SELECTIVE STREAMING OF MULTI-VIEW VIDEO FOR HEAD-TRACKING 3D DISPLAYS

Engin Kurutepe, M. Reha Civanlar, and A. Murat Tekalp

Koç University, School of Engineering, Istanbul, Turkey

ABSTRACT

We present a novel client-driven multi-view video streaming system that allows a user watch 3-D video interactively with significantly reduced bandwidth requirements by transmitting a small number of views selected according to his/her head position. The proposed scheme can be used to efficiently stream a dense set of multi-view sequences (light-fields) or wider baseline multi-view sequences together with depth information. The user's head position is tracked and predicted into the future to select the views that best match the user's current viewing angle dynamically. Prediction of future head positions is needed so that views matching the predicted head positions can be requested from the server ahead of time in order to account for delays due to network transport and stream switching. Highly compressed, lower quality versions of some other views are also requested in order to provide protection against having to display the wrong view when the current user viewpoint differs from the predicted viewpoint. The proposed system makes use of multi-view coding (MVC) and scalable video coding (SVC) concepts together to obtain improved compress ion efficiency while providing flexibility in bandwidth allocation to the selected views. Rate-distortion performance of the proposed system is demonstrated under different experimental conditions.

Index Terms— 3D-TV, Multi-view Coding, Scalable Coding

1. INTRODUCTION

With or without depth information, multi-view representations require large amounts of data, but the correlations between views can be exploited for compression. State of the art in multi-view coding (MVC) is described in [1], where significant compression gains are reported over simulcast coding which compresses each view independently. However, even with the MVC, bit-rates for multi-view video are very high: 38dB PSNR at about 5 Mbps is a common operating point for a 704×480, 30fps, 8 camera sequence with MVC encoding.

In order to reduce the transmission bit-rate of multi-view sequences we observe that for a single-user with a stereoscopic display, only two views are sufficient at any given time to create 3-D perception. Therefore, tracking users' head and selectively transmitting only two required views to render the current viewing angle of the user can save significant bandwidth[1]. An example of a autostereoscopic head-tracking display system with an integrated camera has been presented in [3], although it is also possible to employ a separate head tracking device. Since the transmitted views will vary in time according to user head position in free-view 3D video, we need random access to all views in the bitstream. This requirement cannot be met by MVC because of its complex dependency structure, unless all views are transmitted. When the views are simulcast coded, random access into each stream can be achieved at the cost of reduced compression efficiency. However, this cost is usually more than offset by the reduced number of transmitted views. On the other hand, this approach is sensitive to the delay between the request for the stream and its display. Therefore, in addition to tracking of user's head position, the future positions need to be predicted as well.

There is a wealth of literature on predictive head tracking for 3-D display systems. In [4], Azuma and Bishop formulate a framework to evaluate the performance of prediction algorithms and demonstrate that predictive Kalman filters are suitable for 3-D applications. In light of these results, we decided to use predictive Kalman filtering for our prediction system as described in [5] due to its low computational cost and ease of implementation.

Our goal in this paper is to significantly reduce the bit-rate for transmission of multi-view sequences (with or without additional depth information) in order to enable interactive 3-D TV services over the Internet using a head-tracking display. To this effect, we propose to selectively transmit only low quality versions of those views which would not be needed at the client for display for the purpose of concealing the effects of head prediction errors. The low quality views are delivered in the form of a base layer encoded using the MVC and the high quality views are obtained using specially encoded enhancement layers. This paper is organized as follows: In Section 2, we describe the proposed system in detail. The results are presented in 3 and our conclusions are presented in

Engin Kurutepe is now with Communications System Group, Technische Universität Berlin, Germany.

M. Reha Civanlar is now with DoCoMo Labs, CA, USA.

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

[1] Preliminary works based on this idea have been reported in [2].

4.

2. SYSTEM DESCRIPTION

Suppose that we have a multi-view video with N views on a server. The client-side first determines the user's current head position and a Kalman-filter based predictor predicts the user's head position d frames into the future as described in [5] and requests corresponding M streams from the server, where $2 \leq M \leq N$. The server selectively streams the multi-view video sequence at two quality levels: As a base layer, all M views are encoded using the MVC codec at a lower bit-rate. On top of this base layer, an enhancement layer is encoded for each view independently of other enhancement layers to allow random access in order to improve the quality of the selected views. Since the total bandwidth available to the user is assumed fixed, an increased proportion of the bandwidth needs to be allocated to the base layer as M increases.

If there are no prediction errors, the received high-quality (base + two enhancement) streams are passed on to the display, which shows a high quality view to each eye. The low bit rate base layer MVC enables the user to keep watching 3D video, albeit possibly at a lower quality, when the current user head position differs from the predicted position until correct high quality streams arrive from the server. If there is a prediction error and wrong set of high quality streams arrive, the system displays low quality version of the desired views which may be available in the base layer MVC only. According to subjective quality tests reported in [6], humans perceive high quality 3D video as long as one of the eyes sees a high quality view. Therefore, in the presence of prediction errors, as long as at least one of the required views is delivered in high quality, the viewer might not even notice any loss of quality. If the prediction error is so severe that a required view is not among the M views in the base layer, an error concealment method is employed.

2.1. MVC Base Layer and Simulcast Enhancement Layers

The structure of our coding strategy is depicted in Fig. 1, where small and large squares denote spatially down sampled base-layer frames and high resolution enhancement layer frames, respectively. The arrows indicate prediction reference relationships between the frames. The base layer involves encoding spatially downsampled versions of all views together using the MVC at a lower bit-rate. In addition to this base layer, an enhancement stream is generated for each view as follows: first, the decoded video for each view in the MVC stream is upsampled to the full spatial resolution of the original video and then, the difference between the original and decoded/upsampled MVC videos are encoded using the AVC/H.264. Alternatively, the spatial scalability

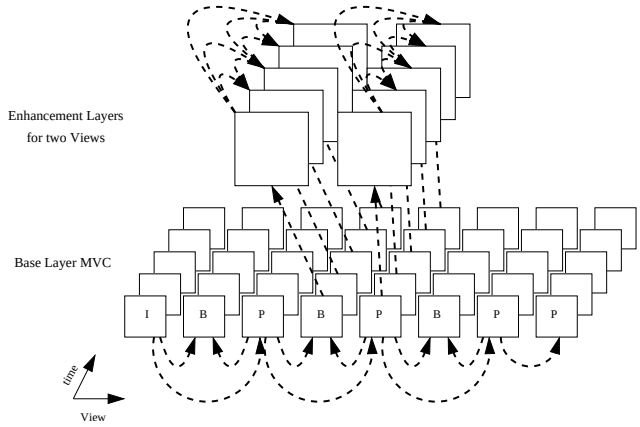


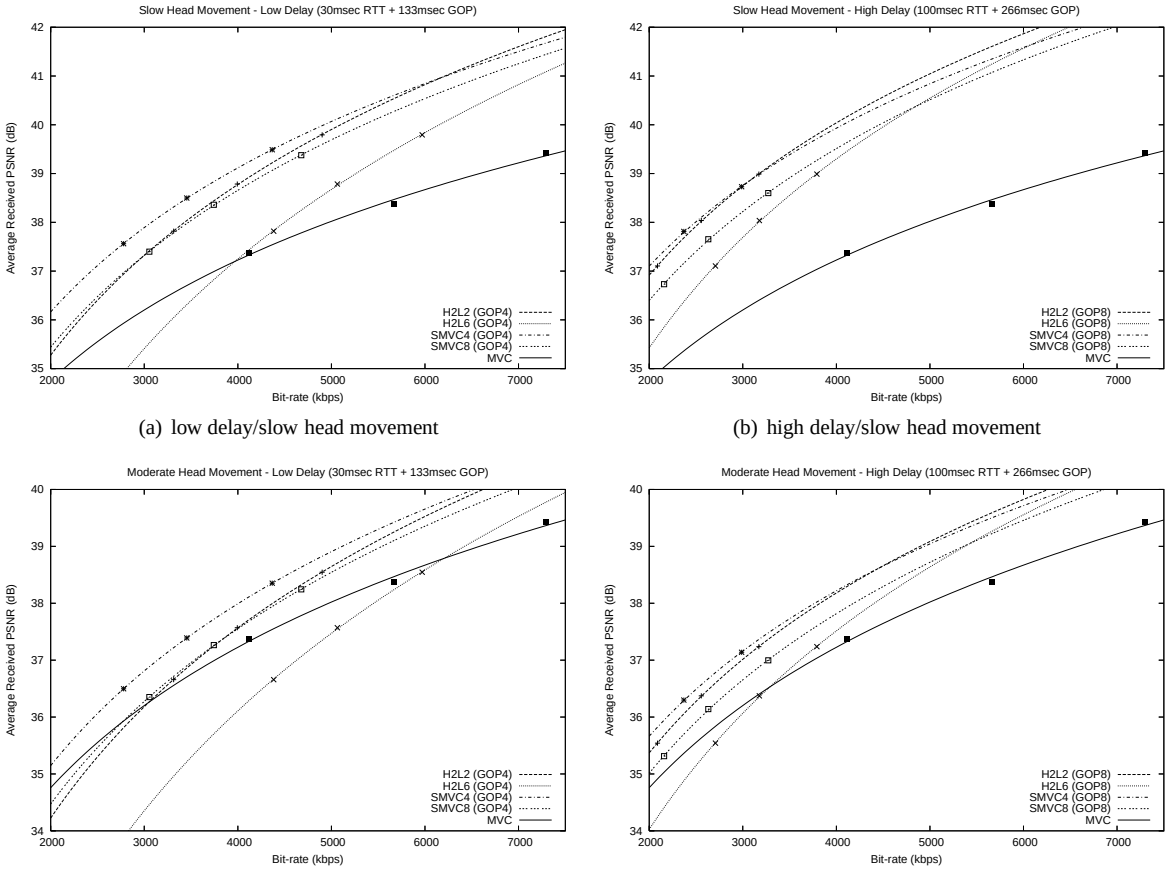
Fig. 1. The coding structure with MVC base layer and simulcast enhancement layers.

features of the emerging SVC standard (extended to multi-view video coding as in [7] and [8]) might be used for this purpose. The enhancement layers are independently coded to provide greater flexibility in switching from one view to the next. This proposed structure benefits from the coding gain offered by MVC, while providing significantly greater flexibility in view selection, such that a user receiving the base layer and the enhancement layer for view n, is able to see it at a high quality, while other views would be available at a lower quality.

The proposed system allows switching base layer and/or enhancement layer streams at the start of each GOP. Selection of the views to follow the user's head position may be achieved by switching only enhancement streams at the beginning of each GOP. It is best to use a shorter GOP size for the enhancement layers because they need to be able to follow the user's head position more closely, whereas a longer GOP may be utilized for the base layer MVC for more efficient compression. In Section 3, we compare the proposed system with a reference system which delivers independently coded streams at two distinct quality levels and present results on the interaction between the prediction distance, GOP size and rate-distortion performance of the resulting sequences for different head motions.

3. RESULTS

The intended application scenario is a single-user receiving multi-view content from a server to watch using a head-tracking stereoscopic display system. In the remainder of this section, we will first present two reference systems implemented for comparison purposes only, and then proceed to rate-distortion results under different network and viewer conditions. The reported results have been obtained using the "Race1" multi-view sequence provided by KDDI.



(a) low delay/slow head movement

(b) high delay/slow head movement

(c) low delay/moderate head movement

(d) high delay/moderate head movement

Fig. 2. Rate-Distortion curves

3.1. Reference Systems for Performance Comparison

We compare the performance of our proposed system with two reference system implementations: i) simulcast encoding and streaming of only the needed views to match the predicted future user viewing angles and a small number of side views; ii) combined MVC encoding and streaming of all views (not using head tracking/prediction data). In the reference system 1, similar to our proposed system, the client determines the required two views using the predicted head position information, and requests the corresponding high-quality streams and a low bitrate version of M adjacent views from the server, where each view is encoded independently at two distinct quality levels. Hierarchical B-pictures were used during encoding to best utilize temporal correlations. The motion search window was fixed at 96 pixels and three different GOP sizes were used: 4, 8 and 16. An IDR-picture was inserted at the beginning of each GOP, such that it becomes a stream switching point. In the reference system 2, the MVC implementation from HHI was used to encode the sequence as reported in MPEG Bangkok meeting configurations [9], and all views are streamed regardless of head tracking/prediction results.

3.2. Rate-Distortion Performance

For the proposed system, the base layer MVC was encoded using modified MPEG Bangkok configurations [9]. Modifications reflect the downsampled resolution and higher QP parameters for higher compression. Additionally, the JSVM SequenceFormatString parameter was modified accordingly for 4-view base layers to preserve the correct reference structure. The enhancement layers were encoded as described in Section 2.1, at various quality settings for each different base layer configuration. Similar to the reference system 1, the enhancement layers used three different GOP sizes: 4, 8 and 16.

In our simulations, at each time instance the client requests the base layer and two enhancement layers for the views corresponding to the d-frame ahead prediction trajectory. After a constant RTT has passed, the requested streams arrive in a decoding buffer. If the viewpoint prediction has failed at

some point between the current frame and beginning of the GOP, some of the packets needed to decode the current frame might not have been delivered, therefore, when the play-out time for a frame arrives, the buffer is checked to determine the decodability of the actually required frames in a recursive fashion. For a frame to be decodable in low quality, its base layer packet(s) must be available, and all reference frames in the base layer MVC prediction structure must be decodable. For a frame to be decodable in high quality, it needs to be decodable in low quality, its enhancement layer packet must have arrived, and its reference frames in the enhancement stream must be decodable as well. If the required frame at a particular time instant is not decodable in high quality, low quality frames are displayed, in the case the frame is not decodable in low quality as well, it repeats the last displayed frame as a simple error concealment method. The output of reference system 1 is generated under the same conditions except that decodability conditions of the simulcast streams are simpler due to the lack of spatial references between views. The reference system 2 is not affected by any missed streams, since all views are available at each time instance in the MVC stream. The PSNR values for all test cases are computed with respect to the "ground truth" stereo sequences, which are generated using the original views corresponding to perfect head position information. Fig. 2(a) through 2(d) demonstrate the Rate-Distortion characteristics of the proposed system when compared to the reference system 1 and reference system 2, where H2Ln denotes the reference system 1 with two high quality and n low quality streams and SMVCn denotes the proposed system with n views encoded in the MVC base layer. As it can be seen from these figures the performance of selective streaming systems, both the reference system 1 and the proposed system, deteriorate with the faster head movement. Additionally, the trade-off between compression efficiency and latency can also be seen: the longer delay architecture is better than the short delay architecture for a slow head motion due to increased compression efficiency and relatively little importance of latency. However for moderate head movements both short delay and long delay options are close to each other. Additionally, the proposed system outperforms the reference system 1 at lower bit-rates. As the operating conditions get worse, the proposed system performs increasingly better when compared to the reference system 1, but the advantage compared to reference system 2 (MVC) starts to decrease.

4. CONCLUSIONS

We introduce a novel view-selective streaming strategy for streaming multi-view video for single-user interactive 3DTV applications. The proposed system features selective streaming of views, such that only the views which are required to display the user's current view are delivered. An integral part of the proposed system is a new multi-view video encoding scheme, which makes use of both MVC and SVC concepts, where the encoded video is composed of an MVC encoded multi-view base layer and simulcast coded individual view enhancement layers. The proposed system also includes methods to predict the user's future head positions. We have shown that the proposed system outperforms MVC in the sense of transmitted bits for most operating conditions and is up to 3dB more efficient in some cases. It has been observed that the low quality neighboring streams are well worth their bandwidth cost, since they allow continuous play-out of the 3D video in cases where the predicted viewing angle differs from the actual current viewing angle.

5. REFERENCES

[1] K. Mueller, P. Merkle, H. Schwarz, T. Hinz, A. Smolic, T. Oelbaum, and T. Wiegand, "Multi-view video coding based on H.264/AVC using hierarchical B-frames," in *Picture Coding Symposium 2006*. PCS, 2006.

[2] E. Kurutepe, M. R. Civanlar, and A. M. Tekalp, "Interactive transport of multi-view videos for 3DTV applications," *Journal of Zhejiang University SCIENCE A: Proc. Packet Video Workshop 2006*, vol. 7, no. 5, pp. 830–836, 2006.

[3] K. Hopf, P. Chojecki, F. Neumann, and D. Przewozny, "Novel autostereoscopic single-user displays with user interaction," in *Proc. of SPIE*, vol. 6392, 2006.

[4] R. Azuma and G. Bishop, "A frequency-domain analysis of head-motion prediction," in *SIGGRAPH '95: Proceedings of the 22nd annual conference on Computer graphics and interactive techniques*. New York, NY, USA: ACM Press, 1995, pp. 401–408.

[5] A. Kiruluta, M. Eizenman, and S. Pasupathy, "Predictive head movement tracking using a kalman filter," *Systems, Man and Cybernetics, Part B, IEEE Trans. on*, vol. 27, no. 2, pp. 326–331, April 1997.

[6] L. Stelmach, W. Tam, D. Meegan, and A. Vincent, "Stereo image quality: effects of mixed spatio-temporal resolution," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 10, no. 2, pp. 188–193, 2000.

[7] N. Ozbek and M. Tekalp, "Scalable multi-view video coding for interactive 3dtv," in *Proceedings of IEEE ICME 2006*. IEEE, 2006, pp. 213–216.

[8] M. Drose, C. Clemens, and T. Sikora, "Extending single-view scalable video coding to multi-view based on h.264/avc," in *Proceedings of IEEE ICIP 2006*. IEEE, September 2006, pp. 2977–2980.

[9] K. Mueller, "Multi-view coding software." [Online]. Available: "http://iphomes.hhi.de/mueller/MVC_SW.htm"