Dynamic adaptive 3D multi-view video streaming over the internet

Conference Paper - October 2013
DOI: 10.1145/2512142.2512143

3 authors, including:
E. Ekmekcioglu
Loughborough University
62 PUBLICATIONS 462 CITATIONS

Some of the authors of this publication are also working on these related projects:

AfectTeach View project
puntos de calor View project
Dynamic Adaptive 3D Multi-View Video Streaming over The Internet

Cagri Ozcinar, Erhan Ekmekcioglu, and Ahmet Kondoz
I-Lab Multimedia Communication Research - University of Surrey
Guildford, UK
{c.ozcinar, e.ekmekcioglu, a.kondoz}@surrey.ac.uk

ABSTRACT
Increasing throughput rates and technical developments in video streaming over the Internet offer an attractive solution for the distribution of immersive 3D multi-view. Nevertheless, robustness of video streaming is subject to its utilisation of efficient error resiliency and content aware adaptation techniques. Dynamic network characteristics resulting in frequent congestions may prevent video packets from being delivered in a timely manner. Packet delivery failures may become prominent, degrading 3D immersive video experience significantly. In order to overcome this problem, a novel view recovery technique for 3D free-viewpoint video is introduced to maintain 3D video quality in a cost-effective manner. In this concept, the undelivered (discarded) views as a result of adaptation in the network are recovered with high quality at the receiver side, using Side Information (SI) and the delivered frames of neighbouring views. The proposed adaptive 3D multi-view video streaming scheme is tested using Dynamic Adaptive Streaming over HTTP (MPEG-DASH) standard. Tests using the proposed adaptive technique have revealed that the perceptual 3D video quality under adverse network conditions is significantly improved thanks to the utilisation of the extra side information in view recovery.

Categories and Subject Descriptors
H.5.1 [Multimedia Information Systems]

Keywords
3D Multi-view coding, 3D video coding, 3D video streaming, adaptive 3D video streaming

1. INTRODUCTION
Video streaming over the Internet is gaining increasingly more popularity and occupies a considerable part of the overall Internet traffic. As the demand for Internet video increases, the networks are becoming increasingly heterogeneous too, which adds more complexity and new challenges to video streaming.

3D stereoscopic services are in the process of changing the landscape of video industry, which has already been provided by several broadcasters. The Wimbledon Championship in 2011 and the opening ceremony of the 2012 London Olympics were broadcast in 3D over Digital Video Broadcasting (DVB). Nevertheless, the necessity to wear special glasses and limited viewing angle are the main issues with 3D stereoscopic video. Multi-view video, on the other hand, enables glasses free 3D from a wider viewing angle on an auto-stereoscopic display, which is much closer to provide a truly immersive 3D media experience. However, more viewpoints require more bandwidth. Since the current Internet architecture suffers to provide enough bandwidth capacity for such services at all times, more efficient compression and adaptive transmission schemes are the key elements in accommodating high quality 3D multi-view video services.

Failures are inevitable in video streaming due to network congestion and latencies. Accordingly, the Quality of Experience (QoE) can be significantly degraded when lost video packets cannot be recovered properly at the receiver. Nowadays, video delivery using HTTP has gained attention [8]. Unlike video services using RTP/UDP, HTTP that is based on TCP, but not as delay-sensitive as RTP/UDP, is not vulnerable to packet losses.

To limit the total load on the network in 3D multi-view video streaming, Depth Image Based Rendering (DIBR) can be utilised at the receiver to estimate additional views. Within Moving Picture Experts Group (MPEG), a reference view synthesis scheme has been developed, called View Synthesis Reference Software (VSRS) [6].

In this work, a novel adaptive 3D multi-view video streaming method for 3D multi-view displays is proposed. The goal is to deliver 3D multi-view video to clients in a cost-effective manner by utilising a small amount of pre-calculated Side Information (SI), which is generated from the statistical correlation of the compressed views. The SI has substantially lower overhead compared to the compressed view streams and is delivered instead of compressed views at periods of adaptation. During bandwidth adaptation, undelivered views are recovered using the SI at the receiver. To test the performance of the proposed adaptive multi-view streaming scheme, the recently emerging Dynamic Adaptive Streaming over HTTP (MPEG-DASH) standard has been used [9]. Nevertheless, the proposed adaptive 3D multi-view streaming concept is applicable to other delivery systems, including P2P video delivery.
The remainder of this paper is organised as follows: Section 2 provides a brief overview of 3D video coding techniques. Section 3 presents the overview of the proposed adaptive 3D multi-view video streaming system. The details of the proposed SI generation method at the transmitter, the view recovery method at the receiver, and the proposed dynamic adaptation method are explained in Section 4. Finally, Section 5 presents the experimental results, which is followed by the concluding remarks in Section 6.

2. BACKGROUND

2.1 3D Video Coding

3D Video coding can be categorised into two main standards, which are explained below:

2.1.1 H.264/MPEG-4 Part 10 Advanced Video Coding (AVC)

This standard has been established jointly by the ISO/IEC MPEG and ITU-T Video Coding Experts Group (VCEG) and Merke et al. [13] have reported nearly 50% bit-rate savings with this standard for the same perceptual quality compared to the prior standard (MPEG-2). More detailed information can be found in [14].

The straightforward solution in 3D multi-view video transmission is to independently encode each view before transmission, which is known as AVC simulcast coding.

2.1.2 Multi-view Video Coding (MVC) extension of AVC

In Free-Viewpoint Video (FVV) systems, more than two views have to be transmitted to the receiver. As the number of views increases, inter-view statistical dependency will increase, because all views represent the same scene from varying perspectives. However, the redundant information between views can be minimised using inter-view prediction. Accordingly, in order to improve 3D multi-view video coding efficiency by taking advantage of inter-view statistical dependencies, Multi-View Coding (MVC) standard has been established as an extension of AVC.

MVC usually yields higher compression efficiency than AVC simulcast coding. However, the bit-rate necessary is still linearly dependent to the number of viewpoints to be delivered [5].

2.2 3D View Synthesis

Multi-view video format is suitable for generic 3D video solutions, where a number of views and their associated depth maps are transmitted, while other views (virtual and/or real) needed for a 3D display are estimated using DIBR technology. The operation of DIBR is based on a coordinate conversion process, in which a reference view is mapped to the target view’s coordinates using the camera parameters. Camera parameters express the relationship between the coordinate system of the camera array and the 3D world.

During the DIBR process, some pixels in the reference view may not be mapped due to rounding of the pixel position to the nearest integer. Accordingly, some pixel positions remain unfilled in the synthesised view, which are usually cause visual distortion. Another visual distortion reason is occlusion, where regions that are observable in the reference view, but not in the target view.

On the other hand, there are some research works for improving view synthesis performance. Cheng et al. [1] proposed a spatially and temporally consistent view synthesis method from video-plus-depth video sequences, where an iterative re-weighing framework was presented by jointly considering texture and depth map temporal consistency. This method does not only achieve temporal consistency, but also reduces noise disturbance.

2.3 3D Video Bit-rate Adaptation

In order to ensure QoE, 3D video streaming applications should employ intelligent methods to minimise the effect of inevitable network related problems, such as re-transmission of lost packets, bit-rate adaptation, scalability, and error concealment.

Packet re-transmission is one of the basic mechanisms to mitigate network failures. However it is not a suitable approach for real-time delay-sensitive video streaming. Additionally, frequent re-transmissions may lead to severe network congestion which is not desirable.

Another approach is the bit-rate adaptation for the varying network conditions. Gurler et al. [3] focused on scaling the quality of multiple views symmetrically and asymmetrically. Also the feasibility of view scaling has been investigated in this work. A set of views were transmitted and then the undelivered views were interpolated using 3D view synthesis at the receiver. To investigate the best rate adaptation strategy in terms of QoE, subjective tests were conducted for this work. Accordingly, the results claimed that intermediate views in 3D multi-view should be compressed as much as possible. When the average Peak-Signal-to-Noise-Ratio (PSNR) threshold for intermediate views is reached, only the edge views should be transmitted. In this work, undelivered views were recovered using 3D view synthesis. However, the performance of view recovery is influenced by the view synthesis distortion factors outlined in Section 2.2.

The recent MPEG-DASH standard allows dynamic adaptive streaming over the Internet [9], using HTTP as its underlying application protocol. The use of HTTP offers several advantages. It is firewall-friendly, since almost all firewalls are configured to support HTTP connections. Besides, it is server-friendly, as the streaming sessions are managed by the DASH client. The video content is divided into equal temporal length segments which can be compressed at different quality/resolution and hence at different bit-rates, is stored in a HTTP server along with a XML-based Media Presentation Description (MPD) file. This file describes a manifest of the available content, including various quality levels, spatial resolutions, viewpoints that are available, and their URL addresses. To play the content, the client starts by downloading the MPD file, and then requests the compressed video segments through the HTTP server. The system is pull-based, where the DASH client has full control over the streaming bit-rate by deciding on which segments to download. Every segment corresponds to a different quality representation of the content.

3. SYSTEM MODEL

A schematic overview of the proposed system is shown in Figure 1. For the experiments to reveal the performance of the adaptive multi-view streaming method, a HTTP server and a DASH client are used. Multi-view with depth is en-
coded at various bit-rates (using AVC simulcast coding or other methods described in Section 2.1) and the produced streams are divided into segments of equal temporal length. All segments are stored in the HTTP server. MPD file is also stored on the HTTP server, which contains the adaptation strategy that will be described in Section 4.3. In addition, the SI, which is utilised to recover undelivered views with high quality at the receiver side, is generated to be delivered to the DASH client using HTTP GET requests [2]. The codebook, which will be described in Section 4.2, is created as a result of coefficient training and downloaded by the DASH client during the start-up buffering period.

In order to generate the SI in the presented work, a DIBR engine has been integrated within the encoding server. The input to the DIBR engine are the encoded and reconstructed segments. For each view that can potentially be discarded as a result of the instantly available network capacity, the DASH client continues to request the segments of the views and also monitors the network throughput. Depending on the network throughput and its quality-cost criterion, the DASH client decides on whether to adapt by requesting a sub-set of views, or not. The information needed to recover the discarded views as a result of adaptation is estimated according to a coefficient estimation and codebook design process, the details of which are given in Section 4.2, and delivered as SI. The SI is requested by the client following the adaptation process.

4. VIEW RECOVERY METHOD FOR 3D - MULTI-VIEW VIDEO STREAMING

The proposed view recovery model is divided into three sections that are described in this section.

4.1 Quadtree based adaptive block-size selection

The proposed system divides the frames into blocks of variable size using a quadtree-based, adaptive block size selection framework, which has been widely used for block partitioning in video coding and image processing applications [11]. The view rendering coefficients, the details of which are going to be described in Section 4.2, are calculated on these adaptively sized blocks.

In the server, different block sizes are adaptively evaluated in a progressive manner. Based on experimental analysis, $8 \times 8$ and $32 \times 32$ block sizes are chosen as the smallest and largest block sizes, respectively. A block is divided into four equal size blocks starting from the largest block size in a top-down approach. An exemplary quadtree structure and the related codeword values are shown in Figure 2.

4.2 Coefficient estimation and codebook design

In order to obtain the optimal quadtree structure, the overall block distortion $D$ is minimised, subject to a limited overall SI bit-budget, using a Lagrangian cost minimisation method. When the cost for the transmission of the quadtree structure is included, the overall cost function becomes

$$J(x) = D(x) + \lambda_1 \cdot l + \lambda_2 \cdot Q(x)$$  \hspace{1cm} (1)

Where $D$ is the Mean Squared Error (MSE) based distortion metric, $x$ is the block number in the view, $\lambda_1$ and $\lambda_2$ correspond to the Lagrangian multiplier values for SI and quadtree structure bit-costs, respectively. $\lambda_1$ and $\lambda_2$ are obtained following subjective trainings. $Q(x)$ denotes the number of bits (quadtree code-length) required for the signalling of the quadtree structure. $l$ is the number of bit required for the SI, which is explained in the following Section 4.2. In this work, MSE based distortion metric have been utilised. However, other perceptual quality metrics (e.g Synthesized View Distortion Change (SVDIC) [12], Spatial Peak Signal to Perceptual-Noise Ratio (SPSPNR) [15]) are equally employable with the proposed method.

Figure 2: Structure of variable block size and quadtree code-word

In order to obtain the optimal quadtree structure for estimating the target view to be discarded from the neighbouring views, a cross correlation method [10] is applied to calculate the best set of weighting coefficients to be used in view recovery. In the recovery process of the discarded view, its neighbouring delivered views from both directions are utilised, as shown in Figure 3.

To calculate a set of weighting coefficients to be used in view recovery, the target view(s) to be discarded during adaptation are estimated in the HTTP server first. Number of discarded views is decided based on the currently available bandwidth, and also which view(s) to be discarded is decided based on a quality-cost criterion. Blocks of each undelivered view are estimated by the weighted summation of all corresponding blocks from the delivered views using:

$$y_N(x,y,t) = \sum_{k=0}^{M} [\tilde{B}_k(x,y,t) \cdot a_k] + B_N(x,y,t_{ref}) \cdot a_N$$  \hspace{1cm} (2)

Where $M$ is the total number of available views in the multi-view video, $y_N(x,y,t)$ represents the estimated pixel luminance at $(x, y, t)$ within the target view $N$, $t$ represents the frame number. $B_N$ is the temporal reference of the target view, which is the last frame in the corresponding previous temporal segment. $\tilde{B}_k$ represents the projected block of the $k_{th}$ view, and $a$ represents the weighting coefficient of each block for each view, which corresponds to the SI.

Each weighting coefficient ($a$) form coefficient vectors in the codebook, which is encoded using an $l$-bit codeword, as...
shown in (3), where coefficient vector \( A \) is chosen from a finite set of coefficient vectors in the codebook of size equal to \( L \), which is \( L = 2^L \). Also, the index number denoted as \( i \) (\( 1 \leq i \leq L \)).

\[
A_i = [a_{N-k} \ a_{N-1} \ ... \ a_{N+k}]
\]  

(3)

In order to optimise the estimated coefficients (i.e., the computed weighting coefficient per block), k-means clustering algorithm [4] has been used in the codebook design. The index numbers of each computed coefficient vector corresponding to each computed block is put into the SI stream. The codebook is downloaded from the HTTP server by the DASH client at the beginning of streaming, during the startup buffering period. The client then parses the codebook index values embedded in the SI stream to recover the corresponding coefficient vectors from the pre-downloaded codebook.

4.3 Dynamic view adaptation

In order to cope with varying network conditions during 3D multi-view video streaming, two possibilities exist in the outlined system architecture. It is possible to symmetrically adapt the bit-rates of all streamed views to match the instantaneous bandwidth [3], since all views are encoded at several quality levels in the HTTP server. However, this approach would lead to a reduction in the reconstruction quality of all received views and depth maps, and hence the reduction in the perceptual 3D multi-video quality. The proposed approach is to selectively request the segments of a subset of views at the highest possible quality, accompanied by the low-overhead SI data, and let the undelivered views be estimated at the receiver using the described view recovery method at high quality. In contrast to the former approach, this would not result in compromising the perceptual quality of all views, but maintain the highest possible quality at the receiver end, owing to the extraction and usage of optimised SI.

The proposed dynamic view adaptation scheme is shown in Figure 4, which is based on the latter approach, i.e. selectively requesting views according to the available bandwidth. First, each view is encoded and divided in multiple temporal segments. Afterwards, the best set of viewpoints to be transmitted is computed for each temporal segment in the priority classification, such that the reconstruction quality of discarded segments using the transmitted SI is maximised, and then these segments are assigned as a higher priority. If the available network bandwidth cannot accommodate the transmission of all segmented views, low priority segments are intelligently discarded to be recovered using delivered views and the SI at the receiver. The goal is to maximise the perceived quality at the client, based on the available bandwidth.

5. RESULTS

In this section, the performance of the proposed view recovery based adaptive 3D multi-view streaming is tested. MPEG’s reference view synthesis tool VSRS (v3.5) has been used as the base to estimate the undelivered views along with the received SI. Subjective as well as objective tests results in terms of PSNR are reported in this section. Double-stimulus continuous quality scale (DSCQS) method has been used in subjective tests [7]. Fifteen observers were asked to participate in the subjective assessment. Observers were asked to grade visual quality using five-grade scale (Excellent, Good, Fair, Poor and Bad). In addition, five consecutive viewpoints were chosen from three multi-view test sequences for the streaming tests, as shown in Table 1.

Table 1: 3D Multi-view Sequences

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Num. of frames</th>
<th>Resolution</th>
<th>Selected views</th>
</tr>
</thead>
<tbody>
<tr>
<td>Café</td>
<td>160</td>
<td>1920×1080</td>
<td>1,2,3,4,5</td>
</tr>
<tr>
<td>Newspaper</td>
<td>160</td>
<td>1024×768</td>
<td>2,3,4,5,6</td>
</tr>
<tr>
<td>Bookarrival</td>
<td>100</td>
<td>1024×768</td>
<td>6,7,8,9,10</td>
</tr>
</tbody>
</table>

Joint Scalable Video Model (JSVM v9.19) was used to encode each viewpoint as a single-layer AVC compatible bitstream. Group of Picture (GoP) size was set to 16, and each coding segment contains a single GoP. Also, each viewpoint has been encoded using six different Quantisation Parameters (QPs): 22, 26, 30, 34, and 38. Coefficient vectors in the codebook were estimated only for the first 100 frames of each test sequence, based on the process explained in Section 4.1 and 4.2.

5.1 View estimation performance

To investigate and validate the proposed view estimation performance, MPEG VSRS was used as a view estimation reference. In this reference, two nearest left and nearest right neighbour views are warped, one from the left view and another one from the right view. The two projected images are then blended to form one synthesised image. The pixel value in the synthesised image is formed by blending the respective pixel values in the projected images with unequal weights, where the weights are inversely proportional to the distance from the target view.

In Figure 5 R-D curves are depicted for the three test sequences, where the same viewpoint was discarded at a time,
and the average estimation quality of all discarded views was calculated. The bit-rate includes the overhead caused by transmitting the SI and the delivered viewpoints. The PSNR of the estimated views was calculated with respect to the uncompressed discarded view. Also, 8 bit codebook was used in the receiver. Accordingly, 13.6%, 8.6%, and 2% SI overhead is obtained respect to overall transmission bit-rate over a wide range of used test points for *Café*, *Newspaper* and *Bookarrival* sequences, respectively.

As can be seen in Figure 5, the proposed view recovery technique outperforms the MPEG VSRS, by 4.6 dB, 7.6 dB and 3 dB on average for the three sequences, over a wide range of the test points.

**5.2 Performance analysis of the proposed adaptive video streaming method**

The performance of the proposed adaptive streaming method is evaluated subjectively as well as objectively in this section. Figure 6 shows the bandwidth availability test patterns related to three test conditions, which are used for two different video resolutions, *Bookarrival*, and *Newspaper* were tested with Figure 6 (a), *Café* was tested with Figure 6 (b). Furthermore, in Figure 6, Test 2 and Test 3 correspond to fixed bandwidth cases throughout the streaming. Test 1 corresponds to the varying bandwidth availability case during the course of streaming.

In this analysis, in order to investigate the performance of the proposed view recovery method with instantaneous changes in the network throughput, 3D multi-view video temporal segments are selectively discarded based on the method described in this paper and SI is transmitted. Furthermore, MPEG VSRS was used as a reference, in which temporal segments were selectively discarded based on their synthesis quality, and then discarded segments were recovered using MPEG VSRS. In order to obtain higher recovery performance, only intermediate views were discarded based on the network throughput. The view discarding order in Test 1 (variable bandwidth case) for both methods, are shown in Table 2.

Furthermore, Table 3 shows the comparison results in terms of PSNR and subjective grades, which were reported respectively. Video streaming was calculated with respect to the reconstructed missing view.

As can be seen in Table 3, the proposed adaptation method consistently outperforms the reference method, both objectively and subjectively.
Table 2: View discarding order for varying bandwidth test case

<table>
<thead>
<tr>
<th>Segment Range (0.5sec/segment)</th>
<th>1-2</th>
<th>2-3</th>
<th>3-4</th>
<th>4-5</th>
<th>5-6</th>
<th>6-7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method</td>
<td>Seq. Discarded viewpoint number(s)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Proposed</td>
<td>Café</td>
<td>4</td>
<td>3</td>
<td>3.4</td>
<td>3.4</td>
<td>4</td>
</tr>
<tr>
<td>Proposed</td>
<td>Newspaper</td>
<td>3</td>
<td>3</td>
<td>3.5</td>
<td>3.5</td>
<td>3</td>
</tr>
<tr>
<td>Proposed</td>
<td>VSRS</td>
<td>8</td>
<td>7</td>
<td>8.9</td>
<td>7.8</td>
<td>7.9</td>
</tr>
</tbody>
</table>

Table 3: Comparison of the adaptation strategy

<table>
<thead>
<tr>
<th>Seq. Method</th>
<th>Quality</th>
<th>Test Conditions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Café</td>
<td>PSNR (dB)</td>
<td>Test 1</td>
</tr>
<tr>
<td>Proposed</td>
<td>39.95</td>
<td>39.58</td>
</tr>
<tr>
<td>VSRS</td>
<td>34.57</td>
<td>31.93</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Newspaper</th>
<th>PSNR (dB)</th>
<th>Sub. Grade</th>
<th>Test Conditions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proposed</td>
<td>39.13</td>
<td>Excellent</td>
<td>Good Excellent</td>
</tr>
<tr>
<td>VSRS</td>
<td>31.1</td>
<td>Good</td>
<td>Excellent Fair</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Bookarrival</th>
<th>PSNR (dB)</th>
<th>Sub. Grade</th>
<th>Test Conditions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proposed</td>
<td>40.77</td>
<td>Excellent</td>
<td>Good Excellent</td>
</tr>
<tr>
<td>VSRS</td>
<td>37.16</td>
<td>Good</td>
<td>Fair</td>
</tr>
</tbody>
</table>

6. CONCLUSIONS

This paper presented a novel dynamic adaptation method for 3D video streaming over the Internet. The method aims to minimize the perceived 3D multi-view video quality degradation resulting from instantaneous changes in the network throughput. In the proposed method, in order to recover undelivered views, extra Side Information (SI) is calculated using neighbouring views in the transmitter side and delivered to the client. Also, the view synthesis process is modified to take into account the received side information for improved view recovery performance. In order to help facilitate a quality-aware bandwidth adaptation mechanism, the best sets of views to discard are computed for various bandwidth conditions, in order to obtain the best overall multi-view reconstruction quality at the client side.

The proposed adaptive 3D multi-view video streaming method was evaluated using a prototype HTTP streaming client and it was compared to the reference technique that uses MPEG’s reference view synthesis software (VSRS). The experimental results have shown that significant quality improvements are obtained compared to the reference technique under challenging network condition. As a future work, the adaptation strategy will be tested on multi-user P2P multi-view video dissemination system. It is also anticipated to integrate the proposed multi-view adaptation framework with perceptual multi-view QoE metrics.

7. ACKNOWLEDGMENTS

This work was supported by the ROMEO project (grant number: 287896), which was funded by the EC FP7 ICT collaborative research programme.

8. REFERENCES