Overview of the Scalable Video Coding Extension of the H.264/AVC Standard

This paper was published in the IEEE Transaction on Circuits and Systems for Video Technology journal, a well respected publication in the field. Published less than three years ago, it was an invited paper and offers a performance enhancing extension to the widely used and newest international coding standard H.264.

The abstract claims the authors have designed a standard that extends H.264 and results in an improvement of coding efficiency and increased degree of scalability. It also says they will provide the details of the basic tools required to implement their idea. Finally, they will provide experimental data, and analyze it for efficiency and complexity.

The introduction begins by talking about the spectrum of todays video transmission and storage systems. Next it brings up the fact that now a days our video receivers range from low powered small screen cell phones to large high definition computer monitors. In addition, our connections are quite variable as well. They define "scalability" to be the "removal of parts of the video bit stream in order to adapt it to the various needs or preferences of end users as well as to varying terminal capabilities or network conditions." They offer Scalable Video Coding (SVC) as a solution to the issue presented. The basic concept behind SVC is to enable high quality video encoding by transmitting a bit stream which contains multiple subset bit streams thats can be decoded on their own as well, all while having a complexity and reconstruction quality on par with that of H.264. Next they give a brief history of scalability solutions in the past, stating that they existed but are rarely used due to significant loss in coding efficiency, and far more complex decoding.

The next section describes the fundamental scalability types, and talks about applications that represent SVC and their requirements. The three normal modes of scalability are temporal (frame rate), spatial (spatial resolution), and quality (fidelity, or signal-to-noise ratio). One of the main benefits of using SVC in applications is that you only need to encode the video once, and it results in a scalable bitstream in which clients can choose varying levels of the aforementioned scalable aspects. Another benefit is error resilience. It helps this because the bit stream contains parts with varying levels of importance in terms of final video quality. By simply using stronger protection of the more important portions of the bit stream, error resilience is noticeably improved. SVC is very useful for surveillance applications, as the video is often viewed on high definition monitors as well as videophones or PDA's, and needs to be stored as well. To further benefit an archive of stored videos, with SVC the application can delete the higher quality parts of the bitstream of older less important videos, while keeping the lower quality parts and still saving a viewable copy of the video. The authors than restate that past scalable methods have existed but were rarely been used, for the same reasons mentioned earlier. In order for SVC to succeed, they believe it must have the following: similar coding efficiency to that of single layer coding, minimal increase in decoding complexity to that of single layer decoding that supports scaling, supports temporal, spatial and quality scalability, supports backward compatible base layer of H.264, and finally supports simple bit stream adaptations after encoding.

The next section covers a short history of SVC.

The following section gives a brief description of H.264, because SVC extends this standard it is important to at least understand the basics of it. The design of H.264 contains two layers, the Video Coding Layer (VCL) and the Network Abstraction Layer (NAL). The VCL makes a coded representation of the original content, while the NAL formats the data and creates header information that helps various systems use the VCL data effectively. The section continues to give a more in depth description of the two layers, which is useful information but not absolutely detrimental to understanding the basic concepts behind SVC.

The next section, which is the bulk of the paper, talks about the basic concepts of extending H.264 to incorporate an SVC standard. This section has three subsections, one for each of the scalable modes starting with temporal. A temporaly scalable bit stream has a set of access units that are partitioned into a base layer with multiple sublayers. Each layer is represented by an identifier T, starting with 0 for the base layer and increasing by one for the following layers. To access layer k, you must remove all access units greater than k. The diagrams of this concept that the authors provide make visualizing it very easy. The focus of the remaining portion of this subsection is Hierarchical Prediction Structures. These are temporal scalability with dyadic enhancement layers, and are typically coded as B-pictures. This reference picture lists 0 and 1 and are restricted to the previous and following picture. This hierarchical prediction structure can be combined with the multiple reference picture concept from H.264. It's efficiency is dependent upon how the quantization parameters are decided for pictures of different temporal layers. The authors then provide a formula for obtaining the quantization parameters with relatively little complexity. Next they analyze the coding efficiency of hierarchical prediction structures. Their results show that providing temporal scalability usually does not have a negative impact on coding efficiency.

The next subsection is about spatial scalability, which in SVC follows the conventional method of multilayer coding. Each layer represents a supported spatial resolution, and is called a spatial layer or dependency identifier D. The base layer's dependency identifier D is 0, and it increases by one between each layer. In addition, SVC implements an inter-layer prediction mechanism. The goal of inter-layer prediction is to enable the use of as much lower layer information as possible to improve rate-distortion efficiency of the enhancement layers. Three inter-layer prediction concepts have been added in SVC, Inter-Layer Motion Prediction, Inter-Layer Residual Prediction, and Inter-Layer Intra-Prediction. The paper than discusses the details of these prediction methods, but limits the description to dyadic spatial scalability, in which the picture height and width is doubled between layers. SVC also supports spacial scalable coding with arbitrary resolution ratio, known as Generalized Spatial Scalability. The single restriction is that the resolution cannot decrease from one layer to the next. It also supports cropping between layers, and even on a picture to picture basis. SVC also supports interlaced sources. As for complexity, the authors claim that under their mandatory restrictions of SVC the overhead in decoder complexity is smaller than prior video coding standards when compared to single-layer coding.

The final subsection is quality scalability, also called coarse-grain quality scalable coding (CGS). This uses the same inter-layer prediction mechanisms as spatial scalability, but with a few modifications. It than gets fairly in-depth about the various aspects of quality scalability, and supplies some useful graphs and diagrams to assist in understanding.

The final section before the conclusion is the high level design of SVC. An SVC bit stream doesn't need to supply all types of scalability, since supporting more scalability typically results in a loss in coding efficiency. This tradeoff can be adjusted by the application according to it's needs. As for the interface, to make bit stream manipulation easier the one bye header of H.264 is extended to an extra three bytes for the SVC unit types (D, Q, and T for dependency identifier, quality identifier, and temporal identifier, accordingly). Another identifier is P, which represents the importance of a NAL unit. An interesting concept is switching between different dependency layers, which is specified to be possible only at well defined points. The authors claim a decoder could be implemented though, that could down-switch at any access unit.

The conclusion simply sums up the benefits of SVC standards in comparison to prior video coding standards relative to single-layer coding.

The organization of this paper is decent. While it touched upon past methods of video scaling, it lacked an explicit section dedicated to past work. For people outside the field of video coding, much of the detailed concepts are difficult to follow. They could have done a better job explaining things in a simpler manor, although the diagrams proved to be quite useful.

The concept of multiple scalable aspects of a video with one single bitstream is a very interesting topic, and beneficial for many situations. The main question is how much of a negative impact implementing such an idea would have on the complexity and efficiency of both encoding and decoding. The authors did an alright job of convincing that it has a small but tolerable increase in these two aspects.

They did not provide a great wealth of evidence to backup their claims in the form of simulation or experiments. While there was some, it was not very conclusive and was not carried out in the most ideal setup. Their concepts seemed very theoretical, lacking a solid implementation to present as a proof of concept.

On the other hand, MPEG and VCEG has accepted SVC as an extension to the H.264 standard, and it has been implemented by multiple companies on various products. In addition, the first company to implement it (Vidyo) licenses a multi-platform H.264/SVC SDK. So clearly the idea of SVC proved to in fact be useful in the real world, and is currently being used today.