5 Fundamental Concepts in Video

In this chapter, we introduce the principal notions needed to understand video. Digital video compression is explored separately, in Chaps. 10–12.

Here we consider the following aspects of video and how they impact multimedia applications:

- Analog video
- Digital video
- Video display interfaces
- 3D video.

Since video is created from a variety of sources, we begin with the signals themselves. Analog video is represented as a continuous (time-varying) signal, and the first part of this chapter discusses how it is created and measured. Digital video is represented as a sequence of digital images. Nowadays, it is omnipresent in many types of multimedia applications. Therefore, the second part of the chapter focuses on issues in digital video including HDTV, UHDTV, and 3D TV.

5.1 Analog Video

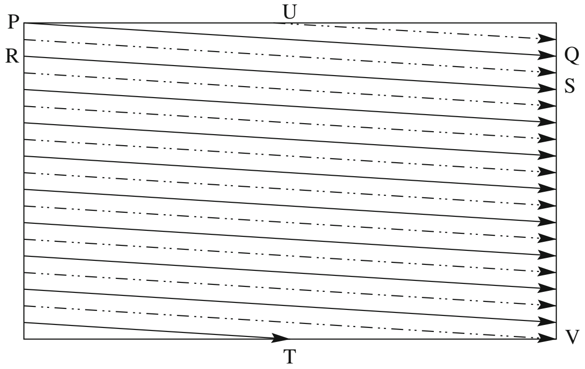
Up until last decade, most TV programs were sent and received as an analog signal. Once the electrical signal is received, we may assume that brightness is at least a monotonic function of voltage, if not necessarily linear, because of gamma correction (see Sect. 4.1.6).

An analog signal $f(t)$ samples a time-varying image. So-called *progressive* scanning traces through a complete picture (a frame) row-wise for each time interval. A high-resolution computer monitor typically uses a time interval of 1/72 s.

In TV and in some monitors and multimedia standards, another system, *interlaced* scanning, is used. Here, the odd-numbered lines are traced first, then the even-numbered lines. This results in “odd” and “even” *fields*—two fields make up one frame.

Z.-N. Li et al., *Fundamentals of Multimedia*,
Texts in Computer Science, DOI: 10.1007/978-3-319-05290-8_5,

Fig. 5.1 Interlaced raster scan



In fact, the odd lines (starting from 1) end up at the middle of a line at the end of the odd field, and the even scan starts at a half-way point. Figure 5.1 shows the scheme used. First the solid (odd) lines are traced—P to Q, then R to S, and so on, ending at T—then the even field starts at U and ends at V. The scan lines are not horizontal because a small voltage is applied, moving the electron beam down over time.

Interlacing was invented because, when standards were being defined, it was difficult to transmit the amount of information in a full frame quickly enough to avoid flicker. The double number of fields presented to the eye reduces perceived flicker.

Because of interlacing, the odd and even lines are displaced in time from each other. This is generally not noticeable except when fast action is taking place onscreen, when blurring may occur. For example, in the video in Fig. 5.2, the moving helicopter is blurred more than the still background.

Since it is sometimes necessary to change the frame rate, resize, or even produce stills from an interlaced source video, various schemes are used to *deinterlace* it. The simplest deinterlacing method consists of discarding one field and duplicating the scan lines of the other field, which results in the information in one field being lost completely. Other, more complicated methods retain information from both fields.

CRT (Cathode Ray Tube) displays are built like fluorescent lights and must flash 50–70 times per second to appear smooth. In Europe, this fact is conveniently tied to their 50 Hz electrical system, and they use video digitized at 25 frames per second (fps); in North America, the 60 Hz electric system dictates 30 fps.

The jump from Q to R and so on in Fig. 5.1 is called the *horizontal retrace*, during which the electronic beam in the CRT is blanked. The jump from T to U or V to P is called the *vertical retrace*.

Since voltage is one-dimensional—it is simply a signal that varies with time—how do we know when a new video line begins? That is, what part of an electrical signal tells us that we have to restart at the left side of the screen?

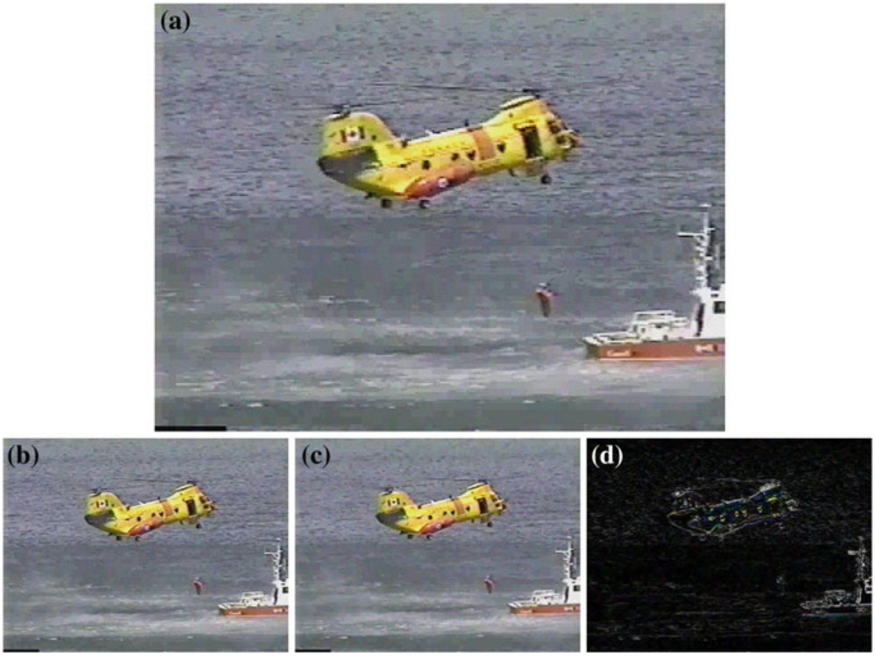
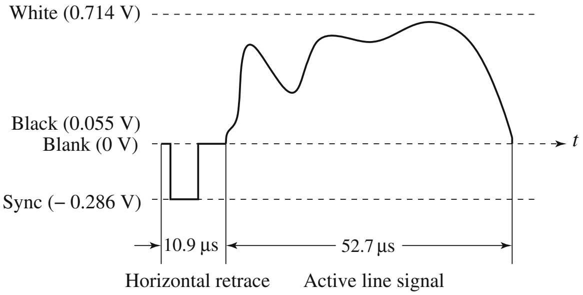


Fig. 5.2 Interlaced scan produces two fields for each frame: **a** the video frame; **b** Field 1; **c** Field 2; **d** difference of fields

Fig. 5.3 Electronic signal for one NTSC scan line



The solution used in analog video is a small voltage offset from zero to indicate black and another value, such as zero, to indicate the start of a line. Namely, we could use a "blacker-than-black" zero signal to indicate the beginning of a line.

Figure 5.3 shows a typical electronic signal for one scan line of NTSC composite video. 'White' has a peak value of 0.714 V; 'Black' is slightly above zero at 0.055 V; whereas Blank is at zero volts. As shown, the time duration for blanking pulses in the signal is used for synchronization as well, with the tip of the Sync signal at approximately −0.286 V. In fact, the problem of reliable synchronization is so important that special signals to control sync take up about 30 % of the signal!

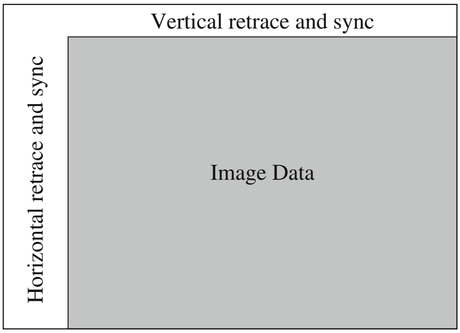


Fig. 5.4 Video raster, including retrace and sync data

The vertical retrace and sync ideas are similar to the horizontal one, except that they happen only once per field. Tekalp [1] presents a good discussion of the details of analog (and digital) video. The handbook [2] considers many fundamental problems in video processing in great depth.

5.1.1 NTSC Video

The NTSC TV standard is mostly used in North America and Japan. It uses a familiar 4:3 *aspect ratio* (i.e., the ratio of picture width to height) and 525 scan lines per frame at 30 fps.

More exactly, for historical reasons NTSC uses 29.97 fps—or, in other words, 33.37 ms per frame. NTSC follows the interlaced scanning system, and each frame is divided into two fields, with 262.5 lines/field. Thus the horizontal sweep frequency is $525 \times 29.97 \approx 15,734$ lines/s, so that each line is swept out in $1/15{,}734\,\text{s} \approx 63.6\,\mu\text{s}$. Since the horizontal retrace takes 10.9 μs, this leaves 52.7 μs for the active line signal, during which image data is displayed (see Fig. 5.3).

Figure 5.4 shows the effect of "vertical retrace and sync" and "horizontal retrace and sync" on the NTSC video raster. Blanking information is placed into 20 lines reserved for control information at the beginning of each field. Hence, the number of *active video lines* per frame is only 485. Similarly, almost 1/6 of the raster at the left side is blanked for horizontal retrace and sync. The nonblanking pixels are called *active pixels*.

Pixels often fall between scanlines. Therefore, even with noninterlaced scan, NTSC TV is capable of showing only about 340 (visually distinct) lines,—about 70 % of the 485 specified active lines. With interlaced scan, it could be as low as 50 %.

Image data is not encoded in the blanking regions, but other information can be placed there, such as V-chip information, stereo audio channel data, and subtitles in many languages.

NTSC video is an analog signal with no fixed horizontal resolution. Therefore, we must decide how many times to sample the signal for display. Each sample

Table 5.1 Samples per line for various analog video formats

Format	Samples per line
VHS	240
S-VHS	400–425
Beta-SP	500
Standard 8 mm	300
Hi-8 mm	425

corresponds to one pixel output. A *pixel clock* divides each horizontal line of video into samples. The higher the frequency of the pixel clock, the more samples per line.

Different video formats provide different numbers of samples per line, as listed in Table 5.1. Laser disks have about the same resolution as Hi-8. (In comparison, miniDV 1/4-inch tapes for digital video are 480 lines by 720 samples per line).

NTSC uses the YIQ color model. We employ the technique of *quadrature modulation* to combine (the spectrally overlapped part of) I (in-phase) and Q (quadrature) signals into a single chroma signal C [3, 1]:

$$C = I\cos(F_{sc}t) + Q\sin(F_{sc}t) \tag{5.1}$$

This modulated chroma signal is also known as the *color subcarrier*, whose magnitude is $\sqrt{I^2 + Q^2}$ and phase is $\tan^{-1}(Q/I)$. The frequency of C is $F_{sc} \approx 3.58$ MHz.

The I and Q signals are multiplied in the time domain by cosine and sine functions with the frequency F_{sc} [Eq. (5.1)]. This is equivalent to convolving their Fourier transforms in the frequency domain with two impulse functions at F_{sc} and $-F_{sc}$. As a result, a copy of I and Q frequency spectra are made which are centered at F_{sc} and $-F_{sc}$, respectively.[1]

The NTSC composite signal is a further composition of the luminance signal Y and the chroma signal, as defined below:

$$\text{composite} = Y + C = Y + I\cos(F_{sc}t) + Q\sin(F_{sc}t). \tag{5.2}$$

NTSC assigned a bandwidth of 4.2 MHz to Y but only 1.6 MHz to I and 0.6 MHz to Q, due to humans' insensitivity to color details (high-frequency color changes). As Fig. 5.5 shows, the picture carrier is at 1.25 MHz in the NTSC video channel, which has a total bandwidth of 6 MHz. The chroma signal is being "carried" by $F_{sc} \approx 3.58$ MHz toward the higher end of the channel and is thus centered at $1.25 + 3.58 = 4.83$ MHz. This greatly reduces the potential interference between the Y (luminance) and C (chrominance) signals, since the magnitudes of higher frequency components of Y are significantly smaller than their lower frequency counterparts.

Moreover, as Blinn [3] explains, great care is taken to interleave the discrete Y and C spectra so as to further reduce the interference between them. The "interleaving" is illustrated in Fig. 5.5, where the frequency components for Y (from the discrete

[1] Negative frequency ($-F_{sc}$) is a mathematical notion needed in the Fourier transform. In the physical spectrum, only positive frequency is used.

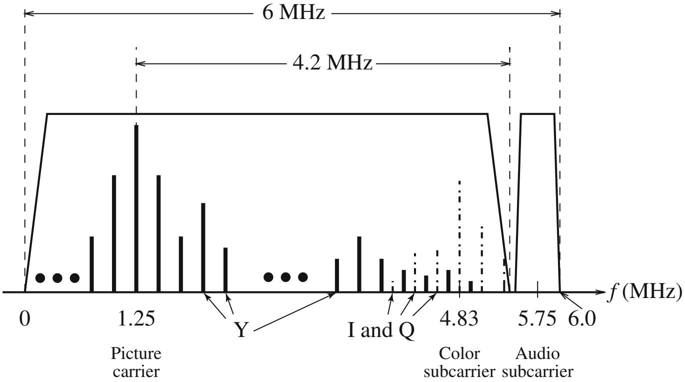


Fig. 5.5 Interleaving Y and C signals in the NTSC spectrum

Fourier transform) are shown as solid lines, and those for I and Q are shown as dashed lines. As a result, the 4.2 MHz band of Y is overlapped and interleaved with the 1.6 MHz to I and 0.6 MHz to Q.

The first step in decoding the composite signal at the receiver side is to separate Y and C. Generally, low-pass filters can be used to extract Y, which is located at the lower end of the channel. TV sets with higher quality also use comb filters [3] to exploit the fact that Y and C are interleaved.

After separation from Y, the chroma signal C can be demodulated to extract I and Q separately.

To extract I:

1. Multiply the signal C by $2\cos(F_{sc}t)$

$$\begin{aligned}C \cdot 2\cos(F_{sc}t) &= I \cdot 2\cos^2(F_{sc}t) + Q \cdot 2\sin(F_{sc}t)\cos(F_{sc}t)\\ &= I \cdot (1 + \cos(2F_{sc}t)) + Q \cdot 2\sin(F_{sc}t)\cos(F_{sc}t)\\ &= I + I \cdot \cos(2F_{sc}t) + Q \cdot \sin(2F_{sc}t).\end{aligned}$$

2. Apply a low-pass filter to obtain I and discard the two higher frequency ($2F_{sc}$) terms.

Similarly, extract Q by first multiplying C by $2\sin(F_{sc}t)$ and then applying low-pass filtering.

The NTSC bandwidth of 6 MHz is tight. Its audio subcarrier frequency is 4.5 MHz, which places the center of the audio band at $1.25 + 4.5 = 5.75$ MHz in the channel (Fig. 5.5). This would actually be a bit too close to the color subcarrier—a cause for potential interference between the audio and color signals. It was due largely to this reason that NTSC color TV slowed its frame rate to $30 \times 1,000/1,001 \approx 29.97$ fps [4]. As a result, the adopted NTSC color subcarrier frequency is slightly lowered, to

$$f_{sc} = 30 \times 1,000/1,001 \times 525 \times 227.5 \approx 3.579545 \text{ MHz}$$

where 227.5 is the number of color samples per scan line in NTSC broadcast TV.

Table 5.2 Comparison of analog broadcast TV systems

TV system	Frame rate (fps)	Number of scan lines	Total channel width (MHz)	Bandwidth allocation (MHz)		
				Y	I or U	Q or V
NTSC	29.97	525	6.0	4.2	1.6	0.6
PAL	25	625	8.0	5.5	1.8	1.8
SECAM	25	625	8.0	6.0	2.0	2.0

5.1.2 PAL Video

PAL (Phase Alternating Line) is a TV standard originally invented by German scientists. It uses 625 scan lines per frame, at 25 fps (or 40 ms/frame), with a 4:3 aspect ratio and interlaced fields. Its broadcast TV signals are also used in composite video. This important standard is widely used in Western Europe, China, India, and many other parts of the world. Because it has higher resolution than NTSC (625 vs. 525 scan lines), the visual quality of its pictures is generally better.

PAL uses the YUV color model with an 8 MHz channel, allocating a bandwidth of 5.5 MHz to Y and 1.8 MHz each to U and V. The color subcarrier frequency is $f_{sc} \approx 4.43$ MHz. To improve picture quality, chroma signals have alternate signs (e.g., $+U$ and $-U$) in successive scan lines; hence the name "Phase Alternating Line."[2] This facilitates the use of a (line-rate) comb filter at the receiver—the signals in consecutive lines are averaged so as to cancel the chroma signals (which always carry opposite signs) for separating Y and C and obtain high-quality Y signals.

5.1.3 SECAM Video

SECAM, which was invented by the French, is the third major broadcast TV standard. SECAM stands for *Systeme Electronique Couleur Avec Memoire*. SECAM also uses 625 scan lines per frame, at 25 fps, with a 4:3 aspect ratio and interlaced fields. The original design called for a higher number of scan lines (over 800), but the final version settled for 625.

SECAM and PAL are similar, differing slightly in their color coding scheme. In SECAM, U and V signals are modulated using separate color subcarriers at 4.25 MHz and 4.41 MHz, respectively. They are sent in alternate lines—that is, only one of the U or V signals will be sent on each scan line.

Table 5.2 gives a comparison of the three major analog broadcast TV systems.

[2] According to Blinn [3], NTSC selects a half integer (227.5) number of color samples for each scan line. Hence, its chroma signal also switches sign in successive scan lines.

5.2 Digital Video

The advantages of digital representation for video are many. It permits

- Storing video on digital devices or in memory, ready to be processed (noise removal, cut and paste, and so on) and integrated into various multimedia applications.
- Direct access, which makes nonlinear video editing simple.
- Repeated recording without degradation of image quality.
- Ease of encryption and better tolerance to channel noise.

In earlier Sony or Panasonic recorders, digital video was in the form of composite video. Modern digital video generally uses component video, although RGB signals are first converted into a certain type of color opponent space. The usual color space is YCbCr [5].

5.2.1 Chroma Subsampling

Since humans see color with much less spatial resolution than black and white, it makes sense to decimate the chrominance signal. Interesting but not necessarily informative names have arisen to label the different schemes used. To begin with, numbers are given stating how many pixel values, per four original pixels, are actually sent. Thus the chroma subsampling scheme "4:4:4" indicates that no chroma subsampling is used. Each pixel's Y, Cb, and Cr values are transmitted, four for each of Y, Cb, and Cr.

The scheme "4:2:2" indicates horizontal subsampling of the Cb and Cr signals by a factor of 2. That is, of four pixels horizontally labeled 0 to 3, all four Ys are sent, and every other Cb and Cr are sent, as ($Cb0$, $Y0$)($Cr0$, $Y1$)($Cb2$, $Y2$)($Cr2$, $Y3$)($Cb4$, $Y4$), and so on.

The scheme "4:1:1" subsamples horizontally by a factor of 4. The scheme "4:2:0" subsamples in both the horizontal and vertical dimensions by a factor of 2. Theoretically, an average chroma pixel is positioned between the rows and columns, as shown in Fig. 5.6. We can see that the scheme 4:2:0 is in fact another kind of 4:1:1 sampling, in the sense that we send 4, 1, and 1 values per 4 pixels. Therefore, the labeling scheme is not a very reliable mnemonic!

Scheme 4:2:0, along with others, is commonly used in JPEG and MPEG (see later chapters in Part II).

5.2.2 CCIR and ITU-R Standards for Digital Video

The CCIR is the *Consultative Committee for International Radio*. One of the most important standards it has produced is CCIR-601 for component digital video. This standard has since become standard ITU-R Rec. 601, an international standard for professional video applications. It is adopted by several digital video formats, including the popular DV video.

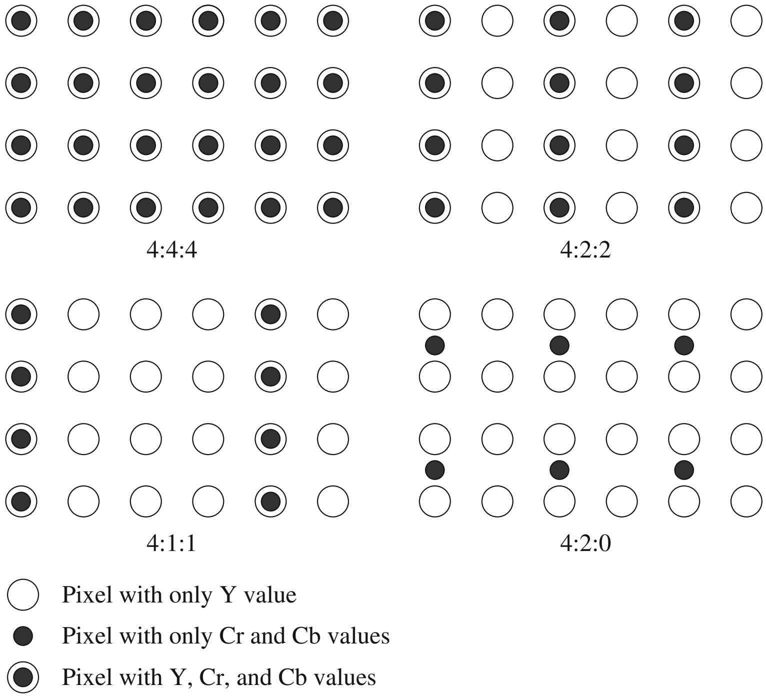

4:4:4
4:2:2
4:1:1
4:2:0
Pixel with only Y value
Pixel with only Cr and Cb values
Pixel with Y, Cr, and Cb values

Fig. 5.6 Chroma subsampling

The NTSC version has 525 scan lines, each having 858 pixels (with 720 of them visible, not in the blanking period). Because the NTSC version uses 4:2:2, each pixel can be represented with two bytes (8 bits for Y and 8 bits alternating between Cb and Cr). The Rec. 601 (NTSC) data rate (including blanking and sync but excluding audio) is thus approximately 216 Mbps (megabits per second):

$$525 \times 858 \times 30 \times 2 \text{ bytes} \times 8 \frac{\text{bits}}{\text{byte}} \approx 216 \text{ Mbps}$$

During blanking, digital video systems may make use of the extra data capacity to carry audio signals, translations into foreign languages, or error-correction information.

Table 5.3 shows some of the digital video specifications, all with an aspect ratio of 4:3. The Rec. 601 standard uses an interlaced scan, so each field has only half as much vertical resolution (e.g., 240 lines in NTSC).

CIF stands for *Common Intermediate Format*, specified by the International Telegraph and Telephone Consultative Committee (CCITT), now superseded by the International Telecommunication Union, which oversees both telecommunications (ITU-T) and radio frequency matters (ITU-R) under one United Nations body. The idea of CIF, which is about the same as VHS quality, is to specify a format for lower bitrate. CIF uses a progressive (noninterlaced) scan. QCIF stands for Quarter-CIF,

Table 5.3 ITU-R digital video specifications

	Rec. 601 525/60 NTSC	Rec. 601 625/50 PAL/SECAM	CIF	QCIF
Luminance resolution	720 × 480	720 × 576	352 × 288	176 × 144
Chrominance resolution	360 × 480	360 × 576	176 × 144	88 × 72
Color subsampling	4:2:2	4:2:2	4:2:0	4:2:0
Aspect ratio	4:3	4:3	4:3	4:3
Fields/sec	60	50	30	30
Interlaced	Yes	Yes	No	No

and is for even lower bitrate. All the CIF/QCIF resolutions are evenly divisible by 8, and all except 88 are divisible by 16; this is convenient for block-based video coding in H.261 and H.263, discussed in Chap. 10.

CIF is a compromise between NTSC and PAL, in that it adopts the NTSC frame rate and half the number of active lines in PAL. When played on existing TV sets, NTSC TV will first need to convert the number of lines, whereas PAL TV will require frame rate conversion.

5.2.3 High-Definition TV

The introduction of wide-screen movies brought the discovery that viewers seated near the screen enjoyed a level of participation (sensation of immersion) not experienced with conventional movies. Apparently the exposure to a greater field of view, especially the involvement of peripheral vision, contributes to the sense of "being there." The main thrust of High-Definition TV (HDTV) is not to increase the "definition" in each unit area, but rather to increase the visual field, especially its width.

First-generation HDTV was based on an analog technology developed by Sony and NHK in Japan in the late 1970s. HDTV successfully broadcasted the 1984 Los Angeles Olympic Games in Japan. MUltiple sub-Nyquist Sampling Encoding (MUSE) was an improved NHK HDTV with hybrid analog/digital technologies that was put in use in the 1990s. It has 1,125 scan lines, interlaced (60 fields per second), and a 16:9 aspect ratio. It uses satellite to broadcast—quite appropriate for Japan, which can be covered with one or two satellites. The Direct Broadcast Satellite (DBS) channels used have a bandwidth of 24 MHz.

In general, terrestrial broadcast, satellite broadcast, cable, and broadband networks are all feasible means for transmitting HDTV as well as conventional TV. Since uncompressed HDTV will easily demand more than 20 MHz bandwidth, which will not fit in the current 6 or 8 MHz channels, various compression techniques are being investigated. It is also anticipated that high-quality HDTV signals will be transmitted using more than one channel, even after compression.

Table 5.4 Advanced digital TV formats supported by ATSC

Number of active pixels per line	Number of active lines	Aspect ratio	Picture rate
1,920	1,080	16:9	60P 60I 30P 24P
1,280	720	16:9	60P 30P 24P
720	480	16:9 or 4:3	60P 60I 30P 24P
640	480	4:3	60P 60I 30P 24P

In 1987, the FCC decided that HDTV standards must be compatible with the existing NTSC standard and must be confined to the existing Very High Frequency (VHF) and Ultra High Frequency (UHF) bands. This prompted a number of proposals in North America by the end of 1988, all of them analog or mixed analog/digital.

In 1990, the FCC announced a different initiative—its preference for full-resolution HDTV. They decided that HDTV would be simultaneously broadcast with existing NTSC TV and eventually replace it. The development of digital HDTV immediately took off in North America.

Witnessing a boom of proposals for digital HDTV, the FCC made a key decision to go all digital in 1993. A "grand alliance" was formed that included four main proposals, by General Instruments, MIT, Zenith, and AT&T, and by Thomson, Philips, Sarnoff, and others. This eventually led to the formation of the Advanced Television Systems Committee (ATSC), which was responsible for the standard for TV broadcasting of HDTV. In 1995, the U.S. FCC Advisory Committee on Advanced Television Service recommended that the ATSC digital television standard be adopted.

Table 5.4 lists some of the standard supported video scanning formats. (For the 50 Hz systems, 60P becomes 50P, 30P becomes 25P, etc.) In the table, "I" means interlaced scan and "P" means progressive (noninterlaced) scan. The frame rates supported are both integer rates and the NTSC rates—that is, 60.00 or 59.94, 30.00 or 29.97, 24.00, or 23.98 fps.

For video, MPEG-2 was initially chosen as the compression standard. As will be seen in Chap. 11, it uses Main Level to High Level of the Main Profile of MPEG-2. For audio, AC-3 is the standard. It supports the so-called 5.1 channel Dolby surround sound—five surround channels plus a subwoofer channel. In 2008, ATSC was updated to adopt the H.264 video compression standard.

The salient difference between conventional TV and HDTV [4,6] is that the latter has a much wider aspect ratio of 16:9 instead of 4:3. (Actually, it works out to be exactly one-third wider than current TV.) Another feature of HDTV is its move toward progressive (noninterlaced) scan. The rationale is that interlacing introduces serrated edges to moving objects and flickers along horizontal edges.

Consumers with analog TV sets will still be able to receive signals via an 8-VSB (8-level vestigial sideband) demodulation box. The services provided include:

- **Standard Definition TV (SDTV)**—the NTSC TV or higher.

- **Enhanced Definition TV (EDTV)**—480 active lines or higher—the third and fourth rows in Table 5.4.
- **High-Definition TV (HDTV)**—720 active lines or higher. So far, the popular choices are 720P ($1,280 \times 720$, progressive scan, 30 fps), 1080I ($1,920 \times 1,080$, interlaced, 30 fps), and 1080P ($1,920 \times 1,080$, progressive scan, 30 or 60 fps).

5.2.4 Ultra High Definition TV (UHDTV)

UHDTV is a new development—a new generation of HDTV! The standards announced in 2012 support 4K UHDTV: 2160P ($3,840 \times 2,160$, progressive scan) and 8K UHDTV: 4320P ($7,680 \times 4,320$, progressive scan). The aspect ratio is 16:9. The bit-depth can be up to 12 bits, and the chroma subsampling can be 4:2:0 or 4:2:2. The supported frame rate has been gradually increased to 120 fps. The UHDTV will provide superior picture quality, comparable to IMAX movies, but it will require a much higher bandwidth and/or bitrate.

In early 2013, the ATSC called for proposals to support the 4K UHDTV (2160P) at 60 fps.

5.3 Video Display Interfaces

We now discuss the interfaces for video signal transmission from some output devices (e.g., set-top box, video player, video card, and etc.) to a video display (e.g., TV, monitor, projector, etc.). There have been a wide range of video display interfaces, supporting video signals of different formats (analog or digital, interlaced or progressive), different frame rates, and different resolutions [7]. We start our discussion with analog interfaces, including Component Video, Composite Video, and S-Video, and then digital interfaces, including DVI, HDMI, and DisplayPort.

5.3.1 Analog Display Interfaces

Analog video signals are often transmitted in one of three different interfaces: *Component video*, *Composite video*, and *S-video*. Figure 5.7 shows the typical connectors for them.

Component Video

Higher end video systems, such as for studios, make use of three separate video signals for the red, green, and blue image planes. This is referred to as *component video*. This kind of system has three wires (and connectors) connecting the camera or other devices to a TV or monitor.

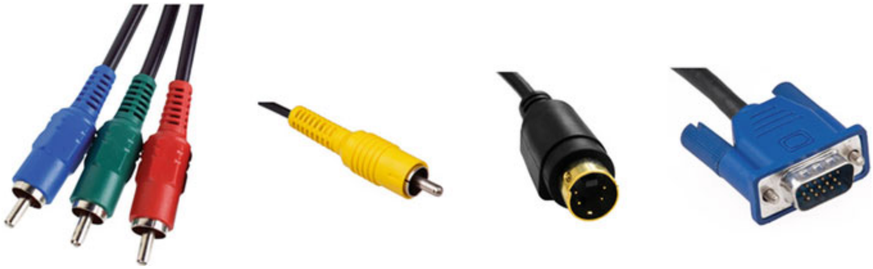


Fig. 5.7 Connectors for typical analog display interfaces. From left to right: Component video, Composite video, S-video, and VGA

Color signals are not restricted to always being RGB separations. Instead, as we saw in Chap. 4 on color models for images and video, we can form three signals via a luminance–chrominance transformation of the RGB signals—for example, YIQ or YUV.

For any color separation scheme, component video gives the best color reproduction, since there is no “crosstalk” between the three different channels, unlike composite video or S-video. Component video, however, requires more bandwidth and good synchronization of the three components.

Composite Video

In *composite video*, color (“chrominance”) and intensity (“luminance”) signals are mixed into a *single* carrier wave. Chrominance is a composite of two color components (I and Q, or U and V). This is the type of signal used by broadcast color TV; it is downward compatible with black-and-white TV.

In NTSC TV, for example [3], I and Q are combined into a chroma signal, and a color subcarrier then puts the chroma signal at the higher frequency end of the channel shared with the luminance signal. The chrominance and luminance components can be separated at the receiver end, and the two color components can be further recovered.

When connecting to TVs or VCRs, composite video uses only one wire (and hence one connector, such as a BNC connector at each end of a coaxial cable or an RCA plug at each end of an ordinary wire), and video color signals are mixed, not sent separately. The audio signal is another addition to this one signal. Since color information is mixed and both color and intensity are wrapped into the same signal, some interference between the luminance and chrominance signals is inevitable.

S-Video

As a compromise, *S-video* (separated video, or super-video, e.g., in S-VHS) uses two wires: one for luminance and another for a composite chrominance signal. As a result, there is less crosstalk between the color information and the crucial grayscale information.

The reason for placing luminance into its own part of the signal is that black-and-white information is most important for visual perception. As noted in the previous chapter, humans are able to differentiate spatial resolution in the grayscale ("black-and-white") part much better than for the color part of RGB images. Therefore, color information transmitted can be much less accurate than intensity information. We can see only fairly large blobs of color, so it makes sense to send less color detail.

Video Graphics Array (VGA)

The *Video Graphics Array* (VGA) is a video display interface that was first introduced by IBM in 1987, along with its PS/2 personal computers. It has since been widely used in the computer industry with many variations, which are collectively referred to as VGA.

The initial VGA resolution was 640×480 using the 15-pin D-subminiature VGA connector. Later extensions can carry resolutions ranging from 640×400 pixels at 70 Hz (24 MHz of signal bandwidth) to $1,280 \times 1,024$ pixels (SXGA) at 85 Hz (160 MHz) and up to $2,048 \times 1,536$ (QXGA) at 85 Hz (388 MHz).

The VGA video signals are based on analog component RGBHV (red, green, blue, horizontal sync, vertical sync). It also carries the *Display Data Channel* (DDC) data defined by *Video Electronics Standards Association* (VESA). Since the video signals are analog, it will suffer from interferences, particularly when the cable is long.

5.3.2 Digital Display Interfaces

Given the rise of digital video processing and the monitors that directly accept digital video signals, there is a great demand toward video display interfaces that transmit digital video signals. Such interfaces emerged in 1980s (e.g., Color Graphics Adapter (CGA) with the D-subminiature connector), and evolved rapidly. Today, the most widely used digital video interfaces include Digital Visual Interface (DVI), High-Definition Multimedia Interface (HDMI), and DisplayPort, as shown in Fig. 5.8.

Digital Visual Interface (DVI)

Digital Visual Interface (DVI) was developed by the *Digital Display Working Group* (DDWG) for transferring digital video signals, particularly from a computer's video card to a monitor. It carries uncompressed digital video and can be configured to support multiple modes, including DVI-D (digital only), DVI-A (analog only), or DVI-I (digital and analog). The support for analog connections makes DVI backward-compatible with VGA (though an adapter is needed between the two interfaces).

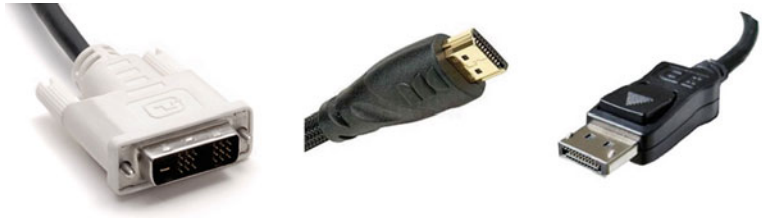


Fig. 5.8 Connectors of different digital display interfaces. From left to right: DVI, HDMI, DisplayPort

DVI's digital video transmission format is based on *PanelLink*, a high-speed serial link technology using *transition minimized differential signaling* (TMDS). Through DVI, a source, e.g., video card, can read the display's *extended display identification data* (EDID), which contains the display's identification, color characteristics (such as gamma level), and table of supported video modes. When a source and a display are connected, the source first queries the display's capabilities by reading the monitor's EDID block. A preferred mode or native resolution can then be chosen.

In a single-link mode, the maximum pixel clock frequency of DVI is 165 MHz, which supports a maximum resolution of 2.75 megapixels at the 60 Hz refresh rate. This allows a maximum 16:9 screen resolution of $1,920 \times 1,080$ at 60 Hz. The DVI specification also supports dual link, which achieves even higher resolutions up to $2,560 \times 1,600$ at 60 Hz.

High-Definition Multimedia Interface (HDMI)

HDMI is a newer digital audio/video interface developed to be backward-compatible with DVI. It was promoted by the consumer electronics industry, and has been widely used in the consumer market since 2002. The HDMI specification defines the protocols, signals, electrical interfaces, and mechanical requirements. Its electrical specifications, in terms of TMDS and VESA/DDC links, are identical to those of DVI. As such, for the basic video, an adapter can convert their video signals losslessly. HDMI, however, differs from DVI in the following aspects:

1. HDMI does not carry analog signal and hence is not compatible with VGA.
2. DVI is limited to the RGB color range (0–255). HDMI supports both RGB and YCbCr 4:4:4 or 4:2:2. The latter are more common in application fields other than computer graphics.
3. HDMI supports digital audio, in addition to digital video.

The maximum pixel clock rate for HDMI 1.0 is 165 MHz, which is sufficient to support 1080P and WUXGA ($1,920 \times 1,200$) at 60 Hz. HDMI 1.3 increases that to 340 MHz, which allows for higher resolution (such as WQXGA, $2,560 \times 1,600$) over a single digital link. The latest HDMI 2.0 was released in 2013, which supports 4K resolution at 60 fps.

DisplayPort

DisplayPort is a digital display interface developed by VESA, starting from 2006. It is the first display interface that uses packetized data transmission, like the Internet or Ethernet (see Chap. 15). Specifically, it is based on small data packets known as *micro packets*, which can embed the clock signal within the data stream. As such, DisplayPort can achieve a higher resolution yet with fewer pins than the previous technologies. The use of data packets also allows DisplayPort to be extensible, i.e., new features can be added over time without significant changes to the physical interface itself.

DisplayPort can be used to transmit audio and video simultaneously, or either of them. The video signal path can have 6–16 bits per color channel, and the audio path can have up to eight channels of 24-bit 192 kHz uncompressed PCM audio or carry compressed audio. A dedicated bi-directional channel carries device management and control data.

VESA designed DisplayPort to replace VGA and DVI. To this end, it has a much higher video bandwidth, enough for four simultaneous 1080P 60 Hz displays, or 4K video at 60 Hz. Backward compatibility to VGA and DVI is achieved by using active adapters. Compared with HDMI, DisplayPort has slightly more bandwidth, which also accommodates multiple streams of audio and video to separate devices. Furthermore, the VESA specification is royalty-free, while HDMI charges an annual fee to manufacturers. These points make DisplayPort a strong competitor to HDMI in the consumer electronics market, as well.

5.4 3D Video and TV

Three-dimensional (3D) pictures and movies have been in existence for decades. However, the rapid progress in the research and development of 3D technology and the success of the 2009 film Avatar have pushed 3D video to its peak. Increasingly, it is in movie theaters, broadcast TV (e.g., sporting events), personal computers, and various handheld devices.

The main advantage of the 3D video is that it enables the experience of immersion—be there, and really Be there!

We will start with an introduction to the fundamentals of 3D vision or 3D percept, emphasizing stereo vision (or stereopsis) since most modern 3D video and 3D TV are based on stereoscopic vision.

5.4.1 Cues for 3D Percept

The human vision system is capable of achieving a 3D percept by utilizing multiple cues. They are combined to produce optimal (or nearly optimal) depth estimates.

When the multiple cues agree, this enhances the 3D percept. When they conflict with each other, the 3D percept can be hindered. Sometimes, illusions can arise.

Monocular Cues

The monocular cues that do not necessarily involve both eyes include:

- Shading—depth perception by shading and highlights
- Perspective scaling—converging parallel lines with distance and at infinity
- Relative size—distant objects appear smaller compared to known same-size objects not in distance
- Texture gradient—the appearance of textures change when they recede in distance
- Blur gradient—objects appear sharper at the distance where the eyes are focused, whereas nearer and farther objects are gradually blurred
- Haze—due to light scattering by the atmosphere, objects at distance have lower contrast and lower color saturation
- Occlusion—a far object occluded by nearer object(s)
- Motion parallax—induced by object movement and head movement, such that nearer objects appear to move faster.

Among the above monocular cues, it has been said that Occlusion and Motion parallax are more effective.

Binocular Cues

The human vision system utilizes effective binocular vision, i.e., *stereo vision*, aka. *stereopsis*. Our left and right eyes are separated by a small distance, on average approximately 2.5 inches, or 65 mm. This is known as the *interocular distance*. As a result, the left and right eyes have slightly different views, i.e., images of objects are shifted horizontally. The amount of the shift, or *disparity*, is dependent on the object's distance from the eyes, i.e., its *depth*, thus providing the binocular cue for the 3D percept. The horizontal shift is also known as *horizontal parallax*. The fusion of the left and right images into single vision occurs in the brain, producing the 3D percept.

Current 3D video and TV systems are almost all based on stereopsis because it is believed to be the most effective cue.

5.4.2 3D Camera Models

Simple Stereo Camera Model

We can design a simple (artificial) stereo camera system in which the left and right cameras are identical (same lens, same focal length, etc.); the cameras' optical axes are in parallel, pointing at the Z-direction, the scene depth. The cameras are placed at $(-b/2, 0, 0)$ and $(b/2, 0, 0)$ in the world coordinate system (as opposed to a local coordinate system based on the camera axes), where b is camera separation, or the length of the *baseline*. Given a point $P(X, Y, Z)$ in the 3D space, and x_l and x_r being the x-coordinates of its projections on the left and right camera image planes, the following can be derived:

$$d = fb/Z, \tag{5.3}$$

where f is the focal length, $d = x_l - x_r$ is the *disparity* or *horizontal parallax*.

This suggests that disparity d is inversely proportional to the depth Z of the point P. Namely, objects near the cameras yield large disparity values, and far objects yield small disparity values. When the point is very far, approaching infinity, $d \rightarrow 0$.

Almost all amateur and professional stereo video cameras use the above Simple Stereo Camera Model where the camera axes are in parallel. The obvious reason is that it is simple and easy to manufacture. Moreover, objects at the same depth in the scene will have the same disparity d according to Eq. (5.3). This enables us to depict the 3D space with a stack of *depth planes*, or equivalently, *disparity planes*, which is handy in camera calibration, video processing and analysis.

Toed-in Stereo Camera Model

Human eyes are known to behave differently from the Simple camera model above. When humans focus on an object at a certain distance, our eyes rotate around a vertical axis in opposite directions in order to obtain (or maintain) single binocular vision. As a result, disparity $d = 0$ at the object of focus, and at the locations that have the same distance from the observer as the object of focus. $d > 0$ for objects farther than the object of focus (the so-called *positive parallax*), and $d < 0$ for nearer objects (*negative parallax*).

Human eyes can be emulated by so-called Toed-in Stereo Cameras, in which the camera axes are usually converging and not in parallel.

One of the complications of this model is that objects at the same depth (i.e., the same Z) in the scene no longer yield the same disparity. In other words, the "disparity planes" are now curved. Objects on both sides of the view appear farther away than the objects in the middle, even when they have the same depth Z.

5.4.3 3D Movie and TV Based on Stereo Vision

3D Movie Using Colored Glasses

In the early days, most movie theaters offering a 3D experience provided glasses tinted with complementary colors, usually red on the left and cyan on the right. This technique is called *Anaglyph 3D*. Basically, in preparing the stereo pictures, the left image is filtered to remove Blue and Green, and the right image is filtered to remove Red. They are projected onto the same screen with good alignment and proper disparities. After the stereo pictures pass through the colored glasses, they are mentally combined (fused) and the color 3D picture is reproduced in the viewer's brain.

The Anaglyph 3D movies are easy to produce. However, due to the color filtering, the color quality is not necessarily the best. Anaglyph 3D is still widely used in scientific visualization and various computer applications.

3D Movies Using Circularly Polarized Glasses

Nowadays, the dominant technology in 3D movie theaters is the RealD Cinema System. Movie-goers are required to wear polarized glasses in order to see the movie in 3D. Basically, the lights from the left and right pictures are polarized in different directions. They are projected and superimposed on the same screen. The left and right polarized glasses that the audience wear are polarized accordingly, which allows one of the two polarized pictures to pass through while blocking the other. To cut costs, a single projector is used in most movie theaters. It has a Z screen polarization switch to alternatively polarize the lights from the left and right pictures before projecting onto the screen. The frame rate is said to be 144 fps.

Circularly (as opposed to linearly) polarized glasses are used so the users can tilt their heads and look around a bit more freely without losing the 3D percept.

3D TV with Shutter Glasses

Most TVs for home entertainment, however, use *Shutter Glasses*. Basically, the liquid crystal layer on the glasses that the user wears becomes opaque (behaving like a shutter) when some voltage is applied. It is otherwise transparent. The glasses are actively (e.g., via Infra-Red) synchronized with the TV set that alternately shows left and right images (e.g., 120 Hz for the left and 120 Hz for the Right) in a Time Sequential manner.

3D vision with shutter glasses can readily be realized on desktop computers or laptops with a modest addition of specially designed hardware and software. The NVIDIA GeForce 3D Vision Kit is such an example.

5.4.4 The Vergence-Accommodation Conflict

Current stereoscopic technology for 3D video has many drawbacks. It is reported that a large number of viewers have difficulties watching 3D movies and/or TVs. 3D objects can appear darker, smaller, and flattened compared to their appearance in the real world. Moreover, they cause eye fatigue and strain. They can make viewers dizzy, causing headache and even nausea.

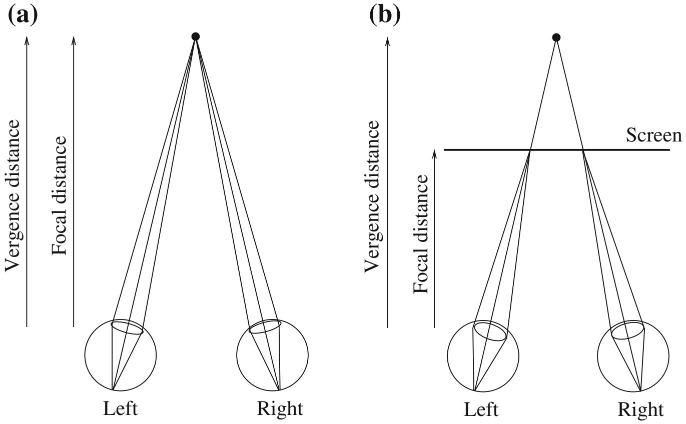


Fig. 5.9 The Vergence-Accommodation Conflict. **a** Real World and **b** 3D Display

Beside many obvious technical challenges in making the left and right images undistorted, synchronized, and separated, there is a more fundamental issue, i.e., the *Vergence-Accommodation Conflict* [8,9].

The word "accommodation" here refers to the physical act of the eye required to maintain a clear (focused) image on an object when its distance changes. As depicted in Fig. 5.9a, human eyes harmonize accommodation and vergence. When we keep our focus on an object of interest, our eyes also converge at the same spot. As a result, Focal distance = Vergence distance. The system is of course dynamic: we change our focus of attention when needed, and adjust our vergence and accommodation accordingly.

In a 3D movie theater, or when we gaze at a 3D display device, the situation is different. We are naturally focusing on the screen at a fixed distance. When our brains process and fuse the left and right images, we are supposed to decouple our vergence from accommodation. This is the *Vergence-Accommodation Conflict*. When the object is supposed to be behind the screen (with positive parallax) as indicated in Fig. 5.9b, Focal distance < Vergence distance; and vice versa.

Most of us seem capable of doing so, except it demands a heavy cognitive load. This explains why we quickly feel visual fatigue and so on. To cite Walter Murch, a distinguished film editor and sound designer, in one of his communications with Roger Ebert, the legendary film critic: "The biggest problem with 3D is the 'convergence/focus' issue. ... 3D films require us to focus at one distance and converge at another, and 600 million years of evolution has never presented this problem before. All living things with eyes have always focused and converged at the same point."

The movie industry has invented many techniques to alleviate this conflict [10]. For example, a common practice is to avoid depth discontinuity between cuts. Within the clips, efforts are made to keep the main object of interest at roughly the screen depth, and to keep its average depth at that level when there must be movements causing depth changes.

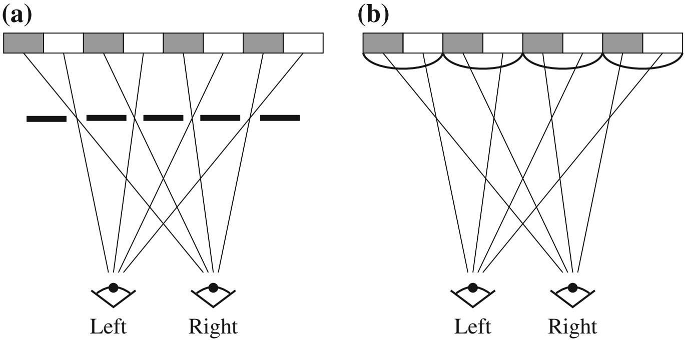


Fig. 5.10 Autostereoscopic display devices. **a** Parallax Barrier and **b** Lenticular Lens

5.4.5 Autostereoscopic (Glasses-Free) Display Devices

Wearing glasses while watching 3D video/TV/movie itself is another major drawback. It is uncomfortable, especially for those who already wear prescription eye glasses. The filters in the glasses inevitably dim the picture by reducing its brightness and contrast, not to mention the color distortion. Figure 5.10 shows two popular glasses-free, so-called *Autostereoscopic Display Devices*.

Figure 5.10a depicts the technique of *Parallax Barrier*, in which a layer of opaque material with slits is placed in front of the normal display device, e.g., an LCD. As a result, each eye only sees half of the columns on the display. By properly arranging the stereo left–right images, separate viewing of the two images in the left and right eyes is realized.

A number of commercial products use the Parallax Barrier technology, such as the portable Nintendo 3DS game console, the screen on the Fujifilm 3D camera FinePix Real 3D W3, and several smartphones.

In order to allow a larger viewing angle so the device can be used from multiple positions, and potentially by multiple users, more than one pair of stereo images can be used, e.g., in one of Toshiba's glasses-free 3D TVs.

Figure 5.10b depicts the technique of using a *Lenticular Lens*. Instead of barriers, columns of magnifying lenses can be placed in front of the display to direct lights properly to the left and right eyes. The same technology has also been applied to *lenticular printing* to generate various 3D pictures and/or animations.

The lenticular technology is a type of *Integral Imaging*, originally proposed by Gabriel Lippmann in 1908 [11]. Instead of cylindrical lenses as shown above, an array of spherical convex microlenses can be used to generate a large number of distinct microimages. These are computer-generated 2D views of the 3D scene, with one view per microlens. Therefore, this technique enables the rendering of multiple views from any directions. The Lytro camera, based on the technology of the *4D light field* [12], is one attempt toward this goal.

5.4.6 Disparity Manipulation in 3D Content Creation

The creation of 3D video content is a major challenge technically, perceptually, and artistically. In postproduction, disparity values are manipulated to create the best 3D percept. Below we will focus on various methods of disparity manipulation where the geometry will be altered. The disparity here is the *image disparity* measured in pixels.

As summarized by Lang et al. in their SIGGRAPH 2010 paper on nonlinear disparity mapping [13], the following are essential concepts:

- **Disparity Range**—When we are asked to look (focus) at the screen, there is a *comfort zone* near the screen distance. Objects in this zone are in the viewing angles of both eyes, and will yield an acceptable range of disparities so they are readily perceived in 3D. In creating 3D video contents it is a common practice to map (often suppress) the original disparities into the range that will fit in the comfort zone of most viewers. To cite [13], "Practical values for disparity on a 30 foot cinema screen, are between +30 (appears behind screen) and −100 (appears in front of screen) pixels, assuming video with a width of 2,048 pixels."
- **Disparity Sensitivity**—Our vision system is more capable of discriminating different depths when they are nearby. The sensitivity to depth drops rapidly with increased viewing distance: it is said to be inversely proportional to the square of the distance. This justifies a nonlinear disparity mapping [13] in which more disparity compression takes place at larger viewing distances. Since the disparity range of nearby objects is better preserved, this alleviates the problem of flattening of foreground objects, which are often more of interest.
- **Disparity Gradient**—This measures the rate of disparity changes within a distance in the stereoscopic images. For example, two points on a frontal surface in the 3D world will yield (approximately) the same disparity in the left and right images due to their identical depth; this will yield a disparity gradient of (near) zero. On the other hand, two points on an oblique surface may yield different disparity values due to their difference in depth, and hence yield a nonzero disparity gradient. Burt and Julesz [14] pointed out that human vision has a limit of disparity gradient in binocular fusion. Beyond this limit, fusion into a single vision is difficult and mostly impossible, which is thus avoided in disparity gradient editing.
- **Disparity Velocity**—When consecutive scenes present little disparity change, we can process the stereoscopic information very quickly. When there are large accommodation and vergence changes (i.e., disparity changes), we will slow down considerably. This is due to the limit of temporal modulation frequency of disparity. As discussed earlier, while focusing on the screen watching 3D video, the rapid change in vergence is a main cause for visual fatigue and must be restricted. We can tolerate some changes of convergence (i.e., disparity) as long as the speed of the changes is moderate.

Some additional technical issues are as follows:

- Most stereoscopic cameras adopt the Simple camera model where the camera optical axes are in parallel. This yields near-zero disparity for far objects and very large disparity for nearby objects, and is very different from the toed-in camera model which better emulates the human vision system. In this case, a conversion of the image disparity values is necessary in the 3D video postproduction stage. A variety of techniques are described in [10] and [15], among them *floating window* where the screen distance can be artificially shifted.

- As stated above, the average *interocular distance* of viewers is approximately 2.5 inches. As a result, in the toed-in camera model, the projected images of a very far object (near infinity), for example, should be about 2.5 inches apart in the left and right images on the screen in order to generate the required positive parallax. Depending on the screen size and screen resolution, a very different *image disparity* will be required. It is therefore a common practice to produce different 3D contents with very different image disparity values targeted for different purposes (large cinema screens vs. small PC or smartphone screens, high resolution vs. low resolution).

The multimedia and movie industries are keenly interested in converting the vast amount of 2D contents into 3D. Zhang et al. [16] provide a good survey on the issues involved in manually and (semi)automatically converting such videos and films.

5.5 Exercises

1. NTSC video has 525 lines per frame and 63.6 μs per line, with 20 lines per field of vertical retrace and 10.9 μs horizontal retrace.
   (a) Where does the 63.6 μs come from?
   (b) Which takes more time, horizontal retrace or vertical retrace? How much more time?
2. Which do you think has less detectable flicker, PAL in Europe or NTSC in North America? Justify your conclusion.
3. Sometimes the signals for television are combined into fewer than all the parts required for TV transmission.
   (a) Altogether, how many and what are the signals used for studio broadcast TV?
   (b) What does S-video stand for? How many and what signals are used in S-video?
   (c) How many signals are actually broadcasted for standard analog TV reception? What kind of video is that called?
4. Show how the Q signal can be extracted from the NTSC chroma signal C [Eq. (5.1)] during demodulation.
5. One sometimes hears that the old Betamax format for videotape, which competed with VHS and lost, was actually a better format. How would such a statement be justified?
6. We do not see flicker on a workstation screen when displaying video at NTSC frame rate. Why do you think this might be?

7. Digital video uses *chroma subsampling*. What is the purpose of this? Why is it feasible?
8. What are the most salient differences between ordinary TV and HDTV/UHDTV? What was the main impetus for the development of HDTV/UHDTV?
9. What is the advantage of interlaced video? What are some of its problems?
10. One solution that removes the problems of interlaced video is to deinterlace it. Why can we not just overlay the two fields to obtain a deinterlaced image? Suggest some simple deinterlacing algorithms that retain information from both fields.
11. Assuming the bit-depth of 12 bits, 120 fps, and 4:2:2 chroma subsampling, what are the bitrates of the 4K UHDTV and 8K UHDTV videos if they are uncompressed?
12. Assuming we use the toed-in stereo camera model, the interocular distance is I, and the screen is D meters away, (a) At what distance will a point P generate a positive parallax equal to I on the screen? (b) At what distance will a point P generate a negative parallax equal to $-I$?

References

1. A.M. Tekalp, *Digital Video Processing* (Prentice Hall PTR, Upper Saddle River, 1995)
2. A. Bovik (ed.), *Handbook of Image and Video Processing*, 2nd edn. (Academic Press, New York, 2010)
3. J.F. Blinn, NTSC: Nice Technology, Super Color. IEEE Comput. Graphics Appl. **13**(2), 17–23 (1993)
4. C.A. Poynton, *A Technical Introduction to Digital Video* (Wiley, New York, 1996)
5. J.F. Blinn, The world of digital video. IEEE Comput. Graphics Appl. **12**(5), 106–112 (1992)
6. C.A. Poynton, *Digital Video and HDTV Algorithms and InterfacesDigital Video and HDTV Algorithms and Interfaces* (Morgan Kaufmann, San Francisco, 2002)
7. R. L. Myers, *Display Interfaces: Fundamentals and Standards* (Wiley, New York, 2002)
8. D.M. Hoffman, A.R. Girshick, K. Akeley, M.S. Banks, Vergence-accommodation conflicts hinder visual performance and cause visual fatigue. J. Vis. **8**(3) (2008)
9. T. Shibata, J. Kim, D.M. Hoffman, M.S. Banks, The zone of comfort: predicting visual discomfort with stereo displays. J. Vis. **11**(8) (2011)
10. B. Mendiburu, *3D Movie Making: Stereoscopic Digital Cinema from Script to Screen* (Elsevier Science, Burlington, 2009)
11. G. Lippmann, La Photographie Integrale. Comptes Rendus Academie des Sciences **146**, 446–451 (1908)
12. M. Levoy, P. Hanrahan, Light field rendering, in *Proceedings of International Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*, 1996
13. M. Lang, A. Hornung, O. Wang, S. Poulakos, A. Smolic, M. Gross, Nonlinear disparity mapping for stereoscopic 3D. ACM Trans. Graph. **29**(4) (2010)
14. P. Burt, B. Julesz, A disparity gradient limit for binocular fusion. Science **208**(4444), 615–617 (1980)
15. R. Ronfard, G. Taubin (eds.), *Image and Geometry Processing for 3-D Cinematography: An Introduction* (Springer, Berlin Heidelberg, 2010)
16. L. Zhang, C. Vazquez, S. Knorr, 3D-TV content creation: automatic 2D-to-3D video conversion. IEEE Trans. Broadcast. **57**(2), 372–383 (2011)