ADAPTIVE STEREOSCOPIC 3D VIDEO STREAMING

C.Göktuğ Gürler1  K.Tolga Bağcı2  A.Murat Tekalp2

College of Engineering, Koc University, Sariyer, Istanbul, Turkey, 34450
{cgurler, kbagci, mtekalp}@ku.edu.tr

ABSTRACT

This paper presents a comparative analysis of scalable stereoscopic video coding strategies for adaptive streaming. In particular, we compare scalable simulcast coding of both views using SVC with scalable coding of one view with SVC and non-scalable coding of the other view using H.264/AVC, and benchmark them against non-scalable dependent coding of both views using the MVC. All of these coding options allow both symmetric and asymmetric coding of stereo videos. In addition, we propose a lightweight and periodic feedback mechanism for rate estimation and a strategy to adapt the total stereo source rate using SNR scalability option of SVC, while minimizing the loss rate of non-discardable packets. Experimental results show that dynamic rate scaling of only one view provides sufficient rate adaptation capability and better overall compression efficiency compared to scaling both of the views.

Index Terms— adaptive streaming, scalable coding, stereoscopic 3D video

1. INTRODUCTION

It is foreseen that 3D video will become available at homes after being introduced in movie theaters and public locations. Therefore, it is expected that, in the near future, today’s popular video streaming applications, such as WebTV, video on demand (VOD), and video conferencing, will all become available in 3D. However, achieving this still requires addressing some critical and challenging issues and problems.

The first one is the bandwidth requirement for transmitting additional data to render the auxiliary view(s). Three main methods have been proposed to reduce the total bit rate to a manageable level. The first is multi-view coding (MVC), which is an extension of H.264/AVC for exploiting redundancies among different views [1], [2]. This work has been initiated in 2006 and the final draft is accepted in 2009. The second method is video-plus-depth coding, where the auxiliary view can be rendered at the client side [3], [4] using the main view and depth information. The cost of this method is lower but there may be rendering problems due to occlusions and transparency in the source video. Lastly, asymmetric coding exploits the human visual system (HVS) which tolerates lack of high frequency components in one of the views. Asymmetric stereo coding is an active research topic where researchers are putting significant effort to achieve the best perceived visual quality using various asymmetry options [5]-[7].

Scalable video coding (SVC) enables serving video at the highest possible quality to all clients depending on their network resources. With the increase in the variety of Internet connections e.g., DSL, cable, satellite, wireless LAN and 3G, scalable video is likely to become more mainstream. While adaptive streaming and rate adaptation are also current research problems for 2D video, 3D video introduces novel possibilities, e.g., using asymmetric rate scaling in rate adaptation to exploit the HVS. Since MVC is not scalable, a plausible method for obtaining scalable stereoscopic video stream is simulcast SVC combined with asymmetry. Then, it becomes possible to achieve rate distortion (RD) performance comparable with the MVC.

In this paper, we extend our non-scalable multi-view streaming architecture [8] and propose a cross-layer adaptive stereoscopic 3D video streaming method based on simulcast SVC combined with asymmetry coding. The solution aims to achieve the best perceived 3D quality by considering features of HVS to determine the best scaling option depending on the bitrate/PSNR as described in [7]. To this effect, in Section 2, we describe our proposed encoding scheme and compare its adaptation capability with an alternative solution. We benchmark both of them against a non-scalable MVC configuration to reveal the cost of achieving scalability. Then, in Sections 3 we describe our proposed cross-layer adaptive streaming framework that aims to adapt the stereo video rate to the available network rate, while minimizing the loss rate for non-discardable packets. We provide details of the test setup which includes a bottleneck link with varying channel capacity in Section 4, which also includes the results of adaptation tests. Finally, in Section 5 we draw some conclusions.

1 This work is partially supported by European Commission FP7 ICT STREP DIOMEDES
2 A. Murat Tekalp acknowledges support from Turkish Academy of Sciences (TUBA).
Table 1: Comparison of Encoding Options: PSNR (dB) and Bitrate (Kbps)

<table>
<thead>
<tr>
<th></th>
<th>Left bitrate</th>
<th>Left PSNR</th>
<th>Right bitrate</th>
<th>Right PSNR</th>
<th>total bitrate</th>
<th>total PSNR</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Adile</strong></td>
<td>max</td>
<td>1027</td>
<td>40.4</td>
<td>370</td>
<td>40.3</td>
<td>1397</td>
</tr>
<tr>
<td></td>
<td>min</td>
<td>1027</td>
<td>40.4</td>
<td>370</td>
<td>40.3</td>
<td>1397</td>
</tr>
<tr>
<td></td>
<td>dif</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Flower Pot</strong></td>
<td>max</td>
<td>1275</td>
<td>36.1</td>
<td>1022</td>
<td>36.5</td>
<td>2298</td>
</tr>
<tr>
<td></td>
<td>min</td>
<td>1275</td>
<td>36.1</td>
<td>1022</td>
<td>36.5</td>
<td>2298</td>
</tr>
<tr>
<td></td>
<td>dif</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Train</strong></td>
<td>max</td>
<td>1812</td>
<td>36.4</td>
<td>1481</td>
<td>36.5</td>
<td>3292</td>
</tr>
<tr>
<td></td>
<td>min</td>
<td>1812</td>
<td>36.4</td>
<td>1481</td>
<td>36.5</td>
<td>3292</td>
</tr>
<tr>
<td></td>
<td>dif</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

2. ENCODING OPTIONS

Three criteria can be enlisted to choose the most suitable encoding scheme for the proposed framework. Firstly, RD performance of the proposed scheme should be comparable to MVC, which has the best compression efficiency for multi-view video (MVV) with the inter-view prediction option. Secondly, the range of scalability should be as much as possible to take full advantage at wider bandwidth range. There is a trade-off between compression efficiency and scalability so the first two criteria enforce conflicting attributes. Lastly, the encoding configuration should take advantage of the HVS to achieve the best perceived visual quality. In [7] a PSNR threshold is defined for the maximum level of asymmetry therefore this criterion determines the lower bound for the range of scalability.

Scalable video coding (SVC) standard is employed to meet the requirements mentioned above. SVC offers highly flexible configurations and allows scaling in three dimensions namely temporal, spatial and SNR. The first two dimensions aim terminal scalability while the last one is more suitable for adapting to variations in available link capacity. SVC generates one H.264/AVC compatible base layer and discardable enhancement layer(s) which can be partially transmitted in case of bandwidth scarcity [9].

We claim that encoding only one of the views using SVC and using H.264/AVC for the other view is the best method to achieve the constraints mentioned above. Table 1 presents the encoding test results for using three encoding schemes: i) MVC ii) SVC and H.264/AVC and iii) both views SVC (both views scalable) for three different test sequences that have different characteristics. For each sequence in Table 1, the first row presents the highest possible PSNR and bitrate results (with the enhancement layer if present) while the second row presents the lowest possible values. Third row presents the difference of these two and intentionally left blank for the non-scalable streams.

2.1 One View Scalable

In order to fully utilize the results in [7] the PSNR range of the scalable view is selected between the range of 32-37 dB as depicted in Fig. 1. In this configuration, if the client’s channel capacity is high enough to receive the entire enhancement layer then the scalable stream becomes the high quality pair of the asymmetry. In the other extreme, if the client cannot receive any enhancement NAL units then scalable stream becomes the low quality pair of asymmetry.

Also notice that even though only a single view is scalable the available bit rate adaptation is very close to its counterpart. Therefore, this configuration both uses asymmetric coding and offers wide scalability range.

2.2 Both Views Scalable

Using scalable coding offers a little wider scalability range but the drawback of this scheme is that due to lower compression efficiency, even when the entire enhancement layer is received the perceived visual quality is lower when compared to the first option. Base layer PSNR thresholds are set to ~32 dB while the upper limit is determined by matching the total bit rate of the first configuration.

Figure 1: One view scalability, PSNR range of views

Figure 2: Two view scalability, PSNR range of views
3. ADAPTIVE STEREOSCOPIC STREAMING

3.1 System Overview

An adaptive streaming model has to minimize the loss of non-discardable packets (which are packets of the non-scalable stream and base layer of the scalable stream), while sending as many enhancement packets from the scalable stream. We note that sending each enhancement NAL unit may increase the risk of congestions in IP layer interfaces; hence, increasing the likelihood for packet losses.

Using forward error correction (FEC) for non-discardable packets provides some protection against packet losses. In our model, we use 10% FEC to protect non-discardable streams, which can erase 3% to 6% packet loss rates creating a safety margin for the transmission of enhancement layer. Optimization of the ratio of FEC [10] is beyond the scope of this work.

3.2 Rate Estimation using Client Feedback

Streaming applications tend to use UDP rather than TCP which nullifies the application layer fragmentation (ALF) [11] concept and causes inevitable fluctuations in data transmission rate due to its additive increase multiplicative decrease policy. On the other hand UDP does not have rate feedback capability while in TCP and DCCP it is easy to determine the state of the channel. In this section we propose a light weight rate estimation method that works based on periodic packet loss feedback from the client. If TCP or DCCP is employed as the transport protocol, then one can use their rate estimation as well.

We define three different cases based on the feedback using two threshold values $\lambda_1$ and $\lambda_2$. $\lambda_1$ separates the rate of packet losses due to lack of link capacity from the transient ones that are due to random early discard (RED) policy it is set to the FEC recovery ratio. That creates a safe region for non-discardable packets. $\lambda_2$ is the threshold for negligible packet losses set to $\sim 1\%$ which is the maximum loss ratio that the system responds gently as opposed to TCP that cuts the rate of transmission drastically.

Moreover a third parameter $\gamma$ is used as memory similar to the TCP’s threshold value to prevent rapid changes in the transmission rate. Based on these parameters, our scaling strategy is as follows: i) If the loss rate at the client side is higher than $\lambda_1$, server decreases the transmission rate by the loss rate and the threshold is set to the new rate. This corresponds to linear decrease in transmission rate instead of drastic cuts forced by TCP; ii) If the loss rate is between $\lambda_1$ and $\lambda_2$, which is the rate that non-scalable NAL units are recovered and while some of the enhancement NAL units also transmitted to the client, the transmission rate is not modified. iii) If the loss is lower than $\lambda_2$, then two sub-cases are considered. If the threshold is higher than the rate that is required for the stream at full quality then the new rate is set to the threshold and the threshold is increased by 10% so that oscillations are avoided. Otherwise, the transmission rate and the threshold are not modified.

3.3 Extraction

If the available bit rate allows only partial transmission of the enhancement layer then the choice of which enhancement NAL units should be discarded in the extraction process becomes important for achieving the highest possible video quality. One basic approach is to discard the enhancement layers of each frame equally until the requested bit rate is achieved. However, this approach is not optimal in terms of rate distortion (RD) performance of the extracted video. We employ an extraction method similar to [12] in which the dependency hierarchy within a GOP is taken into account. In this method, first the enhancement NAL units of I-frames are transmitted. Then, the enhancement layer of B-frame with the highest priority, e.g., frame 8 when GOP size 16, is transmitted and this process continues until the target bit rate is reached.

4. TESTS AND RESULTS

4.1 Test Setup

Adaptive server and stereoscopic client are connected through a link, which is controlled by a Linux box using network emulation (Netem) commands [13]. Netem provides the rate control by adjusting the output rates of two bridged Ethernet interfaces as depicted in Fig. 3. In this scenario, stereoscopic content is streamed over controlled local area network (LAN) in which the available channel capacity is changed randomly and also transient packet additional 1% packet loss is introduced. Then the adaptation of the system is observed.

![Figure 3: Network Structure](image)

4.2 Results

The server responds firmly to the changes in available channel capacity with some delay as depicted in Fig. 4. The delay that occurs in the case of bit rate reduction is due to queueing in the router since it starts dropping packets only when the queue is full. In the other case, the server responds cautiously to the increase in the available network because rapid changes in the send rate leads to oscillations and decrease the throughput of the channel. Notice that the system can maintain a steady rate and also adapts to the available bit rate within a time limit.

Using error correction for the non-discardable streams significantly increases the achieved PSNR rates even though...
it reduces the number of received enhancement NAL units because the absence of base layer nullifies the presence of enhancement layer. The resultant PSNR values averaged over one second for the scalable stream are depicted in Fig 5. The results for the non-scalable view are lightly affected by changing channel capacity, thus omitted for simplicity.

5. CONCLUSIONS

Careful analysis of experimental results leads to the following conclusions.

- For adaptive stereoscopic 3D streaming, scalable coding of only one view achieves better compression efficiency than scaling both views and therefore can deliver better perceived visual quality especially when asymmetric encoding is employed.
- When combined with asymmetric encoding, compression efficiency of the proposed simulcast solution with one view scalable is close to that of non-scalable MVC coding.
- Actual streaming tests over an emulated network environment show that proposed solution with one view scalable offers adequate adaptation capability to match the available network bit rate.
- If scalability is implemented in the reference software of MVC, the adaptation strategy proposed in this paper is still applicable to scalable MVC streaming.

6. ACKNOWLEDGMENTS

The authors would like to thank Burak Görkemli for his help with the Network Emulation module.

7. REFERENCES