1

Video Coding

To develop an accurate video traffic model, we need to understand the statistical characteristics of the video traffic in detail. The model should match the characteristics of a real video sequence, such as the probability density function, mean, variance, peak, autocorrelation and coefficient of variation of the frame sizes. These characteristics can vary greatly depending on the specific video coding algorithms.

Video coding is the process of reducing the amount of data required to represent a digital video signal, prior to transmission or storage. Video data may be represented as a series of still image frames. The sequence of frames contains spatial and temporal redundancy. Video encoding algorithms take advantage of this redundancy to compress video, encoding only the difference between frames. Many different techniques for video coding have been proposed and researched. Hundreds of research papers have been published describing innovative compression schemes. However, commercial video applications use a limited number of standardized techniques for video compression, as standards are well defined and they simplify the interoperability between different manufacturers. In this chapter, we describe in detail the video coding process and also discuss the most popular video coding standards.

1.1. Video coding

A video encoder consists of three main functional units: a prediction model, a spatial model and an entropy encoder [RIC 10]. A block diagram of video encoder is shown in Figure 1.1. The input to the prediction model is an uncompressed (raw) video sequence. The goal of the prediction model is to reduce redundancy by forming a prediction of the data and subtracting this prediction from the current data. There are two types of prediction models:

temporal prediction and spatial prediction. In temporal prediction, the prediction is formed from previously coded frames while spatial prediction uses the previously coded image samples within the same frame. The output of the prediction model is a residual frame, created by subtracting the prediction from the actual current frame, and a set of model parameters indicating the intraprediction type or describing how the motion was compensated, also known as the motion vector.

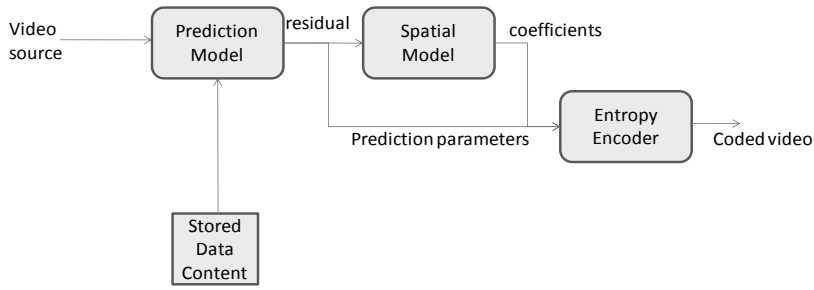

Figure 1.1. *Block diagram of an encoder*

The residual frame is given as an input to the spatial model that makes use of the similarities between local samples in the residual frame to reduce spatial redundancy. Transform coding, such as discrete cosine transform (DCT) is widely used for reducing the spatial redundancy in video coding. The transform converts the residual samples into another domain in which they are represented by transform coefficients. The coefficients are quantized to remove insignificant values, leaving a small number of significant coefficients that provide more compact representation of the residual frame. The output of the spatial model is a set of quantized transform coefficients.

The parameters of the prediction model, i.e. the intraprediction modes or the interprediction modes and motion vectors, and of the spatial model, i.e. the transform coefficients, are further compressed by the entropy encoder.

The entropy encoder removes statistical redundancy in data. For example, it represents commonly occurring vectors and coefficients with short binary codes. The average number of bits per symbol can be reduced if symbols with lower probability of occurrence are assigned longer code words and symbols with higher probability of occurrence are assigned shorter code words. Variable length coding such as the Huffman coding and arithmetic coding are two commonly used entropy coding techniques. The entropy encoder produces a compressed bit stream that may be transmitted or stored. A compressed sequence consists of coded prediction parameters, coded residual coefficients and header information.

Once the data are compressed, the bit stream is packetized and sent over the Internet. The video decoder reconstructs a video frame from the compressed bit stream. The coefficients and prediction parameters are decoded by the entropy decoder after which the spatial model is decoded to reconstruct the residual frame. The decoder used the prediction parameters together with previously decoded image pixels to create a prediction of the current frame and the frame itself is reconstructed by adding the residual frame to this prediction.

1.2. Video coding standards

Various standards have been developed for video encoding, i.e. H.261, H.263, MPEG-1, MPEG-2, MPEG-4 and H.264. Table 1.1 lists the important milestones in history of video coding standards.

Next, we provide a brief overview of some of the most popular coding standards.

Year	Standard	Publisher	Popular implementations
1990	H.261	ITU-T	Video conferencing, videotelephony
1992	MPEG-1	ISO	Video-CD
1994	MPEG-2/H.262	ISO, ITU-T	DVD video, digital video broadcasting, SVCD
1995	H.263	ITU-T	Video conferencing, videotelephony, video on mobile phones (3GP)
1998	MPEG-4	ISO	Video on Internet, DivX
2003	H.264/MPEG-4 AVC	ISO, ITU-T	Blu-ray, HD DVD, digital video broadcasting, HDTV, iPod video, Apple TV
2007	H.264/MPEG-4 SVC	ISO, ITU-T	Polycom video conferencing

Table 1.1. *History of video compression standards*

1.2.1. *The MPEG video coding standard*

The Moving Picture Experts Group (MPEG) was started in 1988 as a working group with the aim of defining standards for digital compression of audiovisual signals. MPEG has produced several standards for video compression that providers can use for Internet protocol (IP)-based services. Newer encoding schemes such as MPEG-4 offers potential bit-rate reductions of two times that of MPEG-2 with comparable quality.

The MPEG layer hierarchy, shown in Figure 1.2, consists of six different layers. A block is an 8×8 matrix of pixels or corresponding DCT information that represents a small chunk of brightness (luma) or color (chroma) within the frame. To ease the computation, four blocks (16×16 pixels)

are grouped together to form a macroblock. A single row of macroblocks in a video frame is called a slice. These slices are then grouped to form a video frame or a picture. Successive video frames are considered together as a group of pictures (GOP), which represents an independent unit in the video scene. The sequence layer is comprised of a sequence of GOPs. A sequence layer can be thought of as a video scene or shot.

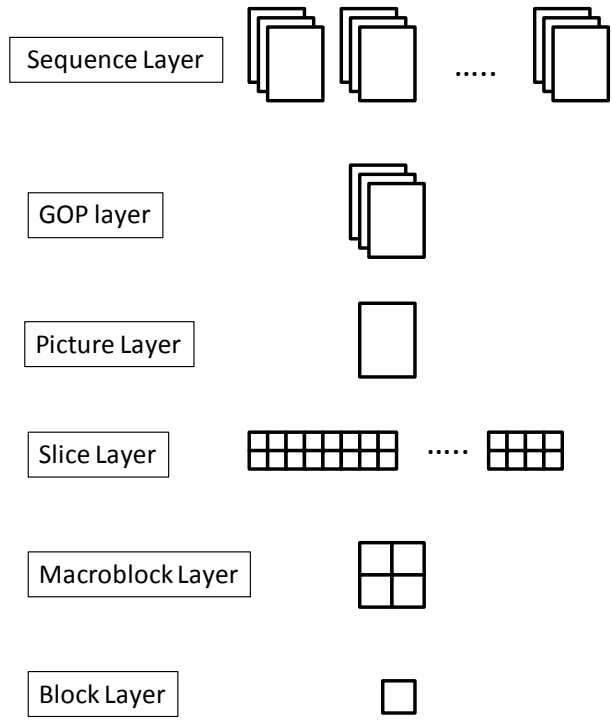

Figure 1.2. *The MPEG video hierarchy*

The MPEG-2/H.262 standard was jointly developed by the ISO and ITU standardization organizations and approved in 1994. MPEG-2 has three frame types that are organized in several possible ways within a GOP. Intra, or I, frames carry a complete video picture. They are coded without reference to other frames and might use spatial compression but do not

use temporal compression. That is, no information from other frames is used in the compression. Therefore, this frame can be decompressed even if other frames in the GOP are lost. Predictive-coded, or P, frames predict the frame to be coded from a preceding I or P frame using temporal compression. P frames can provide increased compression compared to I frames, with a P frame typically 20–70% the size of an associated I frame [GRE 09]. Finally, bidirectionally predictive-coded, or B, frames use the previous and next I frame or P frame as their reference points for motion compensation. B frames provide further compression, which is typically 5–40% the size of an associated I frame.

The frame order within a GOP depends on the interrelationships between frames, i.e. which frame is used as the reference for another frame. The interdependencies between the frames of a GOP are shown in Figure 1.3. The I frame provides direct reference for the B frames immediately preceding it within its GOP. It also provides reference for the first P frame in the GOP. As a result, the I frame directly or indirectly is the source of all the temporal encoding within a GOP. The first P frame in a GOP takes reference from the I frame; subsequent P frames take reference from the preceding P frame. Finally, a B frame takes bidirectional prediction from both P and I frames. In MPEG-2-encoded video, the B frame does not provide reference to any other frame. However, MPEG-4 Part 10 can use B frames as reference frames in hierarchical GOPs. This is one technique that MPEG-4 Part 10 employs to achieve greater compression than MPEG-2.

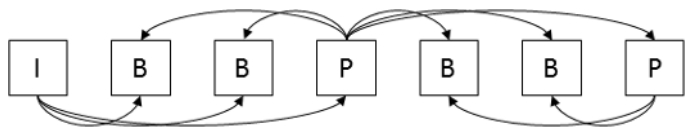

Figure 1.3. *Interframe dependecies within a GOP*

Many possible GOP structures exist and depend on the source video signal's format or on any bandwidth constraints on the encoded video stream (which determine the required compression ratio), and possible constraints on the encoding or decoding delay. One of the common patterns of GOP is G12B2, which means a GOP of size 12 and 2 B frames between successive I and P frames. The encoding order for this sequence is IBBPBBPBBPBB. Another common pattern is G16B3, a GOP of size 16 and 3 B frames between successive I and P frames. The encoding order for this sequence is IBBBPBBBPBBBPBBB. It is also possible to have GOP structures without any P frames such as IBBBBBBBBBBBBBBB, consisting of 16 I frames, with 15 B frames per I frame, denoted by G16-B15.

Due to dependencies between frames, their display order is not the same as their transmission order. Figure 1.4 shows the encoder input order or decoder display order. In Figure 1.4(a), frame I1 provides reference for frame P4. In turn, both frames I1 and P4 provide reference for frames B2 and B3. To decode the B frames, the decoder must have already decoded the associated reference frames. So, the decoder will need to receive and decode frames I1 and P4 before it can decode frames B2 and B3. Figure 1.4(b) shows the corresponding transmission order.

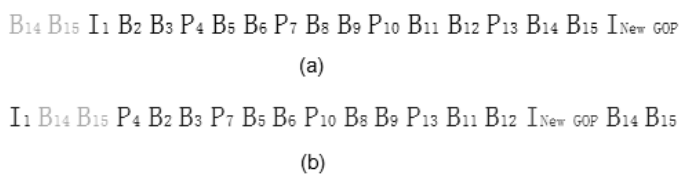

Figure 1.4. *GOP encoding and transmission orders*

To transport MPEG-2-encoded video over IP network, the MPEG frame data are encapsulated within MPEG transport stream (MPEG-TS) packets, which are in turn transported in

real-time protocol (RTP) over user datagram protocol (UDP) over IP packets. MPEG-TS uses a fixed length packet size and a packet identifier identifies each transport packet within the transport stream. A typical IP packet for transporting MPEG video contains seven 188 byte MPEG-TS packets. In Figure 1.5, we can see that several MPEG-TS packets are carried inside an IP packet. An MPEG frame can span multiple IP packets, and a single packet can contain two consecutive frames.

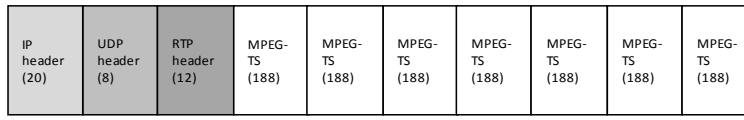

Figure 1.5. *MPEG-TS packetization*

1.2.2. *H.264/MPEG-4 AVC*

H.264/MPEG-4 advance video coding (AVC) represents a big leap in video compression technology with typically a 50% reduction of the average bit rate for a given video quality compared to MPEG-2 and approximately a 30% reduction compared with MPEG-4 Part 2 [OST 04]. Block transforms in conjunction with motion compensation and prediction are still the core of the encoder as in previous standards, but a number of new encoding mechanisms have been added, which improved the performance significantly. H.264 AVC also uses variable block sizes that introduce a different number of square and rectangular macroblock sizes, such as 4×4, 8×8 and 16×8 pixels. These different block sizes permit selecting the optimal block size for motion compensation and prediction.

In previous standards, one reference frame (I or P) from the past was allowed for prediction of P frame blocks, and one reference frame (I or P) from the past and one reference frame (I or P) from the future were allowed for the prediction

of B frame blocks. H.264 makes it possible to create other prediction structures, for example using multiple reference frames for prediction. An example is the IPPPP... structure in which all the previously coded slices are available as reference frames. The encoder can search up to *N* reference frames to find the best match for each P microblock.

Hierarchical B frames are an important new concept that was first introduced in H.264 AVC using generalized B frames and was later found to be the best method to build the scalable video coding (SVC) extension. In the classical B frame prediction structure, each B frame is predicted only from the preceding I or P frame and from the subsequent I or P frame. Other B frames are not referenced since this is not allowed by video standards. The hierarchical B frame structure allows B frames for the prediction of B frames. This can be seen in Figure 1.6.

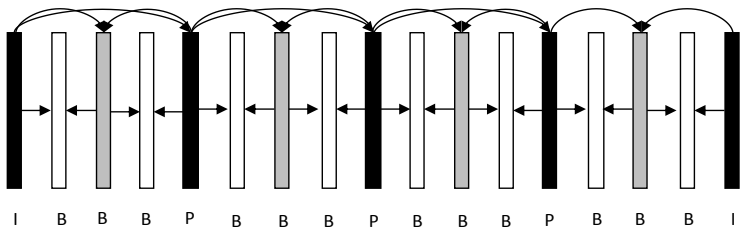

Figure 1.6. *Hierarchal B frames*

Another improvement from previous standards is the introduction of the entropy coding scheme context adaptive binary arithmetic coding (CABAC), which typically gives 10–15% bit rate savings over previous variable length coding schemes used in MPEG-2/4. Since arithmetic coding is computationally intensive, the main profile also supports a scheme called context adaptive variable length coding (CAVLC), which is an improved version of older variable length coding schemes.

The H.264 AVC standard classifies all the video coding mechanisms in the video coding layer (VCL). In addition, the H.264 AVC standard defines a network abstraction layer (NAL), which is responsible for transporting the coded video data over the network. The coded video data are organized into NAL units (NALUs). Each NALU contains an integer number of bytes of video encoded data, as well as a one byte header. The NALUs can then be transported in RTP/UDP/IP packets. In addition, there are also signaling NALUs for carrying the information regarding frame dependencies and the scalability hierarchy, etc.

1.2.3. *H.264 SVC*

Modern video transmission systems and networks support a wide range of connection qualities, network protocols and receiving devices, ranging from smart phones and tablets to high-definition televisions. The varying connection quality is due to the adaptive resource sharing mechanisms of the underlying network that responds to the varying data throughput requirements of a varying number of users. SVC is an attractive solution to the problems due to the distinctive characteristics of these transmission systems. A video bit stream is called scalable when parts of the stream can be removed in a way that the resulting substream forms another valid bit stream for some target decoder. This substream represents the source content with a reconstructed quality that is less than that of the complete original bit stream [SCH 07].

H.264 SVC supports layer-scalable coding. A layer-scalable encoding consists of a base layer and one or several enhancement layers identified by increasing layer identifiers. H.264 SVC provides three types of scalability, i.e. temporal scalability, spatial scalability and quality (signal-to-noise ratio (SNR)) scalability [SCH 07].

A bit stream provides temporal scalability when the set of corresponding frames can be partitioned into a temporal base layer and one or more temporal enhancement layers with the following property. Let the temporal layers be identified by a temporal layer starting from 0 for the base layer. Then, for temporal layer identifier k, the bit stream that is obtained by removing all frames of all temporal layers with a layer identifier T greater than k forms another valid bit stream for the given decode [SCH 07]. The introduction of hierarchical B frames has allowed the H.264 SVC encoder to achieve temporal scalability. It allows B frames for the prediction of B frames. The temporal scalability is shown in Figure 1.6, where the I and P frames constitute the temporal layer 0 while the B frames in light gray color and the B frames in white color constitute the layers 1 and 2, respectively.

Spatial scalability provides different spatial frame resolutions. It provides a mechanism for reusing an encoded lower resolution version of an image sequence for the coding of a corresponding higher resolution sequence. Each spatial layer employs motion compensated prediction and intraprediction [SEE 12]. Figure 1.7 depicts the intra- and interlayer prediction dependencies for two spatial layers (base and enhancement). Here, we should note that the base layer is used for the prediction of the spatial enhancement layer. Quality scalability can be considered as a special case of spatial scalability with identical picture sizes for base and enhancement layer. It is also referred to as coarse-grain quality scalable coding (CGS). H.264 SVC CGS employs same interlayer prediction mechanisms as for spatial scalable coding, but without using the corresponding upsampling operations and the interlayer deblocking for intracoded reference layer macroblocks [SCH 07].

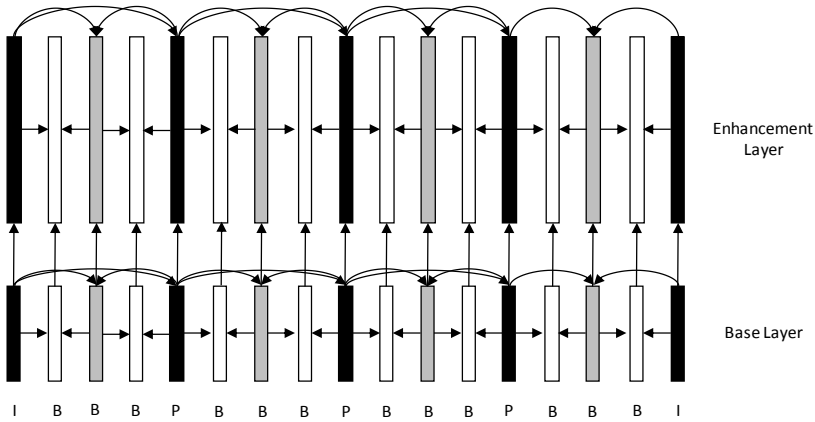

Figure 1.7. *Two-layer spatial scalability intra- and interlayer prediction dependencies*

1.2.4. *H.264 MVC*

Three-dimensional (3D) video has become very popular during the last few years due to the advancement in display technology, signal processing, circuit design and networking infrastructure. One common characteristic of many of these systems is that they use multiple camera views of the same scene, often referred to as multiview video (MVV), which is implemented by simultaneously capturing the video streams of several cameras. Since this approach creates large amounts of data to be stored or transmitted to the user, efficient compression techniques are essential for realizing such applications. An easy solution for this would be to encode all the video signals independently using a video codec such as H.264 AVC. However, MVV contains a large amount of inter-view statistical dependencies, since all cameras capture the same scene from different viewpoints.

The basic approach of most multiview coding (MVC) schemes is to exploit not only the redundancies that exist temporally between the frames within a given view, but also the similarities between frames of neighboring views.

However, it is required for the compressed multiview (MV) stream to include a base view bit stream, which is coded independently from all other views and is compatible with decoders for single-view profiles. This allows the two-dimensional (2D) version of the content to be easily extracted and decoded. For example, in television broadcast, legacy receivers should be able to extract and decode the base view, while newer 3D receivers can decode the complete 3D bit stream.

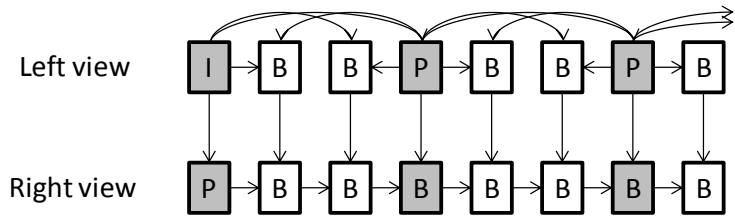

Figure 1.8. *Inter-view prediction in MVC [VET 11b]*

Figure 1.8 shows the frame prediction structure for MVC video that consists of two views: left and right. The left view, being the base view, is coded independently from the right view. The right view is coded with reference to the left view. Each view is transmitted as a separate video stream.

The overall structure of MVC defining the interfaces is shown in Figure 1.9. The MV encoder receives N temporally synchronized video streams and generates one bit stream. The MV decoder receives the bit stream, decodes and outputs the N video signals.

In [PUL 12], the authors analyzed the basic traffic and statistical multiplexing characteristics of 3D videos encoded with the three main representation formats: MV representation, frame sequential (FS) representation and side-by-side (SBS) representation. With MV representation, each view has the same frame rate as the underlying temporal video format. The FS representation merges the two

views to form a single sequence with twice the frame rate and applies conventional single-view encoding. Lastly, the SBS representation halves the horizontal resolution of the views and combines them to form a single-frame sequence for single-view encoding. Typically, the left and right views are subsampled and interleaved into a single frame. The results showed that the MV representation achieves the most efficient encoding, but generates high traffic variability, which makes statistical multiplexing more challenging.

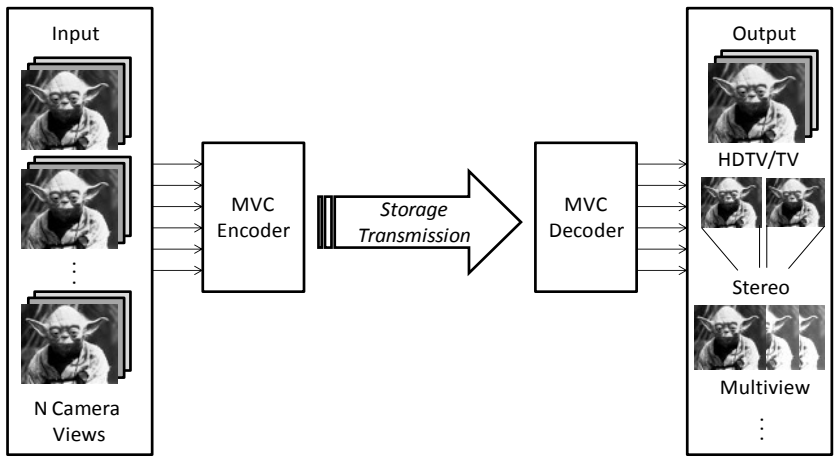

Figure 1.9. *MVC video system*

Vetro *et al.* reviewed the 3D video representation and compression formats and their applicability in different storage and transmission systems in [VET 11a]. The authors have provided a detailed comparison between the frame-compatible format, which merges the left and right views, and MV representation. The frame-compatible signals can work within existing infrastructures and already deployed video decoders. To facilitate and encourage their adoption, the H.264 AVC standard introduced a new supplemental enhancement information (SEI) message that enables signaling of the frame packing arrangement used. Thus, the frame-compatible coding with SEI message

signaling has been selected as the delivery format for initial phases of broadcast, while full-resolution coding with inter-view prediction based on MVC has been adopted for distribution of 3D on Blu-ray disc.

1.3. Rate control

Rate control is an essential part of most video encoders. It determines the number of bits or the quality level of the encoded frame. At lower bit rates, the perceived quality of coded video can be determined by trade-offs between the frame rate and the quantized frame quality. The video sequence contains varying contents and motions, which makes the rate control difficult. An encoder should balance the quality of decoded images against the channel bandwidth.

There are two types of rate control: constant bit rate (CBR) and variable bit rate (VBR). CBR encoding means that the rate at which video should be consumed is constant. CBR is useful for streaming multimedia content on limited capacity channels as this mode generates a CBR that can be predefined by a user. The disadvantage with CBR is that when there is increased activity in a scene, which results in a bit rate that is higher than the target rate, the restriction to keep the bit rate constant leads to a lower image quality and frame rate. Therefore, we see fluctuations in video quality due to scene changes and other video content. For CBR video coding, rate control designers focus on improving the matching accuracy between the target bit rate and the actual bit rate and satisfying the low latency and buffer constraints [HWA 09].

In cases where rate or delay constraints are not so strict, VBR coding is preferred. Using VBR, a predefined level of image quality can be maintained regardless of motion or the lack of it in a scene. This means that bandwidth usage will increase when there is a lot of activity in a scene and will

decrease when there is no motion. An example where VBR is more desirable is video surveillance applications where there is a need for high quality, particularly if there is motion in a scene. Since the bit rate may vary, even when an average target bit rate is defined, the network infrastructure must be able to accommodate high throughputs.

As described earlier, the core part of video coding standards is the motion compensation and the DCT coding. If the coding parameters remain unchanged in the compression process, the number of bits per frame can be significantly different. Using a buffer can make the output bit stream smooth. However, the buffer capacity has certain limitations, for example if the buffer is too large, the propagation delay of real-time communication is longer, which is not desirable. To prevent buffer overflow and underflow, rate control must be used in encoder. This is the reason why the rate control has been widely applied in most video coding standards, such as MPEG-2, MPEG-4, H.263, H.264, etc.

A rate control algorithm dynamically adjusts encoder parameters to achieve a target bit rate. Rate control in VBR video coding is typically accomplished in three steps [WU 11]: (1) Update the target average bit rate in terms of bits per second (bps) for each short time interval, also referred to as the rate update interval. (2) Determine the target bit rate budget for each frame to be coded in this interval. This budget is based on the target average rate for the interval and the current buffer fullness. (3) Determine the quantization parameter (QP) for each macroblock in a frame to meet the target rate for this frame.

1.4. Summary

In this chapter, we have discussed the basics of video coding. We have provided an overview of some video coding standards and discussed the rate control in video coding. The

H.264 AVC and H.264 SVC coding standards lead to significant average bit rate savings with respect to the MPEG-4 Part 2 and earlier standards. At the same time, the variability of the H.264 AVC and H.264 SVC video traffic is significantly higher than the variability of the MPEG-4 Part 2 video traffic. This makes rate control a crucial part of video coding and transmission process.

3D video has become quite popular during the last few years. We have also discussed the basics of MVC and the H.264 MVC coding standard that is used for coding and transmission of 3D video.