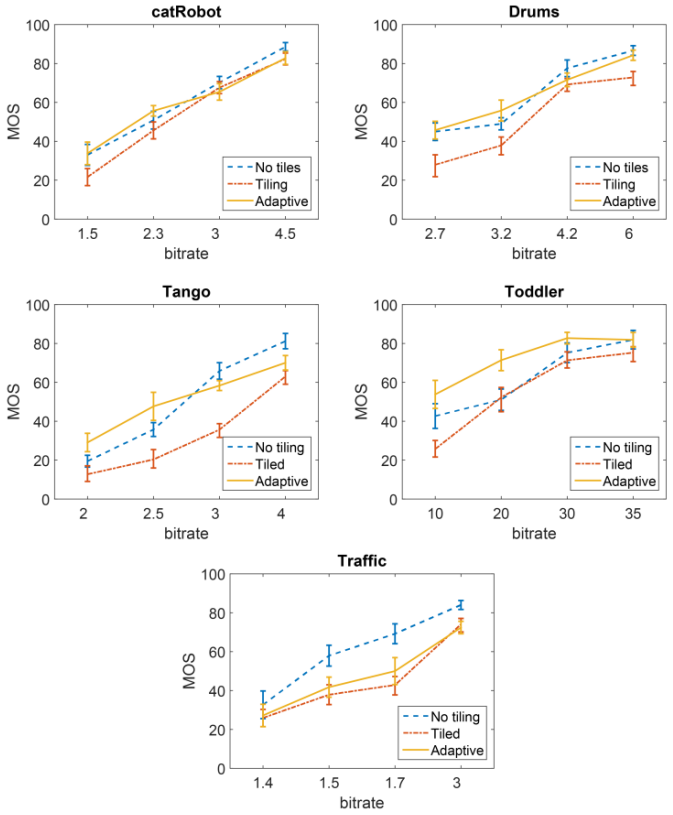

Fig. 4. Rate-MOS curves obtained through the subjective test. For the adaptive tiling curve, we selected for each bitrate δ the giving the best visual quality.

bitrate differences between tiles, producing visible artifacts at Quality-Switch-Edges (QSE), can negatively impact the overall perceived quality. We have also shown that the perceived quality depends on the nature of the content, and proposed a mechanism in MPEG-DASH to help the client in its adaptation logic for spatial adaptation. In future work, we plan to investigate QSE correlation to perceived quality, extending the experiment on video for 360 VR (cube maps or equi-rectangular projections), and designing a tiled-based rate adaptation logic leveraging the results of this study in order to conduct real-time experiments on spatial adaptation for 360 video live broadcast.

Acknowledgment

This work has been achieved within the Live360TV (http://www.live360tv.com/) collaborative research project, supported by three clusters (Images & Réseaux, Imaginove and Cap Digital) and partially funded by the French Government and three local councils (Région Bretagne, Région Rhône-Alpes and Région Ile-De-France).

References

[1] G. J. Sullivan, J. R. Ohm, W. J. Han, and T. Wiegand, "Overview of the high efficiency video coding (hevc) standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 12, pp. 1649–1668, 2012.

[2] L. Itti and C. Koch, "Computational modelling of visual attention," *Nature Reviews Neuroscience*, vol. 2, no. 2, pp. 149–203, 2001.

[3] D. Hepper, "Efficiency analysis and application of uncovered background prediction in a low bit rate image coder," *IEEE Transactions on Communications*, vol. 38, no. 9, pp. 1578–1584, 1990.

[4] A. Vetro, H. Sun, and Y. Wang, "Mpeg-4 rate control for multiple video objects," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 9, no. 1, pp. 186–199, 1999.

[5] A. Vetro, T. Haga, K. Sumi, and H. Sun, "Object-based coding for long-term archive of surveillance," in *Multimedia and Expo, 2003. ICME'03. Proceedings. 2003 International Conference on*, vol. 2. IEEE, 2003, pp. 417–420.

[6] Y. Liu, Z. G. Li, and Y. C. Soh, "Region-of-interest based resource allocation for conversational video communication of h. 264/avc," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 18, no. 1, pp. 134–139, 2008.

[7] Z. Li, S. Qin, and L. Itti, "Visual attention guided bit allocation in video compression," *Image and Vision Computing*, vol. 29, no. 1, pp. 1–14, 2011.

[8] M. Meddeb, M. Cagnazzo, and B. Pesquet-Popescu, "Region-of-interest-based rate control scheme for high efficiency video coding," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, vol. 3, no. 16. IEEE, 2014.

[9] I. Himawan, W. Song, and D. Tjondronegoro, "Impact of automatic region-of-interest coding on perceived quality in mobile video," *Multimedia Tools and Applications*, vol. 76, no. 1, pp. 785–813, 2017.

[10] M. Xu, X. Deng, S. Li, and Z. Wang, "Region-of-interest based conversational hevc coding with hierarchical perception model of face," *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, pp. 475 – 4898, 2014.

[11] L. Bai, L. Song, R. Xie, J. Xie, and M. Chen, "Saliency based rate control scheme for high efficiency video coding," in *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2016 Asia-Pacific*. IEEE, 2016.

[12] W. Lai, X. Gu, R. Wang, W. Ma, and H. Zhang, "A content-based bit allocation model for video streaming," in *IEEE International Conference on Multimedia and Expo (ICME)*, vol. 2. IEEE, 2004, pp. 1315–1318.

[13] C. Concolato, J. Le Feuvre, F. Denoual, E. Nassor, N. Ouedraogo, and J. Taquet, "Adaptive streaming of HEVC tiled videos using MPEG-DASH," *IEEE Transactions on Circuits and Systems for Video Technology*, 2017.

[14] K. Misra, A. Segall, M. Horowitz, S. Xu, A. Fuldseth, and M. Zhou, "An overview of tiles in hevc," *IEEE Journal of Selected Topics in Signal Processing*, vol. 7, no. 6, pp. 969–977, Dec. 2013.

[15] M. Graf, C. Timmerer, and C. Mueller, "Towards bandwidth efficient adaptive streaming of omnidirectional video over http: Design, implementation, and evaluation," in *Proceedings of the 8th ACM on Multimedia Systems Conference*, 2017, pp. 261–271.

[16] Y. SÃąnchez, R. Skupin, and T. Schierl, "Compressed domain video processing for tile based panoramic streaming using hevc," in *IEEE International Conference on Image Processing (ICIP)*. IEEE, Sep. 2015.

[17] ITU-T, "Subjective video quality assessment methods for multimedia applications," *ITU-T Recommendation P.910*, 2008.

[18] F. D. Simone, "Selected contributions on multimedia quality evaluation," Ph.D. dissertation, EPFL, Lausanne, 2012.

[19] ITU-T, "Methodology for the subjective assessment of the quality of television pictures," *ITU-T Recommendation BT.500-11*, 1974.

[20] ITU-R, "The present state of ultra-high definition television," *ITU-R Recommendation BT.2246-6*, 2017.

[21] M. Systems, "w17213 - dash technologies under consideration," 2017.