Efficient Group-Based Packing Strategy for 6DoF Immersive Video Streaming

Soonbin Lee, Jong-Beom Jeong, Eun-Seok Ryu
Department of Computer Science Education, Sungkyunkwan University, Seoul, Republic of South Korea
E-mails: {soonbinlee, uof4949, esryu}@skku.edu

Abstract—MPEG immersive video (MIV) standard technology is designed to support immersive video transmissions by removing the redundancy between multiview videos. Inter-view redundancy removal is achieved by packing the input videos into atlases, which are compact representations of multiview videos. Although this process can significantly reduce the pixel rates, it makes immersive video streaming difficult. The MPEG-I community has proposed group-based encoding as a way to improve the objective rendering performance. However, an efficient packing strategy for groups has yet to be investigated. This paper describes an efficient group-based packing strategy that enables immersive video streaming and parallel processing. The goal of this study was to evaluate the effectiveness of group-based encoding by comparing different packing strategies with MIV-encoding parameters. The experiment results show that the proposed packing method achieves better rendering results with a reasonable pixel rate.

Index Terms—Virtual reality, MIV, 6DoF, immersive video

I. INTRODUCTION

With the current demand and interest in virtual reality (VR), the necessity for an efficient VR technology is critical because of the large number of data that must be processed in the systems [1]. A low latency and high resolution are significant factors that increase the user’s quality of experience (QoE). In addition, the demand for graphical environments and realistic 360° scene representations on head-mounted displays (HMDs) is increasing, as is the need for technologies that provide users with a higher degrees of freedom (DoF). For example, a single 360-degree video does not allow the viewer to walk in a VR environment. In these media markets and technology movements, the Moving Picture Experts Group (ISO/IEC MPEG) has established immersive media standard projects to facilitate the compression, sharing, and distribution of immersive media between various devices and platforms.

In such an effort, the MPEG-Immersive (MPEG-I) subgroup defined three types of DoF for users. First, 3DoF supports only those experiences in which viewers are limited to rotational movements around the pitch, yaw, and roll. Second, 3DoF+ supports restricted movement of the user’s head, which is an intermediate approach to 6DoF. Finally, the 6DoF supports a free viewpoint, which implies full movement of the user [2]–[4]. Motion parallax in 6DoF technologies can be achieved using the depth map-based image rendering (DIBR) technique with depth map information and associated camera parameters. Multiview video plus depth (MVD) representation technologies are still mainly studied for use in immersive videos.

An immersive video includes several videos, including their corresponding depth maps [5]–[7]. Using the DIBR technique, 6DoF can generate the target views of a scene using a fixed number of input images and their corresponding depth maps. Given a target view to generate, the DIBR replaces the texture from the input videos with their new position corresponding to the depth maps. Immersive 6DoF videos should support many target views for user movements, and high computing resources and bandwidth are required to compress and transmit multiple videos acquired from different positions [8]. Although contributions to the compression performance improvement of immersive videos have been proposed in the MIV standard, MVD representation technologies must consider several representations of multiple views in a streaming scenario [9]–[11]. However, there have been only a few studies on immersive video streaming [12], [13]

The goal of this paper is to present a novel packing strategy for a group-based coding approach that can reduce the bandwidth and computing resources without significantly impacting the QoE. Although several techniques have been proposed for group-based MIVs, such techniques have not been investigated from an immersive video streaming perspective. The proposed packing strategy utilizes a group-based approach that separates all input views into groups, considering an efficient video representation. The evaluations included objective assessments based on common test conditions (CTCs), and the experiment results demonstrate the BD-rate gains and pixel rate savings achieved by the packing strategy.
Algorithm 1
MIV Interview Redundancy Removal Algorithm

Require: Multiple texture videos with corresponding depth maps, camera pruned parameters
Ensure: Pruned videos (Atlases)

$V_S : \text{set of all source views, e.g. } S = \{1, 2, 3, \ldots N\}$
$V_B : \text{set of basic views}$
$V_A \leftarrow V_S - V_B : \text{set of additional views}$

Atlas $\leftarrow V_B$ // Initialize atlas

for $i \leftarrow 1$ to $B$ do
    for $j \leftarrow 1$ to $A$ do
        $V_i \odot V_j = Patch_{ij}$ // $\odot$ is pruning operator
        Atlas $\leftarrow$ Atlas $\cup$ Patch$_{ij}$ // Patch packing
    end for
end for

II. Preliminaries

A. MPEG Immersive Video Standardization

The 3DoF+ and 6DoF technologies require the compression and processing of multiple videos to support the user’s head and body movements, which is considerably challenging with high-efficiency video coding (HEVC) [14]. Because HEVC is designed for single video coding, it requires a large bandwidth and several computing limitations when handling multiple videos. MPEG-I proposed a test model for immersive video (TMIV) as a reference software for a 6DoF video compression [15], [16]. The TMIV supports pre-processing and post-processing for immersive video to more efficiently compress multiview videos [17].

Algorithm 1 illustrates the process of the inter-view redundancy removal in an MIV. The TMIV encoder divides the source views into basic and additional views, and the inter-view redundancy is removed using a pruning operation. The TMIV encoder collects the residual patches and merges them into atlases. Consequently, TMIV produces atlases that are pruned videos from all source views. This preprocessing significantly reduces the pixel rate, which also decreases the required decoder instantiations [8]. However, this patch-packing process has several hyperparameters, such as the ratio of basic views and the number of atlases [16], [18].

B. Group-based MPEG Immersive Video

The TMIV provides group-based encoding, which leads to more accurate rendering with locally coherent projections [19]. The subjective and objective results of the TMIV have been improved with group-based encoding, particularly for natural content or at high bitrate levels. The primary feature of group-based encoding is the division of all input source views into subgroups for processing, which enables the TMIV to preserve the relevant regions inside each group. This group division avoids an unnecessary depth estimation from all input source views, effectively avoiding view projection errors and generating better rendering results [20]. Figure 2 illustrates a block diagram of a group-based TMIV encoder. Although the technical motivation for group-based encoding is to increase the rendering quality, the subgroup has no dependence between groups through the processing of TMIV. This feature can lead to sub-bitstream accessibility across groups [21]–[25].

III. Methodology

This section discusses the encoding parameters of the TMIV encoder. Because TMIV considers the pixel rate as well as the coding efficiency for compatibility with existing video codecs, it is important to investigate these factors. The experiments conducted in this study were aimed at evaluating the impact of the encoding parameters on the packing strategy for removing the inter-view redundancy algorithm.

### Table I: Modified parameters in TMIV encoder configuration

<table>
<thead>
<tr>
<th>Encoding parameters</th>
<th>Experiment Configurations</th>
</tr>
</thead>
<tbody>
<tr>
<td>numGroups</td>
<td>Anchor 2G 3G</td>
</tr>
<tr>
<td>maxAtlases</td>
<td>2 2 3</td>
</tr>
<tr>
<td>minNonCodedViews</td>
<td>3 2 1</td>
</tr>
<tr>
<td>maxBasicViewFraction $\alpha$</td>
<td>$[0.25, 0.50, 0.75]$</td>
</tr>
</tbody>
</table>

Table I lists the specifications of the MIV-encoding parameters. The first parameter is the number of groups. With an increase in the number of groups, the scope of the inter-view redundancy removal was constrained, and the total bitrate increased. However, this grouping technique constrains the excessive depth estimation and view projection. As a result, MIV produces better rendering results, particularly when the quality of the depth map is low. The second parameter indicates the number of maximum atlases. The MPEG-I community defined the test condition constraints for the maximum number of simultaneous decoder instantiations in the CTCs. This is because MIVs are intended for use across a variety of platforms and devices. The number of decoder instantiations is a critical factor because parallel decoding is difficult to manage for many edge devices. Four decoder instantiations are recommended, two for texture atlases and two for depth map atlases. The minNonCodedViews parameter limits the number of basic views for each atlas. Determining the number of basic views is important to produce better packing, and this parameter preserves a meaningful objective evaluation of
The pixel rate of packing strategies was also compared with maximum luma sample rate per frame \((4096\times2048\times4=33,554,432)\), when the \(\text{maxBasicViewFraction}\) is set to \(0.5\). The results are summarized in Table IV. Although group-based encoding has a negative impact on the pixel rate owing to the constraint of inter-view redundancy, this method enables sub-bitstream extraction on the decoding side. To ensure compatibility with existing video codecs, TMIV considers the pixel rate as well as the coding efficiency, and it is therefore important to understand these aspects. As illustrated in Figure 4, different packing techniques can significantly reduce the bitrate required for transmission while preserving the objective video quality.
The experiment results show that preserving a more complete view is a significant factor in improving the rendering quality, particularly when using group-based encoding. A comparison of the subjective quality results of the packing method in the SE sequence is shown in Figure 5. The comparison results indicate when the synthesized view was rendered in the test sequences. As described above, it can be observed that the more basic views that are included in the subjective quality, the better the rendering result. As a reason for this, the more basic the views are, the smaller the number of errors occurring in the view synthesis of the patch. In particular, it is possible to observe a view synthesis error at the boundary of an object because the estimated depth map is inaccurate. These results showed the tendency according to the various packing strategies and encoding parameters.

Fig. 5: Synthesized view comparison with enlarged noticeable sections: (a), (b), (c) Frog (NC-E) v4 rendered using TMIV anchor, G2_50, and G3_50 configurations. (d), (e), (f) Fencing (NC-L) v4 rendered using TMIV anchor, G2_25, and G2_75 configurations. (g), (h), (i) Street (NC-U) v6 rendered using TMIV anchor, G3_25, and G3_75 configurations.

V. CONCLUSION

In this paper, an efficient 6DoF immersive video packing strategy for the MIV coding standard was proposed. Specifically, the efficient representation of atlases was discussed in terms of the encoding parameters. It is important to explore these parameters because TMIV considers the pixel rate and coding efficiency as factors for compatibility with the video codecs. In particular, further study is required to develop efficient adaptive streaming based on the group-based packing algorithms discussed in this paper. Several packing strategies for MIV coding were explored in this study, and the experiment results indicate that the better the rendering, the more basic the views that are included. However, further study on the 6DoF streaming scenario in terms of pixel rate is required. Further research will be conducted on immersive video compression based on these observations.

ACKNOWLEDGMENT

This research was supported by the SungKyunKwan University and the BK21 FOUR(Graduate School Innovation) funded by the Ministry of Education(MOE, Korea) and National Research Foundation of Korea(NRF)