QUALITY-AWARE ADAPTIVE DELIVERY OF MULTI-VIEW VIDEO

Cagri Ozcinar[⋆]*, Erhan Ekmekcioglu*[†]*, and Ahmet Kondoz*[†]

[⋆]Institut Mines-Télécom, Télécom ParisTech, Paris, France
[†]Institute for Digital Technologies, Loughborough University London, London, UK
[⋆]cagri.ozcinar@telecom-paristech.fr

ABSTRACT

Advances in video coding and networking technologies have paved the way for the Multi-View Video (MVV) streaming. However, large amounts of data and dynamic network conditions result in frequent network congestion, which may prevent video packets from being delivered on time. As a consequence, the 3D viewing experience may be degraded significantly, unless quality-aware adaptation methods are deployed. There is no research work to discuss the MVV adaptation of decision strategy or provide a detailed analysis of a dynamic network environment. This work addresses the mentioned issues for MVV streaming over HTTP for emerging multi-view displays. In this research work, the effect of various adaptations of decision strategies are evaluated and, as a result, a new quality-aware adaptation method is designed. The proposed method is benefiting from layer based video coding in such a way that high Quality of Experience (QoE) is maintained in a cost-effective manner. The conducted experimental results on MVV streaming using the proposed strategy are showing that the perceptual 3D video quality, under adverse network conditions, is enhanced significantly as a result of the proposed quality-aware adaptation.

Index Terms— 3D, multi-view video, scalable coding, QoE, adaptation of decision strategy, analysis of dynamic network.

1. INTRODUCTION

The 3D viewing experience is enhanced using the form of *Multi-View Video* (MVV) [1], [2] and Super Multi-View (SMV) video [3], which offer an immersive user experience via motion parallax. MVV technology captures more than two views of the same scene from different perspectives. As the number of the captured views increases, the transmission of a potentially significant number of views requires massive bandwidth that is beyond the present Internet capacity.

The new 3D video coding standard, the 3D extension of HEVC (3D-HEVC) [4] addresses the generation of additional views from some color texture views with their associated depth maps. The aim is to use multi-view displays with tens of output views from several camera feeds. However, additional view generation techniques at the receiver side struggle mainly with *view synthesis* artifacts and *occlusion* problems. Furthermore, the increasingly complex prediction dependencies among views reduce the *robustness* of the MVV stream against transmission errors.

Related works: Adaptive streaming is based on adapting the bandwidth required by the video to dynamically changing bandwidth on the network [5], [6]. To produce reliable MVV streaming, Savas *et al.* proposed an adaptive MVV streaming over Peer-to-Peer (P2P) networks [7]. The work focuses on whether reducing the quality of all views or keeping a subset of views and synthesizing the missing views to achieve better results. However, the overall visual quality suffers from severe artifacts on the synthesized views [8]. Similarly, Toni *et al.* in [9] enhance the average user satisfaction by selecting optimal encoding parameters. The selection of the adaptation representations is modeled in such a way that both the compression and spatial scaling artifacts are minimized.

Contributions: Recent works on network adaptation methods have focused on MPEG-DASH [10], [11], and P2P networks [12], [13] and there has been no direct focus on the adaptation of decision strategy and analysis over a dynamic network environment. In this work, we propose a quality-aware adaptation solution to achieve smooth MVV playback quality. In this way, the client can increase/decrease the 3D visual quality incrementally. The MPEG-DASH standard [14] and view reconstruction method [10] architecture have been adopted in our proposed method, however,

1. We designed a new layer-based quality with the view scaling adaptation method at the receiver side. According to the bandwidth capacity, either the coding layers of all views or a subset of views were discarded by the client. The missing views were then reconstructed with high quality at the receiver side.

2. We extended our evaluation using more, and different MVV sequences, and conducted formal subjective testing campaign according to the ITU-T BT.500-13 recommendation for laboratory environment [15].

The rest of this paper is organized as follows: Section

2 presents a detailed description of the proposed adaptation system, including layer-based MVV coding and the designed adaptation method. Section 3 presents the experimental results and discussions, which are followed by the concluding remarks in Section 4.

2. THE PROPOSED ADAPTATION SYSTEM

The HTTP server consists of four distinct elements: the Media Presentation Description (MPD), the chunks, the Look-Up Table (LUT), and the Side Information (SI). The MPD includes the manifest of the adaptation strategy and is discussed in Section 2.1. The chunks comprise the encoded streams in the form of self-decodable transmission packets. The LUT is created in the HTTP server and is downloaded by the client before starting the playback. The SI contains the LUT index values of the optimum weighting factors to estimate discarded stream(s) from delivered ones, as described in Section 2.2.

In the proposed system, layer-based video coding is integrated; so that, the 3D visual performance is further improved. Two different quality coding layers, a base and an enhancement, are generated for each view to achieve smooth streaming under adverse network conditions.

The method is undertaken by either discarding a subset of views or by removing the enhancement layer of views. For instance, MVV streams can be decoded at the lowest quality only if the base layer of MVV sequences are delivered to the client. In contrast, when the enhancement layer is given, the highest MVV quality can be streamed to the client.

First, the client obtains the pre-estimated adaptation strategy (MPD) through HTTP. This data enables the client to select the appropriate set of encoded bit-streams during the streaming session. Depending on the network throughput and its cost-quality criterion, the proposed adaptation strategy focuses on whether to reduce the quality of all views or to keep a subset of views and recover the discarded views. The SI, utilized to recover the discarded streams with high quality, is defined according to the metadata estimation process.

2.1. Integration of the Layer-Based MVV Coding

As illustrated in Figure 1, each view is encoded into a base layer and an enhancement layer, which are indicated as L_b and L_e, respectively. For instance, if the enhancement layer is discarded for adaptation purposes, the remaining base layer is still decodable, but it will produce inferior quality.

L_b comprises all available views' base layers, $L_b = \{ L_b^1, \ldots, L_b^M \}$. L_e also contains all available views' enhancement layers, $L_e = \{ L_e^1, \ldots, L_e^M \}$, where M is the number of available views. Each quality layer is further structured based on view scalable adaptation sets that include various chunk bit-rates. Since the L_e depends on the L_b for decoding, this dependency is defined in the MPD.

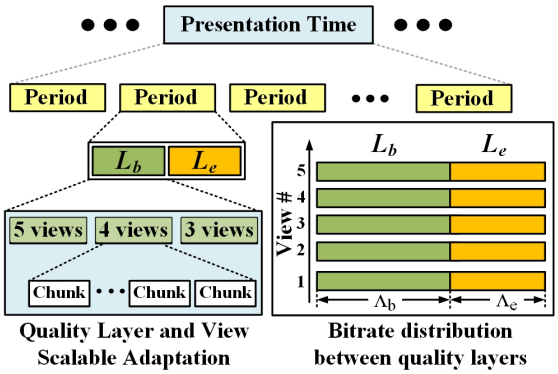

Fig. 1: The proposed MVV adaptation organisation.

In this scheme, Λ_b and Λ_e denote the bit-rate allocations of the L_b and the L_e, respectively. To achieve a significant bit-rate reduction as a result of adaptation, Λ_b is selected as two times Λ_e. Each layer is split and distributed into a different subset of view representations. Therefore, the client can enhance the 3D visual quality of a chunk incrementally by downloading a new bit-stream.

2.2. Metadata Estimation

The cross-correlation method [16] and Depth Image Based Rendering (DIBR) [17] techniques are utilized in the view blending process in such a way that the depth-aided image interpolation quality is superior to that of the synthesised view. This model recovers the discarded views, n where $n \neq \{1, M\}$, with the smallest possible pixel error in relation to its uncompressed original representation, as mathematically described in Equation (1):

$$\begin{bmatrix} w_1 \\ \vdots \\ w_M \end{bmatrix} \cdot \begin{bmatrix} E\left[\widetilde{B}_1 \cdot \widetilde{B}_1\right] & \cdots & E\left[\widetilde{B}_1 \cdot \widetilde{B}_M\right] \\ \vdots & \vdots & \vdots \\ E\left[\widetilde{B}_M \cdot \widetilde{B}_1\right] & \cdots & E\left[\widetilde{B}_M \cdot \widetilde{B}_M\right] \end{bmatrix} = \begin{bmatrix} E\left[\widetilde{B}_1 \cdot \widetilde{B}_n\right] \\ \vdots \\ E\left[\widetilde{B}_M \cdot \widetilde{B}_n\right] \end{bmatrix} \quad (1)$$

where $E[\cdot]$ represents the normalized expectation value, w expresses the weighting factor of each block for each view, $\widetilde{B}_x$ describes the projected block of the x_{th} view.

To design LUT, the K-means algorithm is applied to the estimated weighting factors. Thus, each weighting factor (w) forms coefficient vectors in the LUT, which is encoded using an l-bit codeword. To obtain the optimum coefficient vectors, each vector's reconstruction quality is compared with all candidate vectors in the LUT. Then, the candidate that results in the highest-quality index value (i) is selected. The index value of each computed coefficient vector that corresponds to each computed block is embedded in the SI.

Table 1: Adaptation decision plan

Level	Method
1^{st}	Enhanced quality views (L_b plus L_e)
2^{nd}	L_b only
3^{rd}	L_b with view discarding

2.3. Dynamic Adaptation Using HTTP

Table 1 summarizes the adaptation decision plan, which is obtained as a result of the evaluations in Section 3. Adaptation starts from the maximum quality level, where all color texture views and their depth maps are delivered. Then, in the next level, L_e is discarded for transmission, and only L_b is distributed over the network. As a result of this process, the visual quality is reduced, but it is still acceptable to drive multi-view displays. In the final level, some views are sacrificed for the transmission, and instead, the SI stream is delivered. In doing so, missing views are reconstructed with the help of the SI and all delivered adjacent views at the receiver.

To determine the subset of views to be delivered at the 3^{rd} decision level, priority information is assigned to each view. The priority information is derived from the classification with the aim of minimizing the distortion of discarded chunks subject to the limited overall SI bit-budget (R_a). The overall cost minimization function is described in Equation (2):

$$\underset{k}{\operatorname{argmin}}\, P(k) = D_s(k) + \lambda \cdot R(k), \quad R(k) < R_a \qquad (2)$$

where $R(k)$ is the overall transmitted bit-rate after discarding some of the views (including the bit-rate of the SI) and $D_s(k)$ is the average reconstruction distortion of all discarded view(s), which is estimated using Mean Square Error (MSE). λ is the Lagrangian multiplier, which was set experimentally. After the priority information is assigned to each chunk of the MVV representation, the MPD is formed.

3. EXPERIMENTAL RESULTS AND DISCUSSIONS

3.1. Simulation Setup

The experiments were conducted using five adjacent color texture views and depth maps from five (*M*=5) different MPEG test sequences [18]. The evaluated test sequences were as follows: *BookArrival* (resolution: 1024×768), *Newspaper* (1024×768), *Champagne* (1280×960), *Café* (1920×1080), and *PoznanStreet* (1920×1080).

Each view was encoded with L_b and L_e, using *SHM v6.0* [19]. All layers were coded using hierarchical B-frames with a Group Of Picture (GOP) size of 16 pictures, which is the size of the transmission chunk. In this work, the Quantization Parameter (QP) was set to 26 for L_b, and the QP value for L_e was calculated according to $\Lambda_b = 2 \cdot \Lambda_e$ bit-rate allocation. The depth map bit-rates were fixed at a percentage of 20% of the corresponding color texture bit-rates.

To evaluate the performance of the proposed adaptation system in a congested network environment, MPEG-DASH [20] was utilized in the sever-client setup [21]. Moreover, to verify the performance over HTTP, another view adaptation method, the MPEG View Synthesized Reference Software (VSRS) [22], without additional SI, was incorporated as a reference. This reference adaptation scheme employed the same adaptation method as the proposed adaptive system for unbiased comparison.

PSNR was employed as the criterion to evaluate the objective streaming quality. In addition, subjective tests were undertaken in accordance with ITU-R BT.500-13 [15]. The Single-Stimulus Continuous Quality Evaluation (SSCQE) method was used for subjective assessments with an autostereoscopic display. In total, 18 non-expert observers (12 man and six women) participated in the test. Each test session started after a short training and an instruction session. Each subjective assessment session lasted up to half an hour.

3.2. Adaptation Decision Strategy

The subjective analysis in terms of the Mean Opinion Score (MOS) is presented for each MVV content in subjective assessment sessions. In this analysis, MVV sequences were tested with six different adaptation test cases:

1. All L_e streams were transmitted for each view.
2. Only L_b streams were transmitted for each view (*i.e.,* all L_e streams were discarded).
3. A view was discarded and reconstructed with the help of the proposed view reconstruction method and available L_e streams.
4. A view was discarded and reconstructed with the help of the proposed view reconstruction method and available L_b streams.
5. A view was discarded and reconstructed with the help of the MPEG VSRS and available L_e streams.
6. A view was discarded, and reconstructed with the help of the MPEG VSRS and available L_b streams.

The first adaptation test case represents Case 1 in Table 1, where the bit-rate requirement is sufficiently high to stream MVV content at all times. Moreover, the second adaptation level in Table 1 was tested in Case 2. For the last four test cases, the third adaptation level was utilized to discard a view and to estimate it using the proposed view reconstruction and MPEG VSRS methods.

Figure 2 shows a comparison of the subjective ratings obtained for the adaptation test patterns taken into account,

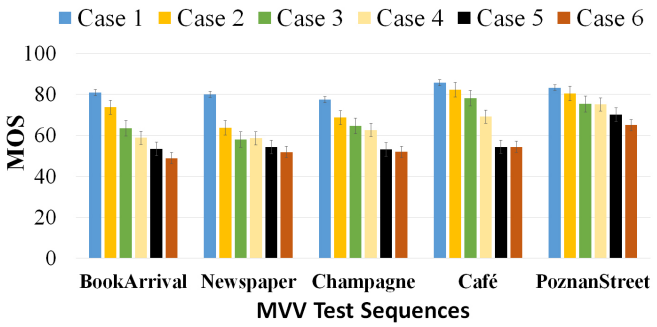

Fig. 2: Comparison of the subjective assessment results of six different adaptation test cases.

along with the 95% confidence interval for each test sequence. As shown in the figure, the highest subjective score is obtained by Case 1, where the enhanced versions of the sequences are streamed at all times. Furthermore, it is apparent that Case 2's score takes the second place for all MVV sequences. In addition, the impact of discarding L_e for Case 2 is less severe for the *Café* and *PoznanStreet* sequences than for the others. Both contents have similar motion activity for color texture views and the same video resolution (1920×1080). It is observed that the compression artifacts are more perceptible at smaller resolutions compared to the larger resolution sequences.

The view-discarding option is evaluated subjectively between Cases 3 and 6. As mentioned previously, this option results in a significant bit-rate reduction. However, more perceptual distortions are caused in the latter case, as demonstrated by the subjective scores. It should be noted that this step is the last operation in the proposed adaptation strategy. Regarding view discarding, the experiment results show that the proposed view reconstruction method provides higher subjective scores compared to the MPEG VSRS for each content.

The overall subjective scores demonstrate a clear pattern of MOS values for all the MVV content. It confirms that the objective observation of the adaptation-decision strategy shown in Table 1 is consistent with the subjective experiment results.

3.3. Analysis over the Dynamic Network Environment

To evaluate the impact of the designed adaptation strategy, a test pattern with instantaneous network throughput changes is used as plotted in Figure 3. The proposed adaptation engine switches between different states according to the available bandwidth. In this test, MPEG VSRS was used as the adaptation reference, where MVV temporal chunks were discarded selectively based on their synthesis quality.

All encoded streams, which were 6.95, 3.99, 5.49, 4.8, and 8.57 *Mpbs* for the *BookArrival*, *Newspaper*, *Champagne*, *Café*, and *PoznanStreet* sequences, respectively, were divided into chunks and stored at the server.

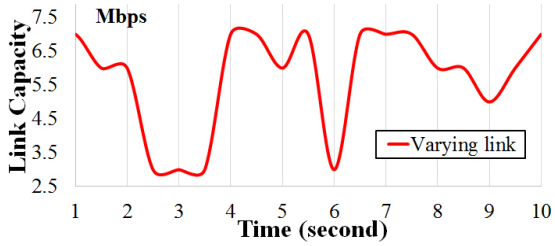

Fig. 3: Bandwidth capacity of the tested client in the network.

Table 2: Objective (PSNR) and subjective (MOS) comparison of the adaptation methods.

Sequence	Method	Quality Measure	
		PSNR (*dB*)	MOS
BookArrival	Proposed method	37.91	73.75
	MPEG VSRS	36.80	58.87
Newspaper	Proposed method	38.91	63.93
	MPEG VSRS	37.82	57.46
ChampagneTower	Proposed method	40.46	78.42
	MPEG VSRS	38.77	56.12
Café	Proposed method	41.36	58.56
	MPEG VSRS	39.14	52.42
PoznanStreet	Proposed method	36.35	64.62
	MPEG VSRS	35.07	59.15

Table 2 provides a comparison of the objective and subjective scores, reported as an average of all the views, including both delivered and discarded/reconstructed ones. The results show that the proposed adaptation method outperforms the reference method consistently, both objectively and subjectively, in all test conditions.

4. CONCLUSION

In this paper, a new quality-aware MVV streaming over HTTP was introduced, to address the adaptation of a decision strategy and provide a detailed analysis in a dynamic network environment. One of the main contributions of the proposed method was a quality-aware adaptation based on a client-server model. To facilitate a quality-aware bandwidth-adaptation mechanism, the proposed adaptation strategy was implemented either by decreasing the number of transmitted views or discarding the enhancement layer of all the available views. The proposed adaptation method was evaluated using a prototype DASH client and MPEG view synthesis method for MVV. Simulation results showed that the proposed adaptation strategy yields a superior and smooth playback performance under dynamic network conditions.

5. REFERENCES

[1] F. Dufaux, B. Pesquet-Popescu, and M. Cagnazzo, *Emerging technologies for 3D video: creation, coding, transmission and rendering*, John Wiley & Sons, 2013.

[2] A. Kondoz and T. Dagiuklas, *3D Future Internet Media*, Springer, 2014.

[3] Q. Wang, Y. Zhang, and L. Yu, "Study on density of viewpoints for super multi-view displays," Tech. Rep. MPEG2014/M33320, ISO/IEC JTC1/SC29/WG11, Valencia, Spain, March 2014.

[4] K. Müller, H. Schwarz, D. Marpe, C. Bartnik, S. Bosse, H. Brust, T. Hinz, H. Lakshman, P. Merkle, F.H. Rhee, G. Tech, M. Winken, and T. Wiegand, "3D high-efficiency video coding for multi-view video and depth data," *Image Processing, IEEE Transactions on*, vol. 22, no. 9, pp. 3366–3378, Sept 2013.

[5] K. Miller, E. Quacchio, G. Gennari, and A. Wolisz, "Adaptation algorithm for adaptive streaming over HTTP," *2012 19th International Packet Video Workshop (PV)*, pp. 173–178, May 2012.

[6] J. Chakareski, "Adaptive multiview video streaming: challenges and opportunities," *Communications Magazine, IEEE*, vol. 51, no. 5, pp. 94–100, May 2013.

[7] S. S. Savas, A. M. Tekalp, and C. G. Gürler, "Adaptive multi-view video streaming over P2P networks considering quality of experience," in *Proceedings of the 2011 ACM Workshop on Social and Behavioural Networked Media Access*, New York, NY, USA, 2011, SBNMA '11, pp. 53–58, ACM.

[8] A Smolić, K. Müller, K. Dix, P. Merkle, P. Kauff, and T. Wiegand, "Intermediate view interpolation based on multiview video plus depth for advanced 3D video systems," in *Image Processing ICIP, 15th IEEE International Conference on*, Oct 2008, pp. 2448–2451.

[9] L. Toni, A.-P. Ramon, G. Simon, A. Blanc, and P. Frossard, "Optimal set of video representations in adaptive streaming," in *Proceedings of the 5th ACM Multimedia Systems Conference*, New York, NY, USA, 2014, MMSys '14, pp. 271–282, ACM.

[10] C. Ozcinar, E. Ekmekcioglu, and A. Kondoz, "Dynamic adaptive 3D multi-view video streaming over the Internet," in *Proceedings of the 2013 ACM International Workshop on Immersive Media Experiences*, Barcelona, Spain, 2013, ImmersiveMe '13, pp. 51–56, ACM.

[11] T. C. Thang, Q. D. Ho, J. W. Kang, and A. T. Pham, "Adaptive streaming of audiovisual content using MPEG DASH," *Consumer Electronics, IEEE Transactions on*, vol. 58, no. 1, pp. 78–85, Feb 2012.

[12] C. Ozcinar, E. Ekmekcioglu, and A. Kondoz, "Adaptive 3D multi-view video streaming over P2P networks," in *Image Processing (ICIP), 2014 IEEE International Conference on*, Oct 2014, pp. 2462–2466.

[13] C. G. Gürler, S. S. Savas, and A. M. Tekalp, "Quality of experience aware adaptation strategies for multi-view video over P2P networks," in *Image Processing (ICIP), 2012 19th IEEE International Conference on*, Sept 2012, pp. 2289–2292.

[14] A. Vetro and I. Sodagar, "Industry and Standards The MPEG-DASH Standard for Multimedia Streaming Over the Internet," *IEEE MultiMedia*, vol. 18, no. 4, pp. 62–67, 2011.

[15] Recommendation, ITU-R BT.500-13, "ITU-R BT.500-13, Methodology for the subjective assessment of the quality of television pictures," Tech. Rep., Jan. 2012.

[16] Y. Sugiyama, "An algorithm for solving discrete-time Wiener-Hopf equations based upon Euclid's algorithm," *Information Theory, IEEE Transactions on*, vol. 32, no. 3, pp. 394–409, May 1986.

[17] C. Fehn, "Depth-image-based rendering (DIBR), compression, and transmission for a new approach on 3D-TV," *Proceeding SPIE*, vol. 5291, pp. 93–104, 2004.

[18] M. Tanimoto, T. Senoh, S. Naito, S. Shimizu, H. Horimai, M. Domanski, A. Vetro, M. Preda, and K. Mueller, "Proposal on a new activity for the third phase of FTV," Tech. Rep. MPEG2011/N12036, ISO/IEC JTC1/SC29/WG11, Vienna, Austria, July 2013.

[19] J. Chen, J. Boyce, Y. Ye, and M. Hannuksela, "Scalable HEVC (SHVC) test model 4 (SHM 4)," Tech. Rep. N13939, ISO/IEC JTC1/SC29/WG11, Geneva, Switzerland, Nov. 2013.

[20] S. Lederer, C. Müller, and C. Timmerer, "Dynamic adaptive streaming over HTTP dataset," in *Proceedings of the 3rd Multimedia Systems Conference*, New York, NY, USA, 2012, MMSys '12, pp. 89–94, ACM.

[21] C. Müller, S. Lederer, and C. Timmerer, "An evaluation of dynamic adaptive streaming over HTTP in vehicular environments," in *Proceedings of the 4th Workshop on Mobile Video*, New York, NY, USA, 2012, MoVid '12, pp. 37–42, ACM.

[22] M. Tanimoto, T. Fujii, K. Suzuki, N. Fukushima, and Y. Mori, "Reference softwares for depth estimation and view synthesis," Tech. Rep. MPEG2008/M15377, ISO/IEC JTC1/SC29/WG11, Archamps, Apr 2008.