

## Chapter 2

# Video Coding Standards and Video Formats

**Abstract** Video formats, conversions among RGB, Y, Cb, Cr, and YUV are presented. These are basically continuation from [Chap. 1](#) and thus complement the topics discussed in [Chap. 1](#).

**Keywords** Video compression • Video coding standards • Sampling formats • Video formats • RGB • YUV • YCbCr • Quality • PSNR • SSIM

### 2.1 Introduction

From analog television to digital television, VHS video tapes to DVDs, cell phones used for only making calls and send text messages to cell phones functioning as cameras, web browsers, navigation systems, social networking devices and barely used to make calls, there has been quite a revolution over the past few years in the way users create, share and watch videos. The continuous evolution of digital video industry is driven by commercial factors and technological advances. The commercial drive comes from the huge revenue potential of persuading consumers and businesses. In the technology field, the factors include better communications infrastructure, inexpensive broadband networks, 4G mobile networks and the development of easy-to-use applications for recording, editing, sharing and viewing videos.

There are a series of processes involved in getting a video from a source (camera or stored clip) to its destination (a display). The key processes in this operation are compression (encoding) and decompression (decoding), which involve in reducing the “bandwidth intensive” raw video source to an optimal size suitable for transmission or storage, then reconstructed for display. For having that commercial and technical edge to a product, the compression and decompression processes should strike a proper balance between three parameters that are odds

with one another: quality of video, encoding time and the size. There is therefore, an acute interest in video compression and decompression techniques and systems.

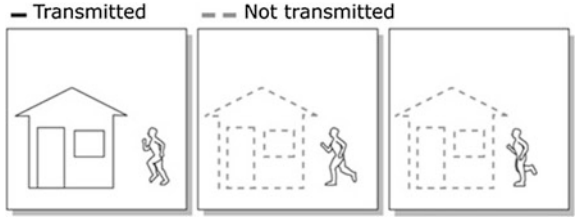
The ever growing market for high bit-rate connections, large storage capacity of hard disks, flash memories and optical media has come a long way to satiate the demands of users. With the price per transmitted or stored bit continually falling, video compression is an absolute necessity and there has been a significant effort to make it better. Imagine a world without video compression; current Internet throughput rates would have been insufficient to handle uncompressed video in real time (even at low frame rates and/or small frame size), a digital versatile disk (DVD) could only store a few seconds of raw video at television-quality resolution and frame rate. Video compression enables an efficient use of transmission and storage resources. For example, if a high bit-rate transmission channel is available, then it is a more attractive proposition to send high-resolution compressed video or multiple compressed video channels than to send a single, low-resolution, uncompressed stream. Even with constant advances in storage and transmission capacity, compression is likely to be an essential component of multimedia services for many years to come [B8].

By definition, compression is the process of removing redundancy from an information carrying signal. In a lossless compression system, statistical redundancy is removed so that the original signal can be perfectly reconstructed at the receiver. Unfortunately, there is a trade off involved here; lossless methods only achieve a modest amount of compression of video signals. Most of the practical video compression algorithms are based on lossy compression, in which greater compression is achieved with the penalty that the decoded video signal is not identical to the original. The goal of a video compression algorithm is to achieve efficient compression while minimizing the distortion introduced by the compression process.

When it comes to video clips, it is possible to compress the data by combining the principles behind lossless and lossy encoding. The simplest ways of building a video clip is to tack together consecutive pictures and refer to them as frames. Inherently, there is a lot of redundancy in a video clip; most of the information contained in a given frame is also in the previous frame. Only a small percentage of any particular frame is new information; by calculating where that percentage of information lies, and storing only that amount, it is possible to drastically cut down the data size of the frame. This compression process involves applying an algorithm to the source video to create a compressed file that is ready for transmission or storage. An inverse algorithm is applied to the compressed video to produce a video that shows nearly the same content as the original video. This pair of algorithms which work together is called a video codec (encoder/decoder).

Video compression algorithms such as MPEG-4 [B8] and H.264 [B8, B18, H44] are highly complex processes which include techniques such as difference coding, wherein only the first image is coded in its entirety. Referring to Fig. 2.1, in the two following images, references are made to the first picture for the static elements, i.e. the house. Only the moving parts, i.e. the running man, are coded using motion vectors, thus reducing the amount of information that is sent and

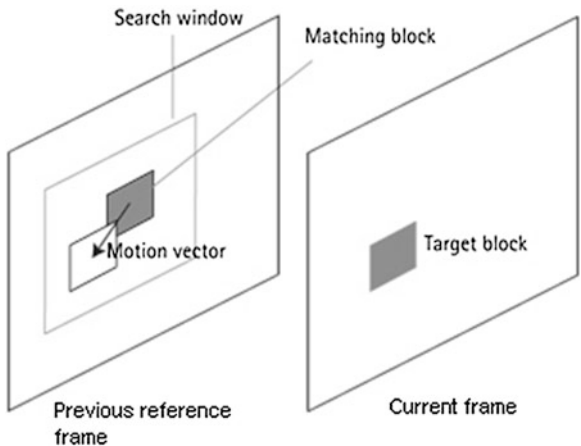
**Fig. 2.1** Inter frame prediction in modern video compression algorithms [V9]



stored. Also, techniques like block-based motion compensation are included to further reduce the data. Block-based motion compensation is based on the observation that a new frame in a video sequence can be found in an earlier frame, but perhaps in a different location. This technique divides a frame into a series of macro-blocks (blocks of pixels). Block by block, a new frame can be predicted by finding a matching block in a reference frame. If there is a match, the encoder codes only the position where the matching block is to be found in the reference frame. This technique takes a lot less number of bits than coding the actual content of a block itself (Fig. 2.2).

H.264 video coding standard is the latest block-oriented motion-compensation-based codec standard developed by the ITU-T Video Coding Experts Group (VCEG) together with the ISO/IEC Moving Picture Experts Group (MPEG) [H44]. The intent of the H.264 standard was to create a video coding standard which could achieve quality equivalent to previous standards at substantially lower bit rates. H.264 provides significantly better compression than any previous standards, it contains a number of built in features to support reliable, robust transmission over a range of channels and networks. Unfortunately, this comes with a cost of increased encoder computational complexity when compared to previous standards. To achieve a practical implementation of H.264/AVC, a significant reduction in encoding complexity must be achieved while maintaining the coding efficiency [H35].

**Fig. 2.2** Illustration of block-based motion compensation [V9]



## 2.2 Complexity Reduction

The requirement of capturing and playing of high definition video applications on devices like smart phones and tablets has led to a challenging scenario of developing efficient video encoders with low complexity. Many techniques have been proposed by researchers around the globe to reduce the complexity in H.264. Different Intra mode complexity reduction approaches like in [H10, H12, H14, H35] have been proposed, but very few approaches achieve efficient encoding. Some approaches reduce the encoding time but, fail to maintain the quality of that of original video clip. It is important to strike a balance between gain and quality. [H27] proposes an efficient intra mode complexity reduction algorithm; wherein the encoding time of a video clip is greatly reduced with negligible quality loss and increase in bit-rate.

The thesis by Muniyappa (see PS at the end) focuses on reducing encoding complexity of H.264 for intra mode selection by making use of JM 18.0 [H30]. It is heavily based on the observation that adjacent macro-blocks tend to have similar properties. Thus, by simple use of directional masks and neighboring modes, the usually tasking RDO (Rate Distortion Optimization) which examines all possible combinations of coding modes process can be reduced significantly [H27]. Results show reduction in complexity in terms of encoding time for different video formats and video context. [Sections 2.3](#) through [2.5](#) give a brief insight about some of the video coding standards and video formats.

## 2.3 Video Coding Standards

There are many video coding techniques proposed and many other researches still ongoing out there. Hundreds of research papers are published each year describing new and innovative compression techniques. However, the commercial video coding applications tend to use a limited number of standardized techniques for video compression. Standardized video coding formats have a number of benefits like [B14]:

- Standards simplify inter-operability between encoders and decoders from different manufacturers.
- Standards make it possible to build platforms that incorporate video, in which many different applications such as video codecs, audio codecs, transport protocols, security and rights management, interact in well defined and consistent ways.
- Many video coding techniques are patented and therefore there is a risk that a particular video codec implementation may infringe patent(s). The techniques and algorithms required to implement a standard are well defined and the cost of licensing patents that cover these techniques, i.e., licensing the right to use the technology embodied in the patents, can be clearly defined.

## 2.4 MPEG and H.26x

The Recommendations or International Standards are prepared jointly by ITU-T SG16 Q.6 (the International Telecommunication Union), also known as VCEG (Video Coding Experts Group) and by ISO/IEC JTC1/SC29/WG11 (the International Organization for Standardization), also known as MPEG (Moving Picture Experts Group). VCEG was formed in 1997 [H44] to maintain prior ITU-T video coding standards and develop new video coding standard(s) appropriate for a wide range of conversational and non-conversational services. MPEG was formed in 1988 [S13] to establish standards for coding of moving pictures and associated audio for various applications such as digital storage media, distribution, and communication. Later on, the ITU-T Video Coding Experts Group (VCEG) and the ISO/IEC Moving Picture Experts Group (MPEG) formed a Joint Video Team (JVT) in 2001 for development of a new Recommendation or International Standard, H.264 Recommendation/MPEG-4 part 10 standard [H2].

### 2.4.1 H.120

H.120 [S1], the first digital video coding standard was developed in 1984 by ITU-T formerly CCITT (the International Telegraph and Telephone Consultative Committee). It evolved into different versions, Version 1 developed in 1984 featured conditional replenishment, differential pulse code modulation, scalar quantization, variable length coding and a switch for quincunx sampling. Version 2 developed in 1988 added motion compensation and background prediction. In 1993, a final edition was published as a result of the creation of the ITU-T to replace the prior CCITT standardization body. H.120 streams ran at 1,544 kbps for NTSC (National Television System Committee) and 2,048 kbps for PAL (Phase Alternating Line) [S11].

H.120 video was not of good quality for practical use since the differential PCM (Pulse Code Modulation) in it worked on pixel by pixel basis, which is good for spatial resolution but the temporal quality was really poor. It was necessary to improve the quality of video without exceeding the target bitrates for the stream. Hence the researchers came up with the block-based codecs that followed H.120, such as H.261 [S4].

### 2.4.2 H.261

H.261 [S4, B3] was the first video codec with the widespread practical success (in terms of product support in significant quantities). The first design of this ITU-T video coding standard was in 1988 and was the first member of the H.26x family. H.261 was originally designed for transmission over ISDN (Integrated Services Digital Network) lines on which data rates are integer multiples of 64 kbps. The

coding algorithm uses a hybrid of motion compensated inter-picture prediction and spatial transform coding with  $16 \times 16$  macro-block motion compensation,  $8 \times 8$  DCT (discrete cosine transform) [B2], scalar quantization, zigzag scan and variable-length coding. All the subsequent international video coding standards have been based closely on the H.261 design [S11]. Figure 2.3 shows an outline block diagram of the H.261 codec.

### 2.4.3 MPEG-1

MPEG-1 (Moving Picture Experts Group) [S3] was developed by ISO/IEC JTC1 SC29 WG11 (MPEG) in 1993 [S1]. MPEG-1 provides the resolution of  $352 \times 240$  (source input format) for NTSC or  $352 \times 288$  for PAL at 1.5 Mbps. MPEG-1 had a superior video quality compared to H.261 when operated at higher bit rates and was close to VHS quality. Its main applications were focused on video storage for multimedia (e.g., on CD-ROM).

### 2.4.4 H.262/MPEG-2

H.262/MPEG-2 [S10] coding standard was jointly developed by ITU-T Video Coding Experts Group and ISO/IEC Moving Picture Experts Group in 1994 [S1]. MPEG-2 video is similar to MPEG-1, but also provides support for interlaced video (the format used by analog broadcast TV systems). MPEG-2 video is not optimized for low bit-rates (less than 1 Mbps), but outperforms MPEG-1 at 3 Mbps and above. For the consistency of the standards, MPEG-2 is also compatible with MPEG-1, which means a MPEG-2 player can play back MPEG-1 video without any modification.

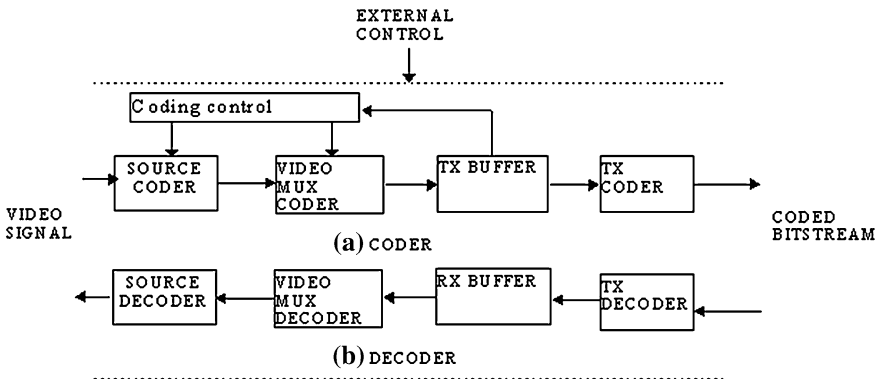


Fig. 2.3 Outline block diagram of H.261 encoder and decoder [S4] © ITU-T 1993

### **2.4.5 H.263, H.263+ and H.263++**

This next generation of video coding overtook H.261 as the most dominant video conferencing codec. H.263 [S6] has superior video quality compared to its prior standards at all bit rates, by a factor of two. H.263 Version 1 was developed by ITU-T in 1995. Features which beat H.261 [S11] are:

- 3-D variable length coding of DCT coefficients
- Median motion vector prediction
- Bi-directional prediction
- Arithmetic entropy coding.

H.263+ or Version 2 was developed in the late 1997 and early 1998 [S7], which included lot of new features like error resilience, custom and flexible video formats, supplemental enhancement information and also there was an improved compression efficiency over H.263v1. H.263++ or Version 3 [S6], developed in 2000 came with significant improvement in picture quality, packet loss and error resilience and additional supplemental enhancement information.

### **2.4.6 MPEG-4**

MPEG-4 [S9] an ISO/IEC standard was developed by MPEG (Moving Picture Experts Group) in late 1998. The fully backward compatible extensions under the title of MPEG-4 Version 2 were frozen at the end of 1999, to acquire the formal International Standard Status early in 2000. To cater to variety of applications ranging from low-quality, low-resolution surveillance cameras to high definition TV broadcasting and DVDs, MPEG-4 Part 2 has approximately 21 profiles. Some of the profiles are listed below [S9]:

- Simple
- Simple Scalable
- Main
- Core
- N-Bit
- Hybrid
- Basic Animated Texture
- Scalable Texture
- Simple FA (face animation)
- Core Scalable
- Advanced Scalable Texture
- Simple FBA
- Advanced Coding Efficiency
- Advanced Real Time Simple.

### 2.4.7 H.264/MPEG-4 Part 10/AVC

In 1998, the ITU-T Video Coding Experts Group (VCEG) started work on a long term effort to draft “H.26L” standard, which would offer significantly better video compression efficiency than previous ITU-T standards. In 2001, the ISO Moving Picture Experts Group (MPEG) recognized the potential benefits of H.26L and the Joint Video Team (JVT) was formed, including experts from MPEG and VCEG with the charter to finalize the new video coding standard H.264/AVC. The “official” title of the new standard is Advanced Video Coding (AVC); however, it is widely known by its old working title, H.26L and by its ITU document number, H.264 [H44, H23, H25, H2].

Figure 2.4 Video coding standardization (courtesy Dr. Nam Ling, Sanfilippo Family Chair Professor, Dept. of Computer Engineering, Santa Clara University, Santa Clara, CA, USA).

H.264 [H44] has brought in a significant increase in compression ratio and also saves up to 50 % bit rate as compared to its prior video coding standards. The standard can increase resilience to errors by supporting flexibility in coding as well as organization of coded data. The increase in coding efficiency and coding flexibility comes at the expense of increase in complexity as compared to the other standards. These features are discussed in much detail in Chap. 4.

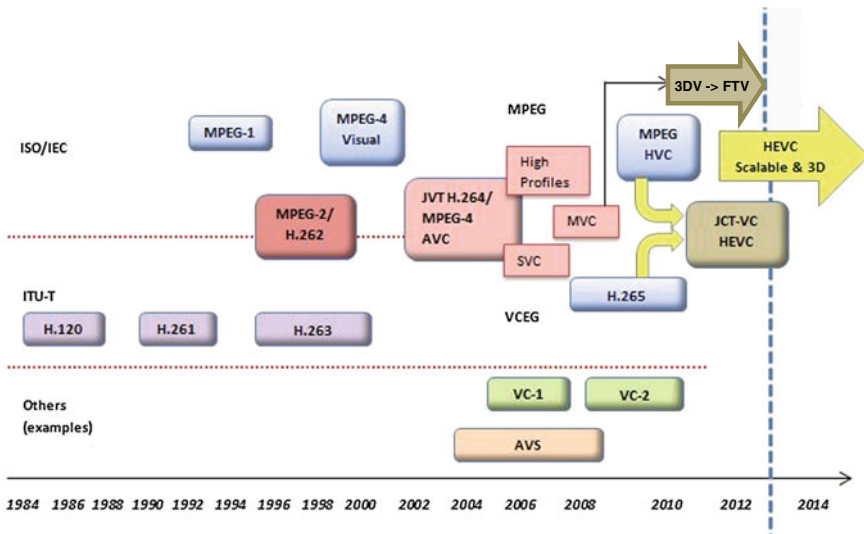


Fig. 2.4 Chronology of international video coding standards



### 2.4.8 H.265/HEVC

High efficiency video coding (HEVC) is the latest video coding standard being developed by ITU-T/ISO-IEC. Three profiles (Main, Main 10 and Main still picture—*intra frame only*) have been approved by ITU-T in January 2013. This is described in detail in [Chap. 5](#).

## 2.5 Video Formats and Quality

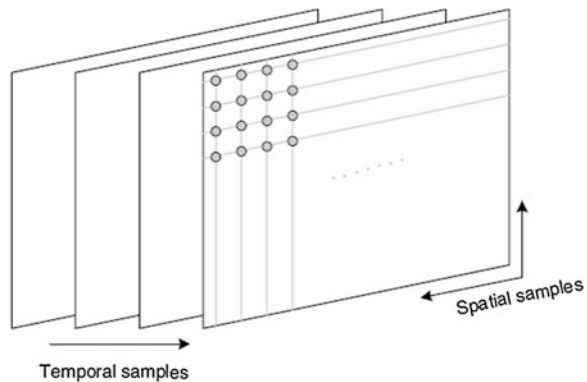
A typical real-world scene is composed of multiple objects with their own characteristic shape, depth, texture and illumination. The spatial characteristics like texture variation within a scene, number and shape of objects, color, etc., and temporal characteristics like object motion, changes in illumination, movement of the camera or viewpoint, etc., of a typical natural video scene are relevant for video processing and compression.

A natural visual scene is spatially and temporally continuous. Representing a visual scene in digital form involves sampling the real scene spatially (usually on a rectangular grid in the video image plane) and temporally (as a series of still frames or components of frames sampled at regular intervals in time) (Fig. 2.5). Digital video is the representation of a sampled video scene in digital form [B18].

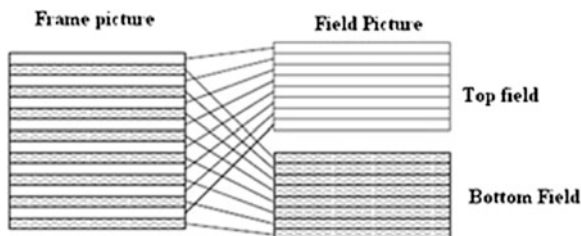
### 2.5.1 Frames and Fields

A video signal can be progressively sampled (series of complete frames) or interlaced (sequence of interlaced fields). In an interlaced video sequence two fields comprise one video frame (Fig. 2.6) and a field consists of either the odd-numbered or even-numbered lines within a complete video frame. The advantage

**Fig. 2.5** Spatial and temporal sampling of a video sequence [B18] © 2010 Wiley



**Fig. 2.6** Interlaced video sequence



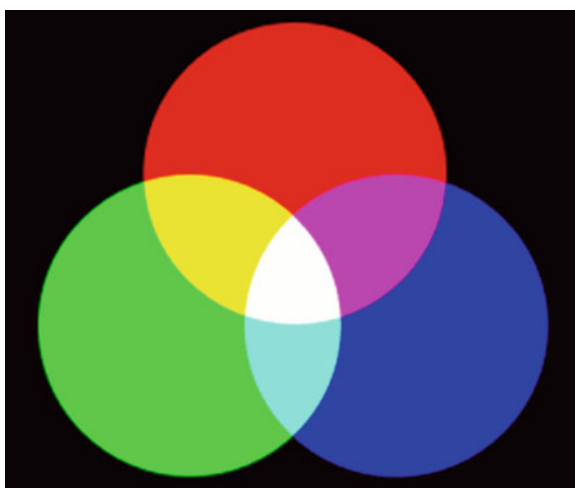
of this sampling method is that it is possible to send twice as many fields per second as the number of frames in an equivalent progressive sequence with the same data rate, giving the appearance of smoother motion [B18].

### 2.5.2 Color Spaces

Almost all digital video applications at present have color displays; hence it becomes a necessity to represent this color information. Color space method comes in handy in symbolizing brightness (luminance or luma) and color (Fig. 2.7).

In the RGB color space, a color image sample is represented with three numbers that indicate the relative proportions of Red, Green and Blue (the three additive primary colors of light). Any color can be created by combining red, green and blue in varying proportions. In the RGB color space the three colors are equally important and so are usually all stored at the same resolution. However, the human visual system has lower acuity for color difference than for luminance. Therefore, the well known color space YUV is used, which represents a color image more

**Fig. 2.7** Red, Green and Blue color space



efficiently by separating the luminance from the color information and representing luma with a higher resolution than color. Y is the luminance (luma) component and can be calculated as a weighted average of R, G and B.

$$Y = k_r R + k_g G + k_b B \quad (2.1)$$

where  $k_r + k_g + k_b = 1$ .

The color difference information (Chroma) can be derived as:

$$C_b = B - Y \quad (2.2)$$

$$C_r = R - Y \quad (2.3)$$

$$C_g = G - Y \quad (2.4)$$

In reality, only three components (Y,  $C_b$  and  $C_r$ ) need to be transmitted for video coding because  $C_g$  can be derived from Y,  $C_b$  and  $C_r$ . As recommended by ITU-R [S12],  $k_r = 0.299$ ,  $k_g = 0.587$  and  $k_b = 0.114$ . The Eqs. (2.2) thru (2.4) can be rewritten as:

$$Y = 0.299R + 0.587G + 0.114B \quad (2.5)$$

$$C_b = 0.564(B - Y) \quad (2.6)$$

$$C_r = 0.713(R - Y) \quad (2.7)$$

$$R = Y + 1.402C_r \quad (2.8)$$

$$G = Y - 0.344C_b - 0.714C_r \quad (2.9)$$

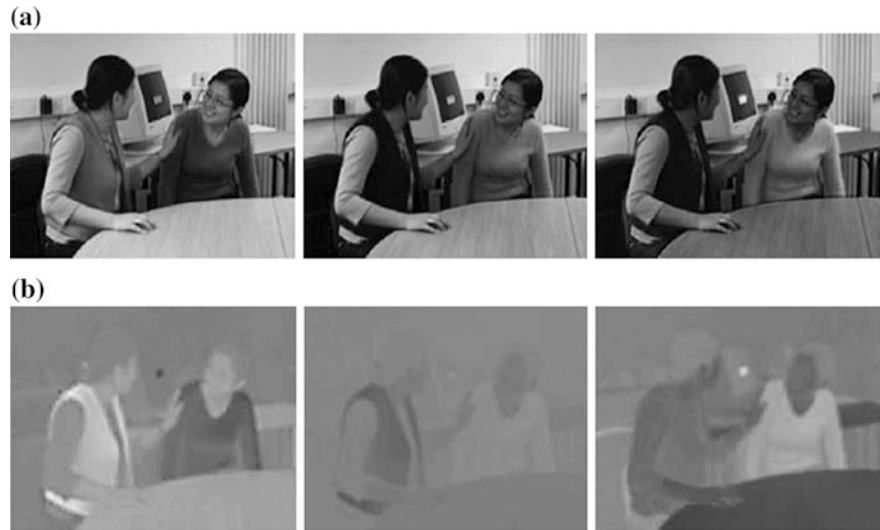
$$B = Y + 1.772C_b \quad (2.10)$$

Figure 2.8a shows the red, green and blue components of a color image in comparison to chroma components  $C_r$ ,  $C_g$  and  $C_b$  of Fig. 2.8b.

### 2.5.2.1 YCbCr Sampling Formats

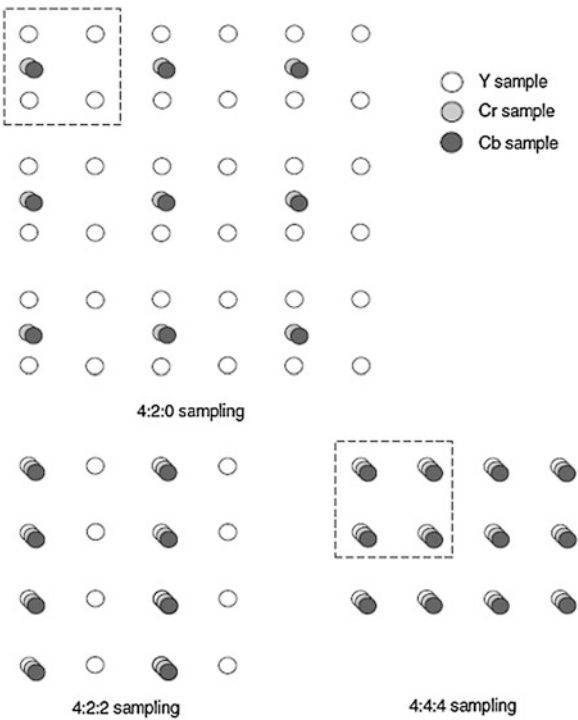
Figure 2.9 shows three sampling patterns for Y,  $C_b$  and  $C_r$  that are supported by modern video coding standards like, MPEG-4 Visual and H.264. 4:4:4 sampling preserves the full fidelity of the chrominance components. The three components Y,  $C_b$  and  $C_r$  have same resolution and for every four luminance samples there are four  $C_b$  and four  $C_r$  samples. In 4:2:2 sampling also referred as YUY2, the chrominance components have the same vertical resolution as the luma but half the horizontal resolution. Meaning, for every four luminance samples in the horizontal direction there are two  $C_b$  and two  $C_r$  samples. 4:2:2 video is usually used for high-quality color reproduction.

The most popular sampling pattern is 4:2:0 also referred as YV12. Here  $C_b$  and  $C_r$  have half the horizontal and vertical resolution of Y, each color difference



**Fig. 2.8** **a** Red, Green and Blue components of color image [B18] © 2010 Wiley. **b**  $C_r$ ,  $C_g$  and  $C_b$  components of color image [B18] © 2010 Wiley

**Fig. 2.9** 4:2:0, 4:2:2 and 4:4:4 sampling patterns (progressive) [B18] © 2010 Wiley



component contains one quarter of the number of samples in the Y component. 4:2:0 YCbCr video requires exactly half as many samples as 4:4:4 or RGB video, hence is widely used for consumer applications such as video conferencing, digital television and DVD [B18].

2.5.3 Video Formats

It is a very common practice to capture or convert to one of a set of “intermediate formats” prior to compression and transmission. Table 2.1 shows some of the popular set of formats.

The choice of frame resolution depends on the application and available storage or transmission capacity. For example, 4CIF is appropriate for standard-definition television and DVD-video; CIF and QCIF are popular for videoconferencing applications; QCIF or SQCIF are appropriate for mobile multimedia applications where the display resolution and the bit-rate are limited. SIF (Source Input Format) is practically identical to CIF, but taken from MPEG-1 rather than ITU standards. SIF on 525-line (“NTSC”) based systems is  $352 \times 240$ , and on 625-line (“PAL”) based systems, it is identical to CIF ( $352 \times 288$ ). SIF and 4SIF are commonly used in certain video conferencing systems [H53].

2.5.4 Quality

It is necessary to determine the quality of the video images displayed to the viewer in order to specify, evaluate and compare. Visual quality is inherently subjective and is influenced by many factors that make it difficult to obtain a completely accurate measure of quality. Measuring visual quality using objective criteria gives accurate, repeatable results but as yet there are no objective measurement systems that completely reproduce the subjective experience of a human observer watching a video display [B18].

Table 2.1 Video frame formats [H53]

Format	Video resolution
Sub-QCIF	$128 \times 96$
Quarter CIF (QCIF)	$176 \times 144$
SIF (525)	$352 \times 240$
CIF/SIF (625)	$352 \times 288$
4SIF (525)	$704 \times 480$
4CIF/4SIF (625)	$704 \times 576$
16 CIF	$1408 \times 1152$
DCIF	$528 \times 384$

### 2.5.4.1 PSNR

Peak signal to noise ratio (PSNR) is the most widely used objective quality measurement. PSNR (Eq. 2.11) is measured on a logarithmic scale and depends on the mean square error (MSE) of between an original and an impaired image or video frame, relative to  $(2^n - 1)^2$  (the square of the highest-possible signal value in the image, where  $n$  is the number of bits per image sample).

$$\text{PSNR}_{\text{dB}} = 10 \log_{10} \left( (2^n - 1)^2 / \text{MSE} \right) \quad (2.11)$$

PSNR can be calculated easily and quickly and is therefore, a very popular quality measure, widely used to compare the ‘quality’ of compressed and decompressed video images.

### 2.5.4.2 SSIM

The structural similarity (SSIM) [Q13] index (see Appendix C) is a method for measuring the similarity between two images. The SSIM index can be viewed as a quality measure of one of the images being compared provided the other image is regarded as of perfect quality.

## 2.6 Summary

This chapter is basically a continuation of Chap. 1. The following chapter describes AVS China in detail.

PS: This chapter is based on the thesis by S.M. Muniyappa, “Implementation of complexity reduction algorithm for intra mode selection,” EE Dept., UTA, Dec. 2011. The thesis can be accessed from [www-ee.uta.edu/dip](http://www-ee.uta.edu/dip) web site. Click on courses and then click on EE 5359. Scroll down and see list of Theses/Projects.

Video coding standards

AVS China, H.264/MPEG-4 PART 10, HEVC, VP6, DIRAC  
and VC-1

Rao, K.R.; Kim, D.N.; Hwang, J.J.

2014, XXIII, 499 p. 335 illus., Hardcover

ISBN: 978-94-007-6741-6