ADAPTIVE STREAMING OF SCALABLE STEREOSCOPIC VIDEO OVER DCCP

Nukhet Ozbek[1], Burak Gorkemli[2], A. Murat Tekalp[2], and Turhan Tunali[1]
[1]International Computer Institute, Ege University, 35100 Bornova, Izmir,Turkey
[2]College of Engineering, Koc University, 34450 Sariyer, Istanbul, Turkey

ABSTRACT

We propose a new adaptive streaming model that utilizes DCCP in order to efficiently stream stereoscopic video over the Internet for 3DTV transport. The model allocates the available channel bandwidth, which is calculated by the DCCP, among the views according to the suppression theory of human vision. The video rate is adapted to the DCCP rate for each group of pictures (GoP) by adaptive extraction of layers from a scalable multi-view bitstream. The objective of the streaming model is to maximize perceived quality of the received 3D video while minimizing the number of possible display interrupts. Experimental results successfully demonstrate stereo video streaming over DCCP on wide area network.

Index Terms— Scalable stereo video coding, adaptive layer extraction, inter-view rate adaptation, streaming over DCCP.

1. INTRODUCTION

3D/multi-view video and free viewpoint video are new types of media for next generation broadcast TV and streaming applications. Multi-view video streaming over the Internet requires effective inter-view rate allocation and rate adaptation strategies in order to maximize the perceived quality of the final 3D presentation while satisfying some transport constraints.

Today, the most widely used transport protocol for media/multimedia is the Real-time Transport Protocol (RTP) over UDP [1]. However, RTP/UDP does not contain any congestion control mechanism and, therefore, can lead to congestion collapse when large volumes of multi-view video are delivered. The Datagram Congestion Control Protocol (DCCP) [2] is designed as a replacement for UDP for media delivery, running directly over the Internet Protocol (IP) to provide congestion control without reliability. DCCP can be thought as TCP minus reliability and in-order packet delivery, or as UDP plus congestion control, connection setup, and acknowledgements.

To this effect, we propose a new adaptive streaming model that utilizes DCCP in order to efficiently stream stereoscopic video over the Internet. The model allocates the available channel bandwidth, which is calculated by the DCCP, among the views according to the suppression theory of human vision. The video rate is adapted to the DCCP rate for each group of pictures (GoP) by adaptive extraction of layers from a scalable multi-view bitstream. The objective of the streaming model is to maximize perceived quality of the received 3D video while minimizing the number of possible display interrupts.

Inter-view rate allocation shall be based on the well-known observation that for appropriate 3D perception from stereo video, the right and left views need not be encoded with full temporal, spatial, and SNR resolutions [3]. Hence, one of the views can be sent with full resolution, whereas spatial, temporal and/or SNR resolution of other view(s) can be dynamically adapted according to video content and network conditions, where scalable encoding can be done once and off-line [4].

In the following, Section 2 gives a brief summary of scalable stereoscopic video coding. Section 3 introduces DCCP together with its interaction with the video streaming model. Section 4 explains the proposed streaming model in detail. Section 5 presents results of streaming experiments over wide area network. Conclusions are drawn in Section 6.

2. SCALABLE STEREO VIDEO CODING

There are many research and standardization activities for stereoscopic video compression based on exploiting inter-view redundancy. Recently, the Joint Video Team (JVT) has started working on standardization of an H.264/AVC based approach for multi-view video coding, where new prediction structures and processing tools are being investigated for efficient multi-view video coding (MVC) [5]. A reference encoder-decoder, called Joint Multi-view Video Model (JMVM) [6] is publicly available, which employs hierarchical B pictures within each view, as well as a hierarchy between views for inter-view prediction. However, JMVM does not support scalable coding.

A scalable multi-view video coder (SMVC) is developed in [7] as an extension of the JSVM reference software [8] by sequential interleaving of the first (right) and second (left) views in each GoP. The prediction structure, where the first view is only temporally predicted, supports adaptive temporal or disparity compensated prediction by using existing SVC MCTF structure without the update steps. Every frame in second view uses past and future frames

from its own view and the same frame from the first view for prediction. In each GoP, the key frame of the first view is encoded as Intra, while the key frame of the second view uses just inter-view prediction to allow receiving any view at some desired temporal resolution.

In order to recover the last temporal layer as the left view the bit stream extractor and decoder modules of the JSVM are modified accordingly. Since we have two views, the effective GoP size reduces to half of the original GoP size of JSVM, namely, number of temporal scalability levels is decreased by one. The spatial and SNR scalability functionalities of the JSVM remain unchanged.

3. THE DCCP

The Datagram Congestion Control Protocol (DCCP) is a transport protocol that implements bi-directional unicast connections of congestion-controlled, unreliable datagrams. Despite of the unreliable datagram flow, DCCP provides reliable handshakes for connection setup/teardown and reliable negotiation of options [2].

Besides handshakes and feature negotiation, DCCP also accommodates a choice of modular congestion control mechanisms. There exist two different congestion control schemes defined in DCCP currently, one of which is to be selected at connection startup time. These are TCP-like Congestion Control [9] and TCP-Friendly Rate Control (TFRC) [10].

TCP-like Congestion Control, identified by Congestion Control Identifier 2 (CCID2) in DCCP, behaves similar to TCP's Additive Increase Multiplicative Decrease (AIMD) congestion control, halving the congestion window in response to a packet drop. Applications using this congestion control mechanism will respond quickly to changes in available bandwidth, but must tolerate the abrupt changes in congestion window typical of TCP.

On the other hand, TFRC, which is identified by CCID3, is a form of equation-based flow control that minimizes abrupt changes in the sending rate while maintaining longer-term fairness with TCP. It is hence appropriate for applications that would prefer a rather smooth sending-rate, including streaming media applications with a small or moderate receiver buffer.

Determination of the TFRC rate: During its operation, CCID3 calculates an allowed sending rate, called TFRC rate, by using the TCP throughput equation given in, which is provided to the sender application upon request. The sender may use this rate information to adjust its transmission rate in order to get better results.

In our video streaming model, we employ CCID3 as the DCCP congestion control mechanism. The TFRC rate that is calculated by CCID3 is used by the sender in extracting the appropriate layers or parts of layers (in case of FGS layers) from a scalable multi-view video that are later sent to the receiver.

4. ADAPTIVE STREAMING OVER DCCP

The proposed scalable stereo video streaming model consists of two nodes, a sender and a receiver, connected through a bottleneck link, as shown in Fig. 1 [11].

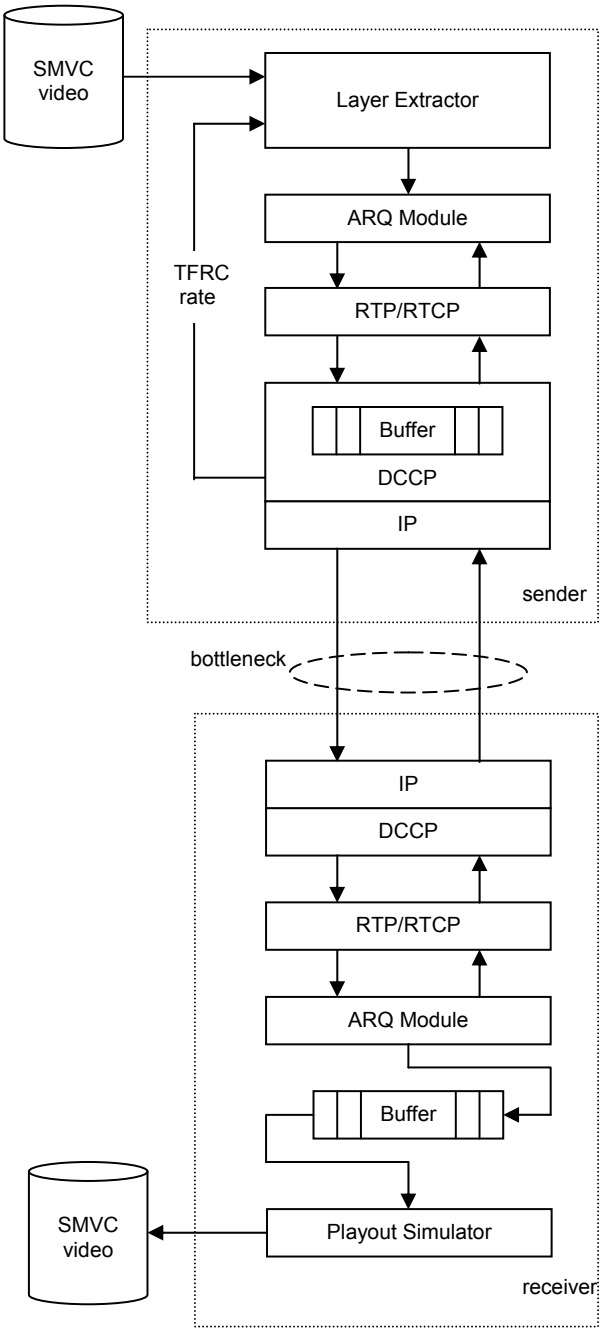

Fig. 1 The proposed scalable stereo video streaming model.

The layer extractor module in the sender takes the TFRC rate from the DCCP module periodically, and extracts those layers from the SMVC coded bitstream at the most appropriate spatio-temporal resolution for each GoP to match the TFRC rate. Next, the sender packetizes extracted

video data according to [12, 13], and sends the RTP packets to the receiver over DCCP.

Upon receiving packets, the receiver depacketizes RTP data and places the video payload to the playout buffer. Once the number of packets in the buffer exceeds a certain threshold, the playout simulator starts fetching packets from the buffer depending on the video frame rate and writes the fetched data to a file. The resulting file is later fed to the SMVC decoder.

The extractor module implements stereo video rate adaptation using combined scalability options and determines the best rate allocation between the right and left views based on the psycho-visual redundancy effect of human stereo vision. That is, the first view is always extracted at full temporal, spatial and SNR resolution, while the rate of the second view is adapted to match the TFRC rate by extracting the desired number of temporal, spatial and FGS layers. In particular, the first view is always extracted at 30 Hz, SIF, and Qp=28, while the second view can be extracted using one of the options given in Table 1.

A particular method for selection of the best rate allocation between views has been described in [4], where a new quantitative measure for 3D video quality was proposed by using a weighted combination of two PSNR values and a jerkiness measure. Experimental results reported in [4] indicate that the seven scalability options can be reduced to three choices (namely Options 1, 3 and 7) without any performance loss.

Table 1: Scaling options for the second view.

Option	Quality for the Option	Corresp. Resolution
OPT1	full spatial, full temporal, full SNR	SIF, 30 Hz, QP=28
OPT2	full spatial, ½ temporal, full SNR	SIF, 15 Hz, QP=28
OPT3	full spatial, full temporal, base SNR	SIF, 30 Hz, QP=34
OPT4	base spatial, full temporal, full SNR	QSIF, 30 Hz, QP=28
OPT5	full spatial, ½ temporal, base SNR	SIF, 15 Hz, QP=34
OPT6	base spatial, ½ temporal, full SNR	QSIF, 15 Hz, QP=28
OPT7	base spatial, full temporal, base SNR	QSIF, 30 Hz, QP=34

We define three total quality layers for our streaming experiments: i) Base total quality layer, includes the first view at full quality together with the second view at the base quality which is the base SNR of base spatial layer. ii) Enhancement1 total quality layer, includes the first view at full quality together with second view at the base SNR of full spatial layer. iii) Enhancement2 total quality layer, corresponds to both the first and second view at full quality. In other words, the minimum quality for the second view is the base SNR-base spatial layer, and enhancement spatial layer is added as bandwidth allows. Further increase in quality is achieved by also adding the enhancement SNR layer.

Since we perform streaming experiments on the real Internet, packet losses are inevitable. In case of a base-layer packet loss, the receiver requests the missing packet from the sender through the ARQ Module. This module talks with its peer at the sender side by using the RTCP Receiver Reports. The sender repeats sending the missing base-layer packets until the playout deadline allowed by the receiver buffer model, after which the packet is declared lost and not resend anymore. Packets that belong to enhancement layers are not resend, in case of loss. Another reason for packet losses may be receiver video buffer overflow. Should the playout buffer be full at an instant, the receiver simply discards the received packets. The same is true for late coming packets. On the other hand, if the playout simulator cannot find any packet to fetch in the buffer, that is, if the buffer is empty, the playout is interrupted and the incident is recorded.

5. EXPERIMENTAL RESULTS

In our experiments, the sender and receiver hosts are located in the cities of Izmir and Istanbul, respectively. They are Linux boxes with kernel version 2.6.20-rc5, rooting from David Miller's 2.6.x networking git tree modified with Ian McDonald's DCCP patches. We observed that packets traveled through an average of 10 hops with an average channel capacity of 1.2 Mbps and 25 millisecond delay.

We have used the stereo sequence *Balloons* that has 240 frames. To increase video duration, we repeated the sequence 20 times yielding a 160 second video. We have conducted four sets of experiments containing three constant bitrate scenarios (Opt 1, 3, and 7), plus adaptive streaming, where the results are given in Table 2. Fig. 2 shows GoP basis TFRC and extracted video rates for a typical adaptive streaming experiment, and the corresponding spatial (SL) and SNR quality (QL) layers are depicted in Fig. 3. We mostly match the TFRC rate coming from DCCP in the extraction process, as seen in Fig.2, with the exception that quality switching is limited to three GoPs. Thus, at least three GoPs are extracted at the same quality, no matter what the TFRC rate is. This limitation avoids quality fluctuations that may disturb the viewer.

During quality switching, there occurs a situation in which the receiver ends up with one low and one high quality key frames leading to PSNR degradation of that particular GoP. In order to solve this problem, enhancement quality packets of the old key frame are accordingly added into the packets of the high quality GoP. This optimization leads to a gain of 0.34 dB in average PSNR for a single switching from Opt 7 to Opt 1 in a single loop.

In order to measure the perceived 3D video quality, the received H264/AVC file is decoded by the scalable multi-

view video decoder developed in [14]. The decoder creates two reconstruction files for the stereo view. For the second view, decoder interpolates the frames reconstructed at the base spatial layer and then writes them to a YUV file. Reconstructed files are viewed at the polarized projection display system at Koc University and quantitatively evaluated by using the 3D video quality metric proposed in [4].

Table 2 gives the quality metrics of the system for each scenario. The PSNR values increase with extracted bitrates, as expected. However, in Opt 1 scenario, where the PSNR value is the highest, the maximum packet delay is observed as 44.3 seconds, which is not acceptable for a 160 seconds video stream since this value determines pre-buffering period when late packets are not discarded. The remaining scenarios result in acceptable maximum packet delays, which are all below 4 seconds. Among these three cases, adaptive streaming scenario produces the best video quality, which is better than Opt 3 by 0.4 dB and Opt 7 by 2.6 dB in average stereo PSNR.

Table 2: Quality metrics of the system.

Scenario	Avg. Extracted Bitrate [Kbps]	Max Packet Delay [sec]	Avg. V1 PSNR [dB]	Avg. Stereo PSNR [dB]
OPT1	1129	44.3	35.5	35.9
OPT3	789	2.5	32.3	34.8
OPT7	738	2.7	28.6	33.6
Adaptive	923	3.5	33.4	35.2

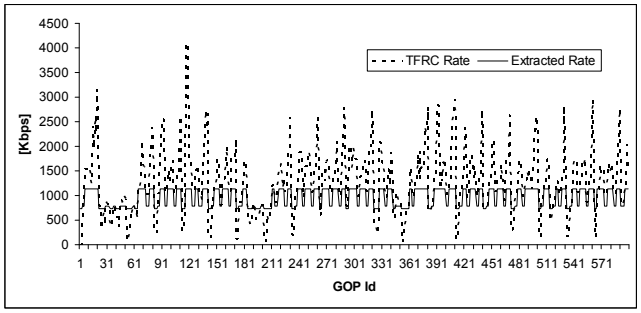

Fig. 2 Observed bitrates for the adaptive streaming experiment.

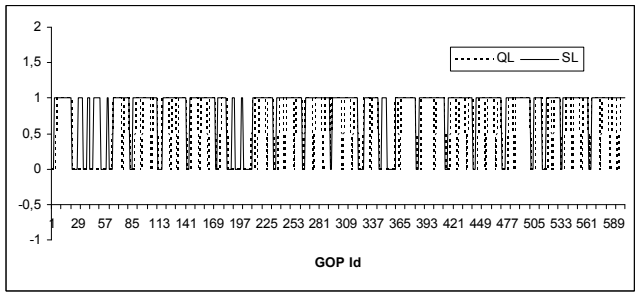

Fig. 3 Observed quality for the adaptive streaming experiment.

6. CONCLUSIONS

We have presented a DCCP based adaptive streaming model for transportation of stereoscopic video over the Internet. Test results demonstrate that we can mostly match the DCCP rate in the adaptive layer extraction process in which best rate visual distortion performance is achieved while keeping maximum packet delay acceptable. Extracting the FGS layer of view 1 with progressive refinement slices to truncate at arbitrary points is future work for utilizing the TFRC rate much more efficiently. Furthermore, packet losses shall be investigated to see the ARQ performance and robustness of the scalable stereo decoder to packet losses.

ACKNOWLEDGEMENTS

This work is supported by EC within FP6 under Grant 511568 with the acronym 3DTV.

7. REFERENCES

[1] H. Schulzrinne, S. Casner, R. Frederick, and V. Jacobson, "RTP: A transport protocol for real-time applications," IETF, July 2003, RFC 3550.

[2] E. Kohler, M. Handley, and S. Floyd, "Datagram Congestion Control Protocol (DCCP)," IETF, March 2006, RFC 4340.

[3] L. Stelmach, W. J. Tam, D. Meegan, and A. Vincent, "Stereo Image Quality: Effects of Mixed Spatio-Temporal Resolution," *IEEE TCSVT*, vol 10, no. 2, pp. 188-193, 2000.

[4] N. Ozbek, A. M. Tekalp, and E. Turhan Tunali, "Rate Allocation between Views in Scalable Stereo Video Coding using an Objective Stereo Video Quality Measure," *ICASSP 2007*, accepted.

[5] A. Smolic, and P. Kauff, "Interactive 3-D Video Representation and Coding Technologies," *Proc. of the IEEE*, vol. 93, No. 1, January 2005.

[6] Joint Video Team of ITU-T VCEG and ISO/IEC MPEG, "Joint Multiview Video Model (JMVM) 1.0," *Doc. JVT-T208*, July 2006.

[7] N. Ozbek, and A. M. Tekalp, "Scalable Multi-View Video Coding for Interactive 3DTV," *IEEE ICME*, Toronto, 2006.

[8] Joint Video Team of ITU-T VCEG and ISO/IEC MPEG, "Joint Scalable Video Model JSVM-4," *Doc. JVT-Q202*, Oct. 2005.

[9] S. Floyd, and E. Kohler, "Profile for Datagram Congestion Control Protocol (DCCP) Congestion Control ID 2: TCP-like Congestion Control," IETF, March 2006, RFC 4341.

[10] S. Floyd, E. Kohler, and J. Padhye, "Profile for DCCP Congestion Control ID 3: TCP-Friendly Rate Control (TFRC)," IETF, March 2006, RFC 4342.

[11] B. Gorkemli, M. R. Civanlar, "SVC Coded Video Streaming over DCCP," 8th IEEE Int. Symp. on Multimedia (ISM'06), 2006.

[12] S. Wenger, Y. Wang and M. M. Hannuksela, "RTP Payload Format for H.264/SVC Scalable Video Coding," In *Journal of Zhejiang University SCIENCE A 2006*, pp. 657-667.

[13] S. Wenger, M.M. Hannuksela, T. Stockhammer, M. Westerlund and D. Singer, "RTP Payload Format for H.264 Video," IETF, February 2005, RFC 3984.

[14] N. Ozbek, and A. M. Tekalp, "Content-Aware Bit Allocation in Scalable Multi-View Video Coding", MRCS, *LNCS 4105*, pp. 691-698, 2006.