

Chapter 2

Video Coding Basic Principle

This chapter gives an overview of basic video coding principles. It consists of five parts. The first part provides the concept of color spaces and the conversion among typical color spaces. In the second part, we describe the typical video formats used in video coding. The third part introduces the basic coding principle and the component of coding tools. The fourth part talks about the quality assessment of the videos briefly, and the last part concludes this chapter.

2.1 Color Spaces

2.1.1 Color Perception

As we know, light is characterized by its wavelength (or frequency) and its intensity, and color is a visual perception of the light arriving at the photoreceptor cells in the retina of human eyes. The ability of the human eyes to distinguish colors is due to the varying sensitivity of different cells to the light of different wavelengths, and there are two kinds of photoreceptor cells in the human eyes, called rods and cones, as shown in Fig. 2.1. Rods are extremely sensitive to the light intensity but insensitive to the colors, while cones are sensitive to the colors and insensitive to the light intensity. At very low light levels, visual experience solely depends on the rods. For example, we cannot recognize the colors correctly in the dark rooms, because only one type of photoreceptor cell is active. For color perception, the retina contains three types of cones. As shown in Fig. 2.2, they can sense light with the spectral sensitivity peaks in short (S, 420–440 nm), middle (M, 530–540 nm), and long (L, 560–580 nm) wavelengths corresponding to blue, green, and red light respectively, also called as blue, green, and red cones respectively. These three kinds of cones comprise a trichromatic color vision system. In the trichromatic color vision system, any color perceived by the human eyes is a weighted sum of stimulus from the three

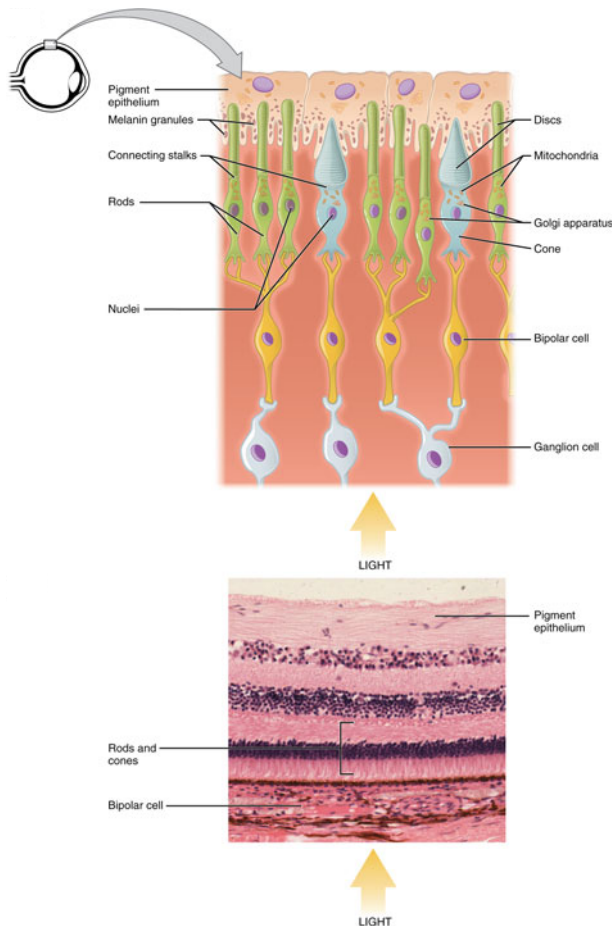


Fig. 2.1 Rods and cones. Attribution: By OpenStax College [CC-BY-3.0 (<http://creativecommons.org/licenses/by/3.0>)] via Wikimedia Commons

types of cones, denoted with three stimulus parameters S , M , and L , which can be indicated using a 3D space, called LMS color space.

The essence of color space is to map the physical color to an objective description in terms of tristimulus values by a trichromatic color model. Instead of using the cone spectral sensitivities defined by LMS color space, the tristimulus values can be conceptually viewed as amounts of three primary colors. Many color spaces have been developed based on the color matching experiments, such as the well-known RGB color space, which will be detailed in Sect. 2.1.2.

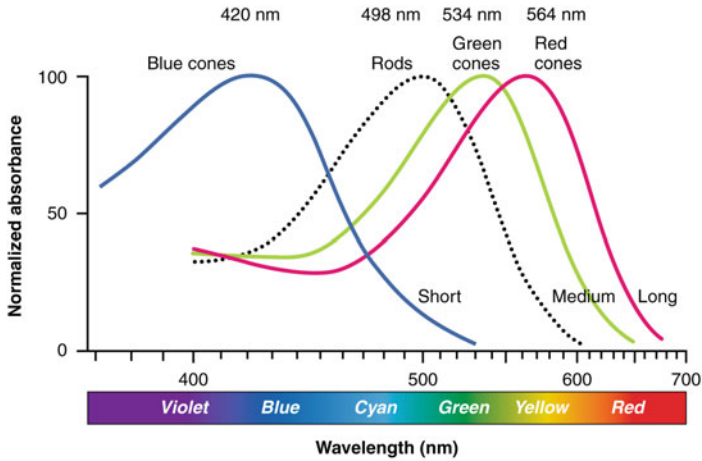


Fig. 2.2 Color sensitivity. Attribution: By OpenStax College [CC-BY-3.0 (<http://creativecommons.org/licenses/by/3.0>)] via Wikimedia Commons

2.1.2 RGB, XYZ, and YUV Color Spaces

RGB color space is based on an additive RGB color model, which describes how to mix red, green, and blue three colors to produce a given color (Fig. 2.3). As the RGB color model is simple for implementation, it has been widely used for sensing, representation, and display of images in electronic systems, such as televisions and computers. However, the RGB color model is device dependent, and different manufacturers may define different RGB color spaces, such as sRGB, created cooperatively by HP and Microsoft, and Adobe RGB, etc. So color conversion among different devices is necessary and one unified color space is needed for reference. Moreover, the additive RGB color model cannot encompass all the colors perceived by the human eyes. The reason is that the spectral sensitivity curves of the three types of cone cells overlap and the perceived light will not stimulate only one type of cone cell. Thus for pure spectral colors, at least one of the three primaries values would be negative in an additive color space, e.g., RGB color space, to match the corresponding true LMS tristimulus values. To avoid these negative RGB values, the International Commission on Illumination (CIE), which is a professional organization working on the science, technology and art in the fields of light and lighting, defined an “imaginary” primary colors-based color space in 1931, also called CIE 1931 XYZ color space. CIE 1931 XYZ color space encompasses all the colors perceived by the human eyes and is usually used as a reference for other color spaces. XYZ is analogous, but not equal to the LMS cone responses of the human eye. They are not true colors and can be divided into two parts, luminance or brightness (Y) and chromaticity (X, Z). Z is quasi-equal to blue stimulation, or the S cone response, and X is nonnegative as a linear combination of cone response curves. Defining Y as luminance has the useful result that for any given Y value, the XZ plane will contain

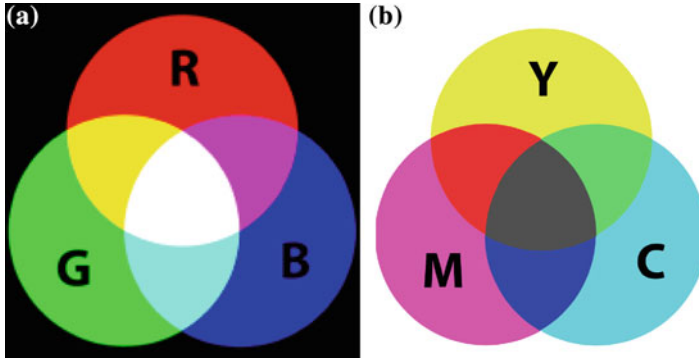


Fig. 2.3 RGB additive color model and CMYK subtractive color model. **a** Attribution: “1416 Color Sensitivity” by OpenStax College—Anatomy and Physiology, Connexions Web site. <http://cnx.org/content/col11496/1.6/>, Jun 19, 2013; **b** Attribution: “SubtractiveColor” by Original uploader was SharkD at en.wikipedia Later version uploaded by Jacobolus, Dacium at en.wikipedia.—Transferred from en.wikipedia. Licensed under Public domain via Wikimedia Commons

all possible chromaticities at that luminance. Figure 2.4 shows the color gamut of sRGB, Adobe RGB and CIE xyY color space. CIE xyY color space is a variation of CIE XYZ color space, where x and y are the normalized values with three tristimulus values X , Y , and Z :

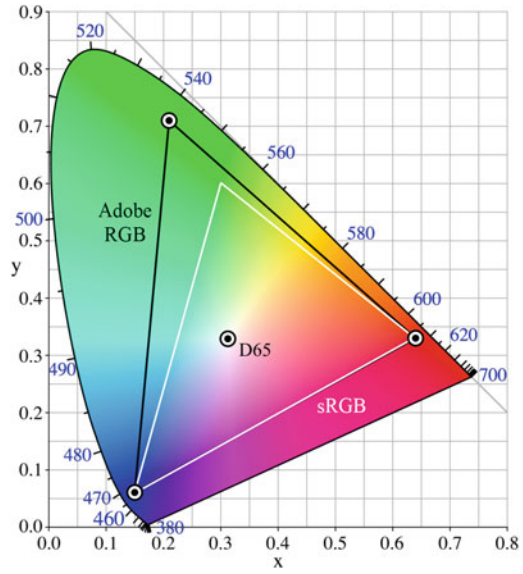
$$x = \frac{X}{X + Y + Z} \quad (2.1)$$

$$y = \frac{Y}{X + Y + Z} \quad (2.2)$$

Besides the above RGB and XYZ color spaces, many other color spaces have been developed for different kinds of applications, e.g., CMYK (cyan, magenta, yellow, and black), HSV (hue, saturation, and value), HSL (hue, saturation, and lightness), CIE Lab, YUV, YIQ, and YCbCr, etc. CMYK is usually used in color printing, which is a subtractive color mixing model and describes what kind of inks need to be applied so the light reflected from the substrate and through the inks produces a given color. RGB and CMYK are oriented to the hardware devices, while HSV and HSL are oriented to the users, which is more intuitive for the users to do color adjustment. CIE Lab is a color-opponent spaces, where L denotes lightness. a and b are the opponent color. CIE Lab is developed for measuring the perceptually uniform color difference, which means that a change of the same amount in a color value should produce a change of about the same visual importance.

In the actual video applications, as the human eyes are sensitive to the luminance and insensitive to the colors, the YUV color model is usually used instead of RGB for the bandwidth reduction, where Y is the luma component, and UV are the chroma component which can be downsampled for data reduction. YUV is

Fig. 2.4 sRGB, Adobe RGB color and CIE xy chromaticity diagram. Attribution: “CIExy1931 AdobeRGB versus sRGB” by Mbearnstein37—own work. Licensed under Creative Commons Attribution-Share Alike 3.0 via Wikimedia Commons



historically proposed for black-and-white/color TV compatible transmission. Y is the existing luminance signal for black-and-white TV broadcasting, and UV signal are added for color information. YIQ is similar to YUV, and the Y component is same in the two color spaces. I and Q represent the chrominance information and can be thought of as a new coordinate system rotated from UV coordinate with 33° . YCbCr is also similar to YUV and YIQ, but it is usually used for digital video. More details about digital video can be found in Sect. 2.2. The above color spaces can be converted from and to each other. Here we list frequently used color spaces conversion in video coding for reference, including YUV and YCbCr converting from or to RGB. YUV and RGB conversion is shown as follows:

$$\begin{bmatrix} Y \\ U \\ V \end{bmatrix} = \begin{bmatrix} 0.299 & 0.587 & 0.114 \\ -0.147 & -0.289 & 0.436 \\ 0.615 & -0.515 & -0.100 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} \quad (2.3)$$

$$\begin{bmatrix} R \\ G \\ B \end{bmatrix} = \begin{bmatrix} 1 & 0 & 1.140 \\ 1 & -0.395 & -0.581 \\ 1 & 2.032 & 0 \end{bmatrix} \begin{bmatrix} Y \\ U \\ V \end{bmatrix} \quad (2.4)$$

YCbCr and RGB conversion defined BT. 601 (BT601) (BT601-5 1995), which is used for digital standard definition TV broadcasting.

$$\begin{bmatrix} Y \\ Cb \\ Cr \end{bmatrix} = \begin{bmatrix} 0.299 & 0.587 & 0.114 \\ -0.169 & -0.331 & 0.500 \\ 0.500 & -0.419 & -0.081 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} \quad (2.5)$$

$$\begin{bmatrix} R \\ G \\ B \end{bmatrix} = \begin{bmatrix} 1 & 0 & 1.403 \\ 1 & -0.344 & -0.714 \\ 1 & 1.773 & 0 \end{bmatrix} \begin{bmatrix} Y \\ Cb \\ Cr \end{bmatrix} \quad (2.6)$$

Here, Y, R, G and B range in [0, 1]. Cb and Cr range in [-0.5, 0.5]. For Y, Cb, Cr, R, G, B ranging in [0, 255], the conversion is done as follows,

$$\begin{bmatrix} Y \\ Cb \\ Cr \end{bmatrix} = \begin{bmatrix} 0.299 & 0.587 & 0.114 \\ -0.169 & -0.331 & 0.500 \\ 0.500 & -0.419 & -0.081 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} + \begin{bmatrix} 0 \\ 128 \\ 128 \end{bmatrix} \quad (2.7)$$

$$\begin{bmatrix} R \\ G \\ B \end{bmatrix} = \begin{bmatrix} 1 & 0 & 1.403 \\ 1 & -0.344 & -0.714 \\ 1 & 1.773 & 0 \end{bmatrix} \begin{bmatrix} Y \\ Cb - 128 \\ Cr - 128 \end{bmatrix} \quad (2.8)$$

2.2 Video Formats

2.2.1 Analog Video and Digital Video

We first introduce two important concepts in video format: analog video and digital video. Analog video is the video transmitted with consecutive analog signal. The early video applications are all analog video-based, such as analog TV broadcasting, analog VCR. In the analog video, the luma and color components can be combined into one channel and transmitted, called composite video, and they can also be carried in two separate channels (luma Y and chroma UV) called S-video, or in three channels (luma Y, chroma U and chroma V) called component video. Composite video has the lowest transmission bandwidth, but it may have color crosstalk artifacts. Whereas component video has the best quality but with the highest bandwidth, S-video is a tradeoff between the quality and the bandwidth. In the actual analog TV broadcasting applications, composite video is widely used and a series of standards has been defined, such as NTSC (National Television System Committee), PAL (Phase Alternating Line) and SECAM (Sequentiel Couleur A Memoire). NTSC standard was widely used in most Americas, Japan, Korea, and some Pacific island nations and territories. PAL was developed to reduce the color artifacts aroused by phase distortion and used by China, India, etc. SECAM was developed by France and used by European countries.

Table 2.1 shows the major parameters of video formats defined by NTSC, PAL and SECAM. Lines/Fields denotes how many scan lines in an image and how many images are displayed in one second. The concept of field originates the interlaced video transmission, which is a technique for doubling the perceived frame rate of a

Table 2.1 NTSC/PAL/SECAM video formats

	NTSC	PAL	SECAM
Lines/fields	525/60	625/50	625/50
Horizontal frequency	15.734 kHz	15.625 kHz	15.625 kHz
Vertical frequency	59.939 Hz	50 Hz	50 Hz
Color subcarrier frequency	3.58 MHz	4.43 MHz	4.25 MHz
Video bandwidth	4.2 MHz	5.0 MHz	5.0 MHz

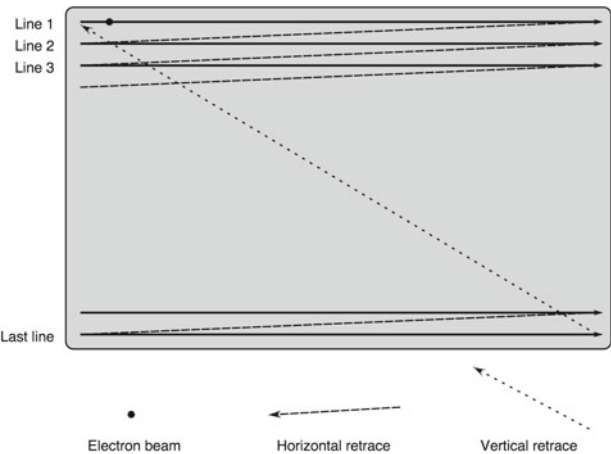


Fig. 2.5 Interlace video scan

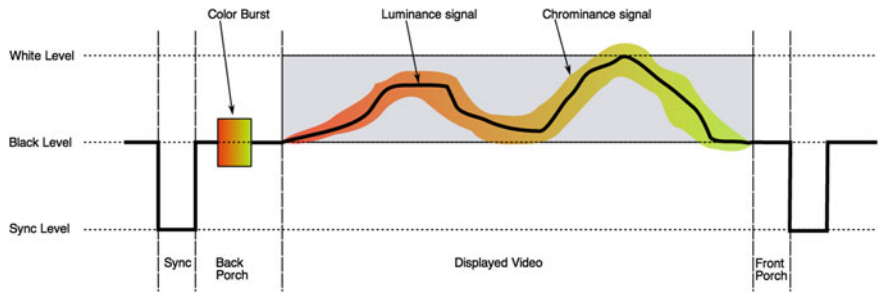


Fig. 2.6 Illustration of timing signal of scan line

video display without consuming extra bandwidth. An example of interlaced video is shown in Fig. 2.5, where the line 1, 3, . . . , will be scanned, transmitted and displayed first, then line 2, 4, . . . , are scanned in the second round. Between the two scan lines, there is an interval called horizontal retrace, which consists of front porch, sync, and back porch, as shown in Fig. 2.6. The front porch is between the end of each transmitted line of picture and the leading edge of the next line sync pulse, which

purposes to allow voltage levels to stabilize and prevent interference between picture lines. The back porch is the portion of a scan line between the rising edge of the horizontal sync pulse and the start of active video. It is used to restore the black level reference in analog video. In a similar manner, after scanning one field, there is a vertical retrace before starting next field to synchronize fields. So for NTSC, 525/60 means 525 lines per frame and 60 fields per second (one frame consists of two fields), which is also called as 525*i* (*i* is for interlace, and use *p* for progressive).

For the parameter of horizontal frequency, it is calculated as the line frequency (number of lines per second) times frame frequency. For instance, the horizontal frequency of PAL 15.625 kHz is equal to 625 lines \times 50 fields/2. The color carrier frequency is a result of 283.75 clock cycles per line plus a 25 Hz offset to avoid interferences, e.g., the color subcarrier frequency 4.43 MHz, which is calculated from $283.75 \times 15.625 \text{ kHz} + 25 \text{ Hz}$.

Along with the coming of digitalization era, analog video signal is also digitalized into digital video. Compared to the analog video, digital video is more convenient for processing and more robust for transmission noise, and has replaced analog video almost everywhere. Instead of continuous scan lines, each picture of the video consists of a raster of discrete samples, called pixels. Assuming the picture has a width of W pixels and a height of H pixels, we say the video has the spatial resolution of $W \times H$. The frame rate of the video is also called as temporal resolution. If each pixel has the color depth 24 bits (8 bits for each color component of RGB), for an hour of 640×480 ($W \times H$) video with frame rate of 25 frames/s, the video size would be up to $640 \times 480 \times 24 \times 25 \times 3,600/8 = 82.8$ Gbytes, and the bitrate is up to 184.25 Mbits/s. It can be seen that after digitalization the high volume of video data is challenging for storage and transmission of digital video. Thus high efficiency video compression is necessary, and how to achieve efficient coding becomes very important for the real applications, which is also what we will talk about in this book.

As analog video, many digital video formats have been defined for various applications. Table 2.2 shows the typical video formats used in video coding. CIF is the abbreviation of Common Intermediate Format and QCIF is a quarter of CIF, which are usually used in early video teleconferencing. SIF is known as Source Input Format defined in MPEG-1, which is used in VCD. In the Table 2.2, the listed frame/field rate value is not fixed but variable for real applications. And for QVGA and VGA kinds of video format defined by computer industry, they can be displayed at any refresh rate the computer can support.

In digital TV broadcasting, based on the analog TV systems, e.g., NTSC, PAL, and SECAM, series of digital television broadcasting standards have been established and used by different countries in the world, including DVB (Digital Video Broadcasting), ATSC (Advanced Television Systems Committee), ISDB (Integrated Services Digital Broadcasting) and DTMB (Digital Terrestrial Multimedia Broadcasting), etc. ATSC is used in US, Canada, South Korea, etc. ISDB is used in Japan and most area of South America. DTMB is developed by China and it is also used in several countries outside China. Most of the other countries use DVB standard. Now the digital TV broadcasting is very common. HDTV (High-definition Television)

Table 2.2 Digital video formats

Name		Spatial resolution	Typical frame/field rate
QCIF (Quarter Common Intermediate Format)		176×144	15, 30
SIF (Source Input Format)		352×240	30
		352×288	25
CIF (Common Intermediate Format)		352×288	30
4SIF		704×480	30
4CIF		704×576	30
SD (Standard Definition)	480i/p	704×480 , 720×480	24, 30, 60i
	576i/p	704×576 , 720×576	25, 50i
ED (Enhanced Definition)	480p	640×480 , 704×480 , 720×480	60
	576p	704×576 , 720×576	50
HD (High Definition)	720p	$1,280 \times 720$	24, 25, 30, 50, 60
	1080i/p	$1,920 \times 1,080$	24, 25, 30, 60, 50i, 60i
UHD (Ultra High Definition)		$3,840 \times 2,160$, $7,680 \times 4,320$	24, 25, 50, 60, 120
QVGA (Quarter Video Graphics Array)		320×240	
WQVGA (Wide Quarter Video Graphics Array)		400×240	
VGA (Video Graphics Array)		640×480	
WVGA (Wide Video Graphics Array)		800×400	
SVGA (Super Video Graphics Array)		800×600	
XGA (Extended Video Graphics Array)		$1,024 \times 768$	
SXGA		$1,280 \times 1,024$	

is replacing SDTV (Standard-definition Television), and even UHDTV (Ultra High Definition Television) maybe become part to the life in the near future.

2.2.2 YCbCr Sampling Formats

As said in Sect. 2.1.2, the human eyes are more sensitive to luminance than to colors. In digital video transmission, using low sampling precision for chrominance is an efficient way to reduce the bandwidth. In video coding, usually YCbCr color space is used, and the typical sampling formats include 4:0:0, 4:2:0, 4:2:2, 4:4:4.

For 4:0:0 format, actually only luminance signal is sampled, which is used for black/white systems. 4:2:0 format is used widely in digital TV broadcasting. In this format the Cb and Cr matrices shall be half the size of the Y matrix in both horizontal

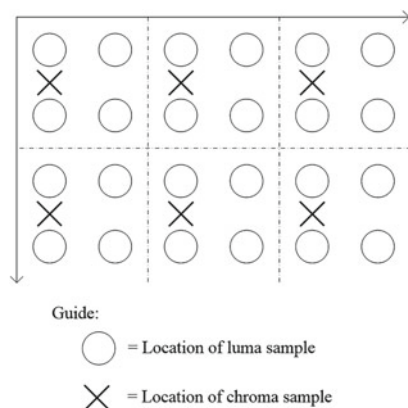


Fig. 2.7 Vertical and horizontal locations of 4:2:0 luma and chroma samples in a picture

and vertical dimensions, and the Y matrix shall have an even number of lines and samples, as shown in Fig. 2.7. 4:2:2 and 4:4:4 formats are usually used for high-end display devices or postproduction in the studio (Fig. 2.8).

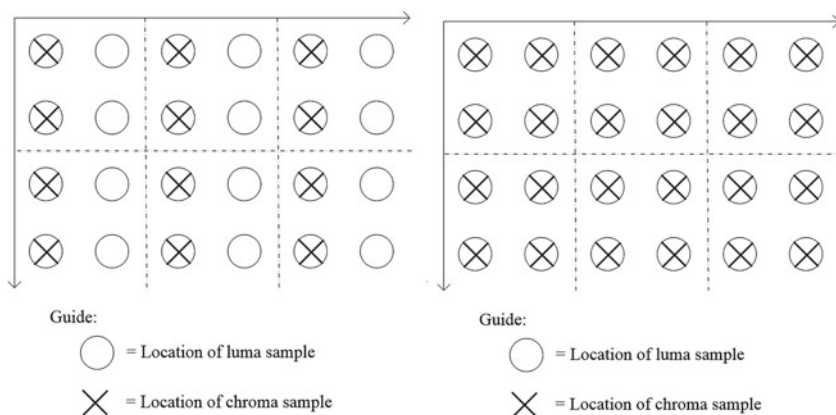


Fig. 2.8 Vertical and horizontal locations of 4:2:2 (left) and 4:4:4 (right) luma and chroma samples in a picture

Besides the sampling rate, the sampling precision of each component of YUV is also an important parameter, called bit depth or color depth, which decides the number of colors represented by YUV. In the earlier systems, usually very low color depth is used due to the hardware limitation, e.g., 4-bit color or 8-bit color (use 4 or 8 bits to represent all the colors directly, not for each color component). Nowadays, 24-bit color (8 bits for each color component) is mainstream, but higher bit depth

would be inescapable. In the industry, HDMI (High-Definition Multimedia Interface) has defined 30, 36, and up to 48-bit color for more vivid color display.

2.3 Video Coding Tools

As referred in Sect. 2.2.1, after digitalization, the data size of digital video increases significantly. However, there are a lot of redundancies in the digitalized video signal. In detail, there exists great correlation among the spatial neighboring pixels and the temporal successive frames. These redundancies can be categorized into spatial redundancy, temporal redundancy, statistical redundancy, and set redundancy. In brief, spatial redundancy means correlation among the neighboring pixels in the picture. And temporal redundancy denotes the correlation between successive pictures. For statistical redundancy, it denotes not only the statistical distribution of the sampled pixels but also the entropy code to represent the video signal. The set redundancy means common information found in more than one image or videos in the set of similar images and videos. The substance of video coding is to reduce these redundancies existing in the video signal, and according to the characteristics of redundancy, many coding tools have been developed for video coding and can be categorized into prediction coding, transform coding, entropy coding, in-loop filter, etc. Figure 2.9 illustrates the overall timeline for the image and video compression techniques. This section will give a brief overview of the following subsections.

2.3.1 Prediction Coding

The basic idea of prediction coding is to transmit a differential signal between the original signal and a prediction for the original signal, instead of the original signal.

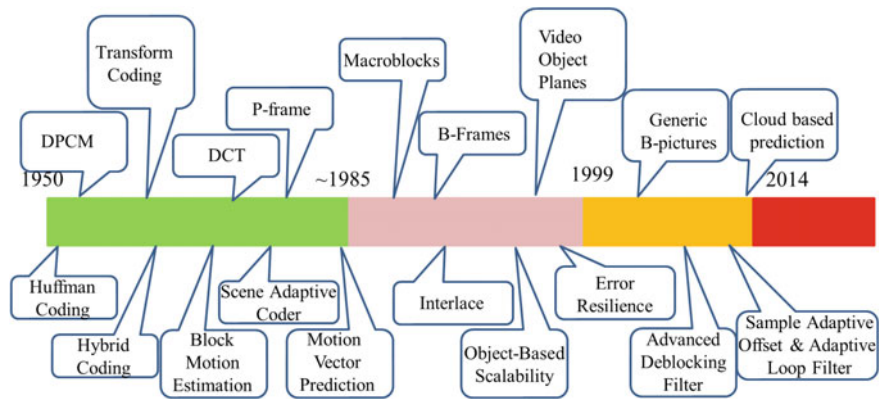


Fig. 2.9 Development of coding tools

The differential signal is also called residual signal and at the receiver side the original signal can be reconstructed by adding the residual and the prediction. Compared to the original signal, the residual signal has lower correlation. Prediction coding is an efficient tool to reduce the spatial, temporal, and set redundancy existing within or among the video signal(s), and many prediction coding tools have been proposed.

The earliest using of prediction coding is pixel-based DPCM (differential pulse coding modulation) coding, where the difference between neighboring two pixels are quantized and transmitted (Cutler 1950). For video prediction, Harrison (1952) proposed the first representative intra-prediction method, which takes the linear combination of reconstructed pixels as the prediction of the current pixel. The modified algorithm, named LOCO-I algorithm (Weinberger et al. 2000), has been adopted in JPEC-LS image compression standard. Afterwards, the AC/DC intra prediction in transform domain (Grgić et al. 1997) and the directional intra prediction in spatial domain (Bjontegaard 1998) have been proposed, and the latter becomes the prevalent prediction method in video coding field. Many popular video coding standards adopt the directional intra-prediction, e.g., AVC/H.264, HEVC/H.265 and AVS.

Later, the unit of prediction is extended from pixel to picture. In Seyler (1962), a picture-based difference coding method was proposed, where only the difference between two pictures is transmitted, and the data redundancy was reduced significantly. Rocca first proposed block-based motion estimation in Rocca (1969), in which an arbitrary-shaped block-based motion-compensated technique was proposed. The basic idea of Rocca's method is to model the scene as a set of constant-brightness zones denoted with arbitrary-shaped block. These zones would move from one frame to the next tracked by motion vectors, and difference values were transmitted for picture reconstruction. Besides these methods, motion-compensating prediction was further improved by employing the long-term statistical dependencies in coded video sequence instead of only the immediately preceding frame used for prediction. Wiegand et al. (1997) proposed a long-term memory scheme that used up to 50 previously decoded frames to determine the best motion vector. In addition, Puri et al. (1990) first proposed the B picture concept, which interpolates any skipped frame taking into account the movement between the two "end" frames, i.e., the forward and backward frames. It can achieve higher compression ratio by more effectively exploiting the correlation between reference pictures and current B picture, especially for coping with occlusion, uncovering problem caused by zooming, nonlinear motion, and so on. The B picture is further generalized by linearly combining motion-compensated signals regardless of the reference picture selection, which is referred to as multihypothesis motion-compensated prediction (Flierl and Girod 2003).

Besides reducing the redundancies within image and video data, the compression performance is further improved by reducing the set redundancies among similar images and videos. Karadimitriou et al. first proposed the set redundancy concept and proposed a series set of similar image compression methods, e.g., Min-Max differential (MMD) method (Karadimitriou and Tyler 1997) and centroid method (Karadimitriou and Tyler 1998). The centroid method generates one central image by averaging the pixel values in the same position among all the images, then the

average image and the difference images between central and non-central images are compressed individually. Yue et al. (2012) proposed to only compress image thumbnail and image local feature descriptor, and reconstruct high quality image with similar image patches retrieved from cloud. Extending it to video compression, Wang et al. (2014) imitated the multi-view coding method to jointly compress several near-duplicate videos by referring the video itself or other coded similar videos.

2.3.2 Transform Coding

Besides the prediction coding, transform coding can reduce the correlation existing in the prediction residual signal through transforming the signal from spatial domain to the frequency domain by orthogonal transform. In 1965, Enomoto and Shibata (1965) first proposed a practical video compression method with one dimensional Hadamard transform. In 1968, Andrews and Pratt (1968) extended the Enomoto's method to two dimensional transform with block-based Fourier transform. However, the Fourier transform has a relative high computational complexity, which is difficult to apply in practical video coding system. In 1973, Chen (1973) proposed Slant transform and the corresponding fast computational algorithm, which has higher energy compaction property than Fourier transform. In 1974, Ahmed et al. (1974) proposed the Discrete Cosine Transform (DCT), which has much lower computational complexity, and is more effective in decorrelation and energy concentration. In 1979, Netravali and Stuller (1979) proposed motion compensation transform framework, which is well known as hybrid prediction/transform coder nowadays and called first generation coding methods usually, and has been widely used in video coding standards since H.261.

In AVC/H.264, integer transform is developed to do DCT-like transform with simple integer arithmetic (Bjontegaard 1997; Malvar et al. 2003), which can also avoid the inverse-transform mismatches. Considering the efficiency of different transform block size, the large transforms can provide a better energy compaction and a better preservation of detail than small transforms but larger transforms introduce more ringing artifacts caused by quantization than small transforms. The adaptive block-size transform (ABT) (Wien 2003) is proposed in AVC/H.264 to improve the coding efficiency, which applies the same transform size as the prediction block size instead of only 4×4 transform. In the development of HEVC/H.265, more adaptive transform schemes were studied, e.g., mode-dependent directional transform (MDDT) (Ye and Karczewicz 2008), rate distortion optimized transform (RDOT) (Zhao et al. 2012). Especially, according to the statistical characteristics of intra prediction residual, Discrete Sine Transform (DST) was proposed for more efficient coding (Saxena and Fernandes 2011).

2.3.3 Entropy Coding

The early image coding methods usually achieve compression by directly exploiting the spatial statistical redundancy in the image, such as Huffman coding (Huffman et al. 1952) and Golomb code (Golomb 1966). After transform coding was invented, considering the coefficients distribution, Tescher and Cox (1976) proposed the famous zig-zag scan which transforms the two dimensional DCT coefficients into one dimensional array. To improve coding efficiency further, MPEG-2/4 use different VLC tables for intra- and inter-prediction residual blocks (MPEG2 1994; MPEG4 1999), and H.263 also adds optional advanced intra coding mode and alternative inter VLC mode to gain some adaptation. In AVC/H.264, a context-based adaptive variable length coding (CAVLC) (Bjontegaard 2002; Au 2002) designed for 4×4 DCT is adopted, which obtains higher coding efficiency further by using multiple contexts, each associated with a VLC table, to adapt to local statistical variations of DCT coefficients.

Compared to the VLC code assigning at least one bit to a symbol, arithmetic coding can represent one symbol with less than one bit in average when the probability of the symbol is higher than 50 % by assigning a codeword to the input symbol sequences instead of each symbol. Although the concept of arithmetic coding has been proposed for a long time, the first practical entropy coding scheme is proposed by Rissanen and Langdon (1979) until 1979. And the first hybrid block-based video coding schemes that incorporate an adaptive binary arithmetic coder capable of adapting the model probabilities to the existing symbol statistics was presented in Gonzales (1989). H.263 is the first block-based hybrid video coding standard that adopts arithmetic entropy coder. However, the arithmetic coder in H.263 encodes the same syntax elements as the VLC method, which makes it difficult to represent symbol with a non-integer length. A real successful arithmetic coding scheme is proposed in AVC/H.264, context-based adaptive binary arithmetic coding (CABAC) (Marpe et al. 2003). In addition, the multiple probability models are assumed to be static, which cannot adapt to local variation of symbols. CABAC does not encode the syntax elements directly, but it only encodes each bin of every binarized syntax elements. The probability model is selected according to previous encoded syntax elements or bins, which is also updated with the value of the encoded symbol. Compared with CAVLC of AVC/H.264, CABAC achieves about 9–14 % bitrate saving. In the latest HEVC/H.265 standard, more efficient CABAC was designed by reducing the Context dependence.

2.3.4 In-Loop Filtering

Block-based coded images and videos usually suffer from annoying artifacts at low bit rates. In hybrid video coding framework, the independent coarse quantization of every block is the main cause of compression artifacts, e.g., blocking artifact and

ringing artifact. In addition, the motion compensated blocks generated by copying interpolated pixel data from different locations of possibly different reference frames may also incur artifacts. The in-loop filtering cannot only improve the visual quality of the current frame, but also significantly improve the coding efficiency by providing high quality reference for subsequent coding frames. Although it is a useful coding tool, it also brings high complexity both in computation and hardware implementation. Therefore, until 1998, an in-loop filter (named as deblocking filter) was first standardized in video coding, H.263v2 Annex J (H263 1998). It was also extensively debated during the development of the AVC/H.264 standard. Although it was finally standardized in AVC/H.264 after a tremendous effort in speed optimization of the filtering algorithm, the filter also accounts for about one-third of the computational complexity of a decoder, which requires lots of conditional processing on the block edge and sample levels.

Thanks to the improvement of computing capability, some more complex in-loop filters can be integrated into video coding systems. In the development of HEVC/H.265 and AVS2, two in-loop filters, Sample Adaptive Offset (SAO) (Fu et al. 2012) and Adaptive Loop Filter (ALF) (Tsai et al. 2013), are extensively discussed. The SAO reduces the compression artifacts by first classifying reconstructed samples into different categories, obtaining an offset for each category, and then adding the offset to each sample. Compared to SAO with only one offset for each sample, the ALF processes one sample with neighboring samples by a multiple taps filter, parameters of which are obtained by minimizing the distortion between the distorted reconstruction frame and the original frame. Many ALF related techniques are proposed during HEVC/H.265 development, e.g., Quadtree-based ALF (Chen et al. 2011) and LCU-based ALF (Tsai 2012).

2.4 Quality Measurement

Video quality measurement is an important issue in video applications, and it also plays an important role in the coding tools development. In general, video quality assessment methods can be classified into subjective and objective quality assessment two categories.

Subjective quality assessment can decide the final quality perceived by the human through a subjective test. There are enormous subjective quality assessment methods. In ITU-R BT.500-13 (2012), double-stimulus impairment scale (DSIS) method and the double-stimulus continuous quality-scale (DSCQS) method as well as alternative assessment methods such as single-stimulus (SS) methods, stimulus-comparison methods, single stimulus continuous quality evaluation (SSCQE) and simultaneous double stimulus for continuous evaluation (SDSCE) method are standardized for the quality assessment of television pictures. However, the subjective test usually costs many human and material resources, thus it cannot be used in real-time applications.

Objective assessment methods usually predict the visual quality by mathematical models which can be quantitatively calculated. PSNR (peak signal noise ratio) is

a widely used quality metric in video coding. But the problem is that sometimes PSNR may not reflect the visual quality well. As the perceptual quality is highly dependent on the human visual system, which is still a too complex problem to model accurately, the research on objective visual quality metric is a hot topic. Some well-known perceptual quality metrics are the Sarnoff JND (just noticeable difference) model, the NASA DVQ (Digital Video Quality) (Watson 1998) model, and SSIM (Wang et al. 2004), but their applications are also very limited.

2.5 Summary

In this chapter, we have introduced some basic knowledge about video coding, including color space, video format, video coding tools, and quality measurement. They are not independent but closely related to each other. Color space provides the representation of the captured image signal. The color space conversion from RGB to YUV with UV downsampled is also an efficient way of data reduction. As the input source of a video codec, it also affects the development of coding tools. Moreover, the quality metric is not only an issue of quality evaluation, but also closely related to the selection of coding tools.

References

- Ahmed N, Natarajan T, Rao KR (1974) Discrete cosine transform. *IEEE Trans Comput* 100(1): 90–93
- Andrews H, Pratt W (1968) Fourier transform coding of images. In: *Proceedings of Hawaii international conference system sciences*, pp 677–679
- Au J (2002) Complexity reduction of CAVLC: ISO/IEC MPEG ITU-T VCEG. JVT-D034
- Bjontegaard G (1997) Coding improvement by using 44 blocks for motion vectors and transform: ITU-T VCEG. Doc Q15-C-23
- Bjontegaard G (1998) Response to call for proposals for H.261. ITU-T/Study Group 16/Video Coding Experts Group, document Q15-F-11
- Bjontegaard LK G (2002) Context-adaptive VLC (CVLC) coding of coefficients: ISO/IEC MPEG ITU-T VCEG. JVT-C028
- BT500-13 IR (2012) Methodology for the subjective assessment of the quality of television pictures. ITU
- BT601-5 IR (1995) Studio encoding parameters of digital television for standard 4:3 and wide-screen 16:9 aspect ratios. ITU
- Chen WH (1973) Slant transform image coding. Technical report, DTIC Document
- Chen Q, Zheng Y, Yin P, Lu X, Solé J, Xu Q, Francois E, Wu D (2011) Classified quadtree-based adaptive loop filter. In: *2011 IEEE international conference on multimedia and expo (ICME)*. IEEE, pp 1–6
- Cutler CC (1950) Differential quantization of communication signals
- Enomoto H, Shibata K (1965) Features of Hadamard transformed television signal. In: *National conference IECE in Japan*, p 881
- Flierl M, Girod B (2003) Generalized b pictures and the draft H. 264/AVC video-compression standard. *IEEE Trans Circuits Syst Video Technol* 13(7):587–597

- Fu CM, Alshina E, Alshin A, Huang YW, Chen CY, Tsai CY, Hsu CW, Lei SM, Park JH, Han WJ (2012) Sample adaptive offset in the hevc standard. *IEEE Trans Circuits Syst Video Technol* 22(12):1755–1764
- Golomb S (1966) Run-length encodings. *IEEE Trans Inf Theory* 12(3):399–401
- Gonzales C (1989) DCT coding of motion sequences including arithmetic coder: ISO/IEC JCT1/SC2/WP8. 89/187
- Grgić M, Zovko-Cihlar B, Bauer S (1997) Coding of audio-visual objects. In: 39th international symposium electronics in Marine-ELMAR 97
- H263 (1998) Video coding for low bitrate communications: Version 2. ITU-T, ITU-T Recommendation H263
- Harrison C (1952) Experiments with linear prediction in television. *Bell Syst Tech J* 31(4):764–783
- Huffman DA et al (1952) A method for the construction of minimum redundancy codes. *Proc IRE* 40(9):1098–1101
- Karadimitriou K, Tyler JM (1997) Min-max compression methods for medical image databases. *ACM SIGMOD Rec* 26(1):47–52
- Karadimitriou K, Tyler JM (1998) The centroid method for compressing sets of similar images. *Pattern Recognit Lett* 19(7):585–593
- Malvar HS, Hallapuro A, Karczewicz M, Kerofsky L (2003) Low-complexity transform and quantization in H. 264/AVC. *IEEE Trans Circuits Syst Video Technol* 13(7):598–603
- Marpe D, Schwarz H, Wiegand T (2003) Context-based adaptive binary arithmetic coding in the H. 264/AVC video compression standard. *IEEE Trans Circuits Syst Video Technol* 13(7):620–636
- MPEG2 (1994) Generic coding of moving pictures and associated audio information c MPEG2 part 2. video ISO/IEC 13818-2
- MPEG4 (1999) Coding of audio-visual objects c part 2. visual ISO/IEC 14496-2 (MPEG-4 visual version 1)
- Netravali A, Stuller J (1979) Motion-compensated transform coding. *Bell Syst Tech J* 58(7):1703–1718
- Puri A, Aravind R, Haskell B, Leonardi R (1990) Video coding with motion-compensated interpolation for CD-ROM applications. *Signal Process: Image Commun* 2(2):127–144
- Rissanen J, Langdon GG Jr (1979) Arithmetic coding. *IBM J Res Dev* 23(2):149–162
- Rocca F (1969) Television bandwidth compression utilizing frame-to-frame correlation and movement compensation. In: Symposium on picture bandwidth compression
- Saxena A, Fernandes FC (2011) Mode dependent DCT/DST for intra prediction in block-based image/video coding. In: 2011 18th IEEE international conference on image processing (ICIP). IEEE, pp 1685–1688
- Seyler A (1962) The coding of visual signals to reduce channel-capacity requirements. *Proc IEE-Part C: Monogr* 109(16):676–684
- Tescher AG, Cox RV (1976) An adaptive transform coding algorithm. Technical report, DTIC Document
- Tsai C (2012) AHG6: Baseline options for ALF: Joint collaborative team on video coding (JCT-VC) of ISO/IEC MPEG and ITU-T VCEG. JCTVC-I0157
- Tsai CY, Chen CY, Yamakage T, Chong IS, Huang YW, Fu CM, Itoh T, Watanabe T, Chujoh T, Karczewicz M et al (2013) Adaptive loop filtering for video coding
- Wang Z, Bovik AC, Sheikh HR, Simoncelli EP (2004) Image quality assessment: from error visibility to structural similarity. *IEEE Trans Image Process* 13(4):600–612
- Wang H, Ma M, Jiang YG, Wei Z (2014) A framework of video coding for compressing near-duplicate videos. In: *MultiMedia modeling*. Springer, pp 518–528
- Watson AB (1998) Toward, a perceptual video-quality metric. In: *Photonics West'98 electronic imaging, international society for optics and photonics*, pp 139–147
- Weinberger MJ, Seroussi G, Sapiro G (2000) The LOCO-I lossless image compression algorithm: principles and standardization into JPEG-LS. *IEEE Trans Image Process* 9(8):1309–1324
- Wiegand T, Zhang X, Girod B (1997) Motion-compensating long-term memory prediction. In: *Proceedings of international conference on image processing, vol 2*. IEEE, pp 53–56

- Wien M (2003) Variable block-size transforms for H. 264/AVC. *IEEE Trans Circuits Syst Video Technol* 13(7):604–613
- Ye Y, Karczewicz M (2008) Improved h. 264 intra coding based on bi-directional intra prediction, directional transform, and adaptive coefficient scanning. In: 15th IEEE international conference on image processing, ICIP 2008. IEEE, pp 2116–2119
- Yue H, Sun X, Wu F, Yang J (2012) Sift-based image compression. In: 2012 IEEE international conference on multimedia and expo (ICME). IEEE, pp 473–478
- Zhao X, Zhang L, Ma S, Gao W (2012) Video coding with rate-distortion optimized transform. *IEEE Trans Circuits Syst Video Technol* 22(1):138–151

Advanced Video Coding Systems

Gao, W.; Ma, S.

2014, XIII, 239 p. 105 illus., 46 illus. in color., Hardcover

ISBN: 978-3-319-14242-5