Multipass Hierarchical View Grouping Method for Efficient 6DoF Video Streaming

Yeongil Ryu
Dept. of Computer Science Education
Sungkyunkwan University
Seoul, South Korea
yeongilryu@skku.edu

Eun-Seok Ryu
Dept. of Computer Science Education
Sungkyunkwan University
Seoul, South Korea
esryu@skku.edu

Abstract—A novel six degree-of-freedom (6DoF) video streaming method based on multipass hierarchical view grouping is proposed herein. The proposed method divides the source views into multiple groups based on the position of each source view and then separately encodes the groups of source views into independent bitstreams. Subsequently, it streams the encoded bitstreams based on the user's position and viewport. This method can increase the number of basic views and atlases for enhanced video quality, while reducing the number of source views to be reconstructed at the decoder side. Experimental results show that the proposed method significantly improves both objective and subjective qualities at the slightly increased bitrate. In addition, the proposed method achieves a decoding speedup of 37.62% owing to the reduced number of source views.

Keywords—6DoF, Video streaming, MIV, View grouping

I. INTRODUCTION

In recent years, industrial and academic institutes have been actively attempting to develop immersive media technologies. However, providing six degrees of freedom (6DoF) immersive videos to users remains challenging as larger storage, powerful computational performance, and an extremely large network bandwidth are required to distribute them in real time. Hence, the Moving Picture Experts Group (MPEG) is developing a standard for coding immersive media, which is known as MPEG-I (Immersive)[1].

The MPEG-I standard encompasses various immersive media such as 360-degree and volumetric videos. In particular, the MPEG Immersive Video (MIV) standard of MPEG-I focuses on providing a framework to compress the representation of actual or virtual three-dimensional (3D) scenes captured by multiple cameras and sensors. The MIV coding framework compresses the visual and depth information from multiple cameras and sensors by pruning inter-view redundancy and packing the preserved data into a few frames, called as atlases [2]. Subsequently, the MIV coding framework codes the atlases using lossy video codecs, such as High Efficiency Video Coding (HEVC) or Versatile Video Coding (VVC), and it delivers the coded bitstream. Fig. 1 illustrates the encoding process of the MIV framework.

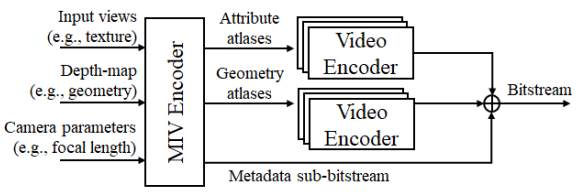

Fig. 1. Encoding process of MIV framework

Although inter-view redundancy is removed, the large amount of the data still remains in multiple atlases. Moreover, decoding multiple coded atlases requires multiple video decoder instances at the decoder side, and the computational complexity of each decoder instance is extremely high. Consequently, real-time MIV encoding and decoding are difficult to perform. Hence, some research institutes, such as Interdigital, have proposed acceleration techniques for the MIV decoder[3]. However, their method focuses on acceleration using a GPU implementation at the decoder side. The proposed GPU-based technique achieves the enhanced decoding time from acceleration of the view synthesis module in the MIV decoder, and it accomplishes meaningful and effective results. Nevertheless, the GPU-based acceleration technique does not consider streaming schemes. A multipass hierarchical view grouping method that not only accelerates the decoding speed, but also supports efficient MIV-based 6DoF video streaming is proposed herein.

II. PROPOSED MULTIPASS HIERARCHICAL VIEW GROUPING METHOD

Two encoding modes exist in MIV encoding: 1) the MIV mode and 2) MIV view mode. In the MIV mode, the MIV encoder selects basic and additional views among all source views and then prunes the inter-view redundancy between additional views. Subsequently, it packs the basic and the pruned views into atlases. Meanwhile, in the MIV view mode, a subset of source views is manually selected for each video sequence, and then the selected views are packed into atlases without pruning. In the MIV view mode, the MIV decoder is not required to reconstruct the pruned views; therefore, its decoding speed is typically faster than that of the MIV mode. However, the bitrate may be higher because pruning is omitted.

The proposed method is based on the MIV mode. However, it takes the advantages of the MIV view mode, as follows: First, the proposed method divides the source views into multiple groups based on the position of each source view. Second, it separately encodes the groups of the source views into independent bitstreams. Subsequently, it streams the encoded bitstreams based on the user's position and viewport. For example, if a user is located in the position within Group 2, then it streams only the bitstream of Group 2. This method can increase the number of basic views and atlases, while reducing the number of source views to be reconstructed at the decoder side. The increased number of basic views and atlases leverages the enhancement in both the objective and subjective video qualities. Although the number of atlases increases, the number of delivered atlases to the decoder side remains unchanged. This implies that the proposed method provides enhanced video quality at a similar bitrate.

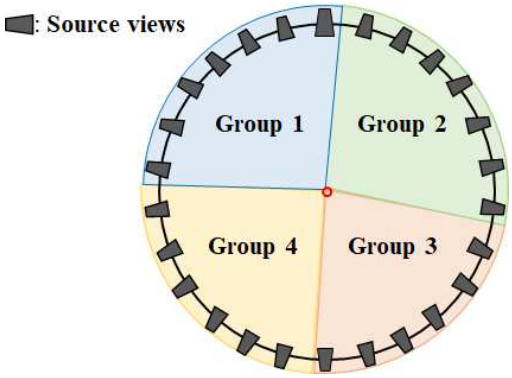

Fig. 2. Conceptual diagram of multipass view grouping

In addition to the improved trade-off between bitrate and quality, the decreased number of source views to be reconstructed at the decoder side provides a higher decoding speed. However, blindly reducing the number of source views per group to achieve a higher decoding speed should be sublated, because fewer source views per atlas causes less compression performance in the MIV mode. Therefore, the number of groups and the number of views per group must be determined precisely. Fig. 2 illustrates the source view grouping based on the position of each source view.

Although the proposed view grouping method offers significant advantages in terms of the objective/subjective video qualities and decoding speedup, it presents a few disadvantages as well. First, the proposed method wastes considerable storage at the encoder side because each group typically includes two atlases per input component (e.g., two texture atlases and two geometry atlases). This implies more atlases and data to be stored. However, the storage cost is decreased, and many video technologies such as MPEG Dynamic Adaptive Streaming over HTTP (DASH) prioritize storage over the performance of video systems. Therefore, this disadvantage is considered acceptable. Second, source view grouping causes video quality degradation when the user moves across groups. In the proposed streaming scheme, the client receives only one group based on the user's position. This implies that the MIV decoder cannot utilize the reconstructed view of other groups when the user moves across groups. This causes the lack of information to synthesize intermediate views between groups, and it degrades the video quality of synthesized intermediate views. Hence, the proposed method employs a hierarchical structure, as shown in Fig. 3. In the case of Fig. 3, users can experience degraded video quality when they move from v3 to v4 or v7 to v11; however, the hierarchical structure prevents degradation by switching from Layer 0 to Layer 1. Groups 1.5 and 2.5 in Layer 1 contain sufficient information for synthesizing the intermediate views between v3-v4 and v7-v8, which improves their video quality when the user moves across the groups.

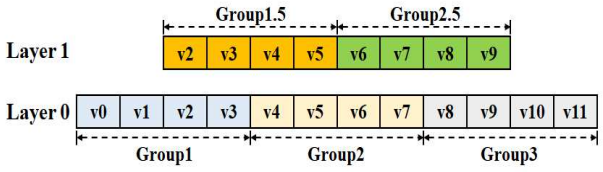

Fig. 3. Hierarchical view grouping

TABLE I. EXPERIMENTAL ENVIRONMENTS

CPU	Intel CoreX-series i9-10980XE (3.0 GHz 18 cores, 36 threads)
Memory	48GB
Graphic card	Nvidia GeForce RTX 3080Ti
OS	Ubuntu 18.04

TABLE II. SPECIFICATIONS OF TEST SEQUENCE AND CODING OPTIONS

Item	**Experimental values**
Sequence	CBAbasketball
Resolution	2048 × 1088
# of source views	30
Frame rate	30
Projection	Perspective
Encoding mode	MIV mode
Intra period	32
# of encoded frames	64
Texture QP	22, 27, 32, 37
Geometry QP	3, 7, 11, 15

III. EXPERIMENTAL RESULTS

To evaluate the proposed method, experiments were conducted to compare the objective quality, subjective quality, and decoding speed. Table 1 lists the environments used in these experiments. The Test Model for Immersive Video (TMIV), which is a reference software of the MIV standard, was used to encode/decode the test sequence. In addition, encoding and decoding were conducted based on the Common Test Condition (CTC) of the MIV standard. As a test sequence, *CBAbasketball* was employed in the experiments. Fig. 4 shows the texture and geometry views of v00 and v20. As shown in Table 2, *CBAbasketball* comprises 30 texture and 30 geometry views with a resolution of 2048×1088.

In the experiments, the test sequence was encoded based on three configurations. Configuration 1 (named *MIV anchor*) is based on the MIV anchor configuration, and Configuration 2 (named *Modified MIV anchor*) partially reflects the MIV anchor configuration. Unlike the MIV anchor, Configuration 2 allows an unlimited number of atlases. When the test sequence is encoded without the restriction of the maximum number of atlases, four texture and four geometry atlases are created. Finally, Configuration 3 (named *Proposed*) employs the proposed method. To employ the proposed method, the number of views per group was set to 10, and the number of groups was set to five (group 1: v00-v09, group 1.5: v05-v14, group 2: v10-v19, group 2.5: v15-v24, and group 3: v20-v29.) In addition, all configurations were encoded with texture QP 22, 27, 32, and 37, and geometry QP 3, 7, 11, and 15. These geometry QPs were derived by mapping from texture QP. The mapping from texture QP (q) to geometry QP (q') is the same for all configurations and is expressed as

$$q' = \max(1, [-14.2 + 0.8q]) \quad (1)$$

Fig. 4. Two input views with corresponding depth-maps views 0 and 20 for *CBAbasketball*

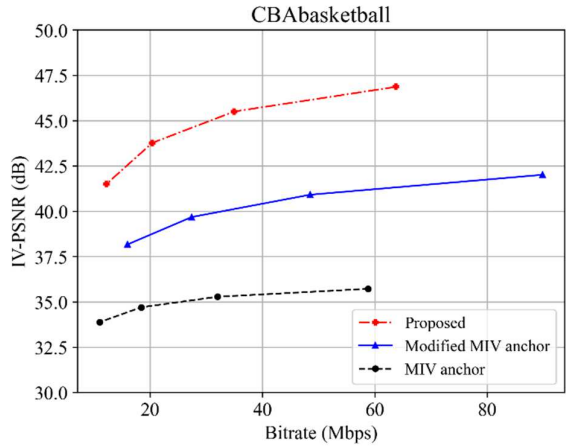

Fig. 5. RD curve: *Proposed* vs. *Modified MIV anchor* and *MIV anchor*

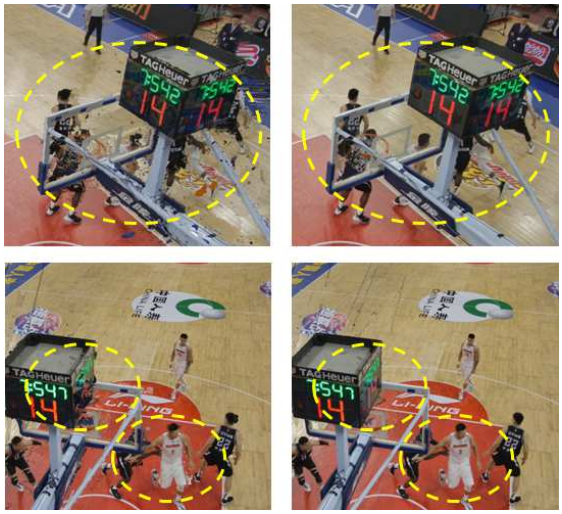

Fig. 6. Comparison of subjective quality between *MIV anchor* (left) and *Proposed* (right) configurations

Fig. 5 shows the experimental results of the objective quality comparison. The test results show that the *Proposed* configuration achieves a significantly enhanced the IV-PSNR at a slightly increased bitrate. Fig. 6 shows a comparison of the subjective quality between the *Proposed* and *MIV anchor* configurations, and those snapshots are synthesized intermediate views between v08-v11 and v18-v21. Like the results of the objective quality comparison, the *Proposed* configuration achieved significantly improved subjective quality.

In addition, the proposed method is advantageous in terms of the decoding speed. In the MIV mode, the decoding speed is dependent on the number of source views. An increase in the number of source views increases the decoding time. Consequently, the *Proposed* configuration achieved a decoding speedup of 37.62% because its number of source views per group was fewer than those of *MIV anchor* and *Modified MIV anchor*.

IV. Conclusion

A novel 6DoF video streaming method using multipass hierarchical view grouping was proposed herein. The proposed method divides the source views into multiple groups based on the position of each source view and then separately encodes the groups of source views into independent bitstreams. Subsequently, it streams the encoded bitstreams based on the user's position and viewport.

This method can increase the number of basic views and atlases, while reducing the number of source views to be reconstructed at the decoder side. The increased number of basic views and atlases leverages the enhancement in both the objective and subjective video qualities at a slightly increased bitrate. Experimental results showed that the proposed method achieved significantly improved both objective and subjective qualities. In addition, the proposed method achieved a decoding speedup of 37.62%.

Acknowledgment

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. 2022R1F1A1074935).

References

[1] M. Wien, J. M. Boyce, T. Stockhammer and W. -H. Peng, "Standardization Status of Immersive Video Coding," in *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 9, no. 1, pp. 5-17, March 2019.

[2] J. M. Boyce *et al.*, "MPEG Immersive Video Coding Standard," in *Proceedings of the IEEE*, vol. 109, no. 9, pp. 1521-1536, Sept. 2021.

[3] J. Fleureau, B. Chupeau, F. Thudor, G. Briand, T. Tapie and R. Doré, "An Immersive Video Experience with Real-Time View Synthesis Leveraging the Upcoming MIV Distribution Standard," *2020 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, 2020, pp. 1-2.