Client-Driven Selective Streaming of Multi-View Video for Interactive 3DTV

Engin Kurutepe, *Student Member, IEEE,* M. Reha Civanlar, *Fellow, IEEE,* and A. Murat Tekalp, *Fellow, IEEE*

***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 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 prefetched in order to account for delays due to network transport and stream switching. The system allocates more bandwidth to the selected views in order to render the current viewing angle. Highly compressed, lower quality versions of some other views are also prefetched for concealment if the current user viewpoint differs from the predicted viewpoint. An objective measure based on the abruptness of the head movements and delays in the system is introduced to determine the number of additional lower quality views to be prefetched. The proposed system makes use of multi-view coding (MVC) and scalable video coding (SVC) concepts together to obtain improved compression 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**

I. Introduction

ALTHOUGH dynamic holography is the ultimate goal in 3-D video and TV, early systems create 3-D viewing experience via stereoscopy by showing a scene from slightly different angles to the left and right eyes of a viewer. Various methods can be employed in order to generate these views.

A popular approach is 3-D warping of an image using the associated depth map. This approach has been recently studied in the European ATTEST project, and it has been reported that the depth map can be compressed to about 10-20% of the video stream [1]. There is an MPEG standardization effort for the transport of video-plus-depth representation [2]. However, the rendering quality may deteriorate due to disocclusions and discontinuities in depth, as the viewer moves away from the original camera angle [3]. The N-view-plus-M-depth representation has been proposed as a promising extension to address the limitations of the video-plus-depth representation [4]. When rendering from this representation, holes due to depth discontinuities can be filled using pixels from neighboring views. However, imperfections in the depth maps may result in ghost-like shadows near the object boundaries, which can be removed by more sophisticated rendering algorithms [5].

Another alternative is using a denser set of views without depth maps, which is called the light field representation [6]. Elimination of the need for depth determination (in scene capture) and view generation (in the client) are significant advantages of this approach. Such a system is demonsrated in [7], where views from 16 cameras are projected onto an autostereoscopic screen. As the user moves through the viewing space, views from appropriate cameras are directed to the eyes with the help of a lenticular lens array on the screen.

Multi-view representations require large amounts of data. State of the art in multi-view coding (MVC) is described in [8], 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 high: 38dB PSNR at about 5 Mbps is a common operating point for a 704 $\times$ 480, 30fps, 8 camera sequence with MVC encoding. Moreover, previous research on single-view-plus-depth sequences [1] suggests that with the addition of depth maps and other auxiliary information such as boundary mask, the bandwidth requirements could increase as much as 20%, which renders 3-D TV service over the current high-speed Internet connections practically impossible.

Our goal in this paper is to significantly reduce the bit-rate for transmission of multi-view sequences in order to enable interactive 3-D TV services over the current Internet using a head-tracking stereoscopic display. To this effect, 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 [11], although it is also possible to employ a separate head tracking device. Since the views to be transmitted will vary in time according to user head position, we need random access to all views in the bitstream. This requirement cannot be met by MVC unless all views are transmitted, because of its complex view-dependency structure. When the views are simulcast coded, random access into each stream can be achieved at the cost of reduced compression efficiency. Hence, we propose a new scalable MVC structure in Section II-A to strike a balance between compression efficiency and random access

Engin Kurutepe is with Telecommunications Institute, Techische Universität Berlin, Germany, M. Reha Civanlar is with DoCoMo Labs, USA, A. Murat Tekalp is with Koc University, Istanbul, Turkey

[1]Preliminary works based on this idea have been reported in [9] and [10].

requirements. Furthermore, the proposed system is sensitive to delay between the request for the stream and its display. Therefore, in addition to tracking of user's head position, the future positions shall be predicted to prefetch the required views. In order to conceal the effect of prediction errors, we propose transmitting low quality versions of all views, in the form of a base layer encoded using the MVC. Higher quality views are selectively provided only for those views which match the predicted viewing direction by transmitting specially encoded enhancement layers. The number of the views contained in the base layer MVC and the bit-rate allocation between the base and the enhancement layers shall be determined for optimal system performance under specific viewer and network conditions. In Section II, we describe the proposed system in detail. The results are presented in III and our conclusions are presented in IV.

II. 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. Then, an error measure is computed at the client to determine the number of views, $M \leq N$, to be requested from the server. 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 intelligent rate allocation scheme between the base layer MVC and enhancement layer streams is necessary.

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 [12] and [13], 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 delivered at all (is not among the M views in the base layer), an error concealment method is employed (e.g., nearest available views are displayed). In the following, details of the server side issues are described in Sections II-A and II-B. Details of the client-side are described in Sections II-C and II-D.

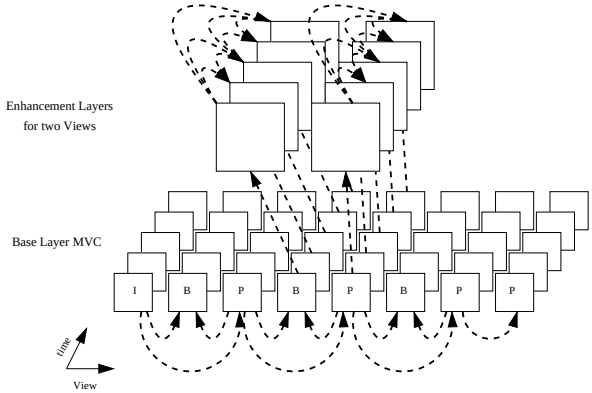


Fig. 2. The coding structure with MVC base layer and simulcast enhancement layers.

A. MVC Base Layer and Simulcast Enhancement Layers

The structure of our coder 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 features of the emerging SVC standard (extended to multiview video coding as in [14] and [15]) 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 all other views would be available at a lower quality.

We assume that the multi-view video server has several previously encoded sets of base layer MVC and enhancement layers. Each of these sets is a complete representation of the original multi-view sequence, but the base layer MVC stream(s) in each set contains a different number of views; hence, different bit allocations between base and enhancmen layers. 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. However, if network conditions and/or error statistics of the head position prediction varies, then the number of views M in the base layer may be adapted as well. Adaptation of the number of views M in the base layer, and bit-rate allocation between base layer and enhancement layers is handled by switching from one multi-view video set to a different set at the start of a GOP. Therefore, it is best to use a shorter GOP size for the enhancement layers because they

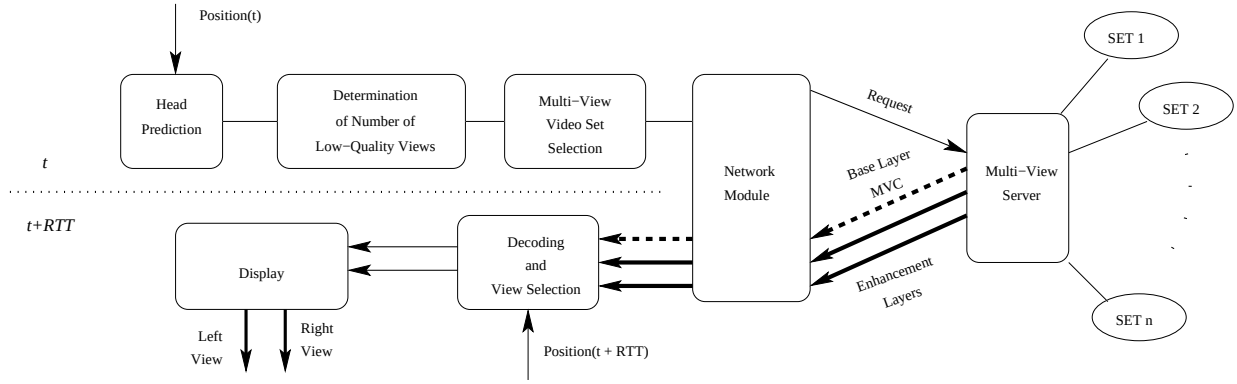


Fig. 1. Overview of the delivery system with two low bit-rate side-streams.

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.

B. Bit-rate Allocation

Determination of the target bitrates for the base layer MVC and enhancement layers can be formulated as an optimization problem, where the expected value of video quality experienced by the user should be maximized. One of the parameters in this bit-rate allocation problem is the number of views contained in the base layer MVC. In Section II-D, we discuss how the receiver selects the number of views M in the base layer MVC; hence, it is assumed given here. The desired quality difference between the base layer and enhancement layers is another parameter which determines the bit-rate allocation between the base and enhancement layers. Hence, the ratio of bitrates of the base and enhancement layers is a parameter which should be used to maximize the quality of the 3D video delivered. The Lagrangian rate-distortion optimization can be formulated as

$$max\{J\}, \text{ where } J = Q + \lambda R, \tag{1}$$

where Q is the average quality and the bit-rate R is given by

$$R = R_{MVC} + 2 \cdot R_{ench}. \tag{2}$$

Here, R_{MVC} and R_{ench} are the bit-rates for the base layer and an enhancement layer respectively. The factor 2 that multiplies R_{ench} accounts for the fact that we only send two enhancements views at a time. The quality of the 3D video is defined as the average of the quality of the right and left pair of views. Hence, the expected quality of the 3D video in the presence of head prediction errors can be formulated as

$$\begin{aligned} Q_{ave} = \quad & p[x=0]Q_{hi} + p[x=1]\tfrac{Q_{hi}+Q_{lo}}{2} + \\ & p[M/2-1 > x > 1]Q_{lo} + \\ & p[x=M/2]\tfrac{Q_{lo}+Q_{err}}{2} + p[x > M/2]Q_{err}, \end{aligned} \tag{3}$$

where $p[x=a]$ denotes the probability of a head prediction error that results in transmission of $\pm a$ views to the right or left of the desired pair of views, Q_{hi} and Q_{lo} denote the average PSNR for both (the base and enhancement) layers and base layer MVC, respectively, and Q_{err} is the quality achieved by error concealment, i.e., showing the previous viewpoint. In (3), the first term denotes the quality achieved when there are no prediction errors and both eyes see the desired high quality views, the second term corresponds to a prediction error by one view when one of the eyes sees a high quality view and the other eye sees a low quality (base layer) view. The third term is the quality expression when both eyes see low quality views, which is the case when the prediction error is greater than one but still inside the M views contained in the base layer MVC. The reason for the $M/2-1$ limit is due to the symmetric nature of the base layer. The fourth term denotes the case when the prediction error is exactly $M/2$, which corresponds to one eye seeing the low quality view and the view for the other eye is displayed using frame repetition from the last available frame. Finally, the fifth term is the case when the prediction error is so large that both eyes see continue to see the old viewpoint. This last case corresponds to error concealment. Since M is adjusted dynamically according to a measure which will be introduced in Section II-D, there is a small chance that the prediction errors will result in a complete miss. Additionally, the PSNR for error concealment is much lower than a high quality or a low quality view. Therefore, the contribution from the last two terms is negligible and they can be dropped to simplify the closed-form average quality expression, which results in

$$\begin{aligned} Q_{ave} = \quad & p[x=0] \cdot Q_{hi} + p[x=1] \cdot \tfrac{Q_{hi}+Q_{lo}}{2} + \\ & p[x>1] \cdot Q_{lo} \end{aligned} \tag{4}$$

that can be further simplified as

$$\begin{aligned} Q_{ave} &= (p[x=0] + 0.5p[x=1]) \cdot Q_{hi} + \\ &\quad (0.5p[x=1] + p[x>1]) \cdot Q_{lo} \\ &= p_{hi} \cdot Q_{hi} + p_{lo} \cdot Q_{lo} \end{aligned} \tag{5}$$

where p_{hi} and p_{lo} denote probability of perceiving high quality 3D video and probability of perceiving low quality 3D video, respectively. The probability distribution of head prediction errors, p, is assumed to be a zero-mean normal distribution. Therefore, variance of the head prediction error can be used to determine the probabilities p_{hi} and p_{lo}.

Although there is no closed-form relationship between the bit-rate and average quality of a video stream, logarithmic models of the form

$$\begin{aligned} Q_{hi} = \quad & a_{hi} + b_{hi} \cdot \ln(R_{ench}) \\ Q_{lo} = \quad & a_{lo} + b_{lo} \cdot \ln(R_{MVC}) \end{aligned} \tag{6}$$

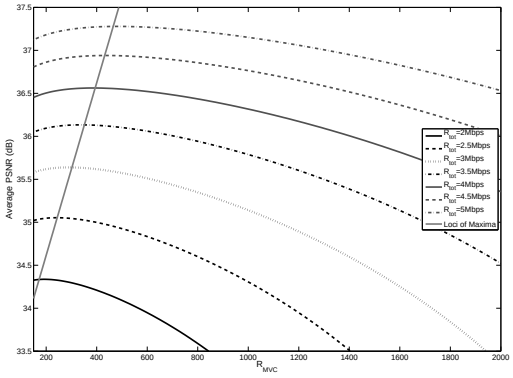


Fig. 3. Projection of the rate-distortion surface on the R_{MVC}-PSNR plane for $p_{lo} = 0.25$. The solid line is the loci of the maxima on the R-D curves as the total available bitrate increases.

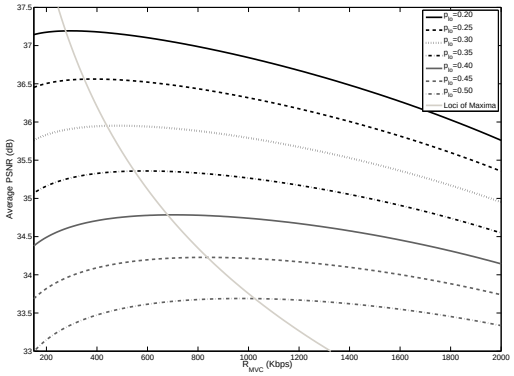


Fig. 4. Curves showing the average PSNR vs. bitrate for base layer MVC at $R_{tot} = 4Mbps$ for different p_{lo}, indicating more bits from the total bit-budget should be spent for the base layer as p_{lo} increases.

can be employed, where a and b parameters can be found by least squares fitting to experimental results. Thus, by substituting the approximations for the quality of high quality and low quality views into 5, we obtain an approximation which relates bitrates for the base and enhancement layers to the received multi-view video quality.

The fixed available bit-rate constraint $R_{MVC} + 2 \cdot R_{enhc} = R_{tot}$ corresponds to a plane in the (R_{MVC}, R_{ench}, Q) co-ordinate system. The projection of the rate-distortion surface on this constraint plane is shown in Fig. 3. It can be seen from the figure that there is an optimal operating point for bits allocated to the base layer MVC for each value of R_{tot}, after which the average 3D video quality starts to deteriorate as more bits are spent on the base layer, because more bits in the base layer means less bits in the enhancement layers. The loci of optimal R_{MVC} values, corresponding to maximum average PSNR value for each R_{tot}, nearly follows a line as R_{tot} increases which is depicted in Fig 3. The other important parameter in bit-rate allocation is the probability p_{lo} of perceiving low quality 3D video. Fig. 4 shows how the average PSNR value changes as p_{lo} varies. Not surprisingly, the average PSNR decreases with increasing p_{lo}. The loci of maximum PSNR values for each p_{lo} (i.e., optimal R_{MVC} values) follows a polynomial curve depicted in Fig. 4. Clearly, this curve indicates more bits should be allocated for the base layer in order to optimize the 3D video quality as p_{lo} increases. In summary, the best R_{MVC} at any R_{tot} and p_{lo} can be determined form a similar plot for the particular value of R_{tot}, which is the approach we employed in our experiments.

C. Viewpoint Prediction for Prefetching

Although a viewpoint in the world is defined by six independent parameters x, y, z for position and the Euler angles θ, ϕ, ψ, we only consider the case where movement of the user is constrained to translation in one (x) dimension. This assumption reflects the one dimensional physical arrangement of cameras for most multi-view sequences. Therefore, the viewpoint prediction is handled by a Kalman filter with three states, in the same fashion as [16], reflecting a piecewise linear acceleration model.

At each time instant t, the viewer's actual viewpoint is observed and a future viewpoint for the time instance $t + d$ is predicted using the Kalman predictor, where d is the prediction distance. The desired value d of prediction distance is related to the total delay between the request for a stream and the arrival of a decodable frame from the requested stream. This delay until a stream begins playing after the request, is a sum of two independent delay components: Network delay and decoding delay. The Round Trip Time (RTT) delay of the connection between the server and the client corresponds to the network delay of the system. The decoding delay is the wait for an independently decodable frame after the first packet of a stream arrives. This depends on the distance between independently decodable frames or the Group of Pictures (GOP) size. A larger GOP size is better for compression efficiency, but results in a longer decoding delay. However, if a frame is received, but not decodable by its play-out time, it cannot be displayed to the viewer. To accommodate for the longer decoding delays, the prediction distance must be increased as well in order to increase the probability that an independently decodable frame is received on time.

It is clear that a longer prediction distance (larger d) results in larger amounts of prediction errors because the probability of an unpredictable head motion increases with the prediction distance. The performance of the prediction depends, in addition to the prediction distance, on the abruptness of the user head movements. This means given the same prediction distance and two viewers watching the same multi-view sequence on two different clients with the same RTT, the viewer who makes more abrupt head movements will experience more prediction errors and possibly a lower perceived 3D video quality.

Since the prediction becomes less reliable with longer prediction distances, a larger GOP will result in more frequent prediction errors and might cause wrong streams to be fetched from the server. In that case, the larger GOP adversely affects the user experience and deteriorate the system's performance. Therefore, there is a tradeoff between the compression efficiency provided by GOP size and the prediction errors caused

by the increased system latency. The increased efficiency offered by a longer GOP can be more than offset by the errors caused due to a long prediction distance. In the presence of frequent prediction errors, redundant neighboring streams greatly improve the system performance, by providing a view to fall back on, when the prediction fails. However, this comes at the expense of increased bitrate which offsets the coding gain provided by a larger GOP size. We present a detailed investigation of these trade-offs in Section III.

D. Determining the Number of Views in the Base Layer

Additional views in the base layer MVC improve the performance of a selective view delivery system. Of course, the additional low-quality views come at a bandwidth cost, but provide insurance against head prediction errors. Logically, the optimal number of low quality views, M, depends on the prediction distance, which in turn depends on the delay in the system, and the abruptness of head movements.

We propose to use a running average of the square of the prediction error as a metric to help determine the number of views in the low bit-rate base layer. The metric for the nth frame can be computed as $R_n = \frac{1}{d}\sum_{i=n-d}^{n}(r_i - p_i)^2$, where r_i and p_i are the real and predicted positions, respectively for frame i and d is the prediction distance. This abruptness metric corresponds to an estimate of the error variance for the last d samples, assuming the prediction error has a zero-mean distribution. This assumption is valid due to the fact the prediction follows the real values and does not introduce any drift in either direction. The rationale behind using d, as the window size over which the variance estimate is computed, is the fact that it corresponds to the unresponsiveness, since it is a function of two delay components in the system: network delay (RTT) and decoding delay (GOP size). Therefore, our proposed metric intuitively combines two most important aspects which affect the prediction performance: head movement abruptness and the delay in fetching streams from the server and displaying them. By thresholding the computed metric, the system can estimate the optimum number of views in the low-quality base layer. In Section III, we present an experimentally determined threshold for the proposed metric in order to select the optimum number of views in the base layer.

III. Results

In this section, we first compare the rate-distortion performance of the proposed system with two reference systems under different network and viewer conditions. Next, determination of the threshold for the selection of the number of views in the base layer and multi-view bitstream switching are demonstrated. The reported results have been obtained using the "Race1" multi-view sequence provided by KDDI.

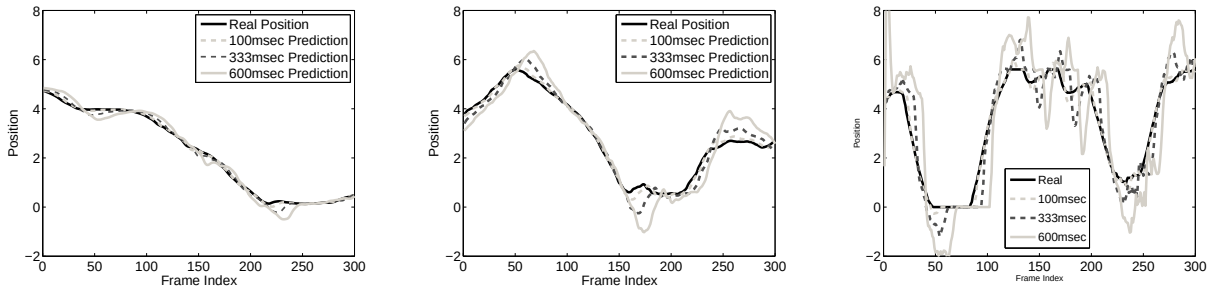
The two reference systems are: 1) simulcast encoding and streaming of only the needed views to match the predicted future user viewing angles and a small number of side views; 2) combined MVC encoding and streaming of all views (not using head tracking/prediction data). In the reference system 1, similar 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 M is determined using the metric defined in Section II-D. The only difference between this system and our proposed system is that here all views are coded independently using AVC/H.264 at two quality levels (two QP values), and the requested high and low quality views are simulcast streamed. 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 [17], and all views are streamed regardless of head tracking/prediction results.

A. Rate-Distortion Performance

The rate-distortion performance of the proposed system has been studied using three 10sec long head motion trajectories, which were recorded using a camera based tracking system. These three trajectories can be classified as slow, moderate and fast head movements. The head position prediction was performed on the recorded trajectories at various prediction distances. The three trajectories and prediction performance at three different prediction distances can be seen in Fig. 5(a) through 5(c). It is clear that the prediction performance decreases as the prediction distance increases and this effect is more pronounced when the head motion is faster.

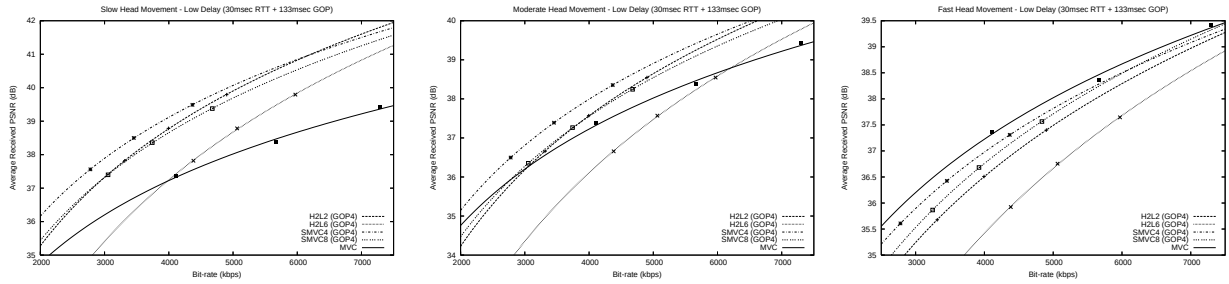
We compare the performance of the proposed system with the reference systems 1 and 2 using the predicted head trajectories. For the proposed system, the base layer MVC was encoded using modified MPEG Bangkok configurations [17]. Modifications reflect the downsampled resolution and higher QP parameters for higher compression. Additionally, the JSVM SequenceFormatString parameter was modified accordingly for 4-view and 6-view base layers to preserve the correct reference structure. The enhancement layers were encoded as described in Section II-A, 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. Our proposed client implementation first checks if the required frame at a particular time instant is decodable in high quality, if not it falls back to low quality frames, 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.



(a) slow head trajectory (b) moderate head trajectory (c) fast head trajectory

Fig. 5. The real head position data, along with prediction results with two prediction distances.



(a) low delay/slow head movement (b) low delay/moderate head movement (c) low delay/fast head movement

Fig. 6. Rate-distortion curves comparing the proposed system to the reference systems.

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 corresponding head tracking (viewpoint) data, assuming no delays, perfect decodability and no encoding distortion, i.e., original views corresponding to perfect head position information is included in the ground truth stereo sequences. Fig. 6(a) through 6(c) 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 increasingly faster head movements. Although not shown here due to space constraints, there is also a trade-off between compression efficiency and latency: 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 and for the fast head movement simulation, the low delay architecture significantly outperforms the longer delay architecture. 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. Reference system 2 outperforms both selective streaming schemes for fast head movements with high delay.

Compression efficiency, latency trade-off: As mentioned in Section II-C, there is a trade-off between the compression efficiency offered by a longer GOP and introduced latency due to the decoding delay. Our results show that a GOP of 16 frames results in a poorer received video rate-distortion performance. The poor performance of 16 frames GOP is caused by two factors: the prediction distance can be increased to compensate for the GOP size, which results in more frequent prediction errors, or a short prediction distance can be used but then some of the frames might be missed if the beginning of the GOP is missing. Although a 4-frame GOP offers better latency performance, it is not enough to offset the compression efficiency lost to frequent IDR-frames, except for very fast head movements where quick stream switching is very important (see. Fig. 6(c)). However it should be noted that in such situations the proposed selective delivery system offers little, if at all, performance gain over sending the whole MVC stream in high quality.

B. Multi-view Stream Set Switching

In light of the results presented in Section III-A, the 3D video quality at the client clearly depends on the abruptness of the head motion and the delay in the system. Hence, in order to achieve best results, the client should request the server to switch between sets of available multi-view streams, e.g., SMVC4, SMVC6 and MVC, according to current user head

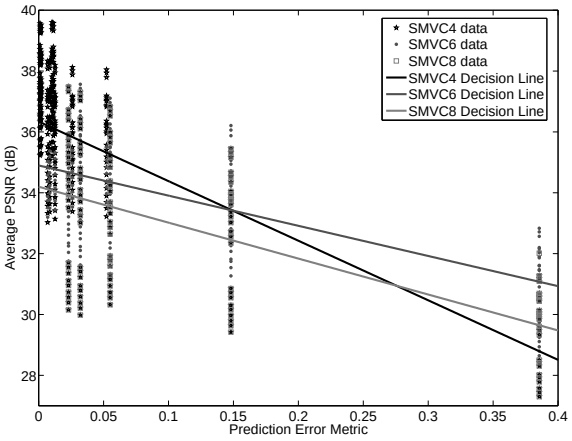


Fig. 7. Average quality vs. prediction error metric. Data points denote that a base layer with fewer views is not worse for that particular prediction error metric value at the same quality and GOP settings. The highest decision line is selected for a given prediction error metric.

motion and total delay, which is supported by the following results. To this effect, we have computed the prediction error variance metric as proposed in Section II-D for the predicted head trajectories. Fig. 7 is a scatter plot of all available multi-view video streaming experiments. It shows three sets of data points for four, six and eight views in the base layer, where we have removed samples which contain more views than necessary, i.e., a data point has been removed if and only if it provides exactly the same average quality with a base layer with fewer views at the same encoding settings. This ensures that the base layer with fewer views, and thus lower bit-rate, is favored at a given quality level. The removed data points correspond to experiments where additional views in the base layer lead to no average PSNR improvement at all, when all other conditions are the same. The decision lines have been fitted to the data points in the least squares sense, which denote the average expected quality for a prediction error measure value. Therefore, for a prediction error measure value, the highest line is the optimum line. As it can be seen from the figure, 0.15 is found to be the decision boundary for switching between SMVC4 and SMVC6. Four views in the base layer are enough, if the prediction error metric is lower than 0.15. Six views are needed otherwise. There is no intersection between the decision line of SMVC6 and SMVC8, which suggests that the bitrate cost of two additional views is not justifiable, since the prediction rarely fails badly enough that the required views are not contained in a base layer with six views.

IV. 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 and to adaptively control the number of low quality views in the base layer according to the prediction error variance. 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.

V. Acknowledgements

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

References

[1] C. Fehn, K. Hopf, and Q. Quante, "Key technologies for an advanced 3D-TV system," in *Proc. of SPIE Three-Dimensional TV, Video and Disp. III*, October 2004, pp. 66–80.

[2] A. Smolic, K. Mueller, P. Merkle, C. Fehn, P. Kauff, P. Eisert, and T. Wiegand, "3D video and free viewpoint video th technologies, applications and MPEG standards," in *Proceedings of IEEE ICME 2006*. IEEE, 2006.

[3] C. Fehn, "Depth-Image-Based Rendering (DIBR), compression and transmission for a new approach on 3D-TV," *Proc. Stereoscopic Displays and Applications*, 2002.

[4] O. Scheer, C. Fehn, N. Atzpadin, M. Muller, A. Smolic, R. Tanger, and P. Kauff, "A flexible 3D TV system for different multi-baseline geometries," in *Proceedings of IEEE ICME 2006*. IEEE, 2006.

[5] C. L. Zitnick, S. B. Kang, M. Uyttendaele, S. Winder, and R. Szeliski, "High-quality video view interpolation using a layered representation," *ACM Trans. Graph.*, vol. 23, no. 3, pp. 600–608, 2004.

[6] M. Levoy and P. Hanrahan, "Light field rendering," in *SIGGRAPH '96*. New York, NY, USA: ACM Press, 1996, pp. 31–42.

[7] W. Matusik and H. Pfister, "3D TV: a scalable system for real-time acquisition, transmission, and autostereoscopic display of dynamic scenes," *ACM Transactions on Graphics (TOG)*, vol. 23, no. 3, pp. 814–824, 2004.

[8] 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.

[9] E. Kurutepe, M. R. Civanlar, and A. M. Tekalp, "A receiver-driven multicasting framework for 3DTV transmission," in *European Signal Processing Conference, Proceedings of*, September 2005.

[10] ——, "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.

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

[12] 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.

[13] A. Aksay, C. Bilen, E. Kurutepe, T. Ozcelebi, G. B. Akar, M. R. Civanlar, and A. M. Tekalp, "Temporal and spatial scaling for stereoscopic video compression," in *EUSIPCO 2006*. EURASIP, 2006.

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

[15] 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.

[16] 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.

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