An Anatomy of SVC for Full HD Video Streaming

Mohammad Reza Zakerinasab  
Department of Computer Science  
University of Calgary  
Calgary, Canada  
Email: mrzakeri@ucalgary.ca

Mea Wang  
Department of Computer Science  
University of Calgary  
Calgary, Canada  
Email: meawang@ucalgary.ca

Abstract—The continuous developments and improvements of network infrastructure along with the growing number of modern smartphones, tablets and smart TVs have led to an increasing popularity of multimedia applications, such as video conferencing, video streaming and mobile TV. Towards delivering video stream to diverse devices over a heterogamous network, scalable video coding (SVC) has received many research attentions. SVC is an extension of H.264/AVC that allows a video streaming service provider to encode a high quality video into a number of scalable layers. The receivers of the stream may decode the video at the appropriate quality level that is suitable for their hardware/software capabilities and network connections. Nevertheless, compared to single layer H.264/AVC, the de facto standard for many commercial streaming service providers such as YouTube, SVC is not widely deployed, mostly due to its overhead in terms of bitrate and complexity. In this paper, we conduct a thorough study on the performance of SVC for full HD video streaming. Our performance analysis identifies good and bad uses of SVC, to quantify the coding overhead, and benchmarks the SVC video quality under different spatial, temporal, and quality settings. Through the use a set of carefully selected and diverse video sequences, we also identify the types of video that can benefit from SVC.

Keywords—H.264/SVC, performance analysis, full HD video streaming.

I. INTRODUCTION

According to a recent study by Cisco [1], real-time entertainment (primarily video streaming) constitutes more than 53% of the Internet traffic. With the wide deployment of broadband access, even now in cellular networks, and the advancement in smartphones and tablets, there is a raising demand for full HD video streaming. Despite the diversity in hardware/software capabilities and variations in network connectivities, end users expect the best possible quality in the video streaming experience. Moreover, increasing traffic from cellular networks poses another challenge in maintaining the playback quality over a streaming session as the network characteristics is much more complicated and may fluctuate significantly.

One solution to this problem is to prepare several versions of a video for a pre-defined set of resolutions. For example, YouTube [2] encodes videos in H.264 format and supports variable bitrates. They offer seven recommended bit rates, from 144p (low resolution) to 2160p (high resolution). A user may manually select the bit rate, and may also leave it to YouTube to send the video in the quality that best suits the device and network conditions. On the one hand, extra storage is needed on the server side to store different versions of the video. On the other hand, users are limited to the bitrates offered by YouTube. Scalable video coding (SVC), an extension of H.264/AVC, has emerged to support full HD video streaming to diverse devices with different network connection qualities. With SVC, the server maintains a single version of each video, but the video content is delivered to end-user devices at different quality levels according to the device capabilities and network conditions.

Moreover, SVC can address another deficiency in the YouTube-like streaming systems. In these systems, if a user views the same video on different devices at different quality levels, a copy at each quality level must be downloaded, which leads to increasing network traffic and high workload demand on the video server. With SVC, the video server may serve one copy of the video in a properly chosen format to the router that is adjacent to the end-user devices. Then the router may serve video streams that best match the characteristics of each device by sending a proper set of layers. This not only alleviates the heavy burden on the streaming server and the Internet routers, but also delivers the video in better quality to each end-user device, thanks to the fine-grained bitrate adaptation provided by the multi-dimensional scalability of SVC [3].

As opposed to conventional single-stream videos, a scalable video consists of multiple video substreams at different quality levels. The substreams are normally referred to as layers [3]. In a nutshell, SVC encodes a video with spatial scalability (layers with different resolutions), temporal scalability (layers with different frame rates), quality scalability (layers with different qualities), or any arbitrary combination of them. Furthermore, SVC can tolerate frame losses, i.e., even if frames are dropped during transmission, the original video can still be rendered with little distortion. For this reason, SVC has received great research attention and has been used in many proposals for improving multimedia streaming systems [4].

Nevertheless, in contrast to H.264/AVC [5], which is the de facto standard for single layer video coding, just a few commercial streaming systems utilize SVC as the video codec. The main reason is the bitrate and computational overhead that is introduced by the multi-stream representation of the video. In this paper, we conduct a systematic study on the use of SVC for full HD video streaming. Our goal is to identify the good and bad uses of SVC, to quantify the coding...
overhead, and to benchmark the video quality under different spatial, temporal, and quality settings. Using a set of carefully selected and diverse video sequences, we also identify the types of video that can benefit from SVC. Our study reveals that the efficiency and computational gap between SVC and H.264/AVC is much less when encoding high quality videos, e.g., in full HD resolution. There are three more interesting observations: (1) Replacing P-frames with I-frames in complex video sequences can decrease the encoding complexity with a very mild increase in the bitrate of the encoded video; (2) When the video is complex and the consecutive frames are considerably different, adding spatial layers may decrease the bitrate of the encoded video; and (3) Increasing the frame size decreases the computational and bitrate overhead of non-dyadic spatial layers, which can be helpful as the diverse screen resolutions of end user devices limits the application of SVC if only dyadic spatial resolution is employed.

The remainder of this paper is organized as follows. Sec. II reviews existing performance analyses on SVC. Sec. III provides an overview of SVC. Sec. IV presents the setup of our study, followed by the performance analysis in Sec. V. Finally, Sec. VI concludes the paper.

II. RELATED WORKS

Due to the diverse hardware/software capabilities and network connectivities of end-user devices, the scalable approach to video streaming has received many research attentions. However, the research work done so far about SVC that is related to this study can be briefly summarized into three directions: comparing SVC with other standards, addressing the impact of different layering configurations in SVC from an objective quality perspective, and the subjective video quality offered by SVC. This section provides a brief review for each direction and relates this work to the current research results.

Comparing SVC with other standards: There are many coding standards used by commercial streaming systems. Therefore researchers have compared the performance of SVC with other coding standards, mainly focusing on the rate-distortion performance, i.e., the objective quality of the video (normally measured by PSNR) as a function of average bitrate. For instance, Wien et al. [6] analyzed the effect of quality scalability and spatial scalability (both dyadic and non-dyadic modes) on the objective quality of eight video sequences in two quality classes, common intermediate format (CIF) (352 × 288 pixels) and quarter CIF (QCIF) (176 × 144 pixels). In this study the rate-distortion performance of SVC is compared with that of H.264/AVC, MPEG-4 Visual [7] and Simulcast. Simulcast is a method for broadcasting video content in which a video is encoded in different settings and the different versions of the video are sent together. Results show that SVC imposes overhead compared to single layer H.264/AVC, and outperforms Simulcast [6], [8]. Furthermore it has been reported that SVC slightly outperforms Google’s VP8 [9] and MPEG-4 Part 2 [10] in terms of rate-distortion but exhibits higher values of rate variability [9], [10]. The rate variability is defined as the covariance of the frame sizes in bytes. From another perspective it has been reported that in terms of time required to extract a substream from a scalable bitstream, SVC bitstream extractor slightly outperforms MPEG-21 DIA, a standard that provides videos in the form of substreams to support interoperable access to them [11].

Impact of different layering configurations in SVC: The effect of temporal, spatial and quality layering on the performance of SVC has been investigated in [12]–[16]. In [12] it has been reported that increasing the number of temporal layers can increase the objective quality of the encoded video in a constant bitrate, while using spatial or quality layers has a negative impact due to the bitrate overhead. Li et al. reported that according to the visual experience of users, SVC inter-layer prediction is more efficient for fast and complex sequences than for slow and simple scenarios [13]. Comparison studies on the performance of different quality scalability modes of SVC (namely CGS and MGS) over five long CIF videos revealed that MGS provides higher objective quality at the cost of higher rate variability [14], [15]. Recently, Slanina et al. [16] studied the impact of the number of temporal and quality layers on the rate distortion performance of SVC, using two full HD video sequences encoded with constant frame rate.

Subjective video quality offered by SVC: In contrast to objective video quality, subjective video quality is the visual quality of the video as perceived by the human viewers. Several research works have investigated the subjective video quality of SVC [17]–[21]. For instance, Politis et al. measured the effect of user mobility and handover on the objective and subjective video quality of SVC and H.264/AVC codecs using two CIF video sequences, and reported that SVC outperforms H.264/AVC in both objective and subjective video qualities [17]. Contrarily, a comparison based on four full HD video sequences concludes that in three out of four video sequences AVC slightly outperforms SVC [18]. However, according to [18] with a bitrate overhead of 10% for SVC, the visual quality should be indistinguishable from that of single layer AVC.

We note that these existing studies have one or several of the following problems: (1) the number of test video sequences is not sufficiently large to represent different types of videos, therefore the conclusion may be biased towards the particular video sequences used, (2) the criteria used to select the test video sequences is not clear, (3) the video resolutions are too small to be relevant in todays applications, (4) the performance of SVC is limited to only few scalability modes. In this paper, we seek to conduct a systematic analysis on the performance of SVC in full HD video streaming. We carefully select a set of video sequences from 29 full HD video sequences. The video sequences represent a variety of content properties, which also allows us to identify any linkage between performance and video content. We also examine the complete range of video resolution, from 288p to 2160p. Furthermore, we conduct the analysis on different aspects of SVC, including the decoding complexity and considering the effect of frame size and all the scalability modes provided by the SVC standard.

Sec. VI concludes the paper, followed by the performance analysis in Sec. V. Finally, Sec. VI concludes the paper.
III. H.264/SVC OVERVIEW

In this section, we provide an overview of SVC, an annex of the H.264/AVC standard, that offers a layered coding approach and provides a framework for scalable video coding. A SVC compliant video stream is scalable in the sense that a valid video stream can be reconstructed at a lower quality level, even in the absence of certain parts of the bitstream. However, the quality is expected to be high considering the smaller bitrate of the substream. This special properly of SVC allows multimedia streaming systems to support diverse devices using just one video stream. In a nutshell, the streaming server encodes the video only once in SVC format, and the devices (e.g., smartphones, tablets, laptops, desktops, TVs, etc.) on the user end may decode the video to the best quality supported by their hardware/software and network connectivities.

SVC supports three modes of scalability, i.e. temporal, spatial and quality (SNR) scalability. Every SVC compliant bitstream contains a H.264/AVC compliant base layer, that contains the lowest temporal, spatial, and quality representation of the video, and several enhancement layers, that provide the scalability in different modes. A block diagram of a SVC encoder for a scalable video stream with two spatial layers is presented in Fig. 1. The base layer is the essential layer needed to playback a video at the lowest possible quality. The quality improves as more enhancement layers become available. The number of enhancement layers available depends on the hardware/software specifications and the network connectivity of the end-user devices. Such a layered design also allows the end-user device to dynamically adjust the playback quality according to the availability of computational and communication resources.

![Fig. 1. Block diagram of a SVC encoder for two spatial layers](image1)

As a part of H.264 standard family, each video sequence in SVC starts with an Instantaneous Decoding Refresh (IDR) access unit, the union of one I frame (e.g., frame 0 of S0 in Fig. 2) with some critical data such as the set of coding parameters, followed by the hierarchical temporal prediction structure defined by the size of group of pictures \(\text{(GOP size)}\) and the distance between two intra-coded pictures \(\text{(Intra Period)}\). GOP size specifies the distance between two key pictures, i.e., I- or P-frames. In the example from Fig. 2 the GOP size is 8.

A. Spatial Scalability

Spatial scalability in SVC is provided by a layered approach. As illustrated in Fig. 2 the base spatial layer S0 encodes lower resolution of frames from only the first three temporal layers \((T0, T1, \text{and } T2)\). The enhancement layer S1 has enhancement frames for the same temporal layers, if not more, as in the preceding layer. SVC supports both dyadic and non-dyadic spatial layering. The dyadic configuration enforces the spatial layers to conform to a 2:1 resolution scale, i.e., lower resolution layers can be scaled up efficiently using bitwise shift operations. Furthermore, with Extended Spatial Scalability (ESS), a class of more complex algorithms for non-dyadic spatial scalability, SVC allows the neighbouring spatial dependency layers to have any arbitrary resolutions. However, the frame resolution of layers with lower spatial dependency identifiers (e.g., S0) cannot be larger than that of the posterior layers (e.g., S1) in height or width.

As shown in Fig. 1 in SVC each spatial dependency layer requires its own prediction module to perform motion-compensated prediction and intra-prediction within the layer. Furthermore, for each dependency layer \(D\) (e.g., layer S1), a reference layer \(D_R < D\) (e.g., layer S0) can be used for inter-layer prediction, where motion vectors, intra texture and residual signals of the reference layer can be used to predict the same data for the predicted layer.

B. Temporal Scalability

To support temporal scalability, SVC relies on a hierarchical temporal prediction mechanism that is extended from H.264/AVC. While previous scalable standards such as H.263 and MPEG-4 Visual basically provide dyadic temporal scalability by segmenting video layers according to different frame types (i.e., I, P and B frame types), in SVC the basis of temporal scalability is found on a hierarchical temporal prediction structure. In the example shown in Fig. 2 there are four temporal layers \((T0, T1, T2, \text{and } T3)\), where \(T0\) is the temporal base layer. Within a spatial layer, frames in layer \(T0\) are predicted only from frames in layer \(T0\). Frames in layer \(T1\) are predicted from layers \(T0\) and \(T1\), where as frames in layer \(T3\), contained in spatial layer \(S1\), are predicted by adjacent frames from any preceding layers.

The hierarchical temporal prediction structure can be characterized by GOP size and Intra Period parameters. Assume that the initial frame rate of a video sequences is 24 fps. With GOP size of 8, there are 3 fps frames for layer \(T0\), 6 fps frames for layer \(T1\), 12 fps frames for layer \(T2\), and 24 fps frames for layer \(T3\). The Intra Period specifies how often a P-frame at the end of a GOP can be replaced by an I-frame. It must be multiples of the GOP size. For example, for Intra Period of 2 and GOP size of 8, the P-frames with frame numbers of products of 16 will be replaced by an I-frame.
H.264/SVC allows the encoder to use complementary data in different layers to generate video streams that provide distinct quality levels for the reconstructed video. As illustrated in Fig. 1, each layer has a SNR refinement module that provides the necessary mechanisms for quality scalability. Three quality scalability modes are supported, namely CGS, MGS and FGS.

CGS (Coarse Grain Scalability): All consecutive layers should have the same resolution. In fact, CGS can be considered as a special case of spatial scaling where upsampling and inter-layer deblocking of the intra-coded macroblocks (the processing unit of frame) of the reference layer are not required, as the predicted macroblocks and the reference ones are the same size. In CGS, the enhancement layer typically contains the residual texture signal that is quantized with a smaller quantization step size relative to that used for the preceding CGS layers [3], hence providing incremental visual information.

MGS (Medium Grain Scalability): Both the base and enhancement layers can be used to predict a layer at the same time, hence improving the coding efficiency when a variety of bitrates are required in a scalable bitstream. To resolve the synchronism offset between encoder and decoder when only the base layer is received, MGS uses periodic key pictures to immediately resynchronize the prediction module. Furthermore, MGS allows switching between different MGS layers at any access unit, hence increasing the flexibility of bitstream adaptation [3].

FGS (Fine Grain Scalability): This scalability mode provides a continuous adaptation of bitstream bitrate by using an advanced bit-plane technique. In this technique different layers contain distinct subsets of bits corresponding to each macroblock. By supporting progressive refinement of transform coefficients, FGS allows the decoder to truncate the bitstream at arbitrary points [3].

Besides these three scalability modes, quantization parameter (QP) is another factor that affects the quality of the encoded layers. The value of QP ranges from 0 to 51. At the beginning of the encoding process where DCT transformation of macroblocks is performed, the result is quantized by dividing the matrices of luma and colour components to two specific matrices of integers. QP directly affects this process as the denominator matrices are multiplied by QP before the division. Therefore, higher value of QP eliminates more coefficients, hence increases coarseness and decreases the bitrate and quality of the encoded video. To optimize the rate-distortion ratio, SVC adjusts the QP of each frame according to its location in the respective group of pictures.

As shown in Fig. 1 all the enhancement layers and the H.264/AVC compliant base layer are merged by the so called multiplex. Thus, different temporal, spatial and quality layers are integrated into a single scalable bitstream. We use a triple (D,T,Q) to specify the number of spatial, temporal and quality layers of a SVC compliant bitstream. Interested readers may refer to [3] for a more detailed discussion.

To conduct a systematic study on SVC for full HD video streaming, a careful design of the experiments is necessary. In this section, we describe the experiment setup and performance metrics.

A. Experiments Testbed

We conduct all the experiments on a server cluster of 10 nodes. Each server node is equipped with four Intel® Xeon® E5640 CPUs with four cores at 2.67GHz and 16GB of 1066MHz memory. To avoid the reciprocal effect of multi-core operation on core performance and CPU cache hit, we utilize only one core on each CPU and at most three CPUs on each machine. We use Joint Scalable Video Model (JSVM-9.19.15) [22], an open-source reference software for SVC, as the SVC encoder. The software is compiled on RedHat® Linux with kernel v2.6.18 and gcc v4.1.2 on each server. In addition, we use EPFL Video Quality Measurement Tool v1.1 [23] to calculate objective video quality metrics.

B. Selecting Source Video Sequences

We have access to 29 raw video sequences with frame size of 1920 × 1080 pixels and frame rate of 24 frames per second (i.e. 1080p24) [24]. This is the minimum frame size and frame rate for full HD in ATSC standards [25]. The raw video sequences are in YUV 4:2:0 format, the standard sampling scheme for H.26x coding families.

We can use all 29 videos to conduct the experiments. However, this approach is not only time consuming, but also requires excessive post processing. To select a proper set of reference video sequences that represent a variety of content types that a video streaming server might encode, we analyze the visual features, including colour, texture and motion using the MPEG-7 Visual Description Tools [26]. More specifically, the colour feature is characterized by the number of dominant colours (with maximum of 8) and the overall spatial coherency of the colour clusters. The texture feature can be captured by the MPEG-7 edge histogram descriptor extracted for each frame. The motion features are best represented by the motion vectors and Motion Activity. The motion vectors provide the gross motion characteristics of a video segment. To extract the motion vectors, each video sequence has been encoded using JSVM-9.19.15 in single layer mode. Then the JSVM-9.19.15 decoder tool has been modified to report the motion vectors for each inter-frame or intra-frame motion compensated macroblock and sub-macroblock prediction. Motion Activity considers the intensity, direction, spatial distribution and temporal distribution of activity in a video sequence. To calculate the motion activity, according to [27], the standard deviation of the magnitudes of all motion vectors of each frame has been quantized between 1 and 5, and the average of the quantized motion activity values over all the frames has been used as the Motion Activity of each video sequence.

1 In this paper, we refer to lossless and uncompressed videos as raw, lossy compressed videos as encoded, and raw video files obtained from decompressing an encoded video as decoded.
According to these visual features, we select seven video sequences that exhibit diverse values for these features among the 29 videos. The seven videos are grouped into three categories: animation, scene, and nature. The selected video sequences and their properties are reported in Table I. The list below further describes each of the seven video sequences used in this study.

- Big Buck Bunny (BB): An animation clip that shows a big rabbit waking up in the morning. The animation has low and high motion activities, and features detailed shading and hair and fur demonstration.
- Elephants Dream (ED): An animation clip that displays a surreal scene wherein two characters are talking, and features a foggy environment.
- Pedestrian Area (PA): The camera is fixed towards a pedestrian area. Pedestrians show diverse contrasts and colours and complex motions.
- Rush Hour (RH): This video is shot at a street in rush hour. The camera is fixed towards the cars passing by.
- Park Joy (PJ): This video is shot from the other side of a river. The camera pans from left to right and follows a group of people running in front of trees.
- Riverbed (RB): The camera is fixed towards a riverbed and records frequent small waves on the edge of the river. Due to the high frequency of the small waves and the reflection of light on the surface of water this scene exhibits high values of motion activity and details.
- Sunflower (SF): The camera pans horizontally and follows a bee on a sunflower.

To ensure that the measurements are feasible on the server cluster, the video sequences were cropped to 240 frames, i.e., 10 seconds of raw video with frame rate of 24 fps. If a video sequence has movie titles at the beginning, the first 1000 frames have been skipped. Otherwise, the frames have been selected from the beginning.

C. Performance Metrics

When evaluating the performance of SVC codec for full HD video streaming, we use different performance metrics to capture coding efficiency, encoding and decoding complexities, and objective quality of the encoded video. For all the experiments, a reference encoding configuration is used, unless otherwise specified. The reference encoding configuration uses the H.264/AVC and SVC codecs in full HD. The GOP is selected to match the configuration being examined. The base quantization parameter is 32; the size of motion prediction buffer, in which the reconstructed frames are kept for motion estimation and decoding, is 16 frames; and fast motion search algorithm has been used, which reduces the encoding time without considerably decreasing the objective quality of the encoded video. For comparison purpose, we define bitrate overhead as the ratio of the bitrate from the configuration being measured and the bitrate from the reference coding configuration. We define the performance metrics in the rest of this section.

Coding efficiency: The bitrate of the encoded video stream is the main metric for measuring the coding efficiency of a codec. Along with the bitrate, MPEG-7 motion activity of the encoded video is used wherever it helps the discussion. Furthermore, to evaluate the performance of SVC codec over extended layering scenarios, the encoder in JSVM-9.19.15 has been modified to increase the number of permissible layer configuration files from 8 to 32.

Encoding and decoding complexity: We measure the CPU time required to encode and decode each video sequence. Because JSVM-9.19.15 implements only the standard coding operations, it is not optimized for the best CPU performance. Considering that SVC is a potential replacement for the current use of H.264/AVC, we compare the performance of SVC with the single layer H.264/AVC and Simulcast, all performed using JSVM-9.19.15 platform. The performance difference in term of encoding and decoding time reflects the difference between SVC and other coding standards, although the JSVM implementation is not optimized.

Objective quality: We focus on the objective quality instead of subjective quality. For any subjective measure to be representative enough and valid, we need a large pool of participants and a large collection of videos. Such a user-based study is orthogonal to our work in this paper. To quantify the objective quality, we calculate the Y-PSNR (the PSNR value of the luma component of the video sequence), SSIM (structural similarity index) [25], MS-SSIM (multi-scale structural similarity index) [29] and the pixel domain version of VIF (visual information fidelity) [30] for each encoded video sequence. However, due to the page limit, only Y-PSNR and VIF metrics are included in Sec. V. Compared to other metrics, VIF is known for its high correlation with the results of an extensive subjective video quality assessment [20].

Throughout the performance analysis, we will use a combination of the aforementioned performance metrics and overhead of the performance metrics. We define the overhead of a performance metric as the ratio of the performance of the configuration being measured and the performance of the reference coding configuration.

V. PERFORMANCE ANALYSIS

In this section, we present the results of our performance study. We begin with an analysis of the effect of frame size, followed by studies on each of the three scalability modes (spatial, temporal, and quality). By default, we configure the SVC encoder to generate a scalable video stream with two dyadic spatial layers (1920 × 1080 pixels and 960 × 540 pixels), five temporal layers (GOP = 16) and two quality layers (using MGS scalability mode). The minimum quantization parameters (QP) is set to 28 with default delta QP of −2 for quality layers to ensure enough spatial detail is preserved. Automatic QP cascading is used for temporal layers. For comparison purpose, we also configured the JSVM to encode H.264/AVC with QP of 28 for both spatial layers. In the analysis, H.264/AVC refers to the single layer encoded version of the video sequence in full HD, and Simulcast refers to the single layer encoded version of the video sequence in both dyadic resolutions. We note that the aforementioned SVC
encoding configuration supports 18 different bitrate points spanning from 119.6kbps to 1.3Mbps, whereas the H.264/AVC and Simulcast support only one and two bitrates, respectively.

A. The Effect of Frame Size

Before examining the scaling factors of SVC, we wish to first understand the effect of frame size (resolution), which is the most observable quality feature by the end users. In this experiment, we downsample the video sequences from full HD resolution (1920 × 1080) to smaller frame sizes using non-normative downsampling [31] with frame heights of 288, 360, 480, 576, 720, and 900 pixels while preserving the 16:9 ratio for the frame width. We observe the similar performance trend among all seven video sequences. Due to page limit, we present the results from video sequence BB in Fig. 3.

In general, as the frame size increases, so do the bitrate, coding complexity, and the objective quality for all three coding standards. However, there are subtle differences among the three standards. According to the bitrates reported in Fig. 3(a) the bitrate overhead, the ratio between the SVC bitrate and the AVC bitrate, decreases from 80.3% to 17.8% when increasing the video frame size from 512 × 288 to full HD. The bitrate overhead in SVC is the tradeoff for supporting 18 different bitrate points spanning from 119.6kbps to 1.3Mbps, while H.264/AVC supports only one bitrate. If less bitrate points are required, this overhead can be decreased by lowering the number of spatial or quality layers. Compared to the two-resolution Simulcast, SVC’s bitrate is much less, especially for frame sizes larger than 576p. This implies that compared with Simulcast, SVC is much more effective in terms of bandwidth utilization.

In terms of coding complexity, as shown in Fig. 3(b) SVC and AVC have similar encoding time for frame sizes less than 720p. Towards large frames, it takes SVC less time to encode than AVC does. When increasing the frame size from 512 × 288 to full HD, the number of 16 × 16 macroblocks in each frame increases from 576 to 8100. Thereby, the probability of finding similar macroblocks for motion compensation increases, which consequently reduces the motion compensated residual error generated for each motion vector, hence resulting in less number of bits required for each motion compensated macroblock. This observation is backed up by our measurement for the video sequence BB, which shows that when increasing the frame size from 512 × 288 to full HD, the average number of bits required to represent each motion compensated macroblock by H.264/AVC and SVC is reduced by 34.8% and 48.7%, respectively. The decoding complexity is different from encoding complexity. Fig. 3(c) shows that the SVC has higher decoding complexity compared to AVC, which is expected due to the growing number of motion vectors needed for layered prediction of motion compensated macroblocks during the decoding process.

Increasing the frame size also improves the Y-PSNR, as the more similar macroblocks allow the rate distortion optimization module of the encoder to select higher quality points when encoding the video sequence. Fig. 3(d) depicts that both SVC and H.264/AVC codecs experience a similar growth in objective video quality measured in Y-PSNR. SVC consistently has higher Y-PSNR values than AVC does.

B. The Effect of Temporal Scalability

As described in Sec. III, SVC temporal scalability is provided by a hierarchical temporal prediction structure among I-, B- and P-frames. The structure can be characterized by GOP size and Intra Period parameters. We vary these parameters to study the effect of temporal scalability.

First, we increase the GOP size from 2 to 16, which adjusts the number of temporal layers from 2 to 5 accordingly. Fig. 3(a) shows that increasing the GOP size from 4 to 8 decreases the bitrate of the encoded video sequences by an average of 3.6%, but increasing the GOP size from 8 to 16 increases the bitrate by an average of 3.9%. This is rather counterintuitive. The bitrate is expected to drop since growing GOP size requires replacing P-frames with B-frames. However, we note that this replacement may increase the residual error of macroblocks that use the P-frame as a reference frame. Thus, more bits are needed to represent the residual error, which yields higher bitrate overall. We also observe that the bitrates of video sequences PJ and RB are noticeably higher than the other video sequences. This is because both videos have high values of Detail and Motion Activity in Table I.

In terms of coding complexity, all video sequences exhibit an increasing trend when the GOP size grows from 2 to 16, as shown in Fig. 3(b) and Fig. 3(c). This observation on full HD videos contradicts with the observation on QCIF and CIF video sequences as reported in [12]. The increase is most due to the extended search domain for motion-compensated predictions when more B-frames are used, as reported in Fig. 3(d). Among the growing coding complexities, there is a slight decrease in decoding complexity for video sequences BB, SF, and ED when the GOP size goes from 2 to 4. According to Table I these three video sequences have the

### Table I

**Selected video sequences and their properties.**

<table>
<thead>
<tr>
<th>Content</th>
<th>Genre</th>
<th>Selected Frames</th>
<th>Avg. Num. of Dominant Colours</th>
<th>Spatial Coherency</th>
<th>Detail</th>
<th>Avg. Num. of Motion Vectors</th>
<th>Motion Activity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Big Buck Bunny (BB)</td>
<td>Animation</td>
<td>1001-1240</td>
<td>7.49</td>
<td>8.85</td>
<td>3.52</td>
<td>10223</td>
<td>1.63</td>
</tr>
<tr>
<td>Elephants Dream (ED)</td>
<td>Animation</td>
<td>1001-1240</td>
<td>3.73</td>
<td>23.97</td>
<td>3.73</td>
<td>10147</td>
<td>2.39</td>
</tr>
<tr>
<td>Pedestrian Area (PA)</td>
<td>Scene</td>
<td>1-240</td>
<td>4.41</td>
<td>22.88</td>
<td>3.15</td>
<td>10116</td>
<td>4.42</td>
</tr>
<tr>
<td>Rush Hour (RH)</td>
<td>Scene</td>
<td>1-240</td>
<td>4.29</td>
<td>25.16</td>
<td>3.17</td>
<td>11041</td>
<td>3.12</td>
</tr>
<tr>
<td>Park Joy (PJ)</td>
<td>Scene/Nature</td>
<td>1-240</td>
<td>6.46</td>
<td>3.49</td>
<td>4.24</td>
<td>17373</td>
<td>3.73</td>
</tr>
<tr>
<td>Riverbed (RB)</td>
<td>Nature</td>
<td>1-240</td>
<td>3.94</td>
<td>24.33</td>
<td>4.72</td>
<td>8612</td>
<td>4.13</td>
</tr>
</tbody>
</table>
lowest Motion Activity, which suggests that replacing P-pictures with B-pictures has little impact on the number of motion compensated predictions.

We also observe quality improvement in terms of Y-PSNR when increasing the GOP size, except for the video sequence RB, which has the highest detail descriptor. As depicted in Fig. 4(e), the increase among video sequences BB, ED, and SF is more noticeable. With larger GOP size, the distance between predicted frames is also larger, which is not a desirable property for the video sequences with high motion activity values. In contrary, for the video sequences with low motion activity values, replacing P-frames with B-frames conserves the bitrate and allows the rate distortion optimizer module to select higher quality points, leading to increased video quality.

Next, we investigate the effect of Intra Period parameter on the performance of SVC codec. In this experiment, we encode all video sequences using three different SVC encoding configurations (0, 2, 2), (1, 3, 1) and (2, 4, 3). The Intra Period parameter is varied from 0 GOP to 4 GOPs, i.e., substituting one motion predicted P-frame with an I-frame for every 0 – 4 GOPs. As Intra Period increases, less I-frames will appear in the video sequence. For Intra Period of 0 GOP, no substitution will take place. Again, due to page limit, only the results from video sequence PA is shown in Fig. 5. As shown in Fig. 5(a), there is a slight increasing trend for bitrate when more I-frames are inserted. Fig. 5(b) and 5(c) show that adding intra-coded frames slightly reduces encoding and decoding complexities, since some motion predicted P-frames are replaced by I-frames that are less computational complex. In terms of objective quality, very little improvement is observed in Fig. 5(d). Although using I-frames improves the quality of the encoded video, it causes the rate distortion optimizer module to select lower quality points due to the increased bitrate.

To compare the effect of Intra Period parameter among all seven video sequences, we present the results when Intra Period parameter is 1 GOP in Table II. We observe that the bitrate overhead is less significant for video sequences with high detail and motion activities, because an additional intra-coded frame can be helpful in providing higher quality reference macroblocks and also resetting the error propagation chain among the predicted macroblocks. This ultimately decreases the residual errors and the number of bits required to represent them. In summary, when using SVC for full HD video streaming, additional intra-coded frames is beneficial to videos with high detail and motion activity values.

### Table II

<table>
<thead>
<tr>
<th>Video Sequence</th>
<th>Bitrate Overhead</th>
<th>Encoding Time Gain</th>
<th>Decoding Time Gain</th>
<th>Y-PSNR Gain</th>
</tr>
</thead>
<tbody>
<tr>
<td>BB</td>
<td>36.0%</td>
<td>1.8%</td>
<td>0.7%</td>
<td>0.7%</td>
</tr>
<tr>
<td>ED</td>
<td>17.3%</td>
<td>1.0%</td>
<td>0.6%</td>
<td>0.4%</td>
</tr>
<tr>
<td>PA</td>
<td>8.6%</td>
<td>1.6%</td>
<td>0.4%</td>
<td>0.2%</td>
</tr>
<tr>
<td>RH</td>
<td>8.1%</td>
<td>4.0%</td>
<td>0.6%</td>
<td>0.1%</td>
</tr>
<tr>
<td>PJ</td>
<td>0.5%</td>
<td>7.8%</td>
<td>0.1%</td>
<td>0.0%</td>
</tr>
<tr>
<td>RB</td>
<td>0.1%</td>
<td>7.4%</td>
<td>2.2%</td>
<td>0.0%</td>
</tr>
<tr>
<td>SF</td>
<td>15.1%</td>
<td>2.6%</td>
<td>0.6%</td>
<td>0.4%</td>
</tr>
</tbody>
</table>

### C. The Effect of Spatial Layering

As described in Sec. III, SVC supports spatial scalability in both dyadic and non-dyadic modes. To investigate the effect of spatial layering on the performance of SVC codec for full HD video streaming, two separate series of experiments were performed, one for the dyadic mode and one for the non-dyadic mode. Since we adjust the GOP in this analysis, we use a smaller GOP size 4 for the reference encoding.

1) **Dyadic Spatial Layering vs. Single Layer Coding:** In this experiment, dyadic spatial layering is applied to the full HD version of each video sequence to create two layered video sequences with two and three dyadic spatial layers, called DY1 and DY2, respectively. For comparison purposes the same
experiment is repeated with single layer H.264/AVC, where the encoder encodes the video in full HD and two dyadic spatial resolutions in parallel. We use SIMC1 to refer to the combination of full HD and one dyadic spatial resolution, and SIMC2 to refer to the combination of full HD and two dyadic spatial resolutions.

As shown in Fig. 5(a), sending two and three different resolutions of the videos in parallel imposes an average bitrate overhead of 28.6% (for SIMC1) and 40.3% (for SIMC2). In contrast, SVC spatial scalability significantly decreases the bitrate overhead to 8.8% (for DY1) and 14.1% (for DY2). The performance gain is due to the intra-texture, motion and residual signals of the lower resolution layers used to predict the higher resolution layers. Interestingly, the video sequence RB in SVC format decreases the required bandwidth by 4.4% and 2.0% compared to the single-layer reference AVC encoding. This is related to the high detail and motion activity value of the video as well as the low number of motion vectors from Table I. These properties indicate that the temporal prediction does not provide enough motion compensated predictions. Compared to the single-layer coding, two additional spatial layers in SVC increases the average number of motion vectors among all test video sequences by 16.4%, and RB experiences a 29.7% increase, which approves the role of spatially predicted macroblocks in the bitrate decrease that is observed for RB. For coding complexity, the use of one dyadic and two dyadic spatial layers in SVC increases the encoding time by an average of 50.6% and 68.4%, respectively. The same trend is observed for all videos, thus, the detailed results are skipped due to lack of space.

![Fig. 5. The effect of Intra Period parameter on (a) coding efficiency, (b) encoding complexity, (c) decoding complexity and (d) objective quality of SVC codec.](image)

![Fig. 6. The effect of SVC spatial layering on (a) the streaming server side and (b) the receiver side.](image)

Fig. 6(b) compares the bitrate required on the client side to receive the video in either AVC or SVC format with the specified dyadic resolutions (480p, 960p, and 1920p). The average bitrate required to receive the base layer of SVC bitstream (SVC-480), which is also AVC compatible, is 70.4% of that of single-layer AVC (AVC-480). This inevitably leads to a lower objective video quality. Our measurements show that using SVC with the specified settings decreases the objective video quality of 480 × 270, 960 × 540 and full HD reconstructed videos by an average of 0.98, 1.17 and 1.05 dBs, respectively. Furthermore, the average bandwidth required to receive the videos in 960 × 540 and full HD resolutions in SVC mode are 10.1% and 15.0% more than that of single-layer AVC mode, respectively. For coding complexity, there is no significant different between SVC-480 and AVC-480, as they are both AVC compatible. The added spatial layers in SVC requires 27.1% (for SVC-960) and 84.6% (for SVC-1920, full HD) more decoding time. Hence, full-HD SVC is not recommended on battery-operated devices or devices with limited CPU power. However, the decoding time of the SVC bitstream can be dramatically decreased in expense of minor limitations in spatial layering capabilities [32].

2) Dyadic vs. Non-Dyadic Spatial Layering: To compare dyadic and non-dyadic spatial layering, we modify the resolution and the frame ratio of the spatial layers so that the number of macroblocks in each layer remains unchanged and the layer resolutions are non-dyadic. This version of coding is referred to as NDY1 and NDY2 for one and two non-dyadic spatial layers, respectively. We repeat the same spatial layering experiment as in Sec. V-C1 with the new non-dyadic layers. Furthermore, to investigate the effect of frame size, the same experiments are repeated with 480 × 270 pixels frames in the highest resolution layer. The average overheads from all video sequences are reported in Table III. We observe that increasing the frame size to full HD in non-dyadic spatial layering significantly reduces the average overhead, and all overheads are less than 7.5%.

<table>
<thead>
<tr>
<th>Video Resolution</th>
<th>Bitrate Overhead</th>
<th>Encoding Overhead</th>
<th>Decoding Overhead</th>
</tr>
</thead>
<tbody>
<tr>
<td>480 × 270</td>
<td>11.2%</td>
<td>16.6%</td>
<td>13.7%</td>
</tr>
<tr>
<td>960 × 540</td>
<td>18.1%</td>
<td>18.1%</td>
<td>6.2%</td>
</tr>
<tr>
<td>Full HD</td>
<td>22.2%</td>
<td>2.2%</td>
<td>4.2%</td>
</tr>
</tbody>
</table>

D. The Effect of Quality Layering

At last, we study the quality scalability in SVC. CGS does not provide the required flexibility for most real world situations, and JSVM-9.19.15 does not allow the configuration of relevant parameters of FGS separately. For these reasons, we study only the MGS mode in this section. Besides these
three quality scalability modes, quantization parameter (QP) is another factor that directly affects the quality of the encoded layers and the overall bitstream. To investigate the effect of quality layers and QP in full HD video streaming with SVC, two separate series of experiments were performed, one for quality layers and one for QP.

1) The Effect of Quality Layers: In JSVM-9.19.15 the number of quality layers and their properties can be specified using the MGSVectorMode parameter and the MGSVector vector defining up to 16 layers. Each element \(i \in \text{MGSVector} \) specifies the quality level of the \( i \)th SNR layer, and the sum of the elements in MGSVector must equal to 16. In this experiment, we vary the number of quality layers from zero to four using a GOP size 4. Table IV presents the MGS configuration for all layer configurations.

<table>
<thead>
<tr>
<th>SNR Layers</th>
<th>MGSV0</th>
<th>MGSV1</th>
<th>MGSV2</th>
<th>MGSV3</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>2</td>
<td>8</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
</tbody>
</table>

As shown in Fig. 7(a), by adding the first quality layer, the SNR refinement module is loaded into the prediction module, resulting in 42.7% – 79.0% increase in bitrate. Moreover, adding the first quality layer introduces more bitrate overhead for less complex video sequences (e.g., BB and ED). Except for the first quality layer, any additional quality layer has almost negligible impact on the bitrate, since the same quality information is divided and placed in different layers. For the same reason, the Y-PSNR is increased by 1.6% on average when adding the first quality layer, and additional quality layers do not improve the quality, as shown in Fig. 7(d). Recall that SVC is used to provide adaptive streaming to allow end-user devices to receive a subset of these quality layers and still be able to receive the video.

According to Fig. 7(b) and Fig. 7(c), additional quality layers introduce little complexity in the encoding process, but do require more decoding time on the receiver side. On the server side, complex video sequences (e.g., RB and PJ) exhibit a decreasing quality trend as more quality layers are added, while less complex video sequence (e.g., BB) require more encoding time. On the receiver side, each quality enhancement layer adds almost 23% more decoding time, i.e., adding four quality layers increases the decoding time by 92.5%. This can be due to the internal structure of JSVM-9.19.15 decoder module, where enhancement layers are decoded and applied to the decoded picture buffer consecutively.

2) The Effect of Quantization Parameter: To investigate the effect of QP, we resort to the three different DTQ configurations used in Sec. V-B (0, 2, 2), (1, 3, 1) and (2, 4, 3). For each configuration the highest value of QP, the QP for the base layer \( (QP_b) \), is varied from 32 to 42, where delta QP is −2 for quality layers. Automatic QP cascading was employed for temporal and spatial layers. The results of experiments for video sequence PA are reported in Fig. 8. All other video sequences share the same performance trend.

Fig. 8 shows that the impact of \( QP_b \) is very little for videos with DTQ configuration (0, 2, 2), where no spatial layering is used. The impact of \( QP_b \) is strongly augmented when for DTQ=(1, 3, 1) or DTQ=(2, 4, 3). According to Fig. 1 when spatial layering is used, the inter-layer prediction is utilized to use intra, motion and residual signal information of the lower spatial layers to predict the macroblocks in the upper layers. The reduced values of the residual signals of the upper layers makes QP a more determining factor in the performance analysis. Therefore, when spatial layering is utilized, increasing the value of QP significantly decreases the bitrate, slightly decreases the encoding time, and negligibly decreases the decoding time. The objective quality of the video stream also considerably decreases. However, we noted that even when Y-PSNR is close to 36 dB in Fig. 8(d), the visual quality of the reconstructed video is very good from a human viewer perspective.

VI. CONCLUSION

In this paper, we conduct a systematic study on the use of SVC for full HD video streaming. We learn that there is a significant coding overhead introduced by SVC, in comparison with single-layer H.264/AVC, but is efficient in serving streaming sessions in multiple quality levels. We identify that SVC is more advantageous in full HD streaming as the efficiency and computational gap between SVC and H.264/AVC is much less when encoding high quality videos. When using SVC for full HD video streaming, additional intra-coded frames are beneficial to videos with high detail and motion activity values. However, full-HD SVC is not recommended on battery-operated devices or devices with limited CPU power due to the increased decoding complexity, unless special optimization is implemented on the receiver side. Using a set of carefully selected and diverse video contents, we also discovered that certain SVC configurations have advantages over H.264/AVC for complex video sequences with high detail and motion activity values. For example, replacing P-frames with I-frames in such video sequences can decrease the encoding complexity without increasing the bitrate of the encoded video, and additional spatial layers may decrease the bitrate of the encoded video.

REFERENCES

Fig. 7. The effect of quality layering on (a) coding efficiency, (b) encoding complexity, (c) decoding complexity and (d) objective quality of SVC.

Fig. 8. The effect of varying quantization parameter on (a) coding efficiency, (b) encoding complexity, (c) decoding complexity and (d) objective quality of SVC. The horizontal axis is the value of the highest quantization parameter used in the layered structure.