Modern Standards for Video Communications: from MPEG-2 and MPEG-4 to H.264/AVC, SVC, MVC, and H.265/HEVC

Patrick Seeling and Martin Reisslein

Abstract

Video encoding for multimedia services over communication networks has significantly advanced in recent years with the development of the highly efficient and flexible H.264 SVC video coding standard. The emerging H.265/HEVC video coding standard as well as 3D video coding will further advance video coding for multimedia communications. This tutorial article gives an overview of these new video coding standards and reviews their implications for multimedia communications. We showcase the effects of the video coding advances from MPEG-2 and MPEG-4 to H.264 SVC and H.265/HEVC on the video traffic for single-layer (non-scalable) video coding for conventional 2D video. Our study is the first to examine the H.265/HEVC traffic variability for long videos. We also illustrate the video traffic characteristics of scalable H.264 SVC video coding as well as video coding for 3D video.

I. Introduction

Network traffic forecasts, such as the Cisco Visual Networking Index, predict strong growth rates for video traffic. Typical predicted annual growth rates are 30 % or higher for wireline IP-based video services and 90 % for Internet TV in mobile networks. Due to these high growth rates, video traffic will account for a large portion of the traffic in communication networks. Estimates by Cisco, Inc. predict that video will contribute close to two thirds of the mobile network traffic by 2014. Network designers and engineers therefore need a basic understanding of video traffic in order to account for the video traffic characteristics in designing and evaluating communication services for this important type of network traffic.

The encoders that are used to compress video before network transport have significantly advanced in recent years. These video encoding advances have important implications for the network transport of encoded video. The purpose of this tutorial article is to give generalists in the communication networks area an overview of the recent developments in video coding and to explain the main implications of these developments for the transport of encoded video in communication networks.

Supported in part by the National Science Foundation through grant No. CRI-0750927.

Please direct correspondence to M. Reisslein.

P. Seeling is with the Dept. of Computer Science, Central Michigan University, Mount Pleasant, MI 48859, Email: pseeling@ieee.org

M. Reisslein is with the School of Electrical, Computer, and Energy Engineering, Arizona State University, Tempe, AZ 85287-5706, http://trace.eas.asu.edu, Email: reisslein@asu.edu

This article covers three main areas of video coding advances: (i) efficient encoding of conventional two-dimensional (2D) video into a non-scalable video bitstream, i.e., a video bitstream that is not explicitly designed to be scaled (e.g., reduced in bitrate) during network transport, (ii) scalable video coding, i.e., video coding that is explicitly designed to permit for scaling (e.g., bitrate reduction) during network transport, and (iii) efficient non-scalable encoding of three-dimensional (3D) video. For the non-scalable and scalable video coding, we consider the standards

- MPEG-2, formally referred to as H.262/MPEG-2 Video
- MPEG-4, formally MPEG-4 Part 2 Visual
- H.264 SVC, formally H.264/MPEG-4 Advanced Video Coding (which is commonly abbreviated as H.264/AVC) with Scalable Video Coding extension
- H.265/HEVC, formally H.265/MPEG-H Part 2.

For 3D video, we consider the Multiview Video Coding (MVC) standard, formally Stereo and Multiview Video Coding extension of the H.264/MPEG-4 AVC standard.

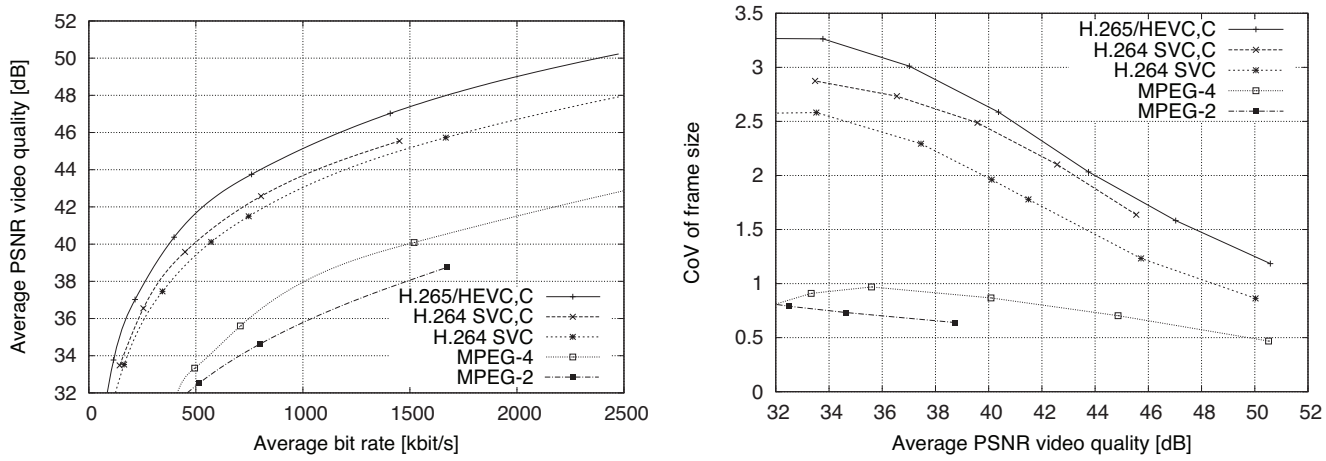
Video coding specific characteristics and performance metrics of these latest video coding standards are covered for non-scalable coding in [1], [2], for scalable video coding in [3], [4], and for 3D video in [5]. The evaluations in this existing literature focus primarily on the rate-distortion (RD) characteristics of the video encoding, i.e., the video quality (distortion) as a function of the mean bitrate of an encoded video stream, for relatively short video sequences (typically up to 10 s). In contrast, this article gives an overview of these video coding standards from a communication networks perspective. This article includes evaluations of the variability of the encoded video traffic, which is a key concern for network transport, for long video sequences (of 10 minutes or more).

Video traces for all three areas of video encoding covered in this article are available from `http://trace.eas.asu.edu`. Video traces characterize the encoded video through plain text files that provide the sizes of encoded frames and the corresponding video quality (distortion) values. The video traces facilitate traffic modeling, e.g., [6], [7], as well as the evaluation of a wide range of emerging video transport paradigms, such as peer-to-peer streaming [8], mobile video streaming [9], and IPTV [10].

II. Non-scalable Video Streams

A. Overview of Video Encoding

We first give a brief overview of the main encoding steps in the major video coding standards and then review the advances in these main encoding steps and their implications for the network transport of encoded video. In the major video coding standards, a given video frame (picture) is



(a) Rate-distortion (RD) curve

(b) Rate variability-distortion (VD) curve

Fig. 1. Comparison of RD and VD curves for MPEG-2, MPEG-4, H.264 SVC in single-layer mode without and with cascading (C) quantization parameters (QPs), and H.265/HEVC with cascading QPs.

divided into blocks. The blocks are then intra-coded, i.e., encoded by considering only the current frame, or inter-coded, i.e., encoded with references to (predictions from) neighboring frames that precede or succeed the current frame in the temporal display sequence of the frames. The inter-coding employs motion compensated prediction, whereby blocks in the reference frames that closely resemble the considered block to be encoded are found; the considered block is then represented by motion vectors to the reference blocks and the prediction errors (differences between the considered block and the reference blocks). The luminance (brightness) and chrominance (color) values in a block, or the corresponding prediction errors from reference blocks, are transformed to obtain a block of transform coefficients. The transform coefficients are then quantized, whereby the quantization is controlled by a quantization parameter (QP), and the quantized values are entropy coded. The entire sequence of encoding steps is commonly optimized through RD optimization, which has advanced along with the individual encoding steps.

MPEG-2 introduced three frame types that are also used in the subsequent coding standards: Intra-coded (I) frames are encoded as stand-alone pictures without references (dependencies) on other frames. Predictive-coded (P) frames are encoded with inter-coding with respect to only preceding I (or P) frames in the temporal frame display order. Bi-directional-coded (B) frames are inter-coded with respect to both preceding (i.e., past) as well as succeeding (i.e., future) I (or P) frames, as illustrated in Fig. 2(a). A group of frames (pictures) from one I frame to the frame immediately preceding the next I frame is commonly referred to as a Group of Pictures (GoP).

We plot the RD curves, i.e., the video quality as a function of the mean bitrate, obtained with reference software implementations for the considered coding standards (and the `ffmpeg` program for MPEG-2 which does not have reference software), in Fig 1(a). We represent the video

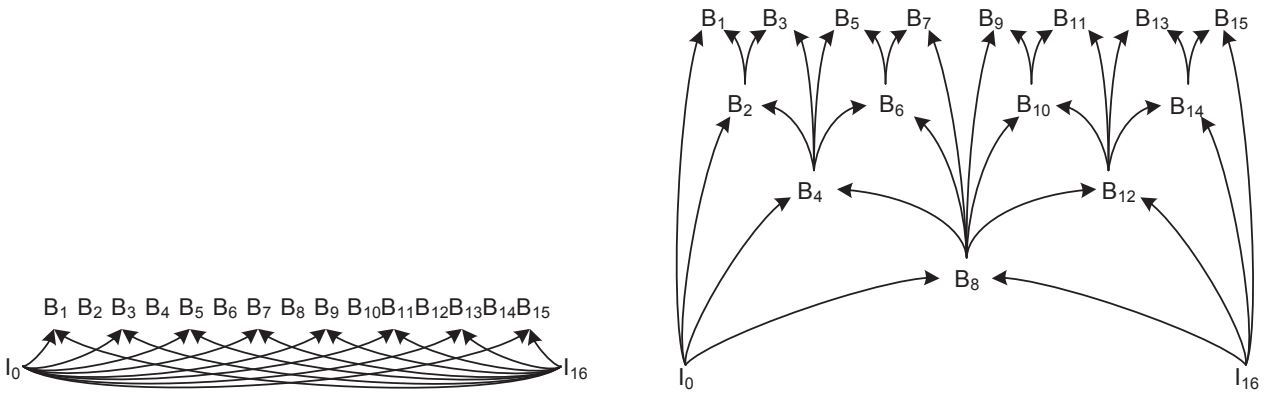
quality in terms of the Peak Signal to Noise Ratio (PSNR) between the luminance values in the sequence of original (uncompressed) video frames and the sequences of encoded (compressed) video frames. The PSNR is an elementary objective video quality metric, for an overview of video quality metrics, we refer to [11]. In Fig. 1(b), we plot the rate variability-distortion (VD) curve defined as a plot of the Coefficient of Variation (CoV) of the encoded frame sizes (in Bytes), i.e., the standard deviation of the frame sizes normalized by the mean frame size. The curves in Fig. 1 are for the 10 minute (17,682 frames) *Sony Digital Video Camera Recorder* demo sequence, which is a widely used video test sequence with a mix of scenes with high texture content and wide a range of motion activity levels. The video for this illustration is in the 352×288 pixel common interchange format and has a frame rate of 30 frames/s, i.e., a frame period of 1/30 s. Video traces and plots for a wide range of other videos are available from `http://trace.eas.asu.edu`.

We observe from Fig. 1 increasing RD efficiency, i.e., higher PSNR video quality for a prescribed mean bitrate, as the video coding standards advance from MPEG-2 through H.265/HEVC. We also observe increasing traffic variability, i.e., higher CoV values for a prescribed PSNR video quality with advancing video coding standards. The observed increased RD efficiencies and increased traffic variabilities are the aggregate effect of the advances in the main coding steps, which we now briefly review.

1) Frame Partitioning into Blocks and Intra-coding of Video Frames: As the video coding standards advanced, the partitioning of a video frame (picture) into blocks has become increasingly flexible to facilitate high RD efficiency in the subsequent coding steps. While MPEG-2 was limited to a fixed block size of 16×16 luminance pixels, MPEG-4 permitted 16×16 and 8×8 blocks, and H.264/AVC introduced block sizes ranging from 4×4 to 16×16. High Efficiency Video Coding (H.265/HEVC) [1] introduces frame partitioning into coding tree blocks of sizes 16×16, 32×32, and 64×64 luminance pixels which can be flexibly partitioned into multiple variable-sized coding blocks.

While preceding standards had very limited intra-coding within a given frame, H.264/AVC introduced spatial intra-coding to predict a block in a frame from a neighboring block in the same frame. H.265/HEVC significantly advances intra-coding through the combination of the highly flexible coding tree partitioning and a wide range of intra-frame prediction modes.

2) Inter-coding (Temporal Prediction) of Video Frames: Advances in inter-coding, i.e., the encoding of frames with motion compensated prediction from other frames in the temporal frame display sequence, have led to highly significant RD efficiency increases in the advancing video coding standards. In MPEG-2 and MPEG-4, B frames are predicted from the preceding I (or P) frame and the succeeding P (or I) frame, see Fig. 2(a). Compared to the motion compensated



(a) Classical B frame prediction structure

(b) Hierarchical B frame prediction structure

Fig. 2. Illustration of classical B frame prediction structure used in MPEG-2 and MPEG-4 (without reference arrows for even frames to avoid clutter) and dyadic hierarchical B frame prediction structure used in H.264 SVC and H.265/HEVC.

prediction at half-pixel granularity in MPEG-2, MPEG-4 employs quarter-pixel granularity for the motion compensated prediction as well as additional RD efficiency increasing enhanced motion vector options and encoding. H.264/AVC and H.265/HEVC employ similarly quarter-pixel granularity for the motion compensated prediction and further improve the motion parameters. Comparisons of RD and VD curves for MPEG-4 and H.264/AVC with the classical B frame prediction from only I (and P) frames, as illustrated in Fig. 2(a), for the different granularities of the motion compensated prediction have indicated that finer granularities improve the RD efficiency at the expense of increased frame size variability [12].

H.264/AVC and H.265/HEVC fundamentally advance inter-coding by predicting B frames from potentially multiple past and/or future B frames. Specifically, in H.264 SVC and H.265/HEVC, the frames in a GoP form typically a dyadic prediction hierarchy illustrated in Fig. 2(b). I frames (and P frames, if present in the GoP) form the base layer of the hierarchy. With β B frames between successive I (or P) frames, whereby $\beta = 2^\tau - 1$ for a positive integer τ for the dyadic hierarchy, the B frames form $\tau = \log_2(\beta + 1)$ layers. For instance, in the GoP structure (without any P frames) illustrated in Fig. 2(b), the $\beta = 15$ B frames between successive I frames form $\tau = 4$ layers. A B frame in a layer $n,\ 1 \leq n \leq \tau$, is inter-coded with reference to the immediately preceding and succeeding frames in lower layers $n-1, n-2, \ldots, 0$, whereby layer 0 corresponds to the base layer. For instance, frame B_3 is encoded through motion compensated prediction with reference to frames B_2 and B_4, while frame B_2 is encoded with reference to frames I_0 and B_4.

3) Quantization, Transform, and Entropy Coding: MPEG-2 and MPEG-4 allow for different quantization parameter (QP) settings for the three different frame types, namely I, P, and B frames. Generally, it is an RD-efficient coding strategy to quantize I frames relatively finely, i.e., with a small QP, since the I frames serve as a reference for the P and B frames. Increasingly

coarse quantization, i.e., successively larger QPs, for P and B frames can increase RD efficiency, since P frames serve only as reference for B frames and B frames are not employed as reference for inter-coding in MPEG-2 and MPEG-4, i.e., no other frames depend on B frames. In H.264 SVC and H.265/HEVC, this principle of increasingly coarse quantization for frames with fewer dependent frames can be pushed further by increasing the QP with each level of the frame hierarchy. This strategy is commonly referred to as QP cascading and is examined quantitatively for H.264 SVC in Fig. 1. We observe from Fig. 1(a) that H.264 SVC with QP cascading slightly improves the RD efficiency, i.e., increases the PSNR video quality for a prescribed mean bitrate, while substantially increasing the traffic variability, as observed in Fig. 1(b). The QP cascading leads to increasing compression for higher levels of the B frame hierarchy, which increases RD efficiency as these B frames in the higher layers are used as references for fewer other B frames. However, the interspersing of more intensely compressed frames inbetween other less compressed frames increases the variability of the encoded frame substantially.

MPEG-2 and MPEG-4 employ the discrete cosine transform (DCT) on blocks of 8×8 samples. H.264 SVC provides more flexibility with 4×4 and 8×8 transforms and H.265/HEVC further significantly increases the flexibility with transforms that match the flexibility of the code tree block structure.

MPEG-2 and MPEG-4 employ a basic variable-length coding of the coefficients resulting from the DCT. H.264/AVC introduced more efficient context-adaptive variable-length coding (CAVLC) and context-adaptive binary arithmetic coding (CABAC). The comparison of RD curves for CAVLC and CABAC in the context of H.264/AVC with classical B frame prediction in [13] showed improved RD efficiency with CABAC. H.265/HEVC employs CABAC with refined context selection.

B. Implications of Coding Standard Advances for Video Network Transport

From the perspective of the transport network there are two main implications from the advances of the video coding standards: (i) increased traffic variability as the RD efficiency increases, and (ii) additional timing constraints due to the frame dependencies in the hierarchical B frame structure.

1) Traffic Variability: As we observe from Fig. 1(b), the frame size variability increases from CoV values near one for MPEG-2 and MPEG-4 to CoV values above two for H.264 SVC (operating in single-layer mode), and to values above three for H.265/HEVC. Investigations, e.g., [12], [14], of this general phenomenon of increasing traffic variability with increasing RD efficiency of the video coding have revealed that the while the video coding advances lead to more efficient compression of the I and P frames, the improvements in B frame compression

are more pronounced, leading to increased variability in the sequences of encoded I, P, and B frames. Overall, the results in Fig. 1 indicate that the modern video coding standards allow for the transmission of higher quality video with lower mean bitrates compared to older standards. However, the network needs to accommodate vastly higher fluctuations of the bitrates required to transport the encoded frame sequence.

In order to translate the gain in RD coding efficiency of the modern video coding standards into increased number or quality of transported video streams it is critical to develop efficient smoothing, buffering, and multiplexing strategies. The goal of these strategies is to mitigate the traffic fluctuations, such that the number of video streams that can be supported by a given transport network is mainly dependent on the mean bitrate of the video. Without such strategies for mitigating the traffic variability, the higher traffic variability of the modern coding standards can result in fewer supported streams compared to an older standard that has a slightly higher mean bitrate, but significantly lower traffic variability.

2) Timing Constraints due to Frame Dependencies: The dyadic hierarchical B frame structure of H.264 SVC and H.265/HEVC imposes additional constraints on the timing of the frame transmissions compared to the classical B frame prediction employed in the preceding MPEG-4 and MPEG-2 standards. Generally, a given frame can only be encoded *after* all references frames have been captured by a video camera and encoded. For instance, in Fig. 2(b), frame B_1 can only be encoded after frames I_0, I_{16}, B_8, B_4, and B_2 have been encoded. In contrast, with classical B frame prediction illustrated in Fig. 2(a), frame B_1 can be immediately encoded after frames I_0 and I_{16} have been encoded. Thus, the dependencies between B frames in the dyadic B frame hierarchy introduce an additional delay of $[\log_2(1+\beta)] - 1$ frame periods, see [14] for details of delay analysis.

III. Scalable Video Streams

A. Layered Video Encoding

MPEG-2 and MPEG-4 provide scalable video coding into a base layer giving a basic version of the video and one or several enhancement layers that improve the video. The quality layering can be done in the dimensions of temporal resolution (video frame frequency), spatial resolution (pixel count in horizontal and vertical dimensions), or PSNR video quality. The layering in the PSNR quality dimension employs coarse quantization (with high QP) for the base layer, and successively finer quantization (smaller QPs) for the enhancement layers that successively improve the PSNR video quality. These layered scalability modes permit scaling of the encoded video stream at the granularity of complete enhancement layers, e.g., a network node can drop an enhancement layer if there is congestion downstream. MPEG-4 has a form of sub-layer PSNR quality scalability referred to as Fine Grained Scalability (FGS). With FGS there is one enhancement layer that

can be scaled at the granularity of individual Bytes of video encoding information. With both MPEG-2 and MPEG-4, the flexibility of scaling the encoded video stream comes at the expense of a relatively high encoding overhead that significantly reduces the RD efficiency of the encoding and resulted in very limited adoption of these scalability modes in practice.

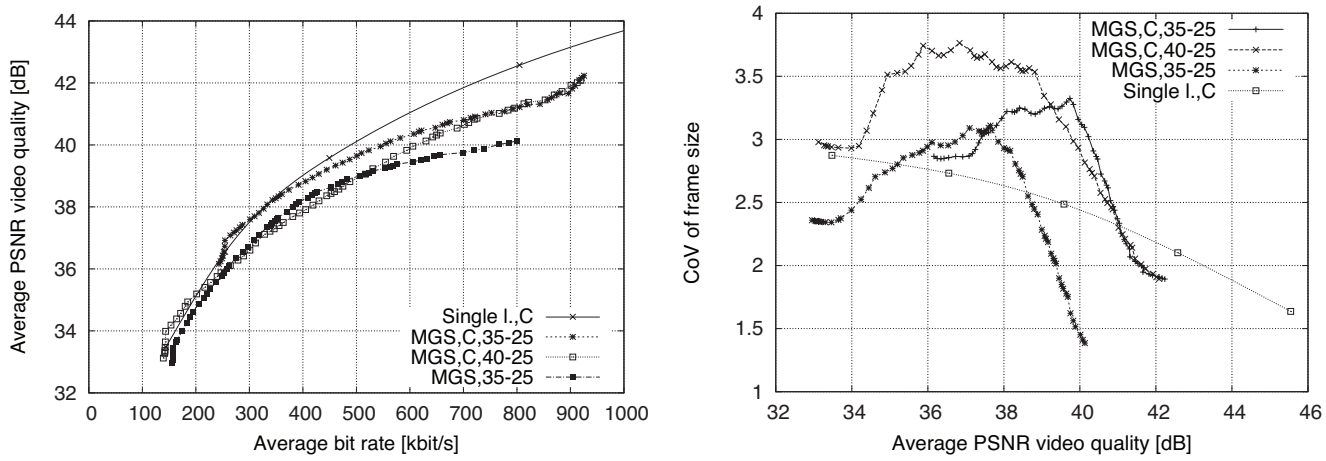
Similar to the preceding MPEG standards, the Scalable Video Coding (SVC) extension of H.264/AVC [3] provides layered temporal, spatial, and PSNR quality scalability, whereby the layered PSNR quality scalability is referred to as Coarse Grain Scalability (CGS). While these H.264 SVC layered scalability modes have reduced encoding overhead compared to the preceding MPEG standards, the overhead is still relatively high, especially when more than two enhancement layers are needed. Scalable encoding is presently not yet available in the H.265 standard.

B. H.264 SVC Medium Grain Scalability (MGS) Encoding

H.264 SVC has a novel Medium Grain Scalability (MGS) that splits a given PSNR quality enhancement layer of a given video frame into up to 16 MGS layers that facilitate highly flexible and RD efficient stream adaption during network transport.

1) MGS Encoding: As for all PSNR quality scalable encodings, the base layer of an MGS encoding provides a coarse quantization with a relatively high QP, e.g., $B = 35$ or 40. MGS encodings have typically one enhancement layer providing a fine quantization with a relatively small QP, e.g., $E = 25$. When encoding this enhancement layer, the 16 coefficients resulting from the discrete cosine transform of a 4×4 block are split into MGS layers according to a weight vector (a similar splitting strategy is employed for larger blocks). For instance, for the weight vector $\mathbf{W} = [1, 2, 2, 3, 4, 4]$, the 16 coefficients are split into six MGS layers as follows. The lowest frequency coefficient is assigned to the first MGS layer $m = 1$, the next two higher frequency coefficients are assigned to the second MGS layer $m = 2$, and so on, until the four lowest frequency coefficients are assigned to the sixth MGS layer $m = 6$, the highest MGS layer in this example. For network transport, the base layer and each MGS layer of a given frame is encapsulated into a so-called Network Adaptation Layer Unit (NALU).

2) Scaling MGS Streams in Network: When scaling down an MGS video stream at a network node, dropping MGS layers uniformly across the frame sequence results in low RD efficiency of the downscaled stream. This is due to the dependencies in the B frame hierarchy. Specifically, dropping an MGS layer from a B frame that other B frames depend on, e.g., frame B_8 in Fig. 2(b) reduces not only the PSNR quality of frame 8, but also of all dependent frames B_1 – B_7 and B_9 – B_{15}. It is therefore recommended to drop MGS layers first from the B frames without any dependent frames, i.e., the odd-indexed B frames in the highest layer in Fig 2(b), then drop MGS layers from the B frames with one dependent B frame, i.e., the B frames in the second highest layer in Fig. 2(b), and so on.



(a) Rate-distortion (RD) curve

(b) Rate variability-distortion (VD) curve

Fig. 3. Comparison of RD and VD curves for H.264 SVC single-layer encoding with cascading QPs and H.264 SVC with medium-grain scalability (MGS) for base layer QPs $B = 35$ and 40 and enhancement layer QP $E = 25$ with and without QP cascading (C).

Alternatively, MGS encodings can be conducted so that each NALU is assigned a priority ID between 0 indicating lowest priority and 63 indicating highest priority for RD efficiency. These priority IDs can be assigned by the video encoder based on RD optimization. For downscaling an MGS stream with priority IDs, a network node first drops MGS layers (NALUs) with priority ID 0, then priority ID 1, and so on.

3) RD and VD Characteristics of H.264 SVC Video Streams: In Fig. 3, we plot the RD and VD curves of the *Sony* video for the priority ID stream scaling. Comparing the RD curves of the MGS streams with cascaded QPs (C) with the RD curve from single-layer encoding, we observe that H.264 SVC MGS provides the flexibility of scaling the stream bitrate in the network with a low encoding overhead from the lower end to the mid region of the quality adaptation range between the base layer only and the base plus full enhancement layer. For instance, the RD curve of MGS, C encoding with $B = 35$, $E = 25$ in Fig. 3(a) is very close to the RD curve of the single-layer encoding from its lower end near 36 dB to the middle of the adaptation region at about 39 dB. Near the lower end of the adaptation range, only the NALUs with the highest priority ID, i.e., the highest ratio of contribution towards PSNR video quality relative to size (in Byte) are streamed, resulting in high RD efficiency that can even slightly exceed the RD efficiency of the single-layer encoding. Towards the upper end of the adaptation range, all NALUs, even those with small PSNR contribution to size ratios are streamed, resulting in reduced RD efficiency. The difference in RD efficiency between the single-layer encodings and the MGS encoding at the upper end of the MGS adaptation range is mainly due to overhead of the MGS encoding.

Similar to the single-layer encoding, we observe from the comparison of MGS streams encoded

without and with cascaded QPs that the cascading increases both the RD efficiency and the traffic variability. We also observe that the cascaded-QPs MGS encoding with the larger adaptation range ($B = 40$, $E = 25$) gives somewhat lower RD efficiency and substantially higher traffic variability than the corresponding $B = 35$, $E = 25$ encoding. That is, the increased adaptation flexibility of the $B = 40$, $E = 25$ encoding comes at the expense of reduced RD efficiency and very high traffic variability reaching CoV values above 3.5. Overall, we observe that the adaptation flexibility of H.264 MGS comes at the expense of increased traffic variability compared to the single-layer encodings.

IV. 3D Video Streams

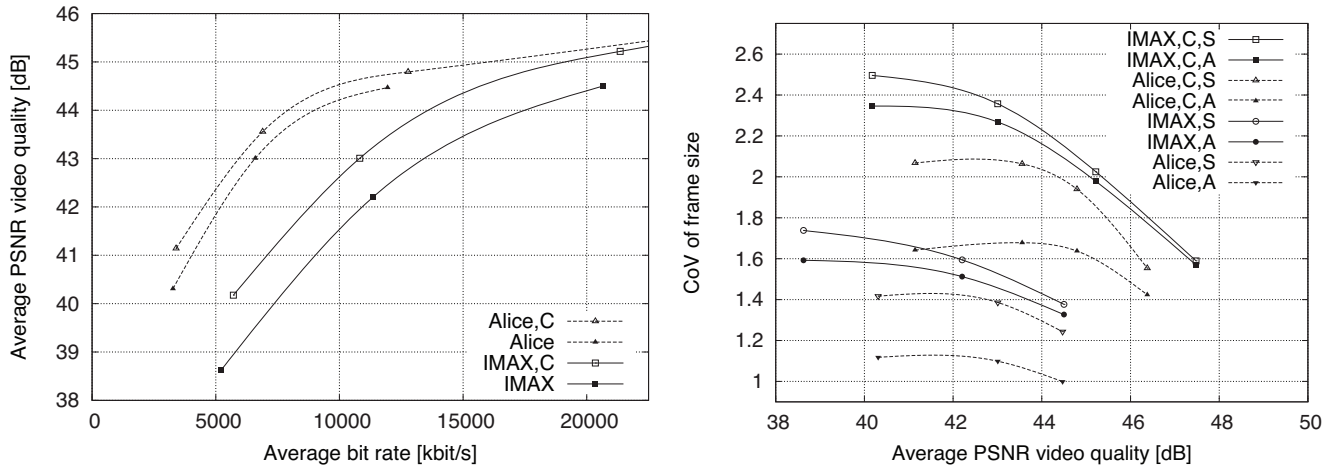
A. Overview of 3D Video

3D video employs views from two slightly shifted perspectives, commonly referred to as the left view and the right view, of a given scene. Displaying these two slightly different views gives viewers the perception of depth, i.e., a three-dimensional (3D) video experience. Since two views are involved, 3D video is also sometimes referred to as stereoscopic video. The concept of employing multiple views from different perspectives can be extended to more than two views and is generally referred to as multiview video.

B. 3D Video Encoding and Streaming

3D video streaming requires the transport of the two sequences of video frames resulting from the two slightly different viewing perspectives over the network to the viewer. Since the two views capture the same scene, their video frame content is highly correlated. That is, there is a high level of redundant information in the two views that can be removed through encoding (compression). The Multiview Video Coding (MVC) standard builds on the inter-coding techniques that are applied across a temporal sequence of frames in single-layer video coding to extract the redundancy between the two views of 3D video. More specifically, MVC typically first encodes the left view and then predictively encodes the right view with respect to the left view.

One approach to streaming the MVC encoded 3D video is to transmit the encoded left and right views as a frame sequence with twice the frame rate of the original video, i.e, left view of first video frame (from first capture instant), right view of first video frame, left view of second video frame, right view of second video frame, and so on. Since the right view is encoded with respect to the left view, it is typically significantly smaller (in Bytes) and the sequence of alternating left and right views result in high traffic variability, as illustrated in the next section.



(a) Rate-distortion (RD) curves

(b) Rate variability-distortion (VD) curves

Fig. 4. RD and VD characteristics of MVC encodings without and with cascaded QPs (C) of 35 minutes each of 3D videos *Alice in Wonderland* and *IMAX Space Station* with full HD 1920×1080 pixel resolution. The encoded left and right views are streamed sequentially (S) or are streamed aggregated (A) into multiview frames.

Another MVC streaming approach is to aggregate the left and right views from a given video frame (capture instant) into one *multiview frame* for transmission. The sequence of multiview frames has then the same frame rate as the original video.

An alternative encoding approach for 3D video is to sequence the left and right views to form a video stream with doubled frame frequency and feed this stream into a single-view video encoder, such as H.264 SVC or H.265/HEVC. This approach essentially translates the inter-view redundancies into redundancies among subsequent frames. Similar to MVC encoding, the two encoded views for a given capture instant can be transmitted sequentially, or aggregated.

Yet another encoding alternative is to down-sample (sub-sample) the left and right views to fit into one frame of the original video resolution. For instance, the 1920×1080 pixels left and right views are horizontally subsampled to 960×1080 pixels so that they fit side-by-side into one 1920×1080 frame. This side-by-side approach permits the use of conventional 2D video coding and transmission systems, but requires interpolation at the receiver to obtain the left and right views at the original 960×1080 pixel resolution.

C. RD and VD Characteristics of 3D Video Streams

In Fig. 4, we plot the RD and VD curves for two representative 3D videos encoded with MVC without QP cascading and with QP cascading. We observe from Fig. 4(a) that (i) for a prescribed mean bitrate, the PSNR video quality is higher for the *Alice* video compared to the *IMAX* video, and (ii) that the QP cascading improves the RD efficiency by up to about 0.5 dB for the *Alice* video and almost 1 dB for the *IMAX* video. These different RD efficiency levels and RD efficiency increases with QP cascading are due to the different content of the two videos.

The *IMAX* video is richer in texture and motion and thus "more difficult" to compress and higher RD efficiency gains can be achieved with improved coding strategies.

We observe from Fig. 4(b) that (i) the more varied *IMAX* video gives higher frame size variability and that (ii) QP cascading increases the frame size variability, mirroring the above observations for single-view (2D) video. We also observe from Fig. 4(b) that the sequential (S) transmission of the encoded left and right views gives substantially higher traffic variability than the aggregated (A) transmission. Thus, the aggregated transmission with its less pronounced traffic fluctuations is typically preferable for network transport.

Recent studies [15] indicate that the frame sequential encoding results in somewhat lower RD efficiency and substantially lower traffic variability than MVC encoding. As a result, when statistically multiplexing a small number of unsmoothed 3D streams, MVC encoding and frame sequential encoding, both with the aggregated (A) transmission strategy, require about the same link bitrate. Only when statistically multiplexing a large number of streams, or employing buffering and smoothing does MVC encoding reduce the required network bitrate compared to frame sequential encoding. The studies in [15] also indicate that the side-by-side 3D video approach gives relatively poor RD performance due to the involved sub-sampling and subsequent interpolation.

V. Conclusion

We have given a tutorial overview of modern video coding standards for multimedia networking. We have outlined the advances in the main video coding standards for non-scalable (single-layer) video, scalable video, and 3D video streams. For single-layer video, we gave an overview of MPEG-2, MPEG-4, H.264/AVC with Scalable Video Coding (SVC) extension, and H.265/HEVC and compared their rate-distortion (RD) and rate variability-distortion (VD) characteristics. This comparison included the first study of the traffic variability of H.265/HEVC encoding for long videos as well as an original study of the effects of cascading of quantization parameters (QPs) for the different levels of the hierarchical dyadic B frame prediction structure of H.264 SVC. We found that the advances in the video coding standards have led to increased RD efficiency, i.e., higher video quality for a prescribed mean video bitrate, but also substantially increased traffic variability. The coefficient of variation (standard deviation normalized by mean), of the encoded frame sizes, a common measure of video traffic variability is around one for MPEG-2 and MPEG-4, but reaches values above two and three for H.264 SVC and H.265/HEVC, respectively.

For scalable video coding, we gave a brief overview of H.264 SVC Medium Grain Scalability (MGS) and the scaling of an encoded H.264 SVC MGS video bitstream in a network node. We compared the RD and VD characteristics of the scaled H.264 MGS stream with corresponding single-layer encodings. We illustrated that H.264 SVC MGS streams can be flexibly scaled in

the network in the lower region of the quality adaptation range while maintaining RD efficiency very close to the unscalable single-layer encodings. For 3D video, we outlined the encoding and streaming of the two views and examined the RD and VD characteristics of the streams.

Overall, we found that the advancing video coding standards have reduced the mean bitrate for a given targeted video quality, i.e., increased the RD efficiency, and made video encoding more flexible through highly RD efficient scalable video coding and coding standards for 3D video. However, the traffic produced by these advanced video coding standards has very pronounced fluctuations that are vastly increased compared to the MPEG-2 and MPEG-4 standards. Thus, effective video traffic management and smoothing are highly important so that the network operator can indeed reap the benefit of the reduced mean bitrates through higher numbers of supported video streams.

ACKNOWLEDGMENT

We are grateful for thorough feedback from Dr. Geert Van der Auwera from Qualcomm, Inc. on an earlier version of this manuscript.

REFERENCES

[1] G. Sullivan, J.-R. Ohm, W.-J. Han, and T. Wiegand, "Overview of the High Efficiency Video Coding (HEVC) standard," *IEEE Transactions on Circuits and Systems for Video Technology, in print*, 2012.

[2] J.-R. Ohm, G. Sullivan, H. Schwartz, T. Tan, and T. Wiegand, "Comparison of the coding efficiency of video coding standards—including High Efficiency Video Coding (HEVC)," *IEEE Transactions on Circuits and Systems for Video Technology, in print*, 2012.

[3] H. Schwarz, D. Marpe, and T. Wiegand, "Overview of the scalable video coding extension of the H.264/AVC standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 17, no. 9, pp. 1103–1120, Sep. 2007.

[4] M. Wien, H. Schwarz, and T. Oelbaum, "Performance analysis of SVC," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 17, no. 9, pp. 1194–1203, Sep. 2007.

[5] A. Vetro, T. Wiegand, and G. J. Sullivan, "Overview of the stereo and multiview video coding extensions of the H.264/MPEG-4 AVC standard," *Proceedings of the IEEE*, vol. 99, no. 4, pp. 626–642, Apr. 2011.

[6] D. Fiems, B. Steyaert, and H. Bruneel, "A genetic approach to Markovian characterisation of H.264/SVC scalable video," *Multimedia Tools and Applications*, vol. 58, no. 1, pp. 125–146, May 2012.

[7] A. Lazaris and P. Koutsakis, "Modeling multiplexed traffic from H.264/AVC videoconference streams," *Computer Communications*, vol. 33, no. 10, pp. 1235–1242, Jun. 2010.

[8] N. Ramzan, E. Quacchio, T. Zgaljic, S. Asioli, L. Celetto, E. Izquierdo, and F. Rovati, "Peer-to-peer streaming of scalable video in future Internet applications," *IEEE Communications Magazine*, vol. 49, no. 3, pp. 128–135, Mar. 2011.

[9] K. Ma, R. Bartos, S. Bhatia, and R. Nair, "Mobile video delivery with HTTP," *IEEE Communications Magazine*, vol. 49, no. 4, pp. 166–175, Apr. 2011.

[10] D. Manzato and N. Fonseca, "A comparison of channel switching schemes for IPTV systems," in *Proceedings of IEEE ICC*, 2011, pp. 1–6.

[11] F. Yang and S. Wan, "Bitstream-based quality assessment for networked video: A review," *IEEE Communications Magazine*, vol. 50, no. 11, pp. 203–209, Nov. 2012.

[12] G. Van der Auwera, P. David, and M. Reisslein, "Traffic characteristics of H.264/AVC variable bit rate video," *IEEE Communications Magazine*, vol. 46, no. 11, pp. 164–174, Nov. 2008.

[13] D. Marpe, T. Wiegand, and G. Sullivan, "The H.264/MPEG–4 advanced video coding standard and its applications," *IEEE Communications Magazine*, vol. 44, no. 8, pp. 134–143, Aug. 2006.

[14] G. Van der Auwera and M. Reisslein, "Implications of smoothing on statistical multiplexing of H.264/AVC and SVC video streams," *IEEE Transactions on Broadcasting*, vol. 55, no. 3, pp. 541–558, Sep. 2009.

[15] A. Pulipaka, P. Seeling, M. Reisslein, and L. Karam, "Traffic and statistical multiplexing characterization of 3D video representation formats," 2012, preprint available from `http://mre.faculty.asu.edu/3drep.pdf`.