# Paper Analysis

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

MPEG-4 is a collection of methods defining compression of audio visual data and was designated a standard for a group of audio and video coding formats (1998-99). This collection is further divided into parts (as agreed upon by the Moving Pictures Expert Group) and the 10$^{th}$ part is referred to as H.264/AVC standard which was completed in 2003. This paper, published in 2007, discusses a Scalable Video Coding extension of the H.264/AVC standard for video encoding. The reasons for choosing H.264/AVC standard was its widespread and further increasing role in the current network systems and devices. Despite its hard readability, the paper is organized well with the authors first giving a motivation for the development of the SVC extension, then describing what scalability is before giving an overview of the H.264/AVC standard. Then the authors describe the concepts for the extension and then finally, its design.

In the introduction section of the paper, the authors give us a very brief 'requirements' of the modern video transmission and storage systems. There are various types of receivers (of audio-visual streams) in existence today and each has its own requirements. For e.g., mobile systems have much less processor power than a PC and thus we need a video encoding scheme which could adapt itself to the various entities' (end points as well as the network itself) requirements. SVC is a good choice as it is designed to provide 'scalable' video encoding where parts of the video stream can be removed so that it can adapt to all kinds of network and system conditions. The encoding of the high quality video bit stream is done in such a way that it contains several subset bit streams that can be decoded with a similar complexity and reconstruction quality. The authors also explain as to why the scalable profiles of earlier standards were not used. The main reasons were the considerable high loss in coding efficiency and the need for much more complex decoders. The authors mention another reason: 'the characteristics of the traditional video transmission system'. However, they do not explain it and it is unclear how this exactly affected the implementation of the encoding schemes. It is a good thing how they mention other alternatives to scalable coding: simulcast, transcoding of a single stream. They also provide a comparison of the SVC's performance and *simulcast* performance later in the paper.

In the next section, the authors systematically describe 'Scalability' and its requirements in detail. A video bit stream is said to be scalable when parts of the stream can be removed in such a way that the resulting stream (called 'substream') forms another valid bit stream which can be decoded by some target decoder. This *substream* is a representation of the source content with a less reconstruction quality. Intuitively, the streams which do not provide this property are the 'single layer bit streams'. The three modes of scalability are:

- Temporal scalability
- Spatial scalability
- Quality scalability (also called as SNR scalability)

Temporal scalability substream represents a reduced frame rate whereas a spatial scalability represents a reduced resolution or picture size. In Quality scalability, the spatio-temporal resolution remains the same, but the SNR ratio (or quality) is reduced. Next the authors tell us the possible benefits of the scheme which are more or less the same things that they mentioned in the previous section. However, there was an interesting point raised here (page 2) when the authors mention that the source content has to be encoded only once for the highest required resolution and bit rate. The receivers can then discard the selected parts from the bit stream and decoding only what they need. An analogy can be drawn here with the *highest required resolution and bit rate* compared to the least common multiple of some numbers where the numbers may represent the requirements of the receivers. Hence, this property can be used when a source has to deliver to a heterogeneous set of receives. The second benefit mentioned by the authors sounds quite clever. The argument that they put forward is that the bit stream usually contains parts that are not equally important (in terms of decoded video quality). Thus, the less important parts can be dropped from the bit stream, thus having to decode a smaller subset of the stream. Thirdly, the scheme is ideal for surveillance cameras where real time (and hence less delay) data is needed. Also, while archiving such data, the high quality parts can be removed from the streams, thus reducing the memory occupation.

There are a few essential requirements for an SVC scheme to be a successful one:

- Similar coding efficiency - The SVC scheme should not have lesser efficiency as compared to the single layer coding.
- Small increase in the decoder complexity – It doesn't make sense to add a multi fold complexity to the decoding system to get the benefits that SVC has to offer. Thus, the complexity increase in the decoder should be small.
- The scheme should support all three mode of scalability

Then the authors describe the long process of arriving at the final design of the SVC scheme (section 3) before plunging into background information for the scheme (section 4). This is where I think the authors lost the plot for the paper. Even though the paper was hard to read, the authors had up till now had described the system in a way which could be understood by a reader with little knowledge in the field. However, beyond this the authors just pour out facts and more facts without explaining even in a single line a multitude of new terminology, which I could not decipher completely even after armed with some background information gained from Wikipedia. Section 4 starts with the conceptual design of the H.264/AVC standard. The standard consists of two layers:

- Network Abstraction Layer (NAL)
- Video Coding Layer (VCL)

In brief, VCL refers to the part where the coding of the source is done to produce a bit stream and NAL formats this VCL data to provide header information to enable customization of the data for the variety of systems. Then a detailed description of NAL follows which tells us that an NAL unit is made of a 1-byte header and a payload. The NAL units are further classified into VCL NAL units and non-VCL NAL units. A set of consecutive NAL units are referred to as an 'access unit'. This access unit represents one picture, that is, the decoding of one access unit will give exactly one picture. Moving up the hierarchy, a set of consecutive access units represent a coded video sequence which an independently decodable part of a NAL unit bit stream.

The video coding layer is the where most of the action takes place, that is the coding of the stream. The scheme followed in the H.264/AVC standard is called as the 'block based hybrid video coding scheme'. Again, the authors throw out new terms now and then without much background information about it. They give an overview of the coding scheme in the next page which was a hard to understand without the aid of any diagrams. A couple of them, showing pictorially how the coding works would have been highly beneficial for the reader. The authors describe the coding method on this page: each picture is partitioned into macroblocks which are organized into slices. The macroblocks are spatially (or temporally) predicted and the resulting prediction error is represented using transform coding. The types of slice coding supported by H.264/AVC standard are:

- I-slice
- P-slice
- B-slice

Next the authors gave details on the various slices, which were terribly hard to understand. Again terms like 'adaptive deblocking filter', 'uniform reconstruction quantifiers' were used, which have not been explained anywhere in the text and thus I was not able to comprehend the matter at all.

The next section is the heart of the whole paper where the authors describe the concepts for extending the H.264/AVC standard towards an SVC standard. After this again it is a barrage of facts and facts without much logic and explanation. I could only catch a very high level idea of the matter. The three modes of scalability are tackled one by one, describing how they were achieved with the various kinds of predictions for all three.

I must admit that even with a little knowledge from class and wikipedia, however hard and however number of times I tried to read and understand the matter beyond this point, it made little sense to me. Though the paper was highly dense and difficult to read, I suppose a lack of necessary background was detrimental in understanding it. I do not have much else to add to the technical aspects of the paper but believe that if it had been written in a more simpler language or perhaps in a more logical (giving explanation of terms and providing more pictures) way, it could have been much easier to understand.