Basic principles 1

This is Chapter 1 of the book *A Technical Introduction to Digital Video*, by Charles Poynton. Copyright © 1996 John Wiley & Sons.

This chapter is a summary of the fundamental concepts of digital video.

If you are unfamiliar with video, this chapter will introduce the major issues, to acquaint you with the framework and nomenclature that you will need to address the rest of the book. If you are already knowledgeable about video, this chapter will provide a quick refresher, and will direct you to specific topics about which you'd like to learn more.

Imaging

The three-dimensional world is imaged by the lens of the human eye onto the retina, which is populated with photoreceptor cells that respond to light having wavelengths in the range of about 400 nm to 700 nm. In an imaging system, we build a camera having a lens and a photosensitive device, to mimic how the world is perceived by vision.

Although the shape of the retina is roughly a section of a sphere, it is topologically two-dimensional. In a camera, for practical reasons, we employ a flat *image plane,* sketched in Figure 1.1 overleaf, instead of a spherical image surface. Image system theory concerns analyzing the continuous distribution of power that is incident on the image plane.

A photographic camera has, in the image plane, film that is subject to chemical change when irradiated by

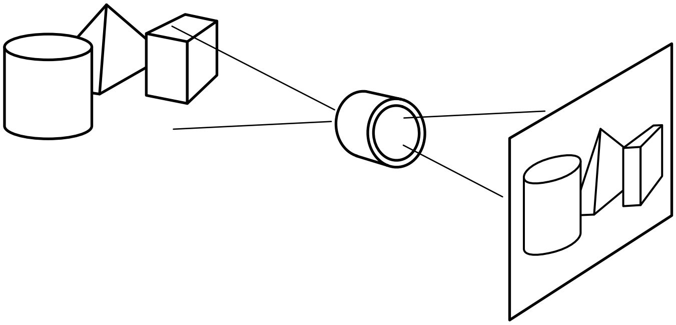


Figure 1.1 **Scene, lens, image plane.**

light. The active ingredient of photographic film is contained in a thin layer of particles having carefully controlled size and shape, in a pattern with no coherent structure. If the particles are sufficiently dense, an image can be reproduced that has sufficient information for a human observer to get a strong sense of the original scene. The finer the particles and the more densely they are arranged in the film medium, the higher will be the capability of the film to record spatial detail.

Digitization

Signals captured from the physical world are translated into digital form by *digitization*, which involves two processes. A signal is digitized when it is subjected to both *sampling* and *quantization*, in either order. When an audio signal is sampled, the single dimension of time is carved into discrete intervals. When an image is sampled, two-dimensional space is partitioned into small, discrete regions. Quantization assigns an integer to the amplitude of the signal in each interval or region.

1-D sampling

A signal that is a continuous one-dimensional function of time, such as an audio signal, is sampled through forming a series of discrete values, each of which represents the signal at an instant of time. *Uniform sampling*, where the time intervals are of equal duration, is ubiquitous.

2-D sampling

A continuous two-dimensional function of space is sampled by assigning, to each element of a sampling grid, a value that is a function of the distribution of intensity over a small region of space. In digital video and in conventional image processing, the samples lie on a regular, rectangular grid.

Samples need not be digital: A CCD camera is inherently sampled, but it is not inherently quantized. Analog video is not sampled horizontally, but is sampled vertically by scanning, and sampled temporally at the frame rate.

Pixel array

A digital image is represented by a matrix of values, where each value is a function of the information surrounding the corresponding point in the image. A single element in an image matrix is a *picture element*, or *pixel*. In a color system, a pixel includes information for all color components. Several common formats are sketched in Figure 1.2 below.

In computing it is conventional to use a sampling grid having equal horizontal and vertical sample pitch – *square pixels*. The term *square* refers to the sample pitch; it should not be taken to imply that image information associated with the pixel is distributed uniformly throughout a square region. Many video systems use sampling grids where the horizontal and vertical sample pitch are not equal.

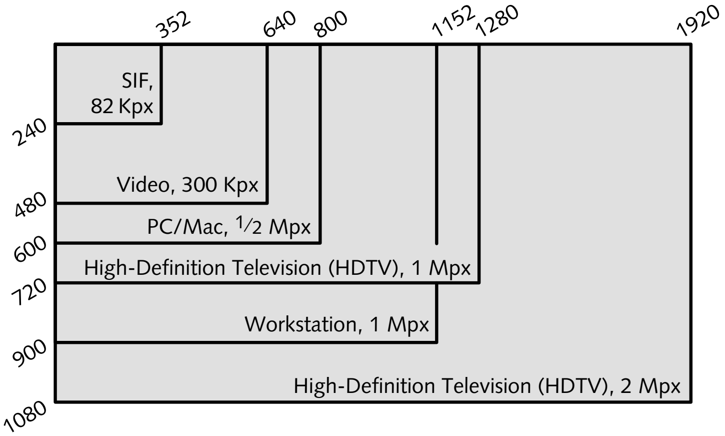


Figure 1.2 **Pixel array.**

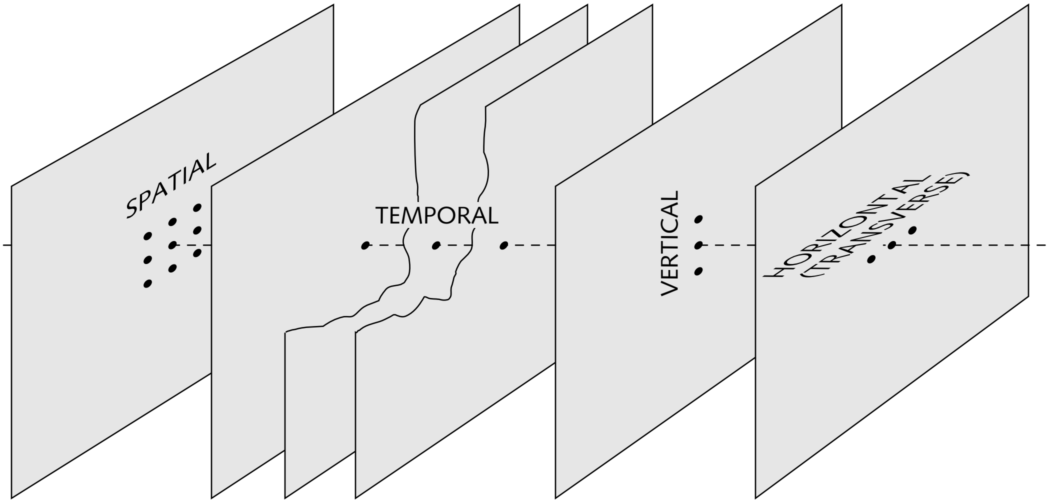


Figure 1.3 **Spatiotemporal domains.**

In computing it is usual to represent a grayscale or pseudocolor pixel as a single 8-bit byte. It is common to represent a truecolor pixel as three 8-bit red, green, and blue (*R'G'B'*) components totaling three bytes – 24 bits – per pixel.

Some framebuffers provide a fourth byte, which may be unused, or used to convey overlay or transparency data.

Spatiotemporal domains

A digital video image is sampled in the horizontal, vertical, and temporal axes, as indicated in Figure 1.3 above. One-dimensional sampling theory applies along each of these axes. At the right is a portion of the two-dimensional *spatial* domain of a single image. Some spatial processing operations cannot be separated into horizontal and vertical facets.

Scanning notation

In computing, a display is described by the count of pixels across the width and height of the image. Conventional television would be denoted 644×483, which indicates 483 picture lines. But any display system involves some scanning overhead, so the total number of lines in the *raster* of conventional video is necessarily greater than 483.

Video scanning systems have traditionally been denoted by their total number of lines including sync and blanking overhead, the frame rate in hertz, and an indication of *interlace* (2:1) or *progressive* (1:1) scan, to be introduced on page 11.

525/59.94/2:1 scanning is used in North America and Japan, with an analog bandwidth for studio video of about 5.5 MHz.

625/50/2:1 scanning is used in Europe and Asia, with an analog bandwidth for studio video of about 6.5 MHz. For both 525/59.94 and 625/50 component digital video according to ITU-R Rec. BT.601-4 ("Rec. 601"), the basic sampling rate is exactly 13.5 MHz. *Bandwidth* and *sampling rate* will be explained in later sections.

1125/60/2:1 scanning is in use for *high-definition television* (HDTV), with an analog bandwidth of about 30 MHz. The basic sampling rate for 1125/60 is 74.25 MHz. A variant 1125/59.94/2:1 is in use. This scanning system was originally standardized with a 1920×1035 image having pixels about 4 percent taller than square.

1920×1080

The square-pixel version of 1125/60 is now commonly referred to as 1920×1080.

1280×720

A progressive-scan one megapixel image format is proposed for advanced television in the United States.

Viewing distance and angle

A viewer tends to position himself or herself relative to a scene so that the smallest detail of interest in the scene subtends an angle of about one minute of arc ($^1\!/_{60}{}^\circ$), approximately the limit of angular discrimination for normal vision. For the 483 picture lines of conventional television, the corresponding viewing distance is about seven times picture height (PH); the horizontal viewing angle is about 11°. For the 1080 picture lines of HDTV, the optimum viewing distance is 3.3 screen heights, and the horizontal viewing angle is almost tripled to 28°. The situation is sketched in Figure 1.4 overleaf.

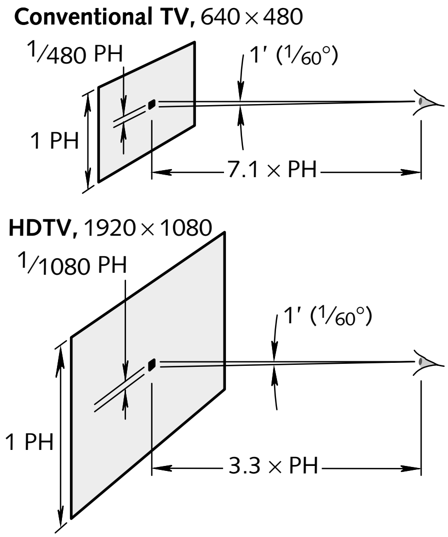


Figure 1.4 **Viewing distance and angle.**

$$distance \approx \frac{3400}{lines} \times PH$$

To achieve a viewing situation where a pixel subtends 1⁄60°, viewing distance expressed in units of picture height should be about 3400 divided by the number of picture lines. A computer user tends to position himself or herself closer than this – about 50 to 60 percent of this distance – but at this closer distance individual pixels are discernible. Consumer projection television is viewed closer than 7×PH, but at this distance scan lines become objectionable.

Aspect ratio

Aspect ratio is the ratio of image width to height. Conventional television has an aspect ratio of 4:3. High-definition television uses a wider ratio of 16:9. Cinema commonly uses 1.85:1 or 2.35:1. In a system having square pixels, the number of horizontal samples per picture width is the number of scanning lines in the picture height times the aspect ratio of the image.

Variants of conventional 525/59.94 systems having 16:9 aspect ratio have recently been standardized, but few are deployed as I write this.

Frame rate, refresh rate

A succession of flashed still pictures, captured and displayed at a sufficiently high rate, can create the illusion of motion. The quality of the motion portrayal depends on many factors.

Take care to distinguish *flicker* from *twitter*, described on page 11.

Most displays for moving images involve a period of time when the reproduced image is absent from the display, that is, a fraction of the frame time during which the display is black. In order to avoid objectionable flicker, it is necessary to flash the image at a rate higher than the rate necessary to portray motion. Refresh rate is highly dependent on the ambient illumination in the viewing environment: The brighter the environment, the higher the flash rate must be in order to avoid flicker. To some extent the brightness of the image itself influences the flicker threshold, so the brighter the image, the higher the refresh rate must be. Since peripheral vision has higher temporal sensitivity than central (foveal) vision, the flicker threshold of vision is also a function of the viewing angle of the image.

Refresh rate is generally engineered into a system. Once chosen, it cannot easily be changed. Different applications have adopted different refresh rates, depending on the image quality requirements and viewing conditions of the application.

In the darkness of a cinema, a flash rate of 48 Hz is adequate. In the early days of motion pictures, a frame rate of 48 Hz was thought to involve excessive expenditure for film stock, and 24 frames per second were found to be sufficient to portray motion. So, a conventional film projector flashes each frame twice. Higher realism can be obtained with specialized cameras and projectors that operate at higher frame rates, up to 60 frames per second or more.

In a dim viewing environment typical of television viewing, such as a living room, a flash rate of 60 Hz is sufficient. Originally, television refresh rates were chosen to match the local AC power line frequency.

In a bright environment such as an office, a refresh rate above 70 Hz might be required.

Motion portrayal

It is conventional in video for each element of an image sensor device to integrate light from the scene for the entire frame time. This captures as much of the light from the scene as possible, in order to maximize sensitivity and/or signal-to-noise ratio. In an interlaced camera, the *exposure time* is usually effectively the duration of the field, not the duration of the frame. This is necessary in order to achieve good motion portrayal.

If the image has elements that move an appreciable distance during the exposure time, then the sampled image information will exhibit *smear*. Smear can be minimized by using an exposure time that is a fraction of the frame time; however, the method involves discarding light from the scene and a sensitivity penalty is incurred.

When the effect of image information incident during a single frame time persists into succeeding frames, the sensor exhibits *lag*. Lag is a practical problem for tube-type cameras, but generally not a problem for CCD cameras.

Charles Poynton, "Motion portrayal, eye tracking, and emerging display technology,"in *Proceedings of the 30th SMPTE Advanced motion imaging conference*, 192–202 (White Plains, New York: SMPTE, 1996).

Flicker is absent in any image display device that produces steady, unflashing light for the duration of the frame time. You might think that a nonflashing display would be more suitable than a device that flashes, and many contemporary devices do not flash. However, if the viewer's gaze is tracking an element that moves across the display, a display with an *on-time* approaching the frame time will exhibit smearing of elements that move. This problem becomes more severe as eye tracking rates increase; for example, with the wide viewing angle of high-definition television.

Raster scanning

In cameras and displays, some time is required to advance the scanning operation – to *retrace* – from one line to the next and from one picture to the next. These intervals are called *blanking intervals*, because in a

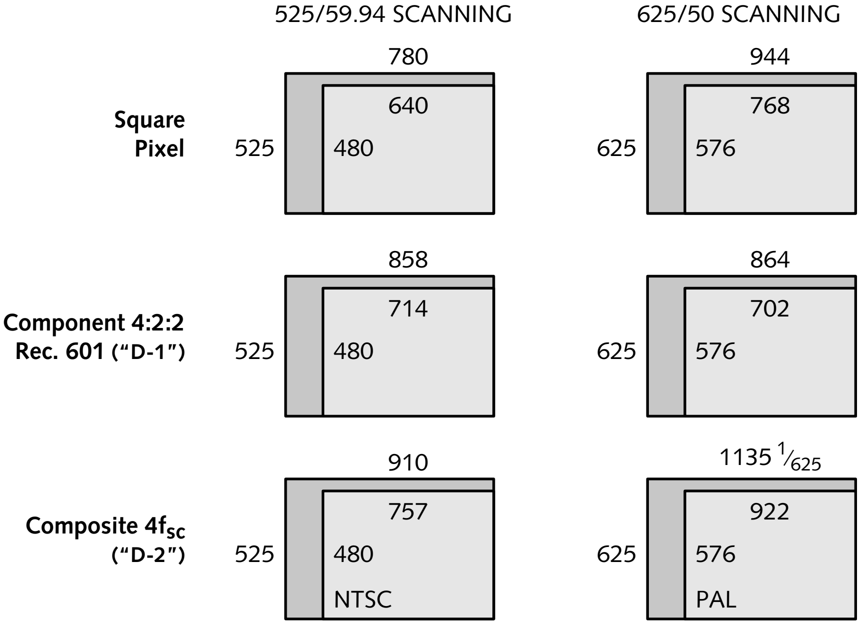


Figure 1.5 **Digital video rasters.** The left column shows 525/59.94 scanning, the right column shows 625/50. The top row shows sampling with square pixels. The middle row shows sampling at the Rec. 601 standard sampling frequency of 13.5 MHz. The bottom row shows sampling at four times the color subcarrier. Blanking intervals are shown with dark shading.

525/59.94 is colloquially referred to as *NTSC*, and 625/50 as *PAL*, but the terms NTSC and PAL properly apply to color encoding standards and not to scanning standards.

conventional CRT display the electron beam must be extinguished (*blanked*) during these time intervals. The *horizontal blanking* time lies between scan lines, and *vertical blanking* lies between frames (or fields). Figure 1.5 above shows the raster structure of 525/59.94 and 625/50 digital video systems, including these blanking intervals. In analog video, sync information is conveyed during the blanking intervals.

The horizontal and vertical blanking intervals required for a CRT display are quite large fractions of the line time and frame time: in 525/59.94, 625/50, and 1920×1035 systems, vertical blanking occupies 8 percent of each frame period. Although in principle a digital video interface could omit the blanking intervals and use a clock having a lower frequency than the sampling clock, this would be impractical. Digital video standards use interface clock frequencies chosen to

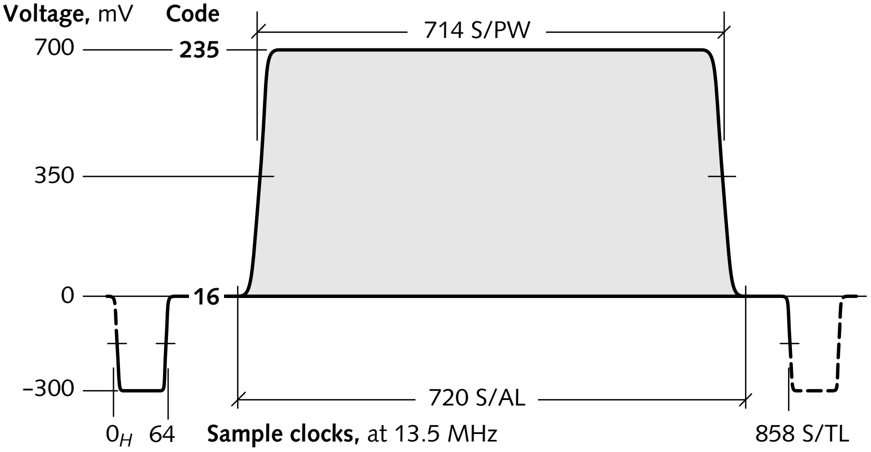


Figure 1.6 **Scan line waveform** for 525/59.94 component video, showing luma. The 720 *active* samples contain picture information. Horizontal blanking occupies the remaining sample intervals.

match the large blanking intervals of typical display equipment. Good use is made in digital systems of what would otherwise be excess data capacity: A digital video interface may convey audio signals during blanking; a digital video tape recorder might record error correction information in these intervals.

In a digital video system it is standard to convey samples of the image matrix in the same order that the image information would be conveyed in an analog video system: first the top line (left to right), then the next lower line, and so on.

In analog video, information in the image plane is scanned uniformly left to right during a fixed, short interval of time – the *active line time* – and conveyed as an analog electrical signal. There is a uniform mapping from horizontal position in the image to time instant in the electrical signal. Successive lines are scanned uniformly from the top of the image to the bottom, so there is also a uniform mapping from vertical position in the image to time instant in the electrical signal. The fixed pattern of parallel scanning lines disposed across the image is the *raster*. The word is derived from the Greek *rake*, from the resemblance of a raster to the pattern left on a newly raked field.

Figure 1.6 above shows the waveform of a single scan line, showing voltage from 0 V to 700 mV in a component analog system (with sync at –300 mV), and code-word value from code 16 to code 235 in an 8-bit component digital system.

Interlace

At the outset of television, the requirement to minimize information rate for transmission – and later, recording – led to *interlaced* scanning. Each frame is scanned in two successive vertical passes, first the *odd field*, then the *even field*, whose scan lines interlace as illustrated Figure 1.7 below. Total information rate is reduced because the flicker susceptibility of vision is due to a wide-area effect. As long as the complete height of the picture is scanned rapidly enough to overcome wide-area flicker, small-scale picture information – such as that in the alternate lines – can be transmitted at a lower rate.

Figure 1.7 exaggerates the slant of a fraction of a degree that results when a conventional CRT – either a camera tube or a display tube – is scanned with analog circuits. The slant is a real effect in analog cameras and displays, although it is disregarded in the design of equipment.

If the information in an image changes vertically at a scale comparable to the scanning line pitch – if a fine pattern of black-and-white horizontal line pairs is scanned, for example – then interlace can cause the content of the odd and the even fields to differ markedly. This causes *twitter*, a small-scale phenomenon that is perceived as extremely rapid up-and-down motion. Twitter can be produced not only from degenerate images such as fine horizontal black-and-white lines, but also from high-amplitude brightness detail in an ordinary image. In computer generated imagery (CGI), twitter can be reduced by vertical filtering.

If image information differs greatly from one field to the next, then instead of twitter, large-scale flicker will

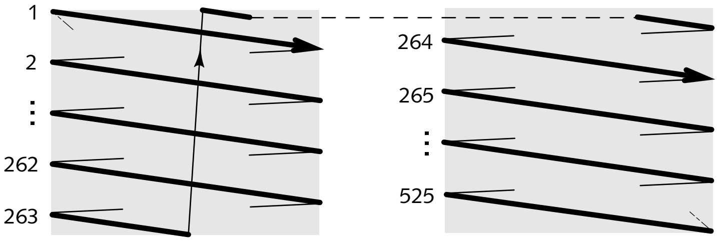


Figure 1.7 **Interlaced scanning** forms a complete picture – the *frame* – from two *fields*, each comprising half the scanning lines. The second field is delayed half the frame time from the first.

result. A video camera is designed to avoid introduction of so much vertical detail that flicker could be produced. In synthetic image generation, vertical detail may have to be explicitly filtered in order to avoid flicker.

Scanning standards

Conventional broadcast television scans a picture whose aspect ratio is 4:3, in left-to-right, top-to-bottom order using interlaced scanning.

A scanning system is denoted by its total line count and its field rate in hertz, separated by a solidus (slash). Two scanning standards are established for conventional television: 525/59.94, used primarily in North America and Japan; and 625/50, used elsewhere. It is obvious from the scanning nomenclature that the line counts and frame rates are different. There are other important differences:

525/59.94 video in Japan uses 10:4 picture to sync ratio and zero setup.

System	*525/59.94*	*625/50*
Picture:Sync ratio	10:4	7:3
Setup, percent	7.5	0
Count of equalization, broad pulses	6	5
Line number 1, and 0_V, defined at	First equalization pulse	First broad pulse

The two systems have gratuitous differences in other parameters unrelated to scanning.

Monochrome systems having 405/50/2:1 and 819/50/2:1 scanning were once used in Britain and France, respectively, but transmitters for these standards have now been decommissioned.

Systems with 525/59.94 scanning usually employ NTSC color coding, and systems with 625/50 scanning usually use PAL, so 525/59.94 and 625/50 systems are loosely referred to as *NTSC* and *PAL*. But NTSC and PAL properly refer to color encoding. Although 525/59.94/NTSC and 625/50/PAL systems dominate worldwide broadcasting, other combinations of scanning and color coding are in use in large and important regions of the world, such as France, Russia, and South America.

The frame rate of 525/59.94 video is exactly $^{60}/_{1.001}$ Hz. In 625/50 the frame rate is exactly 50 Hz. Computer graphics systems have various frame rates with few standards and poor tolerances.

An 1125/60/2:1 high-definition television production system has been adopted as SMPTE Standard 240M and has been proposed to the ITU-R. At the time of writing, the system is in use for broadcasting in Japan but no international broadcasting standards have been agreed upon.

All of these scanning systems are interlaced 2:1, and interlace is implicit in the scanning nomenclature. Noninterlaced scanning is common in desktop computers and universal in computer workstations. Emerging high-definition television standards have interlaced and noninterlaced variants.

John Watkinson, *The Engineer's Guide to Standards Conversion*. Petersfield, Hampshire, England: Snell & Wilcox, 1994.

Standards conversion refers to conversion among scanning standards. Standards conversion, done well, is difficult and expensive. Standards conversion between scanning systems having different frame rates, even done poorly, requires a fieldstore or framestore. The complexity of standards conversion between 525/59.94 scanning and 625/50 scanning is the reason that it is difficult for consumers – and broadcasters – to convert European material for use in North America or Japan, or vice versa.

Transcoding refers to changing the color encoding of a signal, without altering its scanning system.

Sync structure

At a video interface, synchronization (*sync*) is achieved by associating, with every scan line, a line sync datum denoted 0_H (pronounced *zero-H*). In component digital video, sync is conveyed using digital codes 0 and 255 outside the range of picture information. In analog video, sync is conveyed by voltage levels “blacker than black.” 0_H is defined by the 50-percent point of the leading (falling) edge of sync.

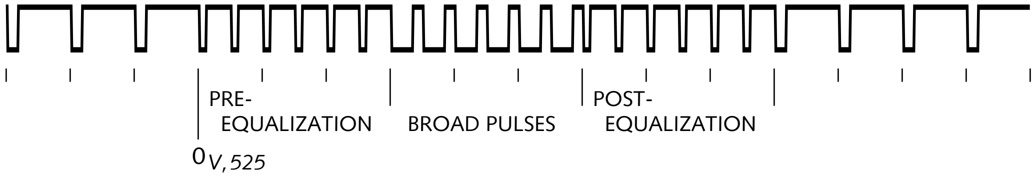


Figure 1.8 **Vertical sync waveform of 525/59.94.**

In both 525/59.94 and 625/50 video the *normal* sync pulse has a duration of 4.7 µs. Vertical sync is identified by *broad pulses*, which are *serrated* in order for a receiver to maintain horizontal sync even during the vertical interval. Narrow *equalization* pulses, half the sync pulse duration at twice the line rate, are present during intervals immediately before and immediately following the broad pulses.

These *equalization pulses* have no relationship with the process of *equalization* that is used to compensate poor frequency response of coaxial cable, or poor frequency or phase response of a filter.

When analog sync separators comprised just a few resistors and capacitors, to achieve stable interlacing required halving the duration of the line syncs and introducing additional pulses halfway between them. Originally the *equalization* pulses were the ones interposed between the line syncs, but the term now refers to all of the narrow pulses. The absence of sync level between the end of a broad pulse and the start of the following sync was called *serration*. If you think of field sync as a single pulse asserted for several lines, serration is the negation of this pulse at twice the line rate.

In digital technology it is more intuitive to consider the pulses that are present rather than the ones that are absent: The term *serration* is now unpopular.

An equalization pulse has half the duration of a normal sync. The duration of a vertical (*broad*) pulse is half the line time, less a full sync width. A 525/59.94 system has three lines of *preequalization* pulses, three lines of vertical sync, and three lines of *postequalization* pulses. A 625/50 system has two and one-half lines (five pulses) of each of preequalization, broad, and postequalization pulses. Figure 1.8 above sketches the vertical sync component of 525/59.94 analog video.

Monochrome 525-line broadcasting originated with a line rate of exactly 15.750 kHz. When color was introduced to NTSC in 1953, the monochrome horizontal frequency was multiplied by exactly $^{1000}/_{1001}$ to obtain the NTSC color line rate of approximately 15.734 kHz. Details are in *Field, frame, line, and sample rates*, on page 199. All 525-line broadcast signals – even monochrome signals – now employ this rate. The line rate of 625/50 systems has always been exactly 15.625 kHz, corresponding to a line time of exactly 64 μs.

Data rate

b = bit
B = Byte

k	10^3	1000
K	2^{10}	1024
SI, datacom:		
M	10^6	1 000 000
disk:		
M	$10^3 \cdot 2^{10}$	1 024 000
RAM:		
M	2^{20}	1 048 576

Data rate of a digital system is measured in bits per second (b/s) or bytes per second (B/s), where a byte is eight bits. The formal, international designation of the metric system is *Système International d'Unités*, SI. The SI prefix *k* denotes 10^3 (1000); it is often used in data communications. The *K* prefix used in computing denotes 2^{10} (1024). The SI prefix *M* denotes 10^6 (1 000 000). Disk storage is generally allocated in units integrally related to 1024 bytes; the prefix *M* applied to disk storage denotes 1 024 000. RAM memory generally has capacity based on powers of two; the prefix *M* applied to RAM denotes 2^{20} or 1024 K (1 048 576).

Data rate of digital video

Line rate is an important parameter of a video system: Line rate is simply the frame rate multiplied by the number of lines per total frame.

The aggregate *data rate* is the number of bits per pixel, times the number of pixels per line, times the number of lines per frame, times the frame rate.

In both analog and digital video it is necessary to convey not only the raw image information, but also information about which time instants (or which samples) are associated with the start of frame, or the start of line. This information is conveyed by signal synchronization or *sync* elements. In analog video and composite digital video, sync is combined with video by being coded at a level *blacker than black*.

All computer graphics systems and almost all digital video systems have the same integer number of sample clock periods in every raster line. In these cases, sampling frequency is simply the line rate times the number of samples per total line (S/TL).

In 625/50 PAL there is not an exact integer number of samples per line: Samples in successive lines are offset to the left a small fraction, $^1/_{625}$ of the horizontal sample pitch. The sampling structure is not precisely *orthogonal*, although digital acquisition, processing, and display equipment treat it so.

The data capacity required for the *active* pixels of a frame is computed by simply multiplying the number of bits per pixel by the number of active pixels per line, then by the number of active lines per frame. To compute the data rate for the active pixels, simply multiply by the frame rate.

Standards are not well established in display systems used in desktop computers, workstations, and industrial equipment. The absence of published data makes it difficult to determine raster scanning parameters.

Linearity

A video system should ideally satisfy the *principle of superposition;* in other words, it should exhibit *linearity*. A function f is linear *if and only if* (iff):

Eq 1.1

$$\mathrm{f}(a+b) \equiv \mathrm{f}(a)+\mathrm{f}(b)$$

The function *f* can encompass an entire system:
A system is linear iff the sum of the individual responses of the system to any two signals is identical to its response to the sum of the two. Linearity can pertain to steady-state response, or to the system's temporal response to a changing signal.

Linearity is a very important property in mathematics, in signal processing, and in video. But linearity in one domain cannot be carried across to another domain if a nonlinear function separates the two. An image signal usually originates in a sensor that has linear response to physical intensity. And video signals are usually processed through analog circuits that have linear response to voltage or digital systems that are linear with respect to the arithmetic performed on the codewords. But a video camera applies a nonlinear transfer function – *gamma correction* – to the image signal. So the image signal is in a linear optical domain, and the video signal is in a linear electrical domain, but the two domains are not the same.

Perceptual uniformity

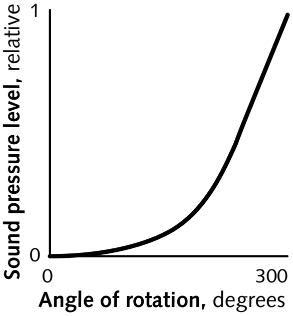


Figure 1.9 **Audio taper.**

A system is *perceptually uniform* if a small perturbation to a component value is approximately equally perceptible across the range of that value. The volume control on your radio is designed to be perceptually uniform: Rotating the knob 10 degrees produces approximately the same perceptual increment in volume anywhere across the range of the control. If the control were physically linear, the logarithmic nature of loudness perception would place all of the perceptual "action" of the control at the bottom of its range. Figure 1.9, in the margin, shows the transfer function of a potentiometer with standard *audio taper*.

The CIE *L** system, to be described on page 88, assigns a perceptually uniform scale to lightness. Video signals are coded according to perceptual principles, as will be explained in Chapter 6, *Gamma*, on page 91.

Noise, signal, sensitivity

Any analog electronic system is inevitably subject to noise that is unrelated to the signal to be processed by the system. As signal amplitude decreases, the noise makes a larger and larger relative contribution. In analog electronics, noise is inevitably introduced from thermal sources, and perhaps also from nonthermal sources of interference.

A distortion product that can be attributed to a particular processing step is known as an *artifact*, particularly if it has a distinctive visual effect on the picture.

In addition to random noise, processing of a signal may introduce distortion that is correlated to the signal itself. For the purposes of objective measurement of the performance of a system, distortion is treated as noise. Depending on its nature, distortion may be more or less perceptible than random noise.

Signal-to-Noise Ratio (SNR) is the ratio of a specified signal, often the reference amplitude or largest amplitude signal that can be carried by a system, to the amplitude of undesired components including noise and distortion. SNR is expressed in units of *decibels* (dB), a logarithmic measure.

Sensitivity refers to the minimum signal power that achieves acceptable (or specified) SNR performance.

Quantization

To make a 50-foot-long fence with fence posts every 10 feet you need six posts, not five! Take care to distinguish *levels* (here, six) from *steps* (here, five).

A signal whose amplitude takes a range of continuous values is *quantized* by assigning to each of a finite set of intervals of amplitude a discrete, numbered level. In *uniform quantization* the *steps* between levels have equal amplitude. The degree of visual impairment caused by noise in a video signal is a function of the properties of vision. In video, it is ubiquitous to digitize a signal that is a nonlinear function, usually a 0.45-power function, of physical (linear-light) intensity. The function chosen minimizes the visibility of noise.

Theoretical SNR for an k-step quantizer:

$$20\log_{10}\left(k\sqrt{12}\right)$$

The effect of quantizing to a finite number of discrete amplitude levels is equivalent to adding *quantization noise* to the ideal levels of a quantized signal. Quantization has the effect of introducing noise, and thereby diminishes the SNR of a digital system. Eight-bit quantization has a theoretical SNR limit of about 56 dB (peak signal to rms noise).

If an input signal has very little noise, then situations can arise when the quantized value is quite predictable at some points, but when the signal is near the edge of a quantizer step, uncertainty in the quantizer is reflected as noise. This situation can cause the reproduced image to exhibit *noise modulation*. It is beneficial to introduce roughly a quantizer step's worth of

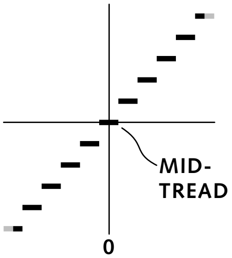


Figure 1.10
Mid-tread quantizer.

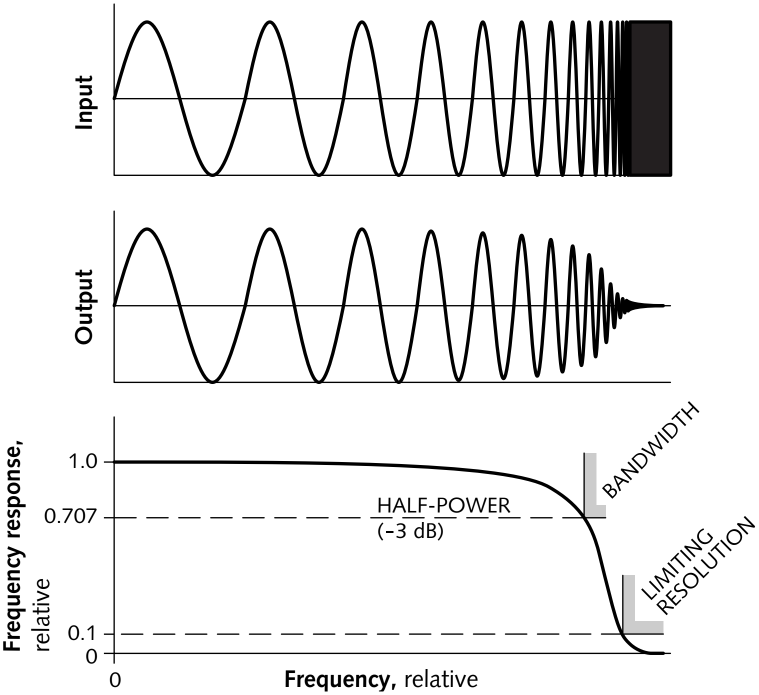
noise (peak to peak) prior to quantization, to avoid this effect. This introduces a very small amount of noise in the picture, but guarantees avoidance of "patterning" of the quantization.

Quantization can be applied to a unipolar signal such as luma. For a bipolar signal such as a color difference it is standard to use a *mid-tread* quantizer, such as the one sketched in Figure 1.10 in the margin, so that no systematic error affects the zero value.

Frequency response, bandwidth

Figure 1.11 below shows a test signal starting at zero frequency and sweeping up to some high frequency. The response of a typical electronic system is shown in the middle graph; the response diminishes at high frequency. The envelope of that waveform – the system's *frequency response* – is shown at the bottom.

Figure 1.11 **Frequency response** of any electronic or optical system falls as frequency increases. Bandwidth is measured at the half-power point (–3 dB), where response has fallen to 0.707. Television displays are often specified at *limiting resolution*, where response has fallen to 0.1.



Loosely speaking, *bandwidth* is the rate at which information in a signal can change from one state to another. The response of an electronic system deteriorates above a certain information rate. Bandwidth is specified or measured at the frequency where amplitude has fallen 3 dB from its value at zero frequency (called *DC*) – that is, to the fraction 0.707 of its value at DC.

The rate at which an analog video signal can change from one state to another, say from white to black, is limited by the bandwidth of the video system. This places an upper bound on *horizontal resolution*. Consumer video generally refers to horizontal resolution, measured as the number of black and white elements (*TV lines*) that can be discerned over a horizontal distance equal to the picture height.

Bandwidth and data rate

Data rate does not apply directly to an analog system, and the term *bandwidth* does not properly apply to a digital system. When a digital system conveys a sampled representation of a continuous signal, as in digital video or digital audio, the bandwidth represented by the digitized signal is necessarily less than half – typically about 0.45 – of the sampling rate.

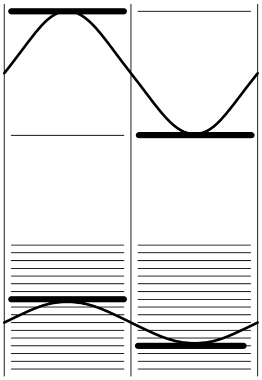


Figure 1.12
Bandwidth and data rate.

When arbitrary digital information is conveyed through an analog channel, as by a modem, the data rate that can be achieved depends on bandwidth, noise, and other properties of the channel. Figure 1.12, in the margin, shows a simple scheme that transmits two bits per second per hertz of bandwidth, or 2400 b/s for a channel having 1200 Hz analog bandwidth. The bottom sketch shows that if each half-cycle conveys one of sixteen amplitude levels, providing the channel has sufficiently low noise, four bits can be coded per half-cycle. The rate at which the signal in the channel can change state – the *symbol rate* or *baud rate* – is constant at 2400 baud, but this modulation method has a *data rate* or *bit rate* of 9600 b/s.

Resolution

As picture detail increases in frequency, the response of an imaging system will eventually deteriorate. In image science and in television, *resolution* refers to the capability of an imaging system to reproduce fine detail in the picture.

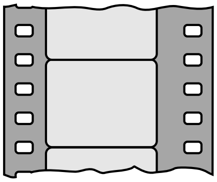
The absolute upper limit to resolution in a digital image system is the number of pixels over the width and height of a frame, and is the way the term *resolution* is used in computing.

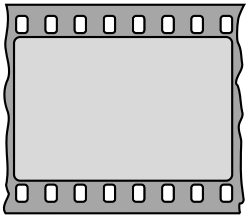
In conventional North American television, 483 scan lines cover the height of the image. High-definition television systems use up to 1080 picture lines. The amount of information that can be captured in a video signal is bounded by the number of picture lines. But other factors impose limits more severe than the number of lines per picture height.

In an interlaced system, vertical resolution must be reduced substantially from the scan-line limit, in order to avoid producing a signal that will exhibit objectionable twitter upon display.

Resolution in film

In film, resolution is measured as the finest pattern of straight, parallel lines that can be reproduced, expressed in *line pairs per millimeter* (lp/mm). A line pair contains a black region and a white region.





Motion picture film is conveyed vertically through the camera and projector, so the width – not the height – of the film is 35 mm. Cinema usually has an aspect ratio of 1.85:1, so the projected film area is about 21 mm × 11 mm, only three-tenths of the 36 mm × 24 mm projected area of 35 mm still film.

The limit to the resolution of motion picture film is not the static response of the film, but judder and weave in the camera and the projector.

Resolution in television

In video, resolution refers to the number of line pairs (cycles) resolved on the face of the display screen, expressed in cycles per picture height (C/PH) or cycles per picture width (C/PW). A *cycle* is equivalent to a *line pair* of film. In a digital system, it takes at least two samples – pixels, scanning lines, or *TV lines* – to represent a line pair. However, resolution may be substantially less than the number of pixel pairs due to optical, electro-optical, and electrical filtering effects. *Limiting resolution* is defined as the frequency where detail is recorded with just 10 percent of the system's low-frequency response.

In consumer television, the number of scanning lines is fixed by the raster standard, but the electronics of transmission, recording, and display systems tend to limit bandwidth and reduce horizontal resolution. Consequently, in consumer electronics the term *resolution* generally refers to horizontal resolution. Confusingly, horizontal resolution is expressed in units of lines per picture *height*, so once the number of resolvable lines is measured, it must be corrected for the aspect ratio of the picture. Resolution in *TV lines per picture height* is twice the resolution in cycles per picture width, divided by the aspect ratio of the picture.

Resolution in computer graphics

In computer graphics, *resolution* is simply the number of discrete vertical and horizontal pixels required to store the digitized image. For example, a 1152×900 system has a total of about one million pixels (one megapixel, or 1 Mpx). Computer graphics is not generally very concerned about whether individual pixels can be discerned on the face of the display. In most color computer systems, an image comprising a one-pixel black-and-white checkerboard actually displays as a uniform gray, due to poor high-frequency response in the cable and video amplifiers, and due to rather large spot size at the CRT.

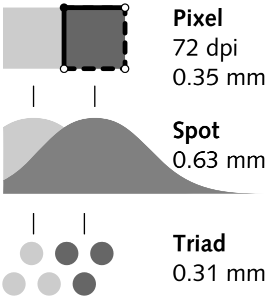


Figure 1.13
Pixel/spot/triad.

Computer graphics often treats each pixel as representing an idealized rectangular area independent of all other pixels. This notion discounts the correlation among pixels that is an inherent and necessary aspect of image acquisition, processing, compression, display, and perception. In fact the rather large spot produced by the electron beam of a CRT and the arrangement of phosphor triads on the screen, suggested by Figure 1.13, produces an image of a pixel on the screen that bears little resemblance to a rectangle. If pixels are viewed at a sufficient distance, these artifacts are of little importance. However, imaging systems are forced by economic pressures to make maximum perceptual use of the delivered pixels, consequently we tend to view CRTs at close viewing distances.

Luma

As you will see in *Luma and color differences*, on page 155, a video system conveys image data in the form of a component that represents brightness, and two other components that represent color. It is important to convey the brightness component in such a way that noise (or quantization) introduced in transmission, processing, and storage has a perceptually similar effect across the entire tone scale from black to white. Ideally, these goals would be accomplished by forming a true CIE luminance signal as a weighted sum of linear-light red, green, and blue; then subjecting that luminance to a nonlinear transfer function similar to the CIE L^* function that will be described on page 88.

There are practical reasons in video to perform these operations in the opposite order. First a nonlinear transfer function – *gamma correction* – is applied to each of the linear R, G, and B. Then a weighted sum of the nonlinear components is computed to form a *luma* signal, Y', representative of brightness.

625/50 standards documents indicate a precorrection of $^1/_{2.8}$, approximately 0.36, but this value is rarely used in practice. See *Gamma* on page 91.

In effect, video systems approximate the lightness response of vision using RGB intensity signals, each raised to the 0.45 power. This is comparable to the $^1/_3$ power function defined by L^*.

Recommendation ITU-R BT.601-4, *Encoding Parameters of Digital Television for Studios*. Geneva: ITU, 1990.

The coefficients that correspond to the so-called *NTSC* red, green, and blue CRT phosphors of 1953 are standardized in Recommendation ITU-R BT. 601-4 of the ITU Radiocommunication Sector (formerly CCIR). I call it *Rec. 601*. To compute nonlinear video *luma* from nonlinear red, green, and blue:

Eq 1.2

$$ {}^{601}Y' = 0.299\, R' + 0.587\, G' + 0.114\, B' $$

The prime symbols in this equation, and in those to follow, denote nonlinear components.

The unfortunate term "video luminance"

Unfortunately, in video practice, the term *luminance* has come to mean *the video signal representative of luminance* even though the components of this signal have been subjected to a nonlinear transfer function. At the dawn of video, the nonlinear signal was denoted *Y'*, where the prime symbol indicated the nonlinear treatment. But over the last 40 years the prime has been elided and now both the term *luminance* and the symbol *Y* collide with the CIE, making both ambiguous! This has led to great confusion, such as the incorrect statement commonly found in computer graphics and color textbooks that in the *YIQ* or *YUV* color spaces, the *Y* component *is* CIE luminance! I use the term *luminance* according to its standardized CIE definition and use the term *luma* to refer to the video signal, and I am careful to designate the nonlinear quantity with a prime symbol. But my convention is not yet widespread, and in the meantime you must be careful to determine whether a linear or nonlinear interpretation is being applied to the word and the symbol.

Color difference coding

In component video, the three components necessary to convey color information are transmitted separately.

The data capacity accorded to the color information in a video signal can be reduced by taking advantage of the relatively poor color acuity of vision, providing full

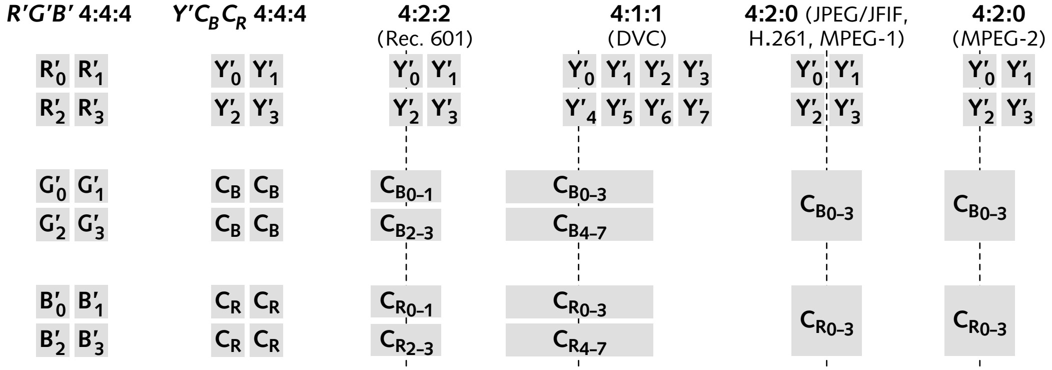


Figure 1.14 **Chroma subsampling.** A 2×2 array of *R'G'B'* pixels can be transformed to a luma component *Y'* and two color difference components C_B and C_R; color detail can then be reduced by subsampling, provided that full luma detail is maintained. The wide aspect of the C_B and C_R samples indicates their spatial extent. The horizontal offset of C_B and C_R is due to cositing. (JPEG, H.261, and MPEG-1 do not use cositing; instead, their C_B and C_R samples are taken halfway between luma samples.)

luma bandwidth is maintained. It is ubiquitous to base *color difference* signals on *blue minus luma* and *red minus luma* (*B'–Y'*, *R'-Y'*). Luma and (*B'–Y'*, *R'–Y'*) can be computed from *R'*, *G'*, and *B'* through a 3×3 matrix multiplication. Once luma and color difference – or *chroma* – components have been formed, the chroma components can be subsampled (filtered).

Y'C_BC_R

In component digital video, C_B and C_R components scaled from (*B'–Y'*, *R'–Y'*) are formed.

Y'P_BP_R

In component analog video, P_B and P_R color difference signals scaled from (*B'–Y'*, *R'–Y'*) are lowpass filtered to about half the bandwidth of luma.

4:4:4

In Figure 1.14 above, the left-hand column sketches a 2×2 array of *R'G'B'* pixels that, with 8 bits per sample, would occupy a total of 12 bytes. This is denoted 4:4:4 *R'G'B'*. *Y'*C_BC_R components can be formed from *R'G'B'*, as shown in the second column; without subsampling, this is denoted 4:4:4 *Y'*C_BC_R.

The use of *4* as the numerical basis for subsampling notation is a historical reference to a sample rate of about four times the color subcarrier frequency.

4:2:2

$Y'C_BC_R$ digital video according to Rec. 601 uses 4:2:2 sampling: Chroma components are subsampled by a factor of 2 along the horizontal axis. Chroma samples are coincident (cosited) with alternate luma samples.

In an 8-bit system using 4:2:2 coding, the 2×2 array occupies 8 bytes, and the aggregate data capacity is 16 bits per pixel. For studio digital video, the raw data rate is 27 MB/s.

4:1:1

A few digital video systems have used 4:1:1 sampling, where the chroma components are subsampled by a factor of 4 horizontally.

4:2:0

JPEG, H.261, MPEG-1, and MPEG-2 usually use 4:2:0 sampling. C_B and C_R are each subsampled by a factor of 2 both horizontally and vertically; C_B and C_R are sited vertically halfway between scan lines. Horizontal subsampling is inconsistent. In MPEG-2, C_B and C_R are cosited horizontally. In JPEG, H.261, and MPEG-1, C_B and C_R are not cosited horizontally; instead, they are sited halfway between alternate luma samples.

H.261, known casually as *p*×64, denotes a video-conferencing standard promulgated by the ITU-T.

MAC

A transmission system for analog components – *Multiplexed Analog Components*, or MAC – has been adopted in Europe for direct broadcast from satellite (DBS). In MAC, the color difference components are not combined with each other or with luma, but are time-compressed and transmitted serially. MAC is not standardized by ITU-R.

Component digital video, 4:2:2

The standard interface for 4:2:2 component digital video is Rec. ITU-R 601-4. It specifies sampling of luma at 13.5 MHz and sampling of C_B and C_R color difference components at 6.75 MHz. This interface is referred to as *4:2:2*, since luma is sampled at four times 3.375 MHz, and each of the C_B and C_R components at twice 3.375 MHz – that is, the color difference signals are horizontally subsampled by a factor of 2:1 with respect to luma. Sampling at 13.5 MHz results in an integer number of *samples per total line* (S/TL) in both

A version of Rec. 601 uses 18 MHz sampling to produce a picture aspect ratio of 16:9.

525/59.94 systems (858 S/TL) and 625/50 systems (864 S/TL). Luma is sampled with 720 *active* samples per line in both 525/59.94 and 625/50.

Component digital video tape recorders are widely available for both 525/59.94 and 625/50 systems, and have been standardized with the designation *D-1*. That designation properly applies to the tape format, not the signal interface.

Transport, electrical, and mechanical aspects of 4:2:2 interface are specified in Rec. 656. See page 248.

Rec. 601 specifies luma coding that places black at code 16 and white at code 235. Color differences are coded in offset binary, with zero at code 128, the negative peak at code 16, and the positive peak at code 240.

Composite video

The terms *NTSC* and *PAL* are often used incorrectly to refer to scanning standards. Since PAL encoding is used with both 625/50 scanning (with two different subcarrier frequencies) and 525/59.94 scanning (with a third subcarrier frequency), the term *PAL* alone is ambiguous. The notation *CCIR* is sometimes used to refer to 625/50 scanning, but that is confusing because the former CCIR – now ITU-R – standardized all scanning systems, not just 625/50.

In composite NTSC and PAL video, the color difference signals required to convey color information are combined by the technique of quadrature modulation into a *chroma* signal using a color subcarrier of about 3.58 MHz in conventional NTSC and about 4.43 MHz in conventional PAL. Luma and chroma are then summed into a composite signal for processing, recording, or transmission. Summing combines brightness and color into one signal, at the expense of introducing a certain degree of mutual interference.

The frequency and phase of the subcarrier are chosen and maintained carefully: The subcarrier frequency is chosen so that luma and chroma, when they are summed, are *frequency interleaved*. Studio signals have coherent sync and color subcarrier; that is, subcarrier is phase-locked to a rational fraction of the line rate; generally this is achieved by dividing both from a single master clock. In industrial and consumer video, subcarrier usually free-runs with respect to line sync.

SECAM sums luma and chroma without using frequency interleaving. SECAM has no application in the studio. See page 254.

Transcoding among different color encoding methods having the same raster standard is accomplished by luma/chroma separation, color demodulation, and color remodulation.

Composite digital video, $4f_{SC}$

The earliest digital video equipment processed signals in composite form. Processing of digital composite signals is simplified if the sampling frequency is an integer multiple of the color subcarrier frequency. Nowadays, a multiple of four is used: *four-times-subcarrier*, or $4f_{SC}$. For NTSC systems it is standard to sample at about 14.3 MHz. For PAL systems the sampling frequency is about 17.7 MHz.

Composite digital processing was necessary in the early days of digital video, but most image manipulation operations cannot be accomplished in the composite domain. During the 1980s there was widespread deployment of component digital processing equipment and component videotape recorders (DVTRs), recording 4:2:2 signals using the D-1 standard.

However, the data rate of a component 4:2:2 signal is roughly twice that of a composite signal. Four-times-subcarrier composite digital coding was resurrected to enable a cheap DVTR; this became the *D-2* standard. The D-2 DVTR offers the advantages of digital recording, but retains the disadvantages of composite NTSC or PAL: Luma and chroma are subject to cross-contamination, and the pictures cannot be manipulated without decoding and reencoding.

The development and standardization of D-2 recording led to the standardization of composite $4f_{SC}$ digital parallel and serial interfaces, which essentially just code the raw 8- or 10-bit composite data stream. These interfaces share the electrical and physical characteristics of the standard 4:2:2 interface, but with about half the data rate. For 8-bit sampling this leads to a total data rate of about 14.3 MB/s for 525/59.94 NTSC, and about 17.7 MB/s for 625/50 PAL.

Analog interface

Video signal amplitude levels in 525/59.94 systems are expressed in IRE units, named after the Institute of Radio Engineers in the United States, the predecessor

In 525/59.94 with setup, *picture excursion* refers to the range from blanking to white, even though strictly speaking the lowest level of the picture signal is 7.5 IRE, and not 0 IRE.

525/59.94 NTSC video in Japan employs zero setup.

of the IEEE. Reference blanking level is defined as 0 IRE, and reference white level is 100 IRE. The range between these values is the *picture excursion*.

Composite 525/59.94 systems have a picture-to-sync ratio of 10:4; consequently, the sync level of a composite 525/59.94 signal is –40 IRE. In composite NTSC systems, except in Japan, reference black is *setup* the fraction 7.5 percent ($^3/_{40}$) of the reference blanking-to-white excursion: Composite 525/59.94 employs a *pedestal* of 7.5 IRE. There are exactly 92.5 IRE from black to white: The picture excursion of a 525/59.94 signal is about 661 mV.

Setup has been abolished from component digital video and from HDTV. Many 525/59.94 component analog systems have adopted *zero setup*, and have 700 mV excursion from black to white, with 300 mV sync. But many component analog systems use setup, and it is a nuisance in design and in operation.

625/50 systems have a picture-to-sync ratio of 7:3, and zero setup. Picture excursion (from black to white) is exactly 700 mV; sync amplitude is exactly 300 mV. Because the reference levels are exact in millivolts, the IRE unit is rarely used, but in 625/50 systems an IRE unit corresponds to exactly 7 mV.

A video signal with sync is distributed in the studio with blanking level at zero (0 V_{DC}) and an amplitude from synctip to reference white of one volt into an impedance of 75 Ω. A video signal without sync is distributed with blanking level at zero, and an amplitude from blanking to reference white of either 700 mV or 714 mV.

High-definition television, HDTV

High-definition television (HDTV) is defined as having twice the vertical and twice the horizontal resolution of conventional television, a picture aspect ratio of 16:9, a frame rate of 24 Hz or higher, and at least two channels of CD-quality sound.

HDTV studio equipment is commercially with 1125/60/2:1 scanning and 1920×1035 image format, with about two megapixels per frame – six times the number of pixels of conventional television. The data rate of studio-quality HDTV is about 120 megabytes per second. Commercially available HDTV cameras rival the picture quality of the best motion picture cameras and films.

NHK Science and Technical Research Laboratories, *High Definition Television: Hi-Vision Technology*. New York: Van Nostrand Reinhold, 1993.

SMPTE 274M-1995, *1920 ×1080 Scanning and Interface*.

Except for their higher sampling rates, studio standards for HDTV have a close relationship to studio standards for conventional video, which I will describe in the rest of the book. For details specific to HDTV, consult the book from NHK Labs, SMPTE 274M and 296M.

Advanced Television (ATV) refers to transmission systems designed for the delivery of entertainment to consumers, at quality levels substantially improved over conventional television. ATV transmission systems based on 1125/60/2:1 scanning and MUSE compression have been deployed in Japan. The United States has adopted standards for ATV based on 1920×1080 and 1280×720 image formats. MPEG-2 compression can compress this to about 20 megabits per second, a rate suitable for transmission through a 6 MHz terrestrial VHF/UHF channel.

The compression and digital transmission technology developed for ATV has been adapted for digital transmission of conventional television; this is known as *standard-definition television* (SDTV). MPEG-2 compression and digital transmission allow a broadcaster to place about four digital channels in the bandwidth occupied by a single analog NTSC signal. Digital television services are already deployed in direct broadcast satellite (DBS) systems and are expected soon in cable television (CATV).

With the advent of HDTV, 16:9 widescreen variants of conventional 525/59.94 and 625/50 component video have been proposed and even standardized. In studio analog systems, widescreen is accomplished by having the active picture represent 16:9 aspect ratio, but keeping all of the other parameters of the video standards. Unless bandwidth is increased by the same 4⁄3 ratio as the increase in aspect ratio, horizontal detail suffers.

In digital video, there are two approaches to achieving 16:9 aspect ratio. The first approach is comparable to the analog approach that I mentioned a moment ago: The sampling rate remains the same as conventional component digital video, and horizontal resolution is reduced by a factor of 3⁄4. In the second approach, the sampling rate is increased from 13.5 MHz to 18 MHz. I consider all of these schemes to adapt conventional video to widescreen be unfortunate: None of them offers an increase in resolution sufficient to achieve the product differentiation that is vital to the success of any new consumer product.