26
Graphics Devices

The first electronic computers were built in the 1940s, during and after World War II, and were used mostly for cryptanalysis (code breaking) and to compute firing tables. These are numeric applications, which require CPU time but use little input or output. Already in the 1950s, after using computers for just a few years, computer designers and users realized that computers can also be used for nonnumeric applications. A computer can compile its own programs, it can process text, and can store and edit images. Such applications generally involve large quantities of input and output data, and this created the need for input and output devices. Initially, the only input devices were the card and paper-tape readers, and the only output devices were the printer and the paper-tape punch. With the advent of computer graphics, however, these devices were insufficient, and new, graphics-oriented devices were developed.

Today, the most-important graphics output device is the display monitor (LCD or CRT) and the most-important graphics input device is the digital camera (although for some it may be a mouse or a scanner). Other important graphics devices are the printer (inkjet or laser), mouse, scanner, and plotter. These devices and others are described in this chapter, some in much detail.

26.1 Displays

A high-resolution, color display monitor is arguably the most important graphics output device. Experts sometimes claim that the absence of such devices in the 1950s and 1960s (and their high prices in the 1970s) were the main reasons for the initial slow progress of computer graphics.

For many years, the CRT was the dominant type of display monitor, but liquid crystal displays (LCDs) became available in the 1970s and have steadily improved since. Today (early 2011), LCDs combine low prices with low weight, high resolutions, high

D. Salomon, *The Computer Graphics Manual*, Texts in Computer Science,
DOI 10.1007/978-0-85729-886-7_26, © Springer-Verlag London Limited 2011

contrast, and low power consumption that together make this type the display of choice of most computer users and graphics professionals.

This section discusses the main features of display monitors, and it is followed by detailed descriptions of CRT and LC displays. Section 6.15 discusses the special autostereoscopic display.

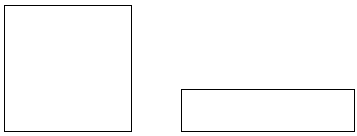
Common Display Standards and Resolutions

Table 26.1 lists many standard resolutions of display monitors (CRT, LCD, TV, and mobile devices). As display technologies mature, production yields rise, and prices plummet, bigger display sizes and higher pixel resolutions can be expected.

Standard	Resolution	Typical Use
CGA (Color Graphics Adapter)	320×200	Aspect ratio 8:5
QVGA (Quarter VGA)	320×240	Telephone displays
VGA (Video Graphics Array)	640×480	Older CRTs
WVGA (Wide VGA)	800×480	LCD projectors
SVGA (Super VGA)	800×600	1987 extension of VGA
PAL (Phase Alternating Line)	768×576	Analog TV, ratio 4:3
WSVGA (Wide SVGA)	1024×600	Mobile PCs, netbooks
HD720	1280×720	Video games, cameras
XGA (Extended GA)	1024×768	15" and 17" CRT, 15" LCD
SXGA (Super XGA)	1280×1024	15" 17" CRT, 17" 19" LCD
SXGA+(Super XGA)	1400×1050	Aspect ratio 4:3
UXGA (Ultra XGA)	1600×1200	19", 20", 21" CRT, 20" LCD
QXGA (Quad XGA)	2048×1536	21" and larger CRT
WXGA (Wide XGA)	1280×800	Wide aspect 15.4" laptops LCD
WXGA+ (Wide XGA+)	1440×900	Widescreen 19" LCD
WSXGA+ (Wide SXGA plus)	1680×1050	Wide 20" LCD
WUXGA (Wide Ultra XGA)	1920×1200	Wide 22" and larger LCD
HD1080 (popular HD format)	1920×1080	Wide aspect ratio 16:9
2K (2000)	2048×1080	Digital projectors 17:9
QWXGA (Ultra-Widescreen)	2048×1152	Wide 23" and larger LCD
QSXGA (Quad Super eXtended GA)	2560×2048	Grayscale monitors

Table 26.1: Current Display Standards and Resolutions.

Common sense demands that display makers will advertise the sizes of each model in their product line. They do that, but in a confusing, ambiguous manner. Instead of specifying the width and height of a display monitor, a manufacturer lists one number, the size of the diagonal. Clearly, one number is not enough to specify the geometry of a rectangle, as the following diagram (where the two rectangles have the same diagonal) aptly illustrates.



Aspect Ratio and Diagonal

A potential user/buyer who wants to know the dimensions of a display monitor before making a purchase, must go to a store and measure the device, or look for reviews on the Internet, or locate the aspect ratio of the monitor in the manufacturer's literature and use it and the diagonal to compute the width and height. (The aspect ratio is the ratio of width to height. It is normally greater than 1, and is specified either as a single number, such as 1.33 or as a ratio such as 4:3.) Older CRT displays in televisions and computers have an aspect ratio of 4:3, while newer, wider displays feature an aspect ratio of $16:9 \approx 1.78$ (and sometimes 16:10 or 15:9).

As if this confusion isn't enough, the meaning of the term "diagonal" depends on the display type (Figure 26.2). The diagonal specified for an LCD is that of the viewable screen, while the diagonal listed for a CRT display is measured from the outside edges of the entire cabinet, not just the screen (this is a tradition from the old days of television, when makers tried to exaggerate the sizes of their products). Thus, a 17" LCD has about the same diagonal as a 19" CRT. For desktop computers, common LCD sizes today are 17–24 inches, while the screens of laptop (or notebook) computers are somewhat smaller. Large screens require higher resolutions, otherwise their pixels become too large and images appear pixelated.

Figure 26.2: LCD and CRT Diagonal Sizes.

Older monitors had a fixed frequency, they could operate only at a single resolution and refresh rate. The computer (specifically, the graphics card, video card, or graphics adapter in the computer) had to generate the precise video signal required by the monitor. Today's monitors are of the MultiSync type, originated by NEC in the 1990s.

They can receive one of several frequencies and vary their resolution and refresh rate depending on the video signal sent from the graphics card. This book is written on a Macintosh computer with a Samsung SyncMaster LCD that can operate at refresh rates of 56.3, 59.9, 60 and 60.3 Hz and can vary its pixel resolutions from a low of 640×480 to a high of $1{,}920 \times 1{,}080$ (although only the highest resolution produces clear, sharp images on this screen).

It should be noted, however, that an LCD monitor has a native resolution, where images look best. Typical native resolutions for LCDs are the following: For 17-inch displays, $1{,}024 \times 768$, for 19-inch LCDs, $1{,}280 \times 1{,}024$, for 20-inch LCDs, $1{,}600 \times 1{,}200$, and for 22–24 displays, $1{,}920 \times 1{,}080$.

Analog and Digital Connections

The video signal sent by the graphics card to the display monitor can be analog or digital. CRTs are analog devices, so they require an analog input signal. The graphics card in the computer must be able to convert the digital image data to analog and send it to the CRT on a VGA (Video Graphics Array) cable. This cable connects to the CRT with a connector that has 15 pins arranged in three rows.

In contrast, LCDs are digital and most of them receive digital information from the graphics card through a DVI (Digital Visual Interface) cable. Special hardware (known as transition minimized differential signaling, or TMDS) in the monitor examines the DVI signal, and based on the current resolution and refresh rate of the monitor, spreads the signal to optimize the image quality on the screen. There are two types of DVI cables and connectors, digital (DVI-D) and integrated (DVI-I, supports both digital and analog transmissions), and each can operate in a single link or a double link mode. The former is for resolutions of up to $1{,}920 \times 1{,}080$ and the latter is for higher resolutions, up to $2{,}048 \times 1{,}536$. Figure 26.3 shows the four types of DVI connectors. (There is also a DVI-A, analog only.)

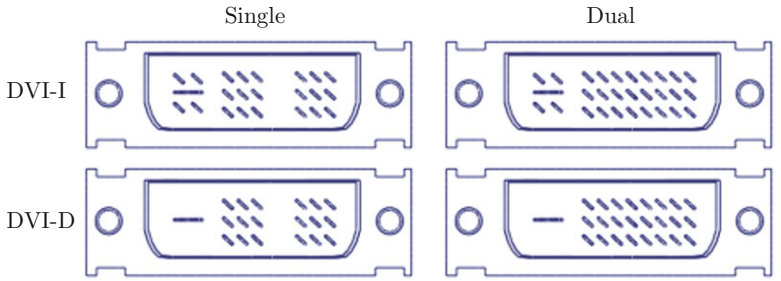


Figure 26.3: Four Types of DVI Connectors.

Color Depth

The term color depth (or color bit depth) refers to the number of bits devoted to each pixel. Each bit added to the color depth doubles the number of colors that can be specified. The following is a list of the color depth currently used with display monitors:

1. A color depth of 1 (one bit per pixel) corresponds to two colors, black and white, or foreground and background. Such a display is monochromatic (i.e., it displays a single color, black, on a white background).

2. With two bits per pixel, four colors can be specified. This color depth is known as CGA (Color Graphics Adapter).

4. Sixteen colors, also known as EGA (Enhanced Graphics Adapter).

8. This is a VGA color depth, which corresponds to $2^8 = 256$ colors. This was common in early personal computers.

16. The number of colors is $2^{16} = 65{,}536$ or 64 K. On most operating systems, this is referred to as "thousands of colors." The formal name of this color depth is XGA (Extended Graphics Array).

24. This is a true color, SVGA (Super Video Graphics Array) color depth, where a pixel can have one of $2^{24} = 16{,}777{,}216$ or about 16.8 million colors. In the operating system jargon this is known as millions of colors and it is the most common color depth today.

32. This color depth is known as True Color + Alpha Channel. Only 24 of the 32 bits are used to specify the color of a pixel, and the remaining eight bits are used to specify one of 256 levels of translucency of the pixel.

Viewing Angle

LCD monitors should be viewed head on as much as possible. When viewed from an angle (as when many people try to watch the same monitor), the image seems dimmer, then becomes unrecognizable, and sometimes even disappears. This limitation was more pronounced in the past, but today LCD makers claim lenient viewing angles of at most 60° (sometimes as much as 85°) off the perpendicular to the screen. This angle should also be the maximum for viewing above and below the display, not just from the left and right.

Other Features of LCDs

Brightness is an important feature of a display and this is especially important in an LCD, where at least half the light's intensity is absorbed by polarizers. Brightness is measured in units called nits (where a nit is one candela per square meter). Typical display brightnesses vary from 250 to 350 nits for general-purpose monitors and up to 500 nits for special monitors designed to play movies.

Contrast is another important feature of displays and is measured as a ratio of the brightest white to the darkest black. Ratios of 1000:1 or 1200:1 are common today and are judged satisfactory by most users.

The term "response rate" indicates how fast the pixels of a display can change their color. A slow response rate produces ghosting artifacts on the screen, especially when a fast animation is displayed.

A good display monitor should be adjustable. The user should be able to vary the height of the display above the table, to tilt the display up and down, swivel it to either side, and even rotate it by 90° from its normal landscape orientation (where the width is greater than the height) to a portrait position (where the height is greater). In addition,

because of their smaller weight, LCD monitors should have special brackets to facilitate wall mounting. Only the high-end, expensive LCDs feature full adjustability.

LCDs vs. CRTs

LCD monitors are popular nowadays because of the following reasons:

- They are thin and lightweight, easier to grab, move, and mount on a wall.
- They are adjustable because of the above reason.
- They consume less than half the power required by a CRT, typically around 40–60 watts depending on size.
- Some people find LC displays easier on the eyes. A CRT always has some flicker because of the movements of its electron beam, while an LCD turns individual pixels on and off as needed.

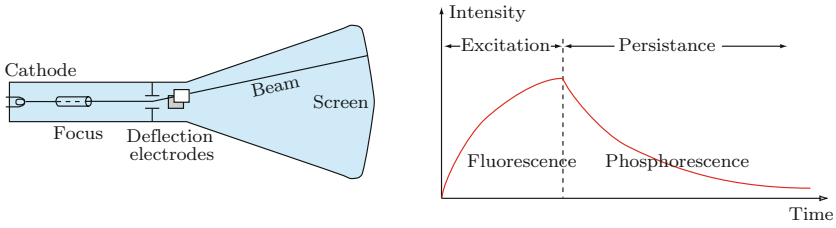
However, CRTs have a number of advantages, which make them a better choice for certain users.

- A CRT is less expensive.
- CRTs are better at displaying vivid colors.
- Refresh rates of LCDs are generally lower than those of CRTs, which may result in annoying ghosting and blurring on an LCD.
- It is easier to vary the resolution of a CRT, because its screen is uniform, whereas the screen of an LCD consists of built-in pixels.
- A CRT is easier to clean and harder to damage than an LCD.

26.2 The CRT

An image can be displayed by a computer on a CRT in two different ways, *raster scan* and *vector scan* (the latter is sometimes called *random scan*). Both methods can use a CRT as the output device (Figure 26.4a), but they differ in many respects. They control the electron beam in the CRT in different ways, they represent the graphics data in memory differently, and they also require different hardware circuits to interface the computer to the CRT.

A CRT (cathode ray tube) is the same kind of tube used in older television sets. It has an electron gun (the cathode) that emits a stream of electrons (Figure 26.4a). The front surface is positively charged, so it attracts the electrons, and is coated with a phosphor compound that emits light when hit by the beam. The flash of light lasts only a fraction of a second, so, in order to achieve a stable, constant display, the picture has to be refreshed about 20 times a second. (The actual refresh rate depends on the *persistence* of the compound (Figure 26.4b). For certain types of work, such as



(a) (b)

Figure 26.4: (a) CRT Operation. (b) Persistence.

architectural drawing, long persistence is acceptable. For animation, short persistence is a must.)

The electron beam can be turned off and on very rapidly. It can also be deflected horizontally and vertically by two pairs (X and Y) of deflection electrodes. Displaying a single point on the screen is done by turning the beam off, moving it to the part of the screen where the point should appear, and turning it on again. These operations are performed by special hardware (the CRT controller or graphics interface) that receives information from the program.

26.2.1 Standard Television

A home television set is based on one of three international standards. The standard used in the United States is called NTSC (National Television Standards Committee), although the new digital standard (Section 26.2.2) is slowly becoming popular. NTSC specifies a television transmission of 525 lines (today, this would be $2^9 = 512$ lines but because television was developed before the advent of computers and binary numbers, the NTSC standard is not based on powers of 2). In practice, though, only 480 lines are visible on the screen. The aspect ratio (height/width) of a television screen is $3:4$, which is why each line is equivalent to $\frac{4}{3}480 = 640$ pixels. The resolution of a standard television set is therefore 480×640. This may be considered, at best, medium resolution. (This is the reason why text is so hard to read on a standard television.) For more information on standard television, see [Pritchard 77].

Many computer graphics applications cannot therefore use a standard television set as a screen. CRTs designed for computer use start at a resolution of $1\text{K} \times 1\text{K}$ and can go much higher.

To display a complex image, the program has to compute and render it, and then store it in memory. It then starts the CRT controller, which displays every element of the image and is also responsible for refreshing it. The specific steps depend on the scan method used.

A word on color. A typical color CRT employs the *shadow mask* technique (Figure 26.5). There are three guns emitting three separate electron beams. Each beam is associated with one color but the beams themselves consist of electrons and have no color. The beams are adjusted such that they always converge a short distance behind the screen. By the time they reach the screen, they have diverged a bit, and they strike three different (but very close) points.

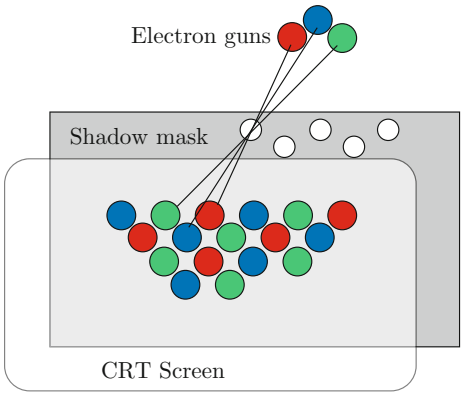


Figure 26.5: A Shadow Mask.

The screen is coated with dots made of three types of phosphor compounds that emit red, green, and blue light, respectively, when excited. At the plane of convergence, there is a thin, perforated metal screen; the shadow mask. When the three beams converge at a hole in the mask, they pass through, diverge, and hit three points coated with different phosphor compounds. The points glow at the three colors and the observer sees a mixture of red, green, and blue whose precise color depends on the intensities of the three beams (see discussion of spatial integration in Sections 21.6 and 2.27). When the beams are deflected a little, they hit the mask and are absorbed. After some more deflection, they converge at another hole and hit the screen at another triplet of points.

26.2.2 High-Definition Television

The original NTSC standard, for black-and-white television transmissions, was created in 1953, after four years of testing. It specifies the shape of the signal sent by a TV transmitter. This is an analog signal, with amplitude that goes up and down during each scan line in response to the black and white parts of the line. Color was later added to this standard, but it had to be added such that black-and-white television sets would be able to display the color signal in black and white. The result was phase modulation of the black-and-white carrier, a kludge (TV engineers call it NSCT "never the same color twice").

With the explosion of computers and digital equipment in the 1970s and 1980s came the realization that a digital signal is a better, more reliable way of sending images over the air. In such a signal, the image is sent pixel by pixel, where each pixel is represented by a number specifying its color. The digital signal is still a wave, but the amplitude of the wave no longer represents the image. Rather, the wave is *modulated* to carry binary information. The term modulation means that something in the wave is modified to distinguish between the zeros and ones being sent. An FM digital signal, for example, modifies (modulates) the frequency of the wave. This type of wave uses one frequency to represent a binary 0 and another to represent a binary 1.

History of DTV: The Advanced Television Systems Committee (ATSC), established in 1982, is an international organization developing technical standards for advanced video systems. Even though these standards are voluntary, they are generally adopted by the ATSC members and other manufacturers. There are currently about 120 ATSC member companies and organizations, representing the many facets of the television, computer, telephone, and motion picture industries.

The ATSC Digital Television Standard adopted by the United States Federal Communications Commission (FCC) is based on a design by the Grand Alliance (a coalition of electronics manufacturers and research institutes) which was a finalist in the first round of DTV proposals under the FCC's Advisory Committee on Advanced Television Systems (ACATS). The ACATS is composed of representatives of the computer, broadcasting, telecommunications, manufacturing, cable television, and motion picture industries. Its mission is to assist in the adoption of an HDTV transmission standard and to promote the rapid implementation of HDTV in the United States.

The ACATS announced an open competition; anyone could submit a proposed HDTV standard, and the best system would be selected as the new television standard for the United States. To ensure a fast transition to HDTV, the FCC promised that every television station in the nation would be temporarily loaned an additional channel of broadcast spectrum.

The ACATS worked with the ATSC to review the proposed DTV standard, and gave its approval to final specifications for the various parts—audio, transport, format, compression, and transmission. The ATSC documented the system as a standard and ACATS adopted the Grand Alliance system in its recommendation to the FCC in late 1995.

In late 1996, corporate members of the ATSC reached an agreement on the DTV standard (Document A/53) and asked the FCC to approve it. On December 31, 1996, the FCC formally adopted every aspect of the ATSC standard except for the video formats. These video formats nevertheless remain a part of the ATSC standard and are expected to be used by broadcasters in the foreseeable future.

HDTV Specifications: The NTSC standard in use since the 1930s specifies an interlaced image composed of 525 lines where the odd numbered lines (1, 3, 5, ...) are drawn on the screen first, followed by the even numbered lines (2, 4, 6, ...). The two fields are woven together and drawn in 1/30 of a second, allowing for 30 screen refreshes each second. In contrast, a noninterlaced picture displays the entire image at once. This *progressive scan* type of image is what's used by today's computer monitors.

The digital TVs that have been available since mid-1998 use an aspect ratio of 16:9 and can display both the interlaced and progressive-scan images in several different resolutions—one of the best features of digital video. These formats include 525-line progressive scan (525P), 720-line progressive scan (720P), 1,050-line progressive scan (1050P), and 1,080-interlaced (1080I), all with square pixels.

The NTSC standard calls for 525 scan lines and an aspect ratio of 4:3. This implies $\frac{4}{3} \times 525 = 700$ pixels per line, yielding a total of $525 \times 700 = 367,500$ pixels on the screen. (This is the theoretical total since only 480 lines are actually visible.) In comparison, a DTV format calling for 1,080 scan lines and an aspect ratio of 16:9 is equivalent to 1920 pixels per line, bringing the total number of pixels to $1{,}080 \times 1{,}920 = 2,073,600$, about 5.64 times more than the NTSC interlaced standard.

◇ **Exercise 26.1:** The NTSC aspect ratio is $4:3 = 1.33$ and that of DTV is $16:9 = 1.77$. Which one looks better?

In addition to the 1,080 × 1,920 DTV format, the ATSC DTV standard calls for a lower-resolution format with just 720 scan lines, implying $\frac{16}{9} \times 720 = 1{,}280$ pixels per line. Each of these resolutions can be refreshed at one of three different rates: 60 frames/second (for live video) and 24 or 30 frames/second (for material originally produced on film). The refresh rates can be considered *temporal resolution*. The result is a total of six different formats. Table 26.6 summarizes the screen capacities and the required transmission rates of the six formats. With high resolution and 60 frames per second, the transmitter must be able to send 124,416,000 bits/sec (about 14.83 Mbytes/sec), which is why this format uses compression. (It uses MPEG-2. Other formats can also use this compression method.) The fact that DTV can have different spatial and temporal resolutions allows for trade-offs. Certain types of material (such as fast-moving horse or car races) may look better at high refresh rates even with low spatial resolution, while other material (such as museum-quality paintings) should ideally be watched in high spatial resolution even with low refresh rates.

Lines × pixels	Total # of pixels	Refresh rate		
		24	30	60
1080 × 1920	2,073,600	49,766,400	62,208,000	124,416,000
720 × 1280	921,600	22,118,400	27,648,000	55,296,000

Table 26.6: Resolutions and Capacities of Six DTV Formats.

Digital television (DTV) is a broad term encompassing all types of digital transmission. HDTV is a subset of DTV indicating 1080 scan lines. Another type of DTV is Standard Definition Television (SDTV) which has a picture quality slightly better than a good analog picture. (SDTV has resolution of 640×480 at 30 frames/sec and an aspect ratio of 4:3.) Since generating an SDTV picture requires fewer pixels, a broadcasting station will be able to transmit multiple channels of SDTV within its 6-MHz allowed frequency range. HDTV also incorporates Dolby Digital sound technology to bring together a complete presentation.

26.2.3 The Light Pen

Sections 2.2.3 and 26.2 explain why a CRT display has to be refreshed often. One of the earliest graphics input devices, the light pen, was based on this feature. The light pen is a stylus (or wand) that is placed by the user at a point of interest on the CRT screen. When the user presses a button on the pen, the pen interrupts the computer, and the graphics controller determines the screen position of the pen and stores it in memory or in a special register.

To a casual observer it seems that the light pen emits light, but in fact it senses the light emitted by the screen. When the light pen is held at a certain location on the screen, it senses the change in brightness when that location is refreshed. The pen interrupts the computer, transferring control of the CPU to the graphics controller. The

controller knows what location on the screen was last refreshed, and it stores the address of that location in a special register or in memory, for use by any graphics software. This interrupt occurs once every refresh as long as the pen is positioned close enough to the screen to detect the change in brightness.

A light pen works best with a CRT monitor, but makers of monitors are currently trying to adapt this concept to LCD displays.

The light pen was first used in the early 1950s, as part of the historically-important Whirlwind project. It became popular in the early 1980s, as an accurate, inexpensive graphical input device, but later fell out of use in favor of the mouse. The light pen is not ergonomic because the user has to raise his arm and hold it in front of the screen for each pointing operation.

26.3 LCDs

Nowadays (in early 2011) liquid-crystal displays (LCDs) are everywhere. We use them with desktop and laptop computers and we see them in digital clocks and watches, microwave ovens, printers, CD players, cell-phone screens, televisions, and electronic billboards. We therefore tend to forget that only a decade ago, this type of display was still in its infancy and was used mostly in wristwatches and calculators. LCD displays are popular because they are lightweight, small and thin, they draw little electrical power, and are reliable.

Before we start with liquid crystals and their use in displays, here are a few facts about light and its properties (see also Section 21.1). Light can be understood as either an electromagnetic wave or as a stream of particles (termed photons) that have energy and momentum, but no mass. An electromagnetic wave consists of electric and magnetic fields that are perpendicular to each other and that propagate at the speed of light. (It should be mentioned that a field is a vector; it has direction and magnitude.)

When we think of light as a wave, it has amplitude (the intensity of the light), frequency (an attribute that our eyes and brain interpret as color), and polarization (the direction in which the electric field vibrates). It is convenient to think of a photon as a particle that always moves at the speed of light and has two attributes, frequency and polarization. Scientists generally agree that there is no way to visualize a photon, and it is therefore preferable to think of it in terms of its attributes.

In general, a beam of light can be considered a stream of photons with different frequencies and different polarizations (i.e., each photon is polarized in a different direction). There are materials, called polarizers, that can transmit only those photons that are polarized in a certain direction, while absorbing all other photons. Light that passes through a polarizer is polarized on one direction and is also normally dimmer than the original light, having lost some of its photons in the polarizer.

Now we are ready for LCDs. From school, as well as from everyday life, we are familiar with three common states of matter, solids, liquids, and gases. We know that crystals are solid and are often hard. The molecules in a crystal maintain their orientation and stay in their positions in the crystal. The molecules in a liquid move around and also change their orientations all the time. Crystals and liquids are very different, so what is a liquid crystal?

Discovered in 1888, liquid crystals are unusual substances that are closer to liquids than to crystals but combine properties of both. There are several types of liquid crystals and the one used in electronic displays is called nematic (more precisely, twisted nematic or TN). This type of liquid crystal has unusual features. When it is grown in the lab, molecule by molecule, each added molecule is twisted by a small angle relative to its predecessor (Figure 26.7). Imagine light that is polarized in the direction of the leftmost molecule (vertically in the figure) entering from the left. As the light propagates to the right through layer after layer of molecules, its polarization is gradually twisted. Given enough layers, the light emerging from the right edge of the liquid crystal has twisted its polarization by 90° (in the figure, it is polarized horizontally).

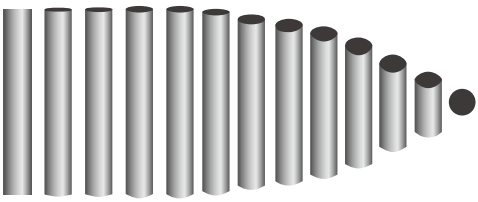


Figure 26.7: A Twisted Nematics Liquid Crystal.

Now comes the important part. When voltage is applied to a TN, its molecules rotate to decrease the angle between them. They become more aligned and thus twist the light's polarization by a smaller angle, depending on the voltage. With high voltage, it is possible to align the molecules perfectly and end up with no polarization twist. It is this *field effect* that turns liquid crystals into light valves and makes them so useful for displays.

With this in mind, it is easy to grasp the principle of LCDs (Figure 26.8). Light is produced by a miniature fluorescent tube (or a cold cathode fluorescent lamp) located at the rear (bottom) of the display. Special optical devices (a lightguide and several diffusers) spread the light evenly over the entire back area of the display. On its way to the front (top) of the display, the light, indicated by (a) in the figure, passes through a bottom vertical polarizer (b), where half of it is absorbed. The other half emerges vertically polarized (c) and enters an array (d) of twisted nematic liquid crystals (the one shown here corresponds to a pixel of the display). The crystals twist the light's polarization by 90° from vertical to horizontal, and the horizontally polarized light then passes through a top horizontal polarizer (e) and finally hits the front (top) of the display, which then turns bright. The front of the display (termed the projection surface or simply the screen) can be glass or plastic, but in large LCDs made for televisions and computers, the front panel is often a plastic film that enhances the display.

The liquid crystals are arranged in rows and columns, such that each corresponds to a pixel of the display, and each can be controlled independently by the display hardware. In order to create a black dot (i.e., a pixel) on the display, electrical voltage must be applied to the appropriate liquid crystal, which causes it to lose its twist property. The narrow beam of light that passes through that crystal is not twisted and emerges from the array vertically polarized, to be absorbed by the top horizontal polarizer. This beam

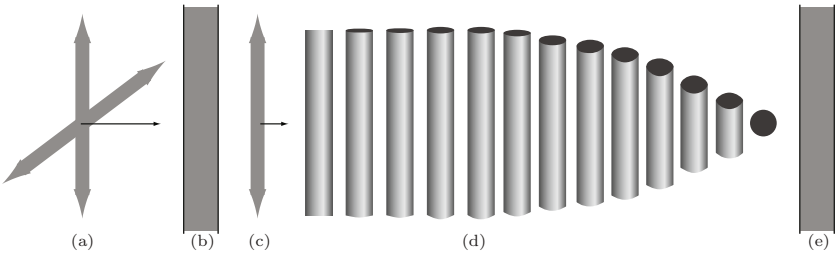


Figure 26.8: Major Components of an LCD.

of light never reaches the screen, which results in a black dot, a pixel, displayed on the screen.

When lower voltage is applied to a liquid crystal, it twists the polarization of the light that passes through it by less than 90°, resulting in a gray pixel.

In a color LCD, each pixel consists of three liquid crystals that correspond to the red, green, and blue components of the color of the pixel. Each crystal (which now corresponds to a subpixel rather than a whole pixel) receives its own voltage and may twist the narrow light beam that passes through it by a different amount. Normally, each liquid crystal subpixel receives one of 256 different voltages and can therefore create one of 256 shades of gray. A color filter (pigment, dye, or metal oxide filter) is placed in front of each crystal, so the light that passes through a set of three liquid crystal subpixels becomes a mixture of the three colors.

Figure 26.9 shows four common subpixel organizations in (from left to right) television CRT, typical LCD, PC CRT, and XO-1 LCD.

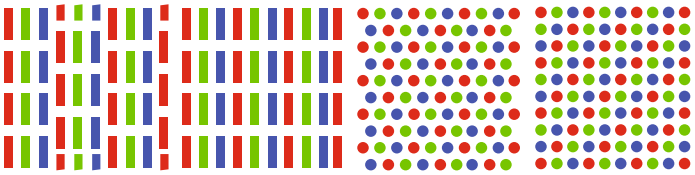


Figure 26.9: Four Common Subpixel Organizations.

A color LCD requires three times the number of liquid crystals than a grayscale LCD. As an example, an LCD with a resolution of $1{,}024 \times 768 = 786{,}432$ pixels must have a total of $786{,}432 \times 3 = 2{,}359{,}296$ liquid crystals and also the extra hardware needed to select each crystal independently. Thus, LCDs constitute an amazing technical achievement and are one of many complex devices that we commonly use and take for granted. It is therefore not surprising that LCDs are notoriously difficult to manufacture and quality control has always been a problem. A pixel may be stuck in an on or off position or in one color, and each LCD maker has its own policy on how many such dead pixels are allowed in a display that leaves the factory. Too many dead pixels, and the

entire display panel has to be rejected, which may happen in up to 40% of the output of an assembly line.

Most large LCDs are of the active matrix type. In this type of display there is a grid consisting of two sets of wires, for the rows and columns of the display. In order to set the pixel at row i and column j to a shade g of gray, the display hardware selects wire i in one set and wire j in the other set, and sends an electrical charge proportional to g (typically, one of 256 possible charges). The charge is saved in a small capacitor (referred to as a thin-film transistor or TFT) and is kept there until the hardware decides to modify the intensity of the pixel.

Many small LCDs do not include a light source. A typical example is a small calculator. Such a device often does not have a battery and it depends on the little electrical power generated by a small, built-in photovoltaic cell. This power is enough to run the electronic circuits of the calculator, but is not enough to also illuminate the display. The display of such a device must use the available external light and it therefore cannot be read in darkness. The LCD of such a device has in its back a mirror instead of a light source. Light enters the display from the front, passes through (or is absorbed by) the polarizers and the liquid crystal array, is bounced by the mirror, and comes back to the front minus those parts that have been absorbed by the polarizers.

Cleaning the screen. In a CRT display, the front panel (screen) is made of glass and can be cleaned like any other glass surface. The screen of an LCD is often made of special plastic that can easily be scratched and damaged if improperly cleaned. (For example, pressing hard on the screen can cause several pixels to burn and die.) It should be cleaned with a soft cloth, of the type used to clean camera lenses and eyeglasses. Paper towels, toilet paper, rugs, and tissue paper should be avoided. If a gentle, dry wipe is not enough, moisten the cloth with a little distilled water.

LCD History

Liquid crystals were discovered in 1888, by Friedrich Reinitzer, a botanist. It was not until 1968, about 80 years later, that the first LCD was made (by RCA). In 1970, a patent was filed by Hoffmann-LaRoche in Switzerland on the twisted nematic liquid crystal field effect, and in 1971 an identical patent was filed by James Fergason in the United States. A year later, in 1972, T. Peter Brody of Westinghouse's Research Labs made the first active-matrix LCD, and this started the LCD revolution. The 4th quarter of 2007 saw more LCD televisions sold than CRT-based televisions, and in 2008 it became clear that CRT-based displays are a thing of the past.

Current displays use variations of the basic twisted nematics (TN) liquid crystals, such as super twisted nematics (STN), dual scan twisted nematics (DSTN), ferroelectric liquid crystal (FLC), and surface stabilized ferroelectric liquid crystal (SSFLC).

26.4 The Digital Camera

A digital camera (or a digicam) is a camera where light is captured by electronic sensors instead of on photographic film. As a result, the images produced by such a camera are digital; they consist of pixels. A digital camera has the basic components of a film camera, most importantly a lens (often with zoom capability), an autofocus mechanism, a diaphragm, a shutter mechanism, and a timer. These components work together to admit the correct amount of light to the light sensing medium, which is an array of electronic sensors. Thus, the chief differences between digital and film cameras are (1) the light sensing medium and (2) the extensive use of electronics, which includes a display screen and a memory card. Digital cameras have the following advantages over film cameras:

- The image is displayed immediately (on a small screen in the back of the camera) after being captured. This enables the photographer to identify, delete, and retake bad pictures.
- Hundreds (often even thousands) of digital images can be stored in the camera on a small reusable memory card. Any image stored in the card can be displayed in the camera, transferred to a computer, and deleted from the card. In contrast, a typical consumer film cartridge has room for only 24 exposures and cannot be reused.
- Most digital cameras can record video (with sound) in addition to still images.
- In addition to the zoom produced by lens movement (optical zoom), a digital camera often features digital zoom, produced inside the camera by software that scales the image. New pixels are computed by interpolating neighboring pixels. However, digital zoom is often judged worthless by expert photographers.
- Higher-quality digital cameras can perform simple image editing operations inside the camera.
- In addition to the consumer market, specialized digital cameras are made for and used in PDAs, notebook computers, cellular telephones, security devices, and telescopes.
- The Polaroid PoGo camera features a built-in printer. After taking a snapshot, the user can crop or edit the image with built-in editing tools, add a fun border, and then print the picture with the built-in ink-free printer. The printing technology, known as ZINK (zero ink), uses a special 2×3-inch paper with three internal layers of cyan, yellow, and magenta dye crystals, sandwiched between a base layer and an overcoat. The printer employs heat to activate and colorize the dye crystals, and a fully printed, durable image emerges out of the printer in about a minute.

The following is a list of the main types of consumer digital cameras:

- Compact cameras. Those are aimed at the casual photographer, the so-called point-and-shoot market. They are small and lightweight (the smallest ones are designated subcompacts and are also very thin). In order to achieve these goals and also reduce costs, compact cameras sacrifice picture resolution and quality, have restricted video and zoom capabilities, and also eliminate most or all advanced features. The pictures are stored only in the lossy JPEG format and the built-in flash is small and weak. At the

time of writing (mid 2010), compact cameras have a capacity of 8–12 Mpixels, but these numbers grow steadily. Higher-quality compact cameras offer features such as larger screens (currently 2.5 in, 2.7 in, and 3 in), wider optical zoom range (currently up to 18 or even 24), image stabilization (sensors that compensate for camera shake by (1) shifting elements within the lens, (2) shifting the sensor array, and (3) adjusting the amount of light), and a wide ISO range (a feature that mimics the sensitivity of film).

An important, but often misunderstood feature of compact cameras is the small physical size of the sensor array. It is typically only six mm on the diagonal, which makes for very small (actually, microscopic) individual sensors. A small sensor simply does not have enough surface area to collect much light, which is why compact cameras do not perform well under low-light conditions. On the other hand, low light implies more depth-of-field (DOF) at a given aperture (Section 26.4.7).

- Hybrid cameras. These have recently been developed for users who want more capabilities and options than are offered by compact cameras, but don't want the size, weight, and price of a DSLR. Hybrid cameras are smaller and lighter than a typical DSLR, they have smaller lenses, and they lack the pentaprism mirror system that characterizes an SLR. As a result, such a camera does not have an optical viewfinder. Instead, it uses an electronic viewfinder from the sensor array. This feature implies that a hybrid camera can shoot video, which gives it an edge over a DSLR.

- Digital single lens reflex cameras (DSLRs). The SLR camera design has been around for decades, so it was natural for camera makers to adopt it to the realm of the digital. A DSLR is bulkier and more expensive than a compact camera, but has features that compensate for this. These cameras have wider optical zoom, interchangeable lenses, and a large sensor array, typically 18–36 mm on the diagonal. A large sensor array implies large sensors, not necessarily more sensors. The advantage of a larger sensor is that it can gather more light, which gives the DSLR better performance in low-light situations, but also reduces the depth-of-field (DOF) at a given aperture (Section 26.4.7).

A minor point is the distinctive clack sound made by a DSLR when it takes a picture. This is caused by the mechanical movement of the mirror which is flipped out of the way and then brought back in.

In addition to these three types, there are other classes of digital cameras, such as bridge cameras, live preview cameras, and line-scan cameras. The latter type contains a single row of light sensors, instead of a two-dimensional array. The camera scans a line on the object, waits for the object to move a bit, and repeats. The pixel data generated by those line scans is sent to a computer where a two-dimensional image is created row by row. This type of camera is suitable for industrial purposes, where products constantly move on a conveyor belt and have to be scanned and automatically checked for defects or routed differently.

26.4.1 History of Digital Cameras

As early as the 1960s, researchers developed and patented arrays of electronic (then referred to as solid-state) light sensors. Those were very bulky and were used for special applications such as tracking spacecraft. The first practical digital camera was built in 1975 at Eastman Kodak by Steven Sasson. The sensor array in this camera consisted of 10,000 CCD devices (developed by Fairchild semiconductors two years earlier). The

camera was big and heavy, and the (black and white) images were saved on a cassette tape. The first image was taken in December, 1975. A picture of this camera is available at [msnbc.camera 09].

It took until 1981 for the first handheld digital camera, the Sony Mavica (Magnetic Video Camera) to make its debut (years later, Sony developed another Mavica that is completely different). The original Mavica had a sensor array, but it did not convert the electrical charges on the sensors to numbers. Instead, they were translated into analog electrical signals (similar to television signals) that were written on a 2-in magnetic floppy disk. As a result, images produced by this camera had noticeable scan lines and were similar to television images. Other analog cameras were developed in the 1980s, but their costs and poor image quality made them unsuitable for consumer applications. They were useful only for special applications such as newspaper and television reporting (the images could be sent on telephone lines, and their resolution was similar to that of newspaper graphics).

It seems that the first fully digital camera was the model DS-1P, made by Fuji in 1988. It was impractical and it probably was never sold commercially. In 1990, Dycam Inc. made and sold its mode 1, a digital camera based on a 376×240 CCD sensor array. This camera produced grayscale (256 levels) images, stored up to 32 pictures internally in a 1 MB memory. The pictures could later be transferred to a computer. Other camera makers soon followed, with the result that size, weight, and price dropped, while resolution and number of colors increased steadily.

The adoption of the JPEG and MPEG compression standards in 1988 and the development of small, inexpensive LCDs also helped to accelerate the development of digital cameras in the 1990s.

The first digital camera that also took videos made its debut in 1995 and the first megapixel cameras appeared in 1997. The first DSLR, the Nikon D1 (2.74 megapixel), was introduced in 1999. It was too expensive for casual users, but was affordable by professional photographers, especially since they could use the same Nikon lenses they already owned.

26.4.2 Camera Resolution

The term resolution is usually understood to mean the total number of light sensors, but should really refer to the width and height of the sensor array. The number of sensors (or pixels) in digital cameras has grown from 2–3 Mpixel in the late 1990s to around 8–12 Mpixel today (mid 2011) although expensive cameras may have up to 60 Mpixels). The sensors are arranged in a rectangular array and the dimensions of the array (measured in sensors) determine both the total number of pixels and the aspect ratio (width over height). Instead of looking only at the total number of sensors, a potential camera user should consider the dimensions of the sensor array. Examples of current sensor array dimensions are $2{,}012 \times 1{,}324$ (a total of 2.74 Mpixel and an aspect ratio of 3:2), $3{,}072 \times 2{,}048$ (a total of 6.3 Mpixel and an aspect ratio of 3:2), and $3{,}648 \times 2{,}736$ (a total of 10 Mpixel and an aspect ratio of 4:3).

Ten million sounds like a large number, but images are two dimensional, which is why doubling the number of pixels of an image does not double the size of the image. Given n^2 pixels, they form a square image of n units on a side. Doubling the number of pixels to $2n^2$ increases each side of the square image to $\sqrt{2n^2} = \sqrt{2}n \approx 1.4n$, approxi-

mately 40% bigger. Thus, once we spread the pixels over rows and columns, there may not be enough of them to cover a single printed page. Here is what a total of 10 Mpixel implies for printing. Suppose we want to print a 10 Mpixel image at the reasonable resolution of 300 dpi (300 dots per inch on the paper). Each square inch of paper will have $300 \times 300 = 90{,}000$ dots, so our 10 mega pixels can cover only 111 square inches, or an area of approximately 12.4×9 inches. This is a little more than the area of a standard American letter-size page.

If the same 10-Mpixel image has to be printed on a poster-size sheet of paper, say 2×3 feet, then two approaches suggest themselves as follows:

- Reduce the printing resolution. An area of 2×3 feet equals 864 square inches, so 10 million pixels provide 11,574 pixels per square inch, for a printing resolution of $\sqrt{11{,}574} \approx 108$ dpi. This sounds low and is certainly low for printing text, but images (grayscale and color, but not line drawings) have noise (i.e., the eye may not notice when several pixels, or even many pixels, have the wrong colors), and experience shows that images printed at such a low resolution do not appear degraded and do not feature ragged edges or other negative effects of low resolution.
- Photograph the original image in several overlapping parts, and then use software to stitch these parts into a single, large image with enough pixels to be printed at a reasonable resolution on a large sheet of paper. This approach is employed by photographers and artists who produce large panoramas.

It is also important to bear in mind that resolution is only one of the factors that affect the quality of a camera, the other factors being the quality of the lens, the physical size of a sensor, the filter array, and the demosaicing algorithm used by the camera (Section 26.4.5). A small sensor simply does not have the surface area to receive many photons and does not have the volume to hold much electrical charge, so when the output of a small sensor is digitized, it always results in a small number.

See also Section 26.9 for a discussion of scanner resolutions.

26.4.3 Light Sensors

An image sensor is a device that converts light energy to electrical energy. More specifically, it converts the energy of the photons impinging on it to electrical charge. Currently (early 2011), image sensors are either charge-coupled devices (CCD) or a complementary metal-oxide-semiconductor (CMOS) devices. These devices operate differently, but the final output is digital; the electrical charges are converted to numbers (normally 12-bit integers). A more detailed discussion of these devices is outside the scope of this book (in fact, of most books), and here they are simply referred to as light sensors, image sensors, or CCDs. (See Page 1226 for more information.)

Virtually all current consumer digital cameras, including DSLRs, are of the single-shot type. In this type, there is a single sensor array with a Bayer filter mosaic (Section 26.4.5). A variation of this type employs three sensor arrays, one for each color component, that are exposed simultaneously via a beam splitter.

Cameras designed for special applications, such as shooting stationary subjects, can be of the multi-shot type, where the sensor array is exposed several times in the same shot. This type cannot be used with moving subjects. One way to implement this type of camera is to have three filters and place a different filter in front of the array during

each exposure. Another option is to employ a single array with a Bayer filter and to physically move it, along the focus plane inside the camera, for each exposure in order to obtain a large number of pixels. In principle, it is possible to combine these versions of multi-shot; to expose a single array three times with different filters, and then move it and again expose it three times.

Sensor arrays are notoriously difficult to manufacture and are never perfect. Some sensors in an array may be more sensitive or less sensitive than others, or may even be dead. It took engineers many years to perfect the processes of making such sensors, and it wasn't until the early 2000s that prices of large (more than 5 Mpixel), reliable sensor arrays dropped to such a level that they began to be installed in inexpensive, compact home cameras.

Digital camera identification.

The fact that CCD sensor arrays are not perfect may be annoying to demanding photographers, but like many annoyances, it may have a useful side. The imperfections in any particular sensor array may be exploited to identify the camera that took a given image.

Given an image and a number of cameras, one of which is suspected to have taken the image, it may be possible to identify that camera by comparing the imperfections in its sensor array (the so called camera pattern noise) to the pixels of the given image. For each candidate camera, several test images are taken and are passed through a denoising filter to create a reference pattern of the camera's noise. This reference is then correlated with the pixel noise of the given image, and the correlation results are judged by a person.

This interesting work, which can also be applied to detecting forgeries in digital images, is described in [Lukáš et al. 06a,b]. The results are not absolute and there may be false alarms, but the technique is certainly intriguing.

26.4.4 Gamma Correction

Since their inception, in the late 1830s and for many years afterwards, cameras were based on film. Even today (early 2011), most digital camera users have used, or have at least seen, film cameras. Therefore, a discussion of digital cameras should mention the most important differences between them and film cameras. The obvious difference is the use of solid-state sensors instead of film to capture the image, but a more basic difference stems from the fact that human perceptions are nonlinear. Here is what this means.

Imagine listening to a whisper (a sound intensity measured at about 20 dB) and immediately afterwards turning on a noisy appliance (such as a vacuum cleaner or a lawn blower) with a sound level of 120 dB. The difference in sound intensity may be a factor of 10,000, but the ear perceives the appliance noise as only about 9–10 times louder than the whisper. The amplitude response of the ear is nonlinear, and the same is true for other human senses, most importantly, weight and vision. When we wake up in a dark room and then walk into bright sunshine, the change in brightness may again be a factor of around 10,000, something that would overwhelm the brain, but the eye and brain perceive it only as a factor of 9 or 10. Our senses protect us by their nonlinear responses, but as a result they are unreliable as measuring instruments.

It has been shown experimentally that the nonlinear nature of human perception is logarithmic and is expressed by an elegant relation, which is known as Weber's law and is expressed as

$$dp = k\frac{dS}{S},$$

where dS is the change in a stimulus S, dp is the perceived change, and k is a constant whose value depends on the particular physical units used to measure the stimulus. Integrating this expression yields $p = k\log_e S$, implying that the perceived stimulus is proportional to the natural logarithm of the actual stimulus.

Thus, human vision is nonlinear, but so is film! When a scene with dark and bright areas is captured on film, the difference between the dark and bright areas on the film is less than the actual, physical difference. The sensors in a digital camera, on the other hand, respond linearly. The output I of the sensors should therefore be transformed to a value O that resembles the actual intensities that would be perceived by film or by the eye.

The electrical charge collected by a sensor in response to photons is converted to an integer. This is normally a 12-bit integer in the range 0–4,095 (enough to express 4,096 levels of gray). A value of 0 indicates black (no photons sensed by the sensor) whereas 4,095 indicates white (the largest number of photons counted by any sensor). The middle value 2,047 corresponds to 50% gray. The eye can certainly sense black and white, so these two values should not be affected by the transform, but what about the values in between? The discussion above implies that these values should become darker, i.e., they should be decreased.

In order to understand the transform (which is referred to as gamma correction), we assume that the output I is a real number in the interval $[0, 1]$ where 0 is black and 1 is white. This is a reasonable assumption because we can simply convert the 12-bit integer $bb\ldots b$ to the real number $0.bb\ldots b$ and such a number is in the interval $[0, 1 - 2^{-12}]$.

The transform $I \rightarrow O$ should be a nonlinear but simple function that decreases all values of I, except 0 and 1, nonlinearly. The simplest such transform has the form

$$O = I^{\gamma},$$

where γ (gamma) is greater than 1 and its precise value is selected experimentally.

This transform leaves the two values 0 and 1 unchanged, and decreases small (dark) values of I less than large (bright) values. As an example of this nonlinearity, consider $\gamma = 2.2$. Increasing I from 0.1 to 0.2 with this gamma, decreases O by 0.0226816, while increasing I from 0.8 to 0.9 decreases O by 0.181045; a much greater amount. Figure 26.10 illustrates this transform. The top part shows a linear variation of grayscale from black to white and the bottom part shows how the values are darkened nonlinearly with a gamma value of 2.2. The *Mathematica* code is also listed, for readers who would like to experiment with this type of transform. (Reference [Schreiber 10] employs animation to illustrate the gamma transform.)

The concepts of gamma and gamma correction are important in all areas of optical electronics, not just in digital cameras. This correction has to be applied to camcorders, CRT monitors, LCD monitors, light detectors, and other devices. Gamma correction is needed because many components of imaging and optical electronics systems respond

```
gamma = 2.2;
Show[Graphics[
 Table[{GrayLevel[x], Rectangle[{x,0},{x+.01,0.1}]},{x,0,1,0.01}]],
 Graphics[{GrayLevel[0], Rectangle[{1,0},{1.001,0.1}]}]]
Show[Graphics[
 Table[{GrayLevel[x^gamma], Rectangle[{x,0},{x+.01,0.1}]},{x,0,1,0.01}]],
 Graphics[{GrayLevel[0], Rectangle[{1,0},{1.001,0.1}]}]]
```

Figure 26.10: The Gamma Transform.

nonlinearly and their response has to be converted to linear. For many years, the CRT was the primary component in computer monitors and televisions, so its performance has been carefully measured and is known in detail. A CRT is driven by a low-voltage video signal and it generates luminance on its screen. It turns out that the luminance (the CRT output) is a nonlinear function of this voltage (the CRT input).

The original NTSC video standard specified a gamma correction function with an exponent of $1/2.2 \approx 0.45$. For practical reasons, this ideal function has been changed and a new standard was proposed and approved as [SMPTE-170M] standard. It defines the two-part function (Figure 26.11)

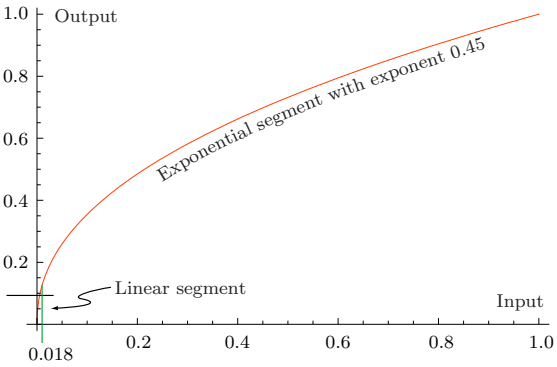
$$\begin{aligned} &\text{if } (V_{in} < 0.018),\ V_{out} = 4.5V_{in}, \\ &\text{if } (V_{in} \geq 0.018),\ V_{out} = 1.099V_{in}^{0.45} - 0.099, \end{aligned}$$

where V_{in} and V_{out} are in the range $[0, 1]$. This is interpreted as follows: For low values of V_{in} (up to 1.8% of the maximum), the output is a linear function of V_{in} with a slope of 4.5. For higher values of V_{in}, the output is a power function with an exponent of 0.45. At $V_{in} = 0.018$, the two functions have the same value 0.081.

26.4.5 Raw Image Format

We start with an analogy. The stone age of photography started in 1839 with Louis Daguerre. His photographic technique, known today as Daguerreotype, created the photograph as a one-of-a-kind image that was not reproducible. At about the same time, Fox Talbot revealed his photographic technique, which was based on a negative, and therefore allowed for easy reproduction of a photograph. Later eras of photography saw the development, among others, of color film and transparencies.

Consider the difference between shooting pictures with a negative and shooting with transparencies. In the former case, the negative has to be developed and a positive is then printed from it. This adds a step to the overall photo production, but also allows for processing of the image in the laboratory. While transferring the negative to the positive, a skilled photographer could create effects such as blurring, zooming, and variations of brightness and contrast. In the case of transparencies, the original film is already positive. The film has to be developed, but little lab processing can be done to improve or vary the resulting image.




```
gamma = 0.45;
Plot[x^gamma, {x, 0, 1}]
```

Figure 26.11: NTSC Gamma Correction Curve.

Today, in the era of digital cameras and images, raw format and JPEG became the modern analogy of negative and transparency film. When an image is saved in a camera (and is later output) in raw format, it has to be processed before it can be printed or viewed. This adds a step to the overall image production, but also gives an experienced user a chance to process the image in a computer in useful ways. (The only camera settings that cannot be changed by software in this case are the ISO speed, the shutter speed, and the aperture.) On the other hand, if a camera compresses and converts each image to JPEG as it is being taken, and immediately discards the raw image data, opportunities for later processing are reduced, because much image information disappears in the lossy JPEG compression. Because of this analogy, raw image files are sometimes referred to as negatives and the process of converting such a file into a viewable/printabe format is referred to as developing.

In principle, a raw image file should contain the dimensions of the image, the number of bits per pixel, a code for the color space used (RGB, CMY, or others), and the three color values for each pixel. In practice, such a file also contains metadata that is generated by the camera for each image. Examples of metadata are the date and time of shooting, the camera model and serial number, the shutter speed and aperture, the focal length, and whether the flash fired in taking the image. This type of metadata is also referred to as EXIF (exchangeable image format). Other important types of metadata are the color filter configuration of the light sensor and GPS information (latitude and longitude).

Currently (early 2010) there are many formats of raw data, developed by makers of digital cameras. Most of these are proprietary. Such a format may not even be raw and may include lossless compression. It is also known or suspected that some raw formats are even encrypted, to prevent an occasional user from processing the image data. Some names of raw image formats are `.3fr` (Hasselblad), `.arw`, `.srf`, `.sr2` (Sony), `.bay` (Casio), `.crw`, `.cr2` (Canon), `.dcs`, `.dcr`, `.drf`, `.k25`, `.kdc`, `.tif` (Kodak), `.dng` (Adobe), `.erf` (Epson), `.fff` (Imacon), `.mef` (Mamiya), `.mos` (Leaf), `.mrw` (Minolta),

`.nef`, `.nrw` (Nikon), `.orf` (Olympus), `.ptx`, `.pef` (Pentax), `.pxn` (Logitech), `.r3d` (Red), `.raf` (Fuji), `.raw`, `.rw2` (Panasonic), `.raw`, `.rwl`, `.dng` (Leica), `.rwz` (Rawzor), and `.x3f` (Sigma).

Raw image format has the following advantages over JPEG images:

- A typical consumer digital camera may have numerous settings—such as cloudy, snow, beach, fluorescent, tungsten lights—to adjust exposure for the available lighting. When a raw image file is processed, the exposure can be modified to any desired values.
- A raw image file makes it possible to change the white balance to the correct value *after* the picture has been taken. The term "white balance" refers to the process of removing or changing wrong colors (or modifying the color temperature, Section 21.3.1), such that white objects will be white in the final image.
- Depending on how the camera creates the JPEG file, a raw file format may provide considerably more dynamic range than a JPEG file. The term dynamic range refers to the range of light to dark that can be captured by a camera before becoming completely white or completely black.
- Each color component in a raw file is normally represented in 12 bits, as opposed to eight bits in a JPEG image file. The larger number of bits makes it possible to correct minor exposure errors and adjust color tones when the raw image is processed outside the camera.
- JPEG compression of an image (Section 24.5) starts by changing the color space to luminance-chrominance. Compression is lossy, which is why trying to change the color space after such a file is decompressed leads to significant loss of visual information. In contrast, a raw image file allows for quick and lossless transformations of the color space.
- JPEG files are small, but excessive JPEG compression results in annoying compression artifacts.

JPEG files, on the other hand, have the following useful features:

- The file is smaller.
- A beginner or an amateur photographer does not have to spend time processing the image files. The pictures are stored in the camera in their final form and can easily be examined, deleted if necessary, or transferred to a computer for printing and storage.
- A JPEG file can easily be exchanged between users because JPEG is a compression standard. Raw formats, on the other hand, are often proprietary.
- If the camera settings are correct, the resulting JPEG file will be as good as a raw file.

◇ **Exercise 26.2:** Discuss the following statement: Digital SLRs can save images in both JPEG and raw formats and the former is preferable for the following reason. Once an image has been saved in JPEG, the camera maker has no further control over it and its future processing. This is because JPEG is a popular format and there is so much software that can open, process, and print this format. In contrast, when an image is saved in raw format, the camera manufacturer can to some extent control how the user

will process and print the image. This is because raw formats are often proprietary and software made by the camera manufacturer is needed for any future processing, conversion, printing, and storage.

In a discussion of raw versus JPEG formats, it is important to explain how the light sensor array inside a camera is organized, what data it captures, and how the raw data is prepared. Most digital cameras have a rectangular array, called a mosaic sensor or color filter array (CFA), of CCD or CMOS sensors, each of which contributes a pixel to the final image. Light of many different wavelengths (corresponding to different colors) falls on each sensor, but the sensor counts only the total number of photons that impinge on it; it does not identify their frequencies (i.e, colors). During exposure, each sensor accumulates electrical charge that is proportional to the intensity (but not the color) of the light it has sensed. Thus, the sensor array generates a grayscale image.

Once this is grasped, it is not hard to figure out how to obtain color data from the sensors. Simply cover each sensor with a filter that lets only one color through. When we look at the world through rose-tinted spectacles, everything looks rosy, because only rose color reaches our eyes. Thus, sensors covered with a red filter output a grayscale value proportional to the red component of the light that strikes them.

> Rose-colored glasses are never made in bifocals. Nobody wants to read the small print in dreams.
>
> —Ann Landers.

A slight problem arises because a color space is three dimensional but the sensor array is rectangular. It is easier to partition a rectangular array into groups of four sensors than into groups of three, but such partitioning can be done and Figure 26.12 illustrates two ways of doing so. The configuration in part (a) of the figure is very common and is called a Bayer pattern color filter [Bayer 76]. Some cameras may filter four colors simply because it is easier to partition the array in groups of four sensors each.

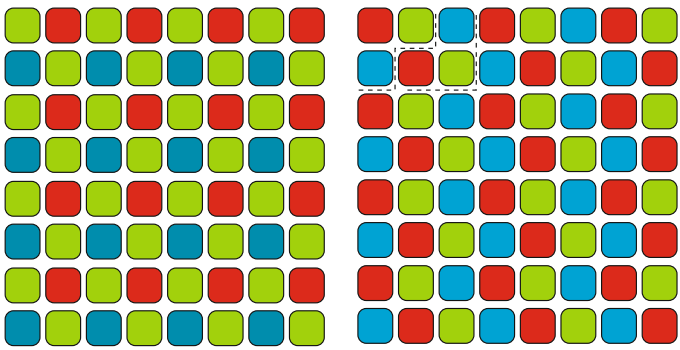


Figure 26.12: Color Filter Arrays.

Notice that half the sensors in the Bayer pattern are covered with green. This is because the eye is more sensitive to green than to red or blue (Figure 21.4). In the pattern of Figure 26.12b there are about the same number of sensors for each color (if the width w of the sensor array is divisible by 3, there are exactly $w/3$ sensors for each color, and the same is true for the height of the array).

After an exposure, the electrical charges in the sensors are converted to numbers (grayscale values) that are written on the raw image file. The file has to be processed later for viewing or printing, a process known as "raw conversion." The first step of raw conversion is to prepare three complete arrays of values for the three color components. Figure 26.13 illustrates this step. Only 25% of the original sensors produce red data, only 25% produces blue data, and only 50% of the sensors produce green data. Each empty position in the three arrays has to be filled up by interpolation from nearby pixels. The figure illustrates the simplest interpolation method. In part (a) of the figure (the red and blue components), each empty position labeled 2 in the top row is set to the average of its two nearest neighbors, while the leftmost position, labeled 1, is set equal to its only neighbor. The third row from the top is interpolated in the same way, and the second row is then computed as the average of its two neighbor rows. In part (b) of the figure (the green color component), each empty position has three or four near neighbors, except two positions (labeled 2) in opposite corners. Such an interpolation is known as demosaicing (or demosaicking), because the three pixel arrays resemble mosaics.

More sophisticated demosaicing methods are possible. Such a method may compute a value for an empty position as a weighted sum of eight (or more) positions, with larger weights assigned to nearby neighbors. However, including many neighbors in an interpolation may lead to blurring or even the complete disappearance of small details. Imagine a detail that occupies a small group of 3×3 pixels centered on the X of Figure 26.13b. It makes sense to compute a value for pixel X by interpolating its four nearest G neighbors, but if we also include in this interpolation the four G neighbors shown in gray (which are located outside the detail), all the visual information of the detail may be lost.

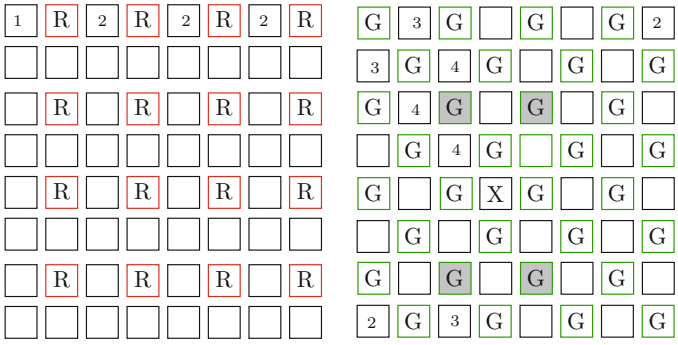


Figure 26.13: Empty Positions in Color Arrays.

Demosaicing by pixel grouping. This is an example of a fast, efficient algorithm that employs interpolation in an original way, depending on relations between neighboring pixels. This simple algorithm, by Chuan-kai Lin [cklin 03], is based on the observation that a continuous-tone image (an image of a natural scene, as opposed to an image of artificial objects) often contains groups of strongly-correlated pixels. Thus, given an empty position X in a Bayer grid, we can best compute a value for it by identifying those neighbors of X that are most similar to it. This principle is applied by the algorithm to the green positions. The red and blue empty positions are computed by simple interpolations based on hue transitions.

We use the following position numbering in a sample 5×5 Bayer grid:

R1	G2	R3	G4	R5
G6	B7	G8	B9	G10
R11	G12	R13	G14	R15
G16	B17	G18	B19	G20
R21	G22	R23	G24	R25

The algorithm computes values for the empty positions in three parts as follows:

Part I. Interpolate the green values in the red or blue positions in two steps.

Step 1. Every blue position (and most red positions) have four green immediate neighbors. In the few red positions that have only two or three immediate green neighbors, we use simple interpolation. For all other red and blue positions, we first compute four differences (or gradients). For position R13, for example, the four gradients are:

$$\begin{aligned}
\Delta N &= 2|R3 - R13| + |G8 - G18|,\\
\Delta E &= 2|R13 - R15| + |G12 - G14|,\\
\Delta W &= 2|R11 - R13| + |G12 - G14|,\\
\Delta S &= 2|R13 - R23| + |G8 - G18|.
\end{aligned}$$

Thus, gradient ΔN expresses the amount of color correlation in the north (up) direction about R13, and similarly for the other three gradients.

Step 2. Select the smallest gradient and compute a value for G13 as a weighted sum of four positions as follows

$$\text{G13} = \begin{cases}
\Delta N \text{ is minimum}, & (3\text{G8} + \text{R13} + \text{G18} - \text{R3})/4,\\
\Delta E \text{ is minimum}, & (3\text{G14} + \text{R13} + \text{G12} - \text{R15})/4,\\
\Delta W \text{ is minimum}, & (3\text{G12} + \text{R13} + \text{G14} - \text{R11})/4,\\
\Delta S \text{ is minimum}, & (3\text{G18} + \text{R13} + \text{G8} - \text{R23})/4.
\end{cases}$$

Part II. Interpolate the blue and red values in the green positions. As an example, we compute B8 and R8 at position G8.

$$\begin{aligned}
\text{B8} &= \text{HueTransit}(G7, G8, G9, B7, B9),\\
R8 &= \text{HueTransit}(G3, G8, G13, R3, R13),
\end{aligned}$$

where function HueTransit is defined as

```
function HueTransit(i3, i2, i3, v1, v3)=
 if(i1<i2<i3 or i1>i2>i3)
  then return v1+(v3-v1)(i2-i1)/(i3-i1)
  else return (v1+v3)/2+(2i2-i1-i3)/4
```

Part III. Interpolate the blue and red values in the red and blue positions. As an example, we compute B13 at position R13.

$$\Delta ne = |B9 - B17| + |R5 - R13| + |R13 - R21| + |G9 - G13| + |G13 - G17|,$$
$$\Delta nw = |B7 - B19| + |R1 - R13| + |R13 - R25| + |G7 - G13| + |G13 - G19|,$$
$$\text{if}(\Delta ne \le \Delta nw)$$
$$\text{then } B13 = \text{HueTransit}(G9, G13, G17, B9, B17)$$
$$\text{else } B13 = \text{HueTransit}(G7, G13, G19, B7, B19)$$

This algorithm is fast because it employs only addition, subtraction, few multiplications, and absolute value. The divisions by 2 and by 4 can be done by right shifts.

(End of algorithm.)

If the camera supports a raw image format, the raw conversion is done in a computer, normally with proprietary software. Such software is either supplied by the camera manufacturer or is implemented (as in the case of Adobe Photoshop) by a software maker. Raw conversion starts with demosaicing, but may also include steps for the following types of processing:

- White balance.
- Colorimetric interpretation. The visual sensation of color is very personal. If you prepare a list of shades of red and ask people to choose the "real," or "best" shade, there may never be complete agreement. Similarly, filters installed in digital cameras differ in the precise shade of red (and any other color) that they transmit. Sophisticated raw conversion done in software may allow the user to correct each color component individually until all the colors of the final image are satisfactory.
- Gamma correction. This is discussed in Section 26.4.4.
- Noise reduction, antialiasing, and sharpening. Demosaicing is based on interpolation, so it necessarily results in a certain amount of blurring. A sharp edge in an image may be lost because of the interpolation, so a raw converter should include a sharpening algorithm. Antialiasing is discussed in Section 3.13.

Now we get to JPEG. A camera that stores and outputs its images in JPEG, includes a raw converter and a JPEG compressor. Each exposure is followed by a blank period of a second or so during which the camera is busy converting the image, compressing it, and storing the resulting JPEG data in its memory card. The raw converter is built into the camera and generally cannot be modified (although in principle the raw converter may be stored in the camera as firmware, and may be updated from time to time). Most cameras permit the user to specify parameters such as the ISO value, the final image

size, amount of loss in compression, several light conditions (cloudy, seaside, fluorescent, tungsten light, nighttime), and aperture. Higher-quality consumer cameras may also offer user-controlled settings for color space, sharpening, contrast, and perhaps others. Obviously, the average user generally leaves these parameters at their default values and simply deletes and retakes any bad images. However, patience, attention to detail, and willingness to experiment with camera settings can work miracles and result in excellent images even when taken under unfavorable conditions. Thus, even the most occasional user is advised to read the camera's manual and experiment with all its settings, because once a bad picture has been taken and converted to JPEG, there is precious little that can be done to improve it.

In addition, most current digital cameras convert the electrical charge in a sensor to a 12-bit number, and raw image files save all 12 bits. A typical JPEG compressor built into a camera, starts by discarding the four least significant bits of a raw value and retaining only the eight most significant bits, thereby losing visual data even before compression begins.

Figure 26.14 lists the main steps taken in a camera to produce either a raw or a JPEG image file.

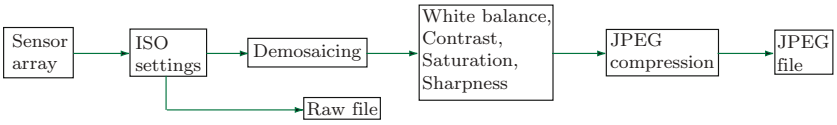


Figure 26.14: Producing JPEG and Raw Image files.

26.4.6 DSLR and Live Preview

A single-lens reflex (SLR) camera is based on a hinged mirror that can swing between up and down positions. In its down position, the mirror enables the user to see precisely what will be captured by the film. In the up position, the mirror is momentarily out of the way, so light can actually reach the film. In pre-SLR cameras, the view from the viewfinder was often different (sometimes even very different) from the final scene captured on the film, and this weakness was corrected by the SLR approach. The principle of SLR has been known even before the development of photography, but it was only in the 1960s that SLR cameras became practical and became the preferred choice of camera designers and users, especially for high-end cameras. A digital SLR (DSLR) follows the basic SLR design with a sensor array instead of film, and with extensive help from sophisticated electronics.

Figure 26.15 illustrates the SLR principle. In part (a), the mirror is down, the shutter is closed, and light travels from the object through the lens and is reflected to the viewfinder screen on top of the camera. Part (b) shows what happens during the short time a picture is snapped. The mirror is swung up, the shutter opens, and the light reaches the film. This SLR design is used in the Hasselblad cameras.

Most SLR cameras operate as illustrated in Figure 26.16. The viewfinder is located at the top rear of the camera, and the mirror reflects the light into a pentaprism, where

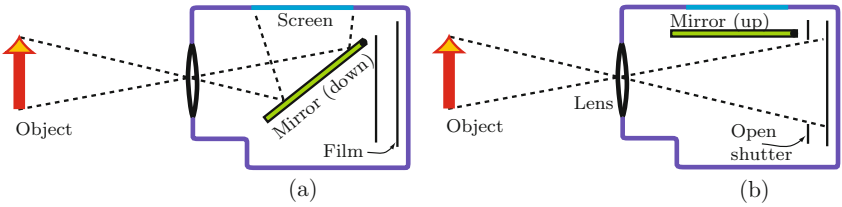


Figure 26.15: Principle of SLR.

the light is reflected twice before it is sent to the viewfinder. A common question asked by many at this point is why a prism and not simply another mirror, as in a periscope? The answer is that the lens projects the image on the mirror (and also on the film/sensor) upside down. With a simple mirror instead of the prism, the image would appear upside down in the viewfinder.

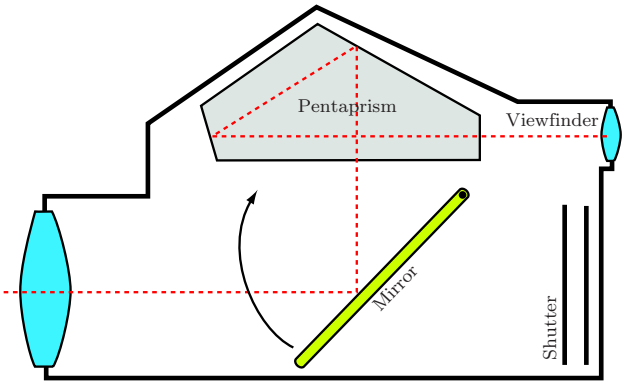


Figure 26.16: Principle of Pentaprism SLR.

Almost all current SLR cameras use focal plane shutters. Such a shutter is located in front of the image plane, and so prevents the light from reaching the film when the lens is removed. Most focal plane shutters consist of two curtains that are commonly made of composite plastic or thin lightweight metal. The curtains are referred to as front and rear. In many cameras, they move vertically, which permits higher shutter speeds (because vertical is the shorter dimension of the image sensor). The front curtain slides open to begin the exposure, and then the rear curtain slides closed in the same direction to end the exposure. The exposure time is counted from the instant the first curtain opens until the moment the second curtain fully closes. (At high speeds, the second curtain may start closing before the first curtain is fully open, which implies that the open area of the shutter is a narrow moving slit.)

If the flash is used for taking a picture, it has to be synchronized with the curtains, and this can be done in two ways as follows:

- Front (or first) curtain sync. The flash fires at the instant the front curtain has fully opened. The flash freezes motion at the beginning of the exposure, which is adequate for most flash photography.

- Rear (or second) curtain sync. The flash fires just before the second curtain closes. This type of sync freezes motion at the end of the exposure and is appropriate for making long exposures.

The type of curtain sync (first or second) is included in the metadata prepared by the camera software.

Over time, minute particles are shed off the shutter curtains and some end up on the image sensor, resulting in dim, blurred images.

Many DSLR cameras made in or after 2000 include a feature called live preview. When the user selects live preview, the mirror is swung up and the light enters the sensor array continuously. Several times a second, the image is sent from the sensor array to a small display screen on the back of the camera. Some cameras, most notably the Olympus E-330, implement live preview differently. The camera has two sensors, a main sensor array for the viewfinder and for capturing the image and an auxiliary array for the live preview. A special beam splitter splits the light entering the camera into two beams that are sent to the two sensor arrays.

Live preview is an advantage because many users prefer to use the display (which can be viewed from a distance) instead of the viewfinder (which has to be held close to the eye). Also, in some situations (such as underwater photography, where the camera is sealed in a waterproof case) it is inconvenient or even impossible to hold the camera close to the user's eye in order to look through the viewfinder.

A minor downside of live preview is that it continuously consumes electrical power for the display and therefore drains the battery quicker.

26.4.7 Appendix: Depth-of-Field

The important term "aperture" has already been mentioned several times. Most cameras control the amount of light that reaches the film/sensor by varying the shutter speed and the diameter of the lens. Behind the lens of the camera there is a diaphragm, similar to the pupil in the eye, that can cover different areas of the lens. When the diaphragm is fully open, more light reaches the film/sensor, but the diaphragm can be closed to reduce the amount of light. The term aperture refers to the diameter of the diaphragm.

When varying the aperture of a camera, the quantity that is used in practice is the f-stop (also referred to as f-number, focal ratio, f-ratio, or relative aperture). This quantity is defined as the focal length of the lens divided by the aperture. Several f-stops—such as f/4, f/5.6, and f/8—are marked on a typical lens to help the user adjust the aperture quickly. Notice that the larger the f-stop, the smaller the aperture (the effective diameter) of the lens.

Thus, the amount of light that reaches the film/sensor is proportional to the effective area of the lens and to the exposure time.

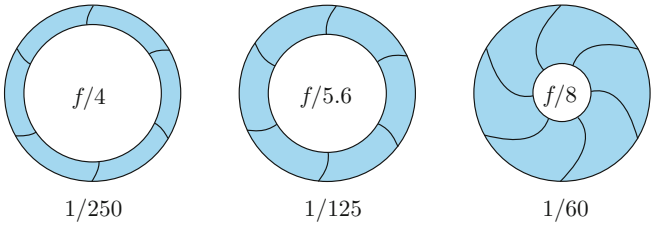


Figure 26.17: Three F-stops and Shutter Speeds.

◇ **Exercise 26.3:** Figure 26.17 shows three different f/stops and shutter speeds. Explain why the same amount of light reaches the sensor in each case.

And now, to the depth of field (Plates P.1 and P.2).

When a beam of light hits an object, it is absorbed, reflected, or refracted. It may even be partly absorbed, partly reflected, and partly refracted. Mirrors are useful because they reflect light, but the magic of lenses is based on light refraction (Section 17.2.2). The light "bends" when it moves from air to glass and bends back when it exits from glass back to air. Refraction happens because the speed of light depends on the density of the medium it travels through.

Figure 26.18a shows how parallel light rays that are perpendicular to a lens are bent because of refraction and converge to the focus at F, but the use of lenses in a digital camera is in focusing an entire image on the sensor array in the focus plane (which is not the same as the focal plane). The problem is that every point on the subject that is being photographed emits light in all directions. To get a sharp image on the sensors, all the rays that are emitted from a certain point x on the subject and that happen to strike the lens have to be bent so that they hit the sensors at the same point y. Figure 26.18b shows a subject to the left of a lens and how three rays leaving point x are bent differently by the lens and end up at point y on the focus plane f. We say that the subject is focused at plane f. If the subject is moved away from the camera, it will be focused in another plane, closer to the focus F. This is why in old cameras focusing was done by moving the lens back and forth, thereby varying the distance between the lens and the film. If the object is moved all the way to infinity, its focus plane becomes the focal plane (the plane containing the focus point F).

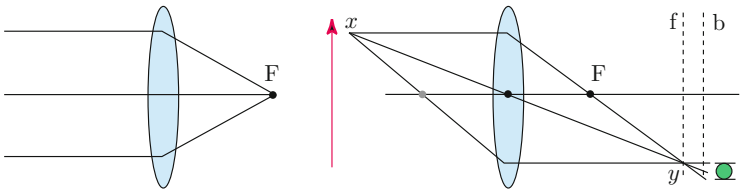


Figure 26.18: Focusing by Refraction.

Now consider plane b in Figure 26.18b. The three rays from point x on the subject diverge and hit this plane at three different (but nearby) points. A three-dimensional diagram showing more rays from point x would show that they form a small circle on plane b, the so-called *circle of confusion* or CoC (more accurately, this is not a circle but the shape of the aperture, which is typically a pentagon or a hexagon). It is now clear that if the subject is located at the precise distance the lens is focused, every point on it will focus to a point on the focus plane. When the subject moves out of focus, the points on the focus plane become circles. The farther out of focus the subject is, the larger these circles become.

(Circle of confusion: A group of photographers sitting around trying to understand depth of field.)

This is in principle. In practice, it is people who look at photographs, and the human eye has limited resolving power (it has evolved to help our ancestors hunt saber tooth tigers, not bacteria). When we look at a small enough circle, we see it as a point, which is why subjects that are in principle out of focus may still look sharp in a photograph. The result is that an image appears to be sharp (in focus) over a range of distances and this range is termed the *depth of field* (DOF).

> In theory, there is no difference between theory and practice, but in practice, there is.
>
> —Jan L. A. van de Snepscheut.

We can therefore define the depth of field as the length of the interval in front of and behind a focused subject in which the photographed image appears sharp.

To measure the DOF of a lens, we first have to decide on the diameter of the CoC. Different diameters yield different DOFs.

The resolving power of the eye depends on the person and on age. It varies widely, but on average we can use one minute of arc as a representative figure. This means that at a normal reading distance of 20 inches, the smallest detail a person with perfect eyesight can see (under ideal conditions) is about 1/16 (or 0.1667) mm. Two dots placed closer than this next to each other will appear as one dot. Obviously, the depth of field depends on what we consider blurred. A person who tolerates larger circles of confusion will claim that his camera has a greater depth of field, while someone less lenient may find that the same camera produces a smaller depth of field. Lens manufacturers often write the depth of field on the lens, and for 35 mm film cameras this specification was based on the following argument:

In a 35 mm camera, 35 millimeters is the size of the diagonal of the negative, so the width of the negative is about 24 mm or 1-in. To enlarge such a negative to a 5×7 print, the enlarging factor is 5. If we want the circles of confusion to be at most 0.1667 mm after the enlargement, they have to be at most $0.1667/5 \approx 0.0333$ mm before the enlargement. This was the CoC size that 35 mm lens manufacturers used when measuring the depth of field of new lenses.

In principle, the depth of field depends on the following factors: aperture size, focus distance, lens focal length, sensor size, sensor array organization, and the final print size.

Of these factors, the easiest for the user to vary is the aperture size, so how does the depth of field depend on the lens size? We provide two answers.

1. The circles of confusion are formed by light that passes through the lens, so less light implies less confusion in the circles, and therefore greater depth of field. This is an intuitive explanation which is easy to illustrate with a camera. An old camera, where it is easy to vary the aperture, is best. Look at a scene that includes objects at different distances, close the diaphragm gradually and you'll see the scene sharper. You'll also see it darker, but you can compensate for that by increasing the exposure time.

2. Figure 26.19 illustrates the effect of aperture on the depth-of-field by tracing light rays. In part (a) of the figure, the aperture is large, and it is obvious that points B and C, which are out of the principal plane of focus, become circles of confusion (indicated by the thick lines) on the image plane. With the much smaller aperture of part (b), the rays from points B and C converge at the same positions in front of and behind the image plane, but the circles of confusion are much smaller because of the narrower angles of convergence and divergence of the light rays.

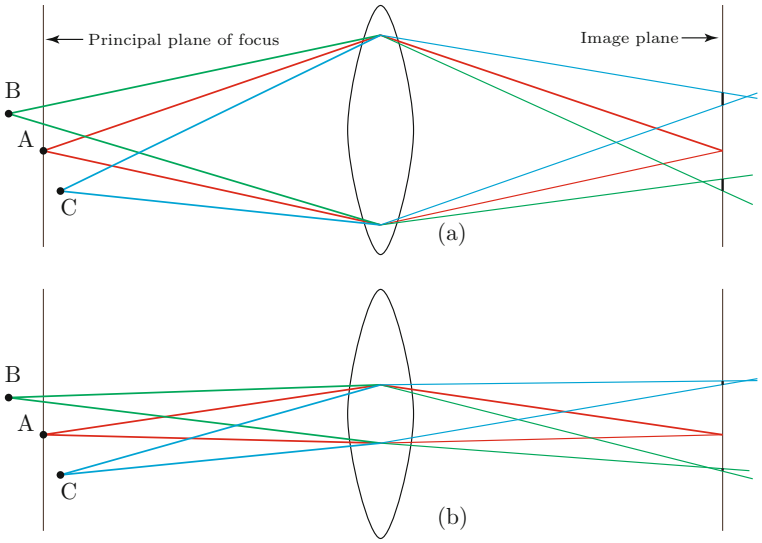


Figure 26.19: Effect of Aperture on Depth-of-Field.

Many advanced DSLRs have a special depth-of-field (DOF) preview button that when pressed, adjusts the aperture to the size it will have when the picture is taken. This allows the photographer to examine the DOF before the picture is taken.

Thus, a small aperture results in a greater DOF, but the application of this fact is limited, because of the effects of diffraction. Some light is always diffracted at the edge of the lens, and for a small lens, this light becomes a significant percentage of the total light, resulting in a poor-quality image.

A detailed (but hard to obtain) reference on DOF is [Blaker 85].

26.5 The Computer Mouse

We are all familiar with the computer mouse, this small, agile, nifty device that points to the computer screen and is slid and sent by computer users left and right, forward and back, all the time. The history of the mouse is one of an early quiet life followed by a sudden roar. Early versions of this device were large, heavy, mechanical, and slow to respond. Then came more buttons and a scroll wheel. Later, the mouse became electronic, losing its moving parts, and finally shedding its cord as well, becoming cordless. Today, the mouse has become an indispensable tool for input. Most computer users agree that the mouse is an ideal input pointing device, although many prefer a trackball (Section 26.6), a graphics tablet, a joystick, a touchscreen, or speech recognition. A typical mouse today (early 2011) is wireless, electronic, and features two buttons and a scroll wheel that also doubles as a third button. Manufacturers (Xerox, Apple, Microsoft, and Logitech are names that come to mind) and inventors continually try to improve this ubiquitous device and have come up with models that track the user's hand movements even if the mouse is held in the air, not touching any surface.

The name mouse occurred naturally to the original inventors of this useful computer peripheral, because early models resembled real mice (especially since they had a cord attached to the rear part of the device, suggesting the idea of a tail). The plural of the English word mouse is mice, but many language experts and dictionaries endorse the term computer mouses, in addition to computer mice. Perhaps the best choice for the plural is mouse devices.

Mouse history

Early computers were not interactive. The user wrote a program, punched its code on cards, and handed the cards to an operator. While the program was running, the user had no access to it. In order to debug a program, users had to rely on error messages printed as part of the output. Early personal computers allowed for some kind of user interaction. The main input device was a keyboard, and over time more and more keyboards featured four arrow keys. Certain programs, such as word processors and spreadsheets, made it possible for the user to move a cursor to any point on the screen with the arrow keys. The mouse became an integral part of personal computers in the early 1980s and especially in 1984, with the release of the Macintosh computer.

The idea of a sliding, easy-to-use pointing device that tracks the user's hand movements occurred to Douglas Engelbart, an inventor and computer pioneer, in 1963. Over the next few years he and his colleague Bill English have built a three-button, two-wheel, palm-sized contraption that they dubbed a mouse. An improved version of this device (with three buttons) was demonstrated by Engelbart on December 9, 1968 as part of a historical demonstration [Sloan 10] of his innovative networked computer system.

In 1972, Jack Hawley and Bill English came up with a mouse based on a single ball, pressed against two rollers, instead of the original two wheels, to track movement. This device also generated digital signals that could be sent directly to the computer. This mouse design remained dominant for many years. A ball-based mouse often requires

a special mousepad in order to roll smoothly, because the ball is relatively heavy and requires more friction than most desk surfaces provide. This type of mouse also becomes clogged with lint and dirt very quickly and has to be cleaned often.

In 1981, Xerox released its 8010 star computer that came with an integral, two-button mouse. This was an innovative machine, but it failed commercially because of its high price. In 1982, an efficient optical mouse was developed by Steve Kirsch. Lacking moving parts, this agile device required a special mouse pad to slide on and proved a commercial success.

In 1984, Apple computer introduced the Macintosh computer that came with an integral mouse (that mouse, incidentally, was developed in Lausanne, Switzerland). The first mouse models were mechanical. The mouse used a tracking ball and had one button, the result of an important design decision that remained controversial for many years. In 1985, a microprocessor was incorporated in this mouse, making it intelligent. In 1986, Apple introduced the Apple Desktop Bus (ADB) standard for connecting keyboards and mice to the computer. ADB remained a standard for 11 years.

Also in 1984, Microsoft started shipping its IBM PC mouse, a low-price device that had two buttons. The first models required a special peripheral card, but later models connected to the PC through a serial port. Also in the same year, Logitech Inc. designed the first cordless mouse. This is a battery-operated device that sends tracking data to the computer with infrared (IR) waves, similar to remote controls today. This technology was never successful because it required a line of sight between the mouse and the IR receiver (base station). In 1991, IR mouse technology was replaced with radio waves (RF).

In 1987, IBM released its PS/2 line of personal computers that featured PS/2 mouse connectors. These remained a standard in the PC world for many years.

In 1993, Honeywell introduced an opto-mechanical mouse, based on two small angled disks at the bottom instead of a rolling ball. The scroll wheel also made its debut in the same year. A scroll wheel (or mouse wheel) is a small rubber or plastic wheel placed perpendicular to the top surface of a mouse. It is often located between the two buttons. It is used mostly to scroll long, vertical windows. In the popular FireFox Web browser, holding down the control key while rolling the scroll wheel increases or decreases the text size. In an image-editing or map-viewing program, the same type of input is used to zoom an image in or out.

In 1998, Apple switched from ADB to USB as its mouse interface standard. New mouse models were introduced. Figure 26.20 (left) shows the Apple Pro Mouse (where the entire top of the mouse serves as the button).

In 1999, the first optical mouse that does not require a special pad was developed by Agilent. This device had an infrared LED that shined light under the mouse and sensed motion by measuring its reflection. The same technology was used and improved in the early 2000s by other mouse makers and adopted in 2000 by Apple. An optical mouse glides over the surface instead of rolling over it, and therefore requires smooth foot covers (or foot pads). These simple attachments decrease the friction between the mouse and the surface and allow the mouse to glide smoothly over many types of surfaces. High quality mice even feature teflon foot pads to reduce friction even further. Generally, an optical mouse cannot function on glossy and transparent surfaces.

In 2005, Apple released its Mighty Mouse (Figure 26.20 right, now called simply

Figure 26.20: Two White Mice.

the Apple Mouse, because of trademark issues with another manufacturer of a device named Mighty Mouse). This device features two buttons in the form of capacitive touch sensors (with a tiny speaker to provide audible clicking feedback). The main innovation was a small clickable scroll ball that lets users scroll in any direction. A wireless version was released in 2006. A new type of mouse, the Apple Magic Mouse was introduced in late 2009. This is a multi-touch device. Its top surface is smooth and seamless and can act (by changing its software preferences) as one or two buttons. Just by touching this surface, sliding or swiping one or two fingers, the user can scroll the cursor on the screen in any direction and at any speed.

Figure 26.21 shows: on the left, the Kensington PocketMouse Pro where (1) indicates a pushbutton that opens a door to a small storage compartment where the USB cable is stowed; and on the right, a Logitech cordless optical mouse (the USB receiver is also shown) where (2) through (6) indicate the on/off switch, reset button, a pushbutton that opens the door of the battery compartment, a footpad, and the IR LED, respectively.

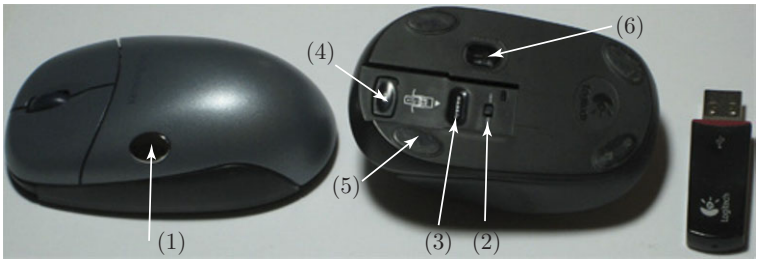


Figure 26.21: Two Black Mice.

Mouse-based user interface

Current computers are interactive. The operating system uses the main input and output devices to allow the user to interact with the operating system and with applica-

tion programs in a natural way. This important feature is referred to as a user interface (more specifically, graphical user interface or GUI). The main input device (which is often a mouse) sends tracking information to the computer.

The principle of a mouse is to follow the user's hand movements in two dimensions and send a signal to the computer, informing the operating system how fast and in what direction the mouse is moving. Based on this data, the operating system (or the user program) maintains a current position on the screen and displays and updates a cursor that indicates to software those parts of the screen that are of interest to the user. Pushing one of the mouse buttons while the cursor is positioned at P tells the software what the user wants done at P.

The user may want (1) to select an object (a file, an input/output volume, a program, or a menu), (2) to open the selection, or (3) to select part of the screen while an application program is running and command the program to perform a certain action. (In a word processor, the user may want to select part of the text and edit it. In a drawing program, the user may want to specify several points and draw a curve through them. In a spreadsheet, the user may want to select a cell and edit its content.)

The following mouse operations are standard:

- A single click (on the primary, left button) selects the object that happens to lie under the cursor.
- A double click (on the left button) opens the selected object.
- A triple click (on the left button) extends the selection to more objects.
- Dragging the mouse over several objects while pressing the left button selects all of them.
- Drag-and-drop moves an object to a new location and may also include copying or deleting the object.
- A single click on the secondary (right) button opens a new menu of options at the cursor location.

These simple operations (also known as gestures) can be combined with pressing keys on the keyboard to provide a mechanism to transmit very complex input commands to the current program. Gestures are especially useful in computer games, where the user must have quick and precise control over the complex objects (and parts of objects) that make up the game.

Mouse speed

The performance (or sensitivity) of a given mouse is normally measured in terms of counts per inch (CPI). This is the number of times the mouse sends data to the computer when it is moved by 1 in. Often, the operating system moves the cursor by one pixel each time the mouse sends data. In such a case, the counts per inch equal the number of dots (or pixels) per inch (DPI). The user can control the mouse sensitivity (and therefore its apparent speed on the screen) by software, making each mouse report (or mouse count) equal more than or less than a pixel. Low mouse speeds translate to better precision and easier selection of small objects on the screen. High speeds make it easier for the user to move the mouse large distances on the screen.

Exotic mice

Over the years, mouse makers have developed several unusual types of this important input device. An air mouse is so called because it does not need a surface to slide on (the terms inertial and gyroscopic mouse are also used). It can be lifted and moved in the air. Such a device employs a gyroscope, an accelerometer, or a tuning fork to sense movement and especially rotation. Thus, small rotations of the wrist are sufficient to control the cursor movement on the screen, thereby eliminating user fatigue and the much publicized carpal-tunnel syndrome and other mouse-related injuries. Such a device should be cordless and typically deactivates itself, to save battery power, after a short interval of not being used.

A three-dimensional mouse (also called a bat, a wand, or a flying mouse) senses motion through ultrasound. The mouse is a small, battery operated device that is worn as a ring on the user's finger. It emits low-power ultrasound that is received by an array of ultrasound microphones. By measuring the time in which each microphone receives the sound, it is possible to determine the location of the mouse in three-dimensional space. Such a mouse can be used as a three-dimensional scanner to scan a solid object and determine the coordinates of points on its surface. It can also be used to track a space curve and enter it into a three-dimensional drawing or illustration program. Such a mouse can also have buttons, pressed by the user's thumb. The main problem with ultrasound-based technology is low resolution, a serious drawback in many applications.

The Wii Remote, from Nintendo, is a hand-held tracking device that does not look like a mouse. This device is the primary controller for the Nintendo Wii game console. It employs an accelerometer and optical sensor to determine its orientation, direction of movement, and acceleration. The game console has a built-in sensor bar with ten infrared LEDs, five at each end of the bar. A camera in the Wii Remote detects the light from the two LED clusters, and a microprocessor in the Wii Remote employs triangulation to calculate the distance between it and the sensor bar. The relative angle of the two clusters of light on the sensor bar is also used to determine the rotation of the Wii Remote with respect to the ground.

A tactile mouse was introduced in 2000 by Logitech. This device contains a small actuator that causes the mouse to vibrate. The vibration provides extra feedback to the user in critical situations such as when the mouse crosses a window boundary.

26.6 The Trackball

A trackball is essentially an upside-down mouse. Instead of moving a mouse on a flat surface, the user rolls a ball which turns two perpendicular wheels mounted under it. The wheels generate movement signals that are sent to the computer.

The trackball device itself is stationary. The user's palm rests lightly on the ball and has to make only small rolling movements. This eliminates the large hand movements required with a mouse, which has the following advantages and makes the trackball the pointing device of choice for many users.

- The small ball movements eliminate fatigue and carpal-tunnel syndrome problems caused by a traditional mouse.

- A stationary pointing device is a better choice for users with a small computer table and for those who like to use a laptop in bed.

- The ball is in contact with the user's hand, not with the table surface. This reduces the amount of dust, lint, and hair that eventually impair the response of the trackball device and require cleaning. In contrast, a mechanical mouse has to be cleaned often.

- A trackball is stationary and can therefore be bigger and heavier than a mouse, which allows for adding pushbuttons and a scroll wheel to the basic trackball.

- Because it is stationary, a trackball can easily be built into a console. This is useful when a computer is used by the public (such as in a library or a coffee shop) where a mouse is easy to steal or vandalize. This is also handy in custom computer consoles, such as the radar consoles used by air-traffic controllers.

- Elderly people may not be able to hold a mouse still while double-clicking. A trackball eliminates this problem.

- More and more mobile devices have a built-in miniature trackball which is operated by the tip of a finger.

Figure 26.22 shows a sophisticated trackball (the Kensington Turbo-Mouse Pro USB) that features, in addition to the main ball, four large buttons, six small buttons, and a scroll wheel. The ball itself has been placed outside the device and the two wheels on which it rotates (plus a third, dummy wheel, for support) are clearly seen inside the ball cavity.

Figure 26.22: A Trackball.

The Apple Mighty Mouse employs a small trackball instead of a scroll wheel.

26.7 The Joystick

A joystick is an input device based on a small vertical lever (the stick) that pivots in two dimensions and can send its orientation to the computer. The orientation consists of two perpendicular angles about the vertical, neutral position of the stick. In addition, the joystick device may feature several buttons that can trigger actions such as shooting.

Joysticks were first developed in the early days of aviation. They were used in World War II to direct missiles, and were adopted by model airplane enthusiasts in the 1960s to control model airplanes by radio. Joysticks were also used by NASA to control space vehicles in the Apollo program.

Figure 26.23 shows a drawing of a basic joystick and a picture of a typical double-joystick transmitter (by Parkzone) for a radio-controlled model airplane.

Figure 26.23: Joysticks.

Today, joysticks can be found in many places. It is especially intriguing to see how a huge container or cruise ship is fully controlled by a small joystick, instead of the traditional, huge, wood and brass steering wheels so familiar from movies. Joysticks are also used to steer airplanes and trucks (often as small sidesticks) and to control a variety of machines, among them cranes, excavators, submersible unmanned research and rescue vehicles, wheelchairs, and surveillance cameras. An unusual application of joysticks is to control a home kitchen or bathroom faucet. Jado, a maker of faucets, has introduced such a faucet, named Cayenne, in early 2010. In addition to its use as a graphics input device, the joystick is often found (in miniature form) in mobile communications devices such as telephones and blackberries.

Many of the early video games of the 1970s and 1980s were designed specifically for joysticks.

A joystick works naturally in two dimensions. It can be pivoted left-right (which corresponds to movement in the x direction) and up-down or forward-backward (the y direction). It is possible to add a third dimension to a joystick by allowing twists of the stick in clockwise and counterclockwise directions (the z direction). When a joystick is

used to control an aircraft, movements in the x, y, and z directions are interpreted as roll, pitch, and yaw (Section 4.4.2).

High-end joysticks may provide haptic feedback to the user, either as vibrations or as a force that resists the pressure of the user's hand. Such a joystick is also an output device, allowing the software in the computer to send the haptic feedback as output to the joystick.

People with certain physical disabilities, such as cerebral palsy, may find the joystick easier to grasp than a standard mouse. Such a person may use the joystick as the main pointing device.

Part (a) of Figure 26.24 shows a miniature joystick (resembling a trackball) built into the blackberry 8800. Part (b) is the combatstick for game playing (courtesy of CH Products).

(a)

(b)

Figure 26.24: Joysticks.

An analog joystick is based on two potentiometers that output voltages proportional to the x and y positions of the stick. A digital joystick is simpler. Instead of two potentiometers, the stick operates two 3-way switches corresponding to the x and y directions. When the stick is in its neutral position, both switches are off. When the stick is pushed to the left or right (the x direction) it switches to one of the two terminals of the x switch. When it is pushed up or down (the y direction) it switches to one of the two terminals of the y switch. Thus, a digital joystick outputs a pair (x, y) of trits (ternary digits). Both x and y can have the values -1, 0, and 1.

For more information and figures, see [joystick 11].

26.8 The Graphics Tablet

A graphics tablet (also known as a digitizing tablet, pen pad, graphics pad, or drawing tablet) is a popular input device. It consists of a flat surface on which the user can move a special pen or stylus. The tablet hardware senses the position of the pen on the surface and sends the coordinates to the computer, for the use of the operating system or an application program. A tablet is commonly used with graphics software (drawing, painting, and illustration programs), although some users employ it as their main pointing device, instead of (or in addition to) a mouse. Some tablets even come with a cordless mouse that works on the tablet surface.

Some computer users, most notably artists, feel that a tablet is a souped-up replacement for a mouse because the user controls the cursor by drawing directly on the tablet. Many graphics programs recognize a tablet as an input device and allow the user to draw curves with the pen. Someone who is used to drawing on paper may quickly get used to the few differences between real pen and paper and a tablet. In fact, current tablet pens are often pressure sensitive; the harder the user presses, the thicker and darker the curve becomes. With experience, such a pen can produce better results than traditional pens. In addition to simply drawing curves, a graphics program can simulate the strokes of a brush as it follows the pen movements on the tablet.

With a tablet, a user can draw images and graphics by hand, similar (but not identical) to how images are drawn with a real pen and paper. A common use of a tablet is to capture a handwritten signature (many delivery persons carry a special tablet on which the receiver signs directly into a hand-held computer).

It is also possible to place on the tablet a sheet of paper with a drawing and trace it with the pen, thus enabling a non-artist to quickly create a reasonable copy of a drawing. Once this is done, the original drawing can be removed and the user may edit, modify, and improve the copy on the screen.

Graphics tablets have been used with computers since 1964, when the RAND Tablet, dubbed the Grafacon, was introduced. This early device employed a grid of wires embedded under the surface, and a pen that sensed weak magnetic fields generated by the wires. These fields were converted to coordinate data that was sent from the pen, on a cable, to the computer.

Figure 26.25 shows the Wacom Graphire ET-0405 tablet, with its pen and mouse.

Several techniques are employed by current tablets to track the pen on the surface. The most important ones are the following:

- In a passive tablet, two large sets, horizontal and vertical, of thin wires are located under the surface. The wires in each set are numbered from zero and these numbers act as coordinates of the surface of the tablet. (Wire resolutions of about 2,500 lines per inch are common.) The wires alternate as transmitters and receivers. The wires first generate an electromagnetic signal that is received by the pen and is stored in an LC circuit (see box below). The same wires then switch to a receiving mode, where they receive the amplified signal from the pen. The two wires, one of each set, closest to the pen receive the strongest signal, and the numbers of those wires become the output of the tablet.

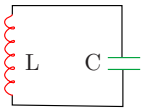
  This technique has important advantages. The electromagnetic signal received by the pen also powers the pen, which therefore does not require batteries or any other

Figure 26.25: A Modern Graphics Tablet.

power source. The pen does not have to actually touch the surface of the tablet and can be held in the air a short distance above it. It is easy to include pressure sensitivity in such a tablet. Varying the pressure on the pen varies the capacitance of its LC circuit and thus varies the resonant frequency. The frequency received by the tablet hardware is digitized and becomes a binary number that is sent to the computer together with each pair of coordinates.

An LC circuit consists of an inductor (a coil) and a capacitor connected in series. In order to start the circuit, the capacitor has to be charged or the coil has to be magnetized through induction. Once the LC circuit has been charged, current alternates between the two components at a certain frequency, determined by the properties of the coil and capacitor, that is referred to as the circuit's resonant frequency.

An LC circuit is used to store energy at its resonant frequency. The energy flows from the capacitor, where it is stored in the electric field between the two plates, to the coil, where it is stored in its magnetic field. The voltage at the capacitor goes down to zero, while the magnetic field of the coil increases. Once the entire circuit's energy resides in the coil, current starts flowing in the opposite direction, charging the capacitor and increasing its voltage (but with the opposite polarity). This process, known as a harmonic oscillator, resembles a pendulum swinging back and forth.



Ideally, the energy flows continually between the two components, but in practice there is always some resistance in the circuit, and its energy dissipates, mostly as heat. However, some energy leaves the circuit as an electromagnetic wave at the resonant frequency, which makes this type of circuit ideal for use in a tablet.

- A similar technique is used in an active tablet. The pen contains a power source and a circuit that generates an electromagnetic wave that is received by the two sets of wires under the surface. The advantage is that the wires don't have to alternate between transmitting and receiving, but the drawback is a larger, bulkier pen.

The two types above are also known as electromagnetic tablets.

■ An optical tablet employs a completely different approach. Such a tablet is based on special paper that has patterns of black dots printed. The patterns are asymmetric and do not repeat. The pen has a small camera that looks at the paper and deduces from the pattern it sees where on the paper it is located. While moving the pen it also writes, as a standard pen, on the paper, so the user can see what graphics data is sent to the computer. Anoto [Anoto 10] is a pioneer of this approach.

The Anoto paper (Figure 26.26) is based on a dense grid of vertical and horizontal lines placed at 0.3 mm (0.012 in) apart. The lines themselves are not printed. At each grid location, a small, black dot is printed. The dot is offset above, below, to the left, or to the right of the grid intersection. Thus, each dot is equivalent to a 2-bit number, and there can be $4^4 = 256$ different dot patterns in each 4×4 grid square. This is not a large number because a typical 8.5×11-in sheet of paper is equivalent to $708 \times 917 \approx 64{,}920$ grid squares. However, the camera sees several grid squares at any time, so it can identify their locations (and consequently, its own location) on the page with high accuracy.

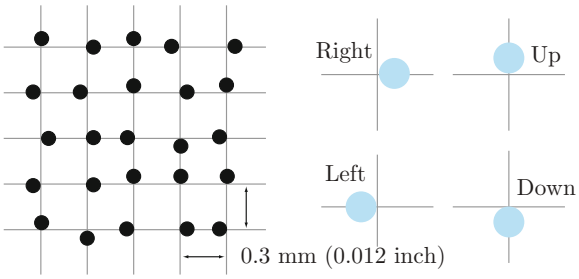


Figure 26.26: The Anoto Asymmetric Grid.

The Anoto pen contains an ink cartridge, a pressure sensor, a microprocessor, a battery, a USB port, and a bluetooth transmitter. It prepares a vector of data (a pair of coordinates, time of writing, page number, and a pressure) about 50 times a second. The pen has enough memory for about four sheets of paper, so it can either store the data and send it in a burst at the end, or send each vector as it is generated. The data can also be output from the pen through its USB port.

■ In an acoustic tablet, the pen emits ultrasound waves that are picked up by two microphones located near the tablet's surface. The microphones receive the sound at different times, and the time difference between them is used to determine the position of the pen on the tablet surface. By adding more microphones, it is possible to locate the position of the pen in three dimensions, on the surface and above it (see the description of the three-dimensional mouse in Section 26.5).

■ A capacitive tablet is based on variations in the electrical charge of a capacitor. The surface of a capacitive tablet is made of glass (an insulator) whose bottom side is coated with a thin layer of indium tin oxide. This material conducts electricity and is also transparent. A voltage is applied to this layer. When the pen touches the glass, it increases the electrical capacity of the layer at that point. This change in capacity is

measured from the four corners of the surface and is used to determine the location of the pen.

This technique is also used in many touch-sensitive screens (Figure 26.27). It is simple and the device is durable, but the resolution is limited and the device is sensitive to false signals (variation of capacity resulting from external sources).

Figure 26.27: A Typical Touch-Sensitive Screen.

Graphics tablets made for home users are generally small. The active surface typically measures 4×5 or 6×8 in. Engineers, architects, and artists may use much larger tablets, but those are very expensive. It takes a while to get used to a small tablet, but it minimizes arm movements, an important feature for those prone to injuries from repetitive movements. A fact that surprises many new tablet owners is that the footprint of a tablet may be much larger than its advertised active surface area. The Wacom tablet of Figure 26.25 has a surface of 4×5 in but its footprint is 9×8 in, 3.6 times bigger!

Most graphics tablets for home use employ the standard USB interface (which is hot swappable, so the tablet can be plugged in and out while the computer is on). For those who prefer a wireless interface, there are a few bluetooth tablet models.

Many current tablets include a switch and a pushbutton on the pen. Those can be used to select objects on the screen (they can serve as a single-click, double-click, or right-button click). Often, the top of the pen serves as an eraser. When the pen is flipped upside down, the tablet sends a different signal to the computer, and the program can use this signal to erase or delete objects or graphics elements in a single swipe of the pen. Low-end tablets offer 256 pressure levels but this is increased to 512 or 1,024 in more advanced models.

The touchpad as a tablet. Most laptop and notebook computers and many personal digital assistants (PDAs) feature a touchpad or trackpad (Section 26.8.2). This is a small (rarely bigger than six square inches or 40 square centimeters), touch-sensitive surface that converts the movement of the user's finger(s) to screen coordinates. A touchpad is not as easy to use as a mouse, but it is useful when there is no room for a mouse on a small desk or when the computer rests on the lap of the user.

Many touchpads double as buttons. When tapping a finger on the surface, the touchpad sends a different signal to the computer, acting as a button. A tap followed by a continuous sliding motion is popularly called a click-and-a-half. Newer touchpad software drivers can also distinguish the movements of two fingers and interpret it either as a button click or a signal to zoom in or out. A touchpad may also have hotspots; locations that can be touched to indicate actions other than pointing. Thus, sliding a finger along one of the edges of the touchpad may indicate vertical or horizontal scrolling, while touching a corner may indicate a pause in playing or the launch of an application.

Most touchpads sense the position of a finger because the human body conducts electricity to some degree and a touch at a point is equivalent to connecting the point to the ground through a resistor. Such a device may not function if touched by a stylus or if the user wears gloves. A touchpad that does not have this drawback can be used as a small tablet. Given a thin stylus and special software that simulates a tablet, such a touchpad can be turned into a small tablet. An example is the pogo sketch stylus [pogo sketch 10] and its accompanying `inklet` software for Macbook computers.

A Typical Tablet

Wacom is the most well-known tablet manufacturer, with a product line for both professionals and home users. The following is a summary of the features of their Bamboo Pen and Touch tablet [wacom 10].

The Bamboo Pen and Touch tablet combines the advantage of a multi-touch screen with the high precision of a modern tablet. The user can apply either finger taps or hand gestures on the surface (the area sensitive to touch is 4.9×3.4 in). Various application programs respond to taps and gestures by scrolling through documents, navigating the Web, zooming in and out of pictures, and rotating images.

When high resolution is needed, the pen can be used. This makes it possible for the user to select small areas on an image, make sketches, and mark and annotate documents in long hand.

There are four keys that can be customized and assigned various functions.

The tablet comes with two graphics programs and it can be plugged into a USB port of a PC or a Macintosh computer.

Features

- Two sensors for precise pen and Multi-Touch input.
- Use a single finger for navigation and multiple fingers for gestures.
- Simple gestures make it easy to scroll, zoom, rotate, and move backward or forward.
- Pressure-sensitive pen tip for natural pen and brush strokes.
- Battery-free, ergonomic pen with two switches.

- Textured work surface for a pen-on-paper feel.
- Quick access to user-defined shortcuts with four keys.
- Attached fabric pen loop conveniently stores pen.
- Easy USB connection to Mac or PC, laptop or desktop.
- Interactive tutorial helps you learn gestures and make the most of your Bamboo.
- Two graphics programs are included.

Specifications

Tablet Dimensions (W × H): 9.8 × 6.9 inches (249 × 175) mm.
Active Area - Touch (W × H): 4.9 × 3.4 inches (124 × 86) mm.
Active Area - Pen (W × H): 5.8 × 3.6 inches (147 × 91) mm.
Pressure Levels: 1024.
Resolution: 2540 lpi.
Max Data Rate: 133 pps.
Accuracy: ±0.02 in (±0.5) mm.
Connectivity: Standard USB.
Orientation: Reversible for right- or left-handed users.

26.8.1 The Digital Pen

The digital pen described here is the Zpen, by Dane-Elec [Zpen 11]. It consists of a pen, a receiver, and software. Once the receiver is clipped to a sheet of paper and is turned on, it follows the movements of the pen much like a tablet and saves (in its 1 GB flash memory) the coordinates of points visited by the pen. While moving and transmitting its position to the receiver, the pen also writes on the paper. When the receiver is clipped to the next page, it starts a new page in its memory. When done, the user plugs the receiver into a USB port in the computer, and special software downloads the data into the computer, where it can be displayed, saved as a pdf file, and passed through OCR software.

The advantages of the Zpen are: (1) The pen actually writes on the paper, so the user can see what will eventually be loaded in the computer. (2) It is easy to replace ink cartridges in the pen. (3) Downloading the data from receiver to computer is fast and simple.

But there are also drawbacks: (1) Users report that the pen has to be held in a special way, so as not to block its transmission to the receiver. (2) Transmission is broken when the pen is held too close to the receiver, so writings at the top of a page are often lost and careful placement of the receiver on the page is crucial. (3) The resolution of the receiver seems to drop when the pen is at the bottom of the page, away from the receiver. (4) The OCR software is not perfect, requiring the user to carefully read and correct the resulting text (but a digital pen may be ideal for drawings).

26.8.2 The Trackpad

A trackpad (also called a touchpad) is a tablet operated with a finger (or two fingers) instead of a stylus. Trackpads are very common with laptop computers. In such a computer, the user can move a finger over a small pad, thereby directing the operating

system to move a cursor on the screen in the same direction and speed. A pushbutton or two often accompany a trackpad, allowing the user to click in order to select an object or open it. Some users don't like the trackpad (mainly because of its small size), which is why many laptop computers can use a mouse instead of or together with a trackpad.

Some trackpads also have "hotspots," locations on the pad that correspond to user commands other than pointing. For example, moving a finger along an edge of the trackpad acts as a scroll wheel. Apple MacBooks respond to two-finger dragging gesture for scrolling. Certain trackpads support tap zones, small regions where a tap indicates a function.

The trackpad of the Apple MacBook is multi-touch, a term that refers to the ability of the trackpad to simultaneously sense several distinct finger touches.

This trackpad has no separate button. Instead, the entire surface of the pad is the button and the user can click anywhere on the pad. The trackpad features a very smooth glass surface that feels comfortable to the fingertips. It supports right-clicking and several multi-touch gestures: pinch to increase font size in a document or to zoom on an image, rotate your fingers to reorient images, and swipe to navigate through Web pages. Figure 26.28 illustrates several possible hand gestures (not necessarily the ones adopted by Apple) and users' comments indicate that interacting with software by means of gestures is easy and comes naturally.

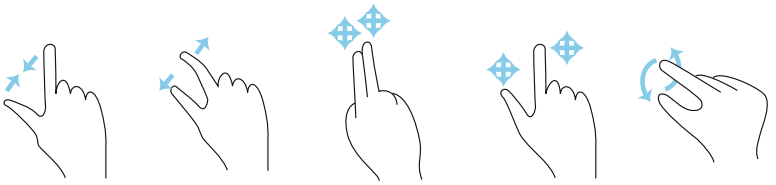


Figure 26.28: Several Hand and Finger Gestures.

Of special interest is the Magic Trackpad, introduced by Apple Computer on 21 July 2010 (the day these words were written). This device is a larger version of the common trackpads found in laptop computers, and it is a stand-alone product that can be moved from one computer to another. Industry watchers immediately justified the Magic Trackpad with the following arguments:

- Apple sells more MacBooks than desktop computers and MacBook users love their small trackpads, which is why Apple decided to gamble on a larger trackpad for desktop computers.
- Apple believes that the mouse was an intermediate step in the development of human-computer interaction and the future of this field is in devices that offer touch gestures, such as in the iPhone and iPad.

At the time of writing, not much is known about the Magic trackpad, not even its dimensions (later found out to be 5.2 in wide by 5.1 in tall). What is known is that

it offers gestures such as pinch-to-zoom, inertial scrolling, and either tap or touch to simulate a mouse click. The Magic Trackpad connects to the Macintosh via Bluetooth wireless technology and requires two AA batteries.

The immediate question on the mind of many Macintosh users is will the Magic Trackpad be the ultimate mouse trap? Will it convince users that the mouse is unnecessary? The following is a list of reasons of why users will likely keep their mice even if they have a trackpad:

- Most of the fingers have to be kept off a trackpad to avoid triggering unwanted gestures, and this may be the reason why many users feel that their fingers get tired after using a trackpad for a while. This ergonomic reason works for the mouse and against a trackpad.

- Certain gestures needed for game playing either cannot be performed on a trackpad or feel very artificial, while with a mouse they happen naturally.

- When a mouse arrives at the edge of its pad (or other region where it moves) it can be lifted and dropped at another point. This is especially important when the mouse button is pressed. This simple maneuver cannot be performed on a trackpad.

- Many computer users may simply hate a trackpad and be nostalgic about their old, trusty mice. Old habits die slowly, and currently no one can guess how important this psychological reason is.

On the other hand, the new Magic Trackpad may signal the demise of the mouse. Here are some reasons for this prediction:

- If trackpads become common for desktop computers, new users who "grow" with trackpad-based computers may not even be familiar with a mouse and may consider the trackpad the only pointing and gesturing device.

- Gestures let the user interact with the data on the monitor screen, so the user ends up feeling closer to the data, whether it is text or images. Flipping pages on the screen with finger gestures feels as natural as flipping pages in a real book. Inertial scrolling senses the speed of the fingers and generates response that feels natural.

- The large area of the Magic Trackpad may offer users a sense of freedom. Clicking may be done at any point on the trackpad, and the trackpad itself lies at the same height and angle as the Apple wireless keyboard. All this may help users get used to the new Magic Trackpad and its future successors.

26.9 Scanners

A scanner is a graphics input device that converts a document to a digital image. The document may be a thin sheet of paper, cloth, plastic, or parchment, a page in a thick book, or even a solid object (although the latter type requires a special, three-dimensional scanner). The scanner may operate at one of several resolutions and may save the resulting image in several formats (normally in compressed form). The user may specify only part of the document to be scanned, and also inform the scanner what type of data is to be scanned (a black-and-white document, grayscale image, color image, high-gloss photo, etc.).

Currently (in early 2010), the most common type of scanner is the flatbed (or desktop). The document is placed upside down in the scanner on a glass pane (the scanner bed), it is covered by the top of the scanner to block any ambient light, and a carriage slides under the bed, shining light from below on the document and scanning it slowly, in small steps, by measuring the light reflection and generating a row of pixels in each step. Two techniques are used to scan color documents as follows:

- Three arrays of CCD (charge-coupled device) act as sensors that determine the three color components of individual pixels. Each array is covered by a transparent sheet that lets only one primary color through. White light is provided by a xenon or cold cathode fluorescent bulb.
- A single array of contact image sensor (CIS) devices sense reflected light. Illumination is provided by three sets of red, green, and blue LEDs that are strobed for each scanning step.

Figure 26.29 shows the HP Scanjet 4600 flatbed scanner. Notice how the top of this device comes off completely, to make it easier to place thick books on the scanner bed. This scanner is also unusual in that the light source and the sensors are mounted on a carriage inside the top of the scanner. Thus, the scanned document must be placed rightside up.

Figure 26.29: The HP Scanjet 4600.

Current high-end flatbed scanners can scan up to 5,400 dpi, and this figure increases every year. This is considered high resolution and is satisfactory (in fact, it is way too high) for most applications. When even higher scan resolutions are required, drum scanners may be the ideal choice.

If the document is larger than the scanner bed, it has to be scanned in sections that are later matched by software and stitched into a large image. This kind of matching may prove problematic, as demonstrated by the little-known and interesting case of the *hunt of the unicorn* tapestries (see story in [unicorn 10a,b]).

In the 1970s and 1980s, scanners were important graphics input devices. With film-based cameras, the best way to input a photograph into the computer was to scan it. Today, with high-resolution, high-precision digital cameras, flatbed and other types of scanners are used only in cases were high-quality results are needed and cannot be obtained with a camera because of reflections, shadows, or low contrast. Scanning a document by shooting it with a camera is quick and convenient, especially if the object to be scanned is a page in a thick book or an ancient, rare document that should be handled as little as possible.

A current, 12 Mpixel camera can capture an entire page at a fairly good resolution. A typical letter-size, 8.5×11-in page has a surface area of 93.5 square in. When captured on 12 million pixels, each square inch contains 128,342 pixels, which translates to a scanning resolution of 358 dpi. If the page has 1-in margins on all sides, an area of only 6.5×9 in has to be shot by the camera, which increases the effective resolution to 453 dpi.

Other types of scanners are drum, film, and hand.

A drum scanner is based on a transparent, rotating drum. The document is attached securely to the drum, the drum is spun at high speed, and a carriage is moved slowly over the drum. A lens in the carriage focuses a small, sample area of the document into an array of photomultiplier tubes which convert the sample area into pixels (grayscale or color) with exceptional sensitivity. The size of the sample area can be varied by changing the optics, which proves a distinct advantage in scan jobs requiring very high resolutions (up to 12,000 dpi) and many color gradations. If the document is transparent, light is shined from inside the drum and is transmitted by the document. Otherwise, light is shined from above and is reflected by the document. Currently, the main use of drum scanners is in applications that require high resolution, such as scanning artwork in museums and scanning film (because a small, 35 mm film requires high-resolution scanning to remain sharply defined after being enlarged).

A film scanner can scan positive (slides) or negative film. Generally, a strip of several negatives or mounted slides is inserted into the scanner and is moved in small steps under a lens that sends a focused image to a CCD sensor array

A hand-held scanner is a small device that is dragged across a document to scan a narrow stripe at a time. To make sure the scanner is moved on a straight line, it can be slid along a ruler placed on the document. A roller on the bottom of the scanner measures the scanning speed and generates a pulse for each scanning step. The size of a step is the scanner's resolution and is controlled by the user. This type of scanner generally has a start button that should be held down while the scanner is moved. There is also a warning light that comes on when the scanner is moved too fast. While the document is scanned, part of it is displayed and scrolled on a small LCD screen.

Choosing scan resolution. Before scanning an image, the user should specify the scanning resolution. This simple, basic decision is usually made by the user based on intuition or on past experience, but the discussion that follows tries to convince the reader that this decision should be based on the future use of the scanned image. Generally, an image is scanned in order to be displayed or printed, but there are big differences between display monitors and printers. Following is a short comparison:

Features of Printed Images	Features of Displayed Images
On paper, image size is measured in inches or centimeters.	On a screen, image size is measured in pixels.
Image size is independent of scanned resolution.	Image size depends on scanned resolution.
On paper, image size is modified by scaling.	On the screen, image size is modified by zooming (interpolation).
On paper, image pixels are spaced using specified scaled resolution (dpi).	On the screen, image pixels are located at pixel locations, one location per pixel.
Several printer dots create the color or grayscale of one image pixel.	On a screen, each location displays one image pixel.

Thus, if a w-inch-wide image is scanned at 300 dpi, then every linear inch of the image will generate 300 pixels. Currently, a typical display resolution is 96 dpi, so when this image is displayed, each group of 96 consecutive pixels will occupy one linear inch of screen space and 300 pixels will occupy 3.125 linear inches; the image will look large. When the image is printed on a monochromatic laser printer, its width on the paper will be w in and each linear inch will consist of 300 halftone square grid of dots (Section 2.27). When the same image is printed on an inkjet printer, its width on the paper will also be w in, but each image pixel will consist of a square grid of color ink droplets that together approximate the pixel's color as much as possible. The resolution of these dots may be 1,220 or 1,440 per inch, but they will simulate 300 pixels per inch.

Scanning for a display. The following examples may shed more light on this topic. Given a 6×4 in photograph, assume that we want to scan and display it (but not print it). If we have a 640×480 display monitor, we can scan this photo at a scanner resolution of 110 dpi. The resulting image size is $(6 \cdot 110) \times (4 \cdot 110) = 660 \times 440$ pixels, enough to more or less fill up the entire screen. If we have a bigger, 800×600 monitor, then scanning the photo at a resolution of 140 dpi will result in an image of $(6 \cdot 140) \times (4 \cdot 140) = 840 \times 560$ pixels, again enough to more or less fill up the entire screen. With a $1{,}024 \times 768$ monitor, scanning at 180 dpi will also produce an image that fills up the screen. When we scan our photo at the screen resolution (which nowadays is often 96 dpi), the scanner will generate $(6 \cdot 96) \times (4 \cdot 96) = 576 \times 384$ pixels and this will occupy approximately 6×4 in on the display monitor. However, a monitor advertised as a 96-dpi-display may actually have only about 90–92 dots per vertical inch, and a similar number for a horizontal inch.

We therefore conclude that displaying an image on the entire screen requires only low-resolution scanning. Scanning at higher resolutions is needed only if we want to print the image or to display it bigger than the screen, so we can scroll and examine individual

parts of it on the screen. With a display monitor, scanning resolution determines the size of the displayed image, not its quality. With a printer, it's the opposite.

Scanning for a printer. When a document is scanned for future printing, higher resolutions are needed. Grayscale and color images are said to have much noise. The eye cannot perceive the precise color of every pixel, so modifying the color of several pixels (or even many pixels) may not be noticeable. Experience indicates that scanning such a document at 300 dpi results in a print of acceptable quality. Discrete-tone images (Section 23.2) feature smooth curves and straight lines and edges, and are therefore an exception, because straight lines and edges reduce the noise of an image and require high scan resolution.

Text and line drawings are different and require higher scan resolutions. Placing a ruler on a printed page and measuring the width of words verifies that most texts occupy about 1/16 of an inch per character (some characters, such as `i` and `j` are narrower, while `m` and `w` are wider). Assuming that 20% of this figure is the spacing between characters, we conclude that the average width of a character of text (at 10 printer's points) is $0.8 \cdot (1/16) = 0.05$ in. Scanning text at 300 dpi translates to 15 pixels per the width of a typical character. This may be enough for most characters of text, but experience shows that wide characters, upper-case letters, and text printed at 12 points and larger may suffer from low printing quality when scanned at 300 dpi.

For most people, a line drawing is an image and is closer to a picture than it is to text. However, as far as scanning is concerned, a line drawing has less noise than a painting or a photograph and therefore should be scanned at a higher resolution.

Long experience with scanning and printing documents suggests scan resolutions of 400–600 dpi for text and line drawings. Higher scan resolution may be needed in the following cases: (1) text that is going to be magnified before printing, (2) text that will be converted (by OCR software) from pixels back to ASCII or Unicode, and (3) text in large fonts or complex scripts such as Chinese.

Notice that a printer can print a scaled image, but this does not affect the original image in memory. The pixels of the scaled image are computed by the printer driver, sent to the printer, printed, and stay only on the paper. In contrast, when an image is resampled (stretched or shrunk) on order to be displayed (not printed) at a different size, the original pixels in memory are modified, but there is normally a copy saved on a disk or a CD that remains unchanged. The new image can also be saved from memory to a disk.

See also Section 26.4.2 for a discussion on how to print big images.

Scanning for a monochromatic printer. A monochromatic (black-and-white) printer can print only black dots, but such printers print grayscale images using a technique termed halftoning (Section 2.27). With halftoning, a gray pixel with p% gray is printed as a small grid of dots, p% of which are black. When text is printed, there is only black and white, and so the full printer resolution (typically 600 dpi) is used, resulting in sharply-defined text. If halftoning is used with grids of 6×6 printer dots, then 37 shades of gray can be printed, but the resolution is only $600/6 = 100$ lines per inch (where "line" means a row of halftone grids). With 7×7 halftone grids, a printed image can have 50 shades of gray, but only $600/7 = 85$ lines per inch (lpi) are possible. A larger grid of 8×8 dots increases the number of shades to 65 but decreases the resolution to $600/8 = 75$ lpi. At present, magazines typically print grayscale images at 133 or 150

lpi, newspapers make do with the much lower figure of 85 lpi, while high-quality books demand about 200 lpi.

◇ **Exercise 26.4:** A 6×6 grid has 36 boxes, so why can it specify 37 shades of gray and not just 36?

Scanning for an inkjet printer. A color printer works differently. We discuss only inkjet printers, because nowadays they are by far the most common type of color printer. They are inexpensive (although the ink cartridges represent a considerable investment over time, see "ink wars" on Page 1268) and they produce excellent results if the right paper and the correct scan resolution are used. A typical inkjet printer has three or four ink cartridges (some have six or even more) and can therefore print dots with only those colors. Such a printer cannot mix the colors (inside the printer or on the paper) to print more colors. When a color C other than the three or four basic colors is needed, the printer employs dithering (Section 2.28 and especially Page 115) to print dots of the basic colors that will create in the viewer's brain the same sensation as C. Figure 26.30 shows how violet is obtained by dithering red and blue.

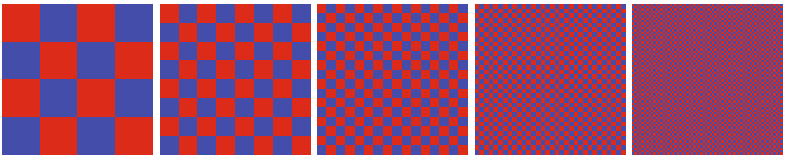


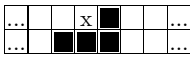
Figure 26.30: Dithering Red and Blue.

(The two additional ink cartridges in higher-quality inkjet printers contain light magenta and light cyan. They allow the printer more choices, especially in image regions with light colors. Such inkjet printers, while more expensive to operate, may sometimes benefit from higher image resolutions.)

Thus, an inkjet printer may have an advertised resolution of 1,440 dpi and may be able to generate and spray 1,440 submicroscopic droplets of ink on each linear inch of paper, but dithering reduces this fantastic resolution by a substantial factor, because dithering a single color pixel may require many ink droplets. Scanning an image at 1,440 dpi and trying to print it at that resolution may limit the printer to one ink droplet per pixel, resulting in very wrong colors. Scanning the image at 300 dpi gives the printer a chance to use $1{,}440/300 = 4.8$ droplets per pixel (actually a square of $4.8 \times 4.8 \approx 23$ droplets), thereby ending up with much better colors.

The software that drives an inkjet printer expects the image to employ the RGB color space, but the ink in the printer has the CMYK colors (see Chapter 21 for color spaces). Thus, the printer driver has to convert RGB to CMYK and then apply dithering to obtain the best possible color. The resulting color C' often differs from the required color C, and there is a color error, or difference $\Delta = C - C'$ each time a pixel is printed. The driver normally uses an error-diffusion dithering algorithm that carries this error to those nearest neighbor pixels that haven't yet been printed. Those are (the black pixels in the figure) the neighbor to the right and the three neighbors centered below

the current pixel x. If any of those pixels should have color D, the driver tries to print it in color $D+\Delta$. This normally results in another error that is carried over in the same way.



This complex process should be compared with displaying an image on a monitor screen. The resolution of a display monitor is typically 96 dpi, much lower than the 1,440 dpi of the inkjet printer, but each pixel of the image is displayed on a pixel of the monitor in its original RGB color.

We therefore conclude that paintings and photographs intended for printing on an inkjet printer should be scanned at relatively lower resolutions, such as 240 to 300 dpi. This applies even to the best printers, those with advertised resolutions of $1{,}440 \times 720$, $2{,}880 \times 720$, and $1{,}200 \times 1{,}200$ dpi.

Text and line drawings are again an exception. Text is black and white, so no dithering is needed, each pixel corresponds to a single ink droplet, and the printer can use its full resolution and produce sharply-defined small characters of text. The same applies to black and white drawings. If a drawing is in grayscale, it makes sense to scan it at half the printer's resolution (or even smaller), which gives the printer a chance to print a grid of at least 2×2 droplets for each pixel and thus simulate at least 17 shades of gray.

Printer makers offer a wealth of literature in their websites. Epson recommends scanning color photos at 240 dpi for its printers and to increase this to 300 dpi if very sharply-defined images are desired (and also on its 6-cartridge models). HP employs a proprietary technique that can print several ink droplets at the same location and thus blend their colors to some extent. There is no dithering, but error diffusion is still used. This is an attempt to achieve one printer dot for one pixel, but color errors may often become very high.

Tip. Often, a look at the histogram before the final scanning can result in greatly improved scanning. Many scanner drivers ask the user to start with a preview scan of the entire image. The user then selects the image area (a rectangle) for the final scan, and the software displays the histogram of that area. Figure 26.31 shows an image and its histogram. The image is in color, and the histogram displays the luminance values of the pixels in 256 steps, with black, as 0, on the left. It is easy to see that some of the lowest and the highest luminance values are missing from this histogram because the image does not have any pure white or black pixels.

If the scanner displays such a histogram, it may have sliders under it that can be moved by the user. Just move the left slider to the leftmost nonzero luminance value and similarly for the right slider. This tells the scanner to consider the gray shade under the left slider as pure black and the value under the right slider as pure white. This simple operation can significantly improve the resulting image.

It should also be mentioned that the physical resolution of the scanner itself is fixed. A flatbed scanner has a moving carriage with one or three rows of sensors and the density of these sensors (their number per inch) can be considered the natural resolution of the scanner. Current scanners have typical natural resolutions of 600–1,200 dpi. If we want to scan at a different (lower or higher) resolution, the scanner itself must employ

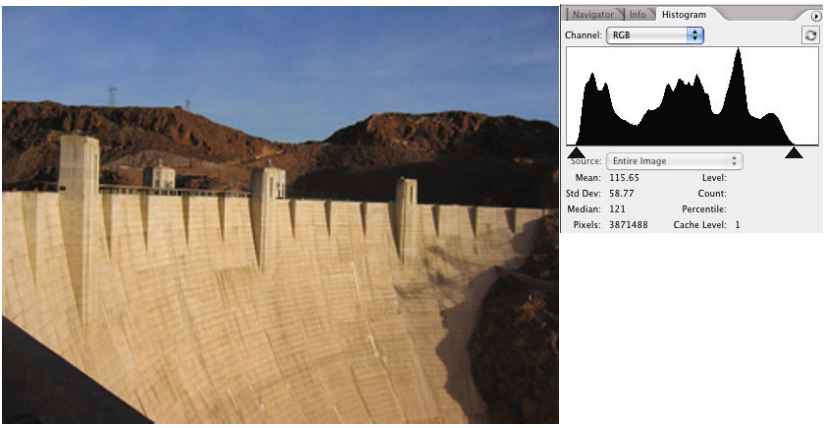


Figure 26.31: An Image and its Histogram.

interpolation to generate the number of pixels per inch specified by the user. Thus, some experts recommend always using the scanner's natural resolution and employ software, such as Adobe Photoshop, to change the resolution later to the desired value.

Suppose that the natural resolution of a scanner is r pixels per inch and the user wants to scan the image at a lower resolution l. If $r = nl$ where n an integer, the scanner can simply average the colors sensed by n adjacent sensors and use the result as a single pixel.

Sometimes, the scanner resolution is advertised as a pair, such as $1{,}200 \times 2{,}400$. In such a case, the smaller number is the scanner's natural resolution and the larger number is the number of steps per inch of the carriage. In the example above, the carriage is moved (by a stepper motor) 2,400 steps per inch, and it scans 1,200 pixels in each step.

Scanning for fax. So far we assumed that an image is scanned in order to display it or print it. Sometimes, an image is scanned in order to fax it directly from the computer. Naturally, it should be scanned at the standard fax resolution. Fax machines are made by many manufacturers, so in order for them to work together, they must all conform to a standard. Two such standards, known as Group 3 and Group 4 (or T3 and T4), were developed starting in 1980 and 1984, respectively, by the International Telecommunications Union (ITU) and are currently used by virtually all fax machines. Among other things, these standards refer to the pixels resulting from the scan as pels, and also specify the scanning resolution of these machines.

The horizontal resolution of fax is always 8.05 pels per millimeter (about 205 pels per inch). An 8.5-in-wide scan line is therefore converted to 1,728 pels. The T4 standard recommends to scan only about 8.2 in of the width of the paper, thereby producing 1,664 pels per scan line (these numbers are to within $\pm 1\%$ accuracy). The vertical resolution is either 3.85 scan lines per millimeter (in the standard mode) or 7.7 lines/mm (in the fine mode). These are equivalent to about 98 and 196 dpi. Many fax machines have also a very-fine mode, where they scan 15.4 lines/mm.

Scanning for fax should therefore be done at 200 dpi. If the scanner accepts different horizontal and vertical resolutions, then the best choice is a 200×100 scan resolution.

Finally, we turn to the question of saving a scanned image. There is again a difference between color and grayscale paintings and photos (continuous-tone images), on one hand, and text, line drawings, and pictures of artificial objects, (discrete-tone images) on the other hand. For best results, both types of images should be saved in TIFF or PNG formats because these are lossless. If file size is an important consideration, then color photos should be saved in JPEG (but keep in mind that this format is lossy), while text and drawings should be saved in TIFF, PNG, or GIF.

26.9.1 Three-Dimensional Scanners

A three-dimensional scanner is a graphics input device that can scan solid objects. Such a scanner looks and works differently from the common, two-dimensional variety, because three-dimensional space is so much more complex than two-dimensional space (a good example of the extra complexity introduced by adding one dimension is the three-dimensional transformations described in Chapter 4, which are so much more complex than the two-dimensional transformations).

A three-dimensional scanner measures and collects the coordinates (and sometimes also the color) of many points on the scanned object. These coordinates are later used to construct an accurate, three-dimensional model of the object in the computer. The model can be used to manufacture a copy of the original object, to vary the shape of the object so it becomes a prototype of a new object, to modify its surface texture and reflectivity so it can be viewed in a new light, or to use the object as part of a computer animation such as a video game.

Three-dimensional scanners can be classified in different ways as follows:

- Moving and stationary. Most three-dimensional scanners are stationary. Such a scanner works somewhat like a camera. Its field of vision is a rectangular pyramid, and it cannot measure those parts of the scanned object that are hidden from its view. With such a scanner, an object has to be scanned several (even many) times in different orientations and the results aligned into a common coordinate system. A moving three-dimensional scanner (also known as a coordinate measuring machine or CMM) has a turntable. The object is positioned on the turntable and is rotated slowly as it is scanned. If the object is large and heavy (such as a sculpture), the CMM itself must rotate around the object. Such a scanner consists of a fixed central station and a probing beam or a moving small probe (or stylus). The user moves the probe manually over the object and touches each point whose coordinates need to be measured and saved by the central station. This is a slow, tedious process, but the advantage is that only one scan is needed and any point, on any part of the object, can be reached by the probe and measured.

- Contact and non-contact. In a contact scanner, there is a probe that actually slides over the object and measures points. A non-contact scanner employs a beam of laser light, infrared radiation, or X rays that are reflected from points on the scanned object. This type is appropriate for delicate or precious objects and for objects that are too big or too far away to reach with a probe. Non-contact scanners are further divided into active and passive scanners. The former sends a beam to the object, while the latter uses ambient radiation that is emitted from the object.

Here is a short list of the chief technologies employed by active three-dimensional scanners.

A laser scanner sends a beam of laser light that is reflected from a point P on the object. The scanner measures the roundtrip time of the beam and converts it to the distance of P from the detector. Notice that the distance is not enough to determine the coordinates of P, so such a scanner also measures the direction in which the beam was sent (by means of two angles). This is the principle of a rangefinder and of radar, but the distances that a scanner measures are relatively short, which is why the scanner must be able to measure very short time intervals, on the order of picoseconds (10^{-12} seconds, or the time it takes light to travel about 0.3 mm). Because of this, a laser scanner is accurate for long distances (up to kilometers) and can measure the coordinates of 10,000–100,000 points each second. Such a scanner may be the ideal type to scan an entire building, a job that may require millions of samples. The object scanned must be stationary, and the scanner itself must be protected from movement, vibrations, and even temperature changes that may stretch its parts.

A laser triangulation scanner also sends a beam of laser light to a point on the scanned object, but instead of receiving its reflection, the scanner employs a camera to locate the dot of light on the object. The camera C, the dot D, and the laser source S are located at three different points, so triangulation can be used to determine the location of the dot. The triangle edge between C and S is a known constant, and the two angles $\angle DCS$ and $\angle DSC$ can be measured accurately. The edge and the two angles are sufficient to solve the triangle and determine the coordinates of D. This type of three-dimensional scanner has limited range and is generally used with distances of a few meters or less. On the other hand, its accuracy is on the order of tens of micrometers (10^{-6} m), but it cannot scan moving objects.

A hand-held laser scanner is also based on triangulation, but the position of the scanner must be taken into account each time a sample is taken. The sample (the coordinates of a point on the object) is determined relative to point S (the location of the light source in the scanner). If the location of S relative to a fixed point G on the ground is known at all times, then the coordinates in each sample can be corrected and computed relative to G.

A structured-light three-dimensional scanner projects a known pattern of light on the object and examines its reflection for deformations caused by the shape of the object. Perhaps the simplest scanner of this type is the one described in [bouguetj 10], which uses the deformation of shadows to extract three-dimensional coordinate information (Figure 26.32). A lamp illuminates the object. The user holds a thin pencil that casts a shadow on the object. The shadow is deformed by the varying heights of points on the object. A camera snaps a picture of the object and the shadow, and the user slides the pencil a short distance to repeat the process.

At each position of the pencil, an entire row of points on the object is sampled. Also, several objects can be scanned together.

A modulated light scanner shines a modulated beam of light at the object. Often, the attribute being modulated is the amplitude of the light, which is varied in a sinusoidal pattern. A lens and a light sensor detect the reflected light and determine its roundtrip time by comparing the reflected amplitude to the amplitude at the time of detection. Once the time interval is known, the distance can be computed. Because its light is

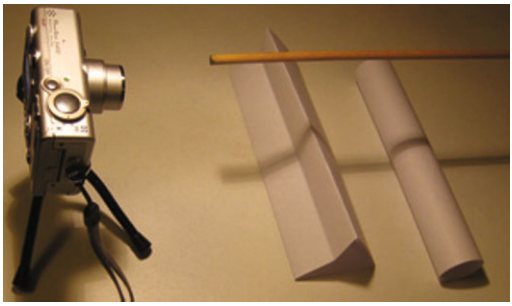


Figure 26.32: Deformed Shadows for Scanning.

modulated, this type of scanner is not confused by light from other sources that happens to be reflected from the point that's being measured.

Computed tomography (CT) represents a different approach to three-dimensional scanning. It employs X rays to create a two-dimensional slice of the scanned object, and then moves the X-ray source slightly and repeats the process. The various slices can be combined by software into a single three-dimensional model and viewed from various directions. This technique is commonly used in medicine, where it is fast and painless, but involves the well-known risk of exposure to X rays.

Magnetic resonance imaging (MRI) is another approach to medical scanning. It employs a powerful magnetic field to align the magnetic poles of atoms (usually hydrogen) in various solutions in the body. MRI does not involve the risk of radiation and produces more contrast between the different soft tissues of the body than is obtainable with CT.

A non-contact passive three-dimensional scanner works by detecting the reflection of light and infrared radiation that do not originate at the scanner. A few techniques are described here.

A stereoscopic scanner consists of two cameras that look at the same point on the object. Because of the distance between them, the cameras see slightly different images and exploit this difference to determine their distances from the point. (See Section 6.13 for stereoscopic images.)

A photometric scanner consists of a camera that takes several (n) snapshots of the object under different lighting conditions. Reference [Woodham 80] discusses the surprising fact that the vector of n reflected intensities from a point p can be used to determine the surface normal at p.

A silhouette-type scanner starts by taking a set of photographs around the object. This must be done against a background with a different contrast. Software then prepares silhouettes of outlines from each photo and extrudes and combines them to build a hollow three-dimensional skin (or hull) of the object. The coordinates of any point on the skin can then be determined by interpolation.

26.10 Inkjet Printers

An inkjet printer prints a digital image (that may consist of text and bitmaps) by spraying minute, colored ink droplets on the paper. Inkjet printing technology has developed rapidly since the 1970s and today it boasts a wide range of printers, from portable, inexpensive models aimed at the home market, to large, heavy-duty machines that can print huge posters.

Figure 26.33 shows the HP PhotoSmart 335 Printer, designed specifically to print photographs. The printer is small (notice the 7-in ruler placed for comparison) and weighs 1170 g. The maximum paper size is 4×12 in. The printer is designed to be portable and easy to use. It works with either a power supply or a battery and features a front panel USB port plus five built-in memory card slots for direct printing from many types of flash cards. There is one ink cartridge (shown through the open door).

Figure 26.33: The HP PhotoSmart 335 Printer.

The ink droplets in a typical inkjet printer are extremely small, normally 50–60 microns in diameter (where a micron equals 10^{-6} m). For comparison, the diameter of a human hair is about 70 microns. The droplets must be placed very accurately, because typical current inkjet printer resolutions are $1{,}440 \times 720$ dpi (while high resolutions hover around $9{,}600 \times 2{,}400$ dpi). The droplets come from different ink tanks, they have different colors and are combined on the paper by various dithering methods to create photo-quality images (see Page 1256 for details). Most inkjet printers employ four ink tanks, for the CMYK colors (Figure 26.34), but high-end models boast six, eight, and even 10 ink tanks for vivid, true colors.

(The Canon PIXMA Pro 9500 Mark II inkjet printer features 10 individual ink tanks with matte black, photo black, cyan, magenta, yellow, photo cyan, photo magenta, red, green, and gray inks. The Canon PIXMA Pro 9000 Mark II photo inkjet printer employs the eight colors black, cyan, magenta, yellow, photo cyan, photo magenta, red, and green.)

Figure 26.34: Four Ink Cartridges.

Inkjet (and also laser) printers are non-impact. No part of the printhead comes in contact with the paper. This is in contrast to the old dot matrix and daisywheel printers, where parts of the head had to impact the paper for each dot or character printed.

Inkjet printing is not limited to paper. The technique of ink jet material deposition is quickly finding many applications. The idea is to deposit (with high precision) various liquid materials, not just ink, on many types of substrates. Typical examples are: (1) Inkjet printing on a CD or DVD, (2) inkjet printing of a color photograph on a cake with edible "ink" (United States patent 6,319,530), (3) inkjet printing on ceramic tiles for decoration, and (4) spraying narrow stripes of conductive liquid on a thin sheet of plastic to create a printed circuit board.

The idea of printing by spraying drops of ink dates back to 1867, but its application to digital computers started in the early 1950s. The first models suffered from low resolution (large ink droplets) and ink that dried in its container when not used for a while. It was only in the 1970s that manufacturers learned how to create extremely small ink droplets, how to route them to the precise locations on the paper, and how to keep the liquid ink from clogging the printing nozzles and drying in its reservoir when not in use.

Today, most inkjet printers are made and sold by Canon, Hewlett-Packard, Epson, and Lexmark. These and other printer makers continually introduce new models with higher droplet resolution, higher printing speeds, and more ink tanks for better, vivid colors.

Piezoelectricity

When pressure is applied to certain materials, most notably crystals (especially lead zirconate titanate crystals), they generate an electric field in response. This surprising feature is known as piezoelectricity or the piezoelectric effect. The same materials also exhibit the inverse piezoelectric effect; when voltage is applied to a piezoelectric crystal, it deforms itself and varies its dimensions slightly.

The piezoelectric effect was discovered by the brothers Pierre and Jacques Curie in 1880. The inverse piezoelectric effect was first deduced theoretically by Gabriel Lippmann (color photography inventor) in 1881, and then demonstrated by the Curie brothers.

Both the direct and inverse effects are employed today in common appliances such as inkjet printers and igniters for cigarette lighters, stoves, and gas grills. They are also used to generate sound waves, high voltages, and in various sensors.

The three chief inkjet printing technologies—continuous inkjet, piezoelectric drop-on-demand, and thermal drop-on-demand—are listed in Figure 26.35.

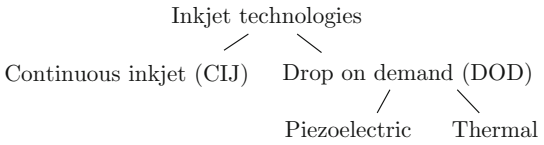


Figure 26.35: Inkjet Printing Technologies.

Continuous inkjet. The principle of continuous inkjet is to create and maintain a continuous stream of ink from the ink tank, to deflect some of the ink and route it to the paper, and collect and recycle the rest. The ink is pumped at high pressure into a gunbody, where a vibrating piezoelectric crystal creates a high-frequency sound wave that breaks the ink into a stream of small droplets. The density of this stream is between 50,000 and 150,000 droplets per second, so the gaps between consecutive drops are extremely small. The ink drops are then charged electrostatically (the amount of charge varies between drops), but between each pair of charged drops there are several neutral drops, to prevent electrostatic repulsion between consecutive drops.

The drops then pass through a microscopic hole on their way to the paper and the charged drops are deflected by an electric field and are sent to the paper (drops with more charge are deflected more and so hit the paper at different points). The uncharged drops are collected in a gutter and are recycled. When printing has to stop momentarily (to allow the paper to move), the stream of ink continues, but no drops are electrically charged.

This technology was invented by William Thomson (Lord Kelvin) and patented by him in 1867. It was revived in the 20th century because the state of the art of electronics now allows full control over charging and deflecting individual drops.

The advantages of continuous inkjet are as follows: (1) The high speed of the jet (about 50 m/s), which makes it possible to position the paper away from the print head. This is important when printing on large items, such as packages. (2) The large number of drops per second allows for high printing speed. (3) There is no clogging of nozzles and drying of the ink, because the jet of ink is continuous. This makes it possible to use fast-drying liquids such as ketones and alcohols, instead of traditional printing ink.

The two DOD printer technologies are based on generating ink droplets on demand. Software directs the printing head to where ink is needed on the page, and the head can propel several droplets (of different colors) onto to the same pixel on the paper.

Thermal DOD inkjet. A sudden pulse of heat is used in this type of printer to squeeze an ink droplet out of a small nozzle and direct it onto the paper. The printing head is a replaceable cartridge with little chambers, each with a tiny heating element (a resistor) and a small quantity of ink. When a short pulse of current is sent through a heater, some of the ink in the chamber evaporates and becomes an expanding vapor bubble. This increases the pressure in the chamber and results in an ink droplet ejected from the nozzle toward the paper (Figure 26.36). When the bubble later shrinks, the vacuum thus created sucks fresh ink from the ink tank into the chamber. This principle was developed by an engineer at Canon in 1977, which is why that company started advertising its printers as bubble jets. A typical bubble jet print head can have between 300 and 600 tiny nozzles that can fire droplets simultaneously.

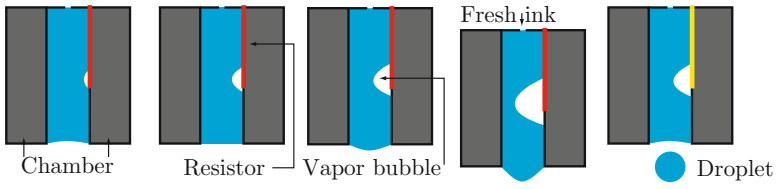


Figure 26.36: A Thermal DOD inkjet.

The advantage of this technology is that the printing head with the chambers is easy to manufacture by photolithography and no piezoelectric crystals are required.

The ink itself consists of a coloring agent (minuscule particles of pigments or dyes) suspended in a carrier fluid (water). A volatile agent such as alcohol must be added to create a large enough vapor bubble.

Notice that thermal DOD is not the same as the old, thermal printers, common in the 1970s and 1980s (and still used today by some old fax machines).

Piezoelectric DOD inkjet. Instead of a heating element to create a vapor bubble, the piezoelectric DOD technique employs a special crystal as a miniature pump to create and propel the ink droplets. The printing head again consists of ink-filled chambers, each with a piezoelectric crystal (often PZT, lead zirconium titanate) located right behind a nozzle. When a short electric pulse is applied to the crystal, it deforms momentarily and this change of shape is used to collect and force a small amount of ink through the nozzle on its way to the paper (Figure 26.37).

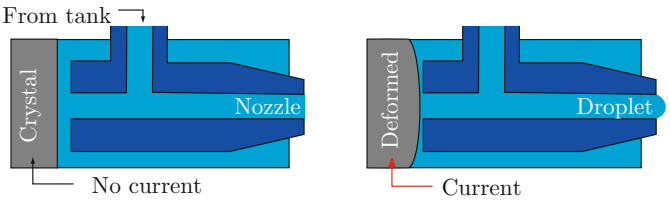


Figure 26.37: A Piezoelectric DOD inkjet.

Such a print head is more expensive to make than a thermal-DOD head, but piezoelectric-DOD printers offer the following advantages: (1) The ink does not require a volatile agent, which allows for more types of ink. (2) The print head lasts longer. (3) The gap between the head and paper can be wider than in thermal DOD. (4) The operating costs are lower.

This technique is used mostly in large, industrial inkjet printers (although Epson makes several piezoelectric-DOD printer models for home use).

Types of ink. The perfect ink for an inkjet printer must include a coloring agent that will quickly stick to the paper (or the material being printed) and a carrier fluid that will dry fast and leave no marks or wetness. Much detail about the composition and manufacture of inkjet inks can be found in [Magdassi 10].

The ink that is typically used in home inkjet printers is known as aqueous. It employs dyes or pigments as coloring agent and a mixture of water and derivatives of paraffin or glycol as carrier fluid. Pigments resist ultraviolet radiation and result in prints that take longer to fade. Water is a necessary component of ink for a thermal DOD printer because the bubbles require water vapor. Printer manufacturers always recommend the use of special glossy or coated paper, where the dyes generate sparkling colors. Figure 26.38 shows how an ink drop on glossy paper (left) retains its ideal circular shape, while the same drop on standard printing paper (right) soaks into the paper and becomes irregular.

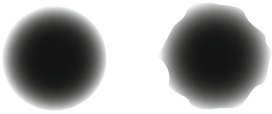


Figure 26.38: Ideal and Deformed Ink Drops.

Large banners, billboards, and posters spend their lives outdoors and have to be water proof and resistant to ultraviolet. They are normally printed on vinyl, where better results are achieved with solvent inks. This type of ink employs color pigments that are immersed in liquid organic compounds, for high vapor pressure and fast printing. Fast vinyl printers often employ special heaters and blowers to quickly dry the vinyl after printing.

When printing on fabrics with a high content of polyester fibers, the best choice is dye sublimation ink. Sublimation is a printing technique that employs heat to transfer solid dye particles onto the paper (or other media such as fabric, plastic, or cardboard) and convert them to gas, without going through a liquid state. The gas diffuses into the paper, solidifies inside, and becomes durable.

Inkjet head design. Over the years, printer makers settled on two approaches to printhead design, fixed and disposable.

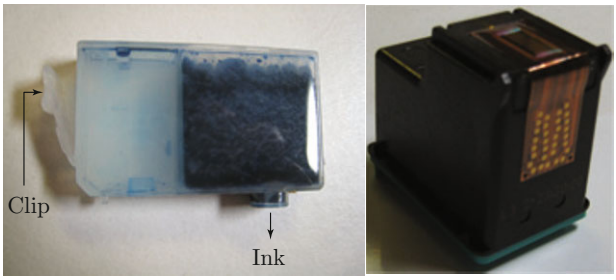


Figure 26.39: Fixed (Left) and Disposable (Right) Ink Cartridges.

In the former type (Figure 26.39, left), the head is built into the printer and lasts the life of the printer. The ink tank (or cartridge) contains just the ink. This design reduces the cost of the ink cartridges and allows for a higher-quality print head. However, if the head gets damaged or clogged, the entire printer may have to be replaced. (Canon makes fixed print heads that are supposed to last the life of the printer but can also be replaced by the user if needed.) Because of these features, fixed heads are built mostly into large, industrial printers (but Canon uses such heads in its successful Pixma line of printers).

A clip is clearly visible on the left of the fixed printhead in the figure. This serves to fix the cartridge in the printer. There is also an opening on the bottom, where the ink is released into the thermal chambers as needed. The printhead shown in the figure is empty and a large pad of absorbent material can clearly be seen inside it, right above the opening. The pad absorbs ink and releases it at a constant rate regardless of how much ink remains in the tank.

A disposable head (Figure 26.39, right) is an integral part of a replaceable ink cartridge. When the ink runs out, the cartridge has to be replaced, which brings in a new print head. The front of the head in the figure has electrical contacts through which the head receives image information. The small bright area on the top contains the microscopic nozzles. Such cartridges are more expensive and the internal components of the heads may not be precisely made and aligned. However, when the head malfunctions, it is easy to replace.

Opponents of the disposable print head philosophy claim that its main advantage is to make it difficult to manufacture and thus harder to compete with the original maker of the printer.

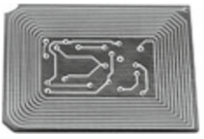
It is also possible to have a disposable print head connected to a disposable ink tank. This seems a good compromise and is used by many HP models.

Ink wars. Today, more and more men use electric shavers, but up until the 1970s, the razor blade was king (or rather, was sold by a Mr King). Early razors were thick and had to be sharpened periodically. Around 1900, a businessman named King Camp Gillette hit on the idea of disposable razor blades. He separated the razor from the razor blades, made the blades thin and inexpensive, so users simply replaced a dull blade instead of sharpening it. A while later, Gillette had the brilliant idea of increasing sales by dropping prices. He started giving away his razors, charging only for the blades. This unusual pricing strategy (referred to as the razor and blades business model) proved its worth quickly. Within a decade, Gillette's company dominated the razor and shaving market and his innovation has since been adopted by many businesses, most notably inkjet printers and cell telephones.

Printer makers were quick to adopt the Gillette strategy. Inkjet printers for home use are generally inexpensive, but the ink cartridges are not. Thus started the ink wars. Third-party vendors (also referred to as aftermarket) started competing by selling cheaper cartridges.

A typical ink cartridge contains 5 ml of ink and sells for about $5. This comes to $1,000 per liter or $3,800 per gallon. At such prices, it is easy and profitable for the aftermarket to compete, especially since the ink is made of cheap ingredients such as dyes, water, and glycol or alcohol.

The big printer makers fought back by installing integrated circuits (chips) in their cartridges and making printers that refused to use a cartridge without the chip. This practice is referred to as product tying. The aftermarket competitors reverse engineered those chips and installed them in their own, moderately-priced products. The big manufacturers, HP, Lexmark, and Epson among them, retaliated by filing patents on their cartridges and suing, citing the Digital Millennium Copyright Act (DMCA). The aftermarket vendors responded by launching counter class-action lawsuits, citing anti-trust laws.



At the time of writing (mid 2010) most courts have ruled that aftermarket products do not violate the DMCA, but if a manufacturer receives a patent that covers every possible way of reverse engineering its products, then no one else has the right to duplicate them.

A refreshing step was taken in 2007 by Eastman Kodak. This company entered the inkjet market with a line of All-In-One printers where the printer itself costs more but the ink cartridges cost less. At this point, it is not clear how successful this strategy is, but at least it has avoided costly litigation.

One result of the ink wars has been a price increase of old printers, from the days before the cartridges had chips. Online auction sites are full of offers of new (but no longer made) printers which often fetch high bids of more than their original price.

Cleaning mechanisms. Inkjet printers were made as early as the mid 1980s, but this technology really took off about five years later, when its main problem, ink drying in the printhead's nozzles, was finally solved. If the ink dries off, the pigments become

solid and plug the microscopic nozzles. The first solution tried by printer makers was to cover the printhead with a rubber cap as soon as the printer finishes its job. This solution is not ideal because the small rubber caps do not provide a perfect seal, causing ink to dry over a period of days to weeks.

Once this was understood, a cleaning function was included in inkjet printers. The printer-driver software offers such an option and can automatically clean the nozzles by spraying a small amount of ink to moisten and perhaps completely flush any solid ink deposits. Following this, a wiper blade is swept across the nozzles in the printhead. These steps can be repeated several times.

The ink sprayed in this process must be collected somewhere, so printers often have a collection tray, known as a spittoon, with a large absorption pad, located at the bottom for this purpose. When the tray is full, the printer stops working, and the tray has to be drained (either by the user or by a technician).

Special cleaning kits are also available for many inkjet printers. Such a kit replaces an ink cartridge with an identical cartridge filled with a cleaning solution. The driver software then prints an entire page with this liquid, and this solves most clogging problems.

Another point to consider is that the precise composition of the "official" ink brand made by a printer manufacturer is a trade secret. A manufacturer always claims that its ink is formulated to match its printing mechanism, and such claims may have some merit. Therefore, generic aftermarket inks are somewhat different from the original ink and some may contribute to printhead clogging.

Ink drying is affected by another source. While ink flows from the reservoir to the printhead, air must be let into the reservoir. Every ink cartridge has a long, narrow tube or channel that wraps back and forth around the tank and carries air inside, to replace the ink. Figure 26.40 shows such a channel (after the tape that normally covers it has been removed). Ink evaporates slowly through this channel, which causes the ink to dry from the inside of the tank out.

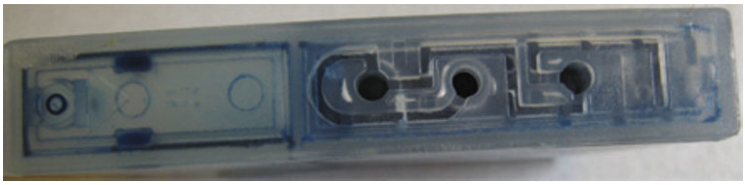


Figure 26.40: Air Channel (Exposed).

The following printer cleaning technique was related to me in 2008 by Tim Taylor, a computer technician.

> "I bought two identical printers and I swap them every six months or so. I wash one printer and let it dry for several months while I use the other printer. I put the entire printer into the sink and hose it out. The heads can be flushed separately with hot water under the sink faucet. There is nothing in the printer that can be damaged by the water as long as it is dry before putting power to

it. This also cleans all the other parts of dust and ink that may get damaged over time if left dirty."

Inkjet chief features. The advantages of inkjet printers are obvious to anyone who has used one. Currently, these printers feature high resolutions and many colors. With the right kind of paper, an inexpensive, little inkjet printer can produce a sharp, glossy color image comparable in quality to the best-printed photographs. The printers are also quiet and lightweight (compared to the old dot matrix and daisywheel printers and the modern laser printers). The cost of ink per page (especially ink from third-party suppliers) is higher than that of a laser printer, but much lower than the cost of developing and printing a traditional photograph. A laser printer takes a while to warm up each time it is turned on, but the overhead time of an inkjet printer is negligible.

The shortcomings of inkjet printers are less obvious and take longer to find out. Here is a list of the main ones.

- Often the chip installed in the ink cartridge is "intelligent." It tries to estimate the amount of ink required for the next page. If it decides that there is not enough ink, it declares the cartridge empty and stops the printer until a new cartridge is inserted. However, the estimate of the chip may often be off by a wide mark, thereby causing the user to remove and throw away a nonempty cartridge; an annoying, costly, and unintelligent feature.

- The shelf life of an inkjet cartridge (especially those using volatile inks) is limited because of evaporation and is shorter than the life of a laser cartridge. (But I have several ink cartridges for the Brother MFC240C all-in-one that are 3–4 years old and yet work fine.)

- An ink cartridge is prone to clogging. Once installed in the printer, the cartridge should be used at least every few days (by printing a special multi-color test pattern) to prevent drying of the ink and clogging. The printer can clean itself, but this process wastes ink.

- The colors on the printed page change over time and tend to fade. This is especially true for aqueous inks. Important documents should be printed with special archival inks.

- A drop of water on a freshly-printed paper may dissolve the aqueous ink and cause runs and serious print damage. An inkjet-printed page should be left to dry for about 24 hours before it is handled. Water-based highlighter markers can also damage the print.

Figure 26.41 shows a Canon PIXMA inkjet printer, a typical printer made for home use. Like other inkjet printers, this model is driven by proprietary software (in firmware). Over time, hackers have disassembled this software and have discovered how to program the printer to print in duplex mode and to print on CDs and DVDs (see [photo.net 11] and [pixma.ulmb 11]).

Figure 26.41: A Canon PIXMA Inkjet Printer.

26.11 Solid-Ink Printers

Perhaps the most serious drawback of an inkjet printer is the tendency of liquid ink to dry up and clog the narrow nozzles. This problem does not exist in a solid-ink printer, which is one reason why such printers have been developed since the early 1990s, in parallel with liquid-ink (inkjet) printers.

A solid-ink printer uses wax-like ink (also referred to as phase-change ink or hot-melt ink) that is solid at room temperature. The ink is melted as soon as the printer is turned on, and is sprayed on the paper with a piezoelectric pump, much like liquid ink. The chief advantages of this printing technology are (1) the ink dries very quickly on the page and (2) it does not dry up in the nozzles because it re-solidifies quickly when the printer is shut down or is put to sleep.

The first technically (but not commercially) successful solid-ink printer was developed in 1991 by Tektronix. It was large, heavy, expensive, had low resolution, and was slow (it took about two minutes to print a single `A`-size, 8.5×11 in sheet). This printer and its immediate successors were based on a simple approach to printing. The print head moved left-to-right across the page, spraying ink droplets of various colors from 16 nozzles to print a 16-pixel-tall (i.e., a very narrow) stripe. The paper was then moved up and the head moved back (right-to-left), spraying another narrow stripe.

In order to speed up the printer, the printing head had to accelerate and decelerate rapidly, and so had to be very light in spite of having 16 nozzles. An even greater problem was the placement of the ink droplets. Recall that two primary colors have to overlap in order to print a pixel in a secondary color. If the head sprays a magenta droplet on top of a cyan droplet in its first pass, then it sprays these droplets in the reverse order when it is going back, in its second pass. This causes slight differences in the hues of the resulting blue pixels, and the eye is very sensitive to hue.

These early solid-ink printers suffered from another problem. The gap between the paper and printing nozzles had to have a fixed width. When heavy, thick paper was used, the gap narrowed, which resulted in lower print quality. Clearly, a different, innovative approach to printing was needed, and Tektronix engineers came up with such an approach in 1995 [solid-ink 11]. The principle (Figure 26.42) was to spray ink droplets on a spinning drum from a printing head that moves in steps along the drum. (This is a little similar to a cutting head moving slowly in a lathe, cutting a spiral in a steel cylinder to make a bolt.) The circumference of the drum equals the height of the paper. In each drum revolution, a stripe of pixels in the form of a ring is sprayed on the drum and then the printing head moves to the next drum area. It takes several revolutions to spray an entire page-worth of ink, as a set of stripes (or rings) of pixels, on the drum. By then, the ink has cooled and become solid but soft. The image is then transferred (offset) from the drum to the paper. This simplifies the paper path, which can be a straight line, but complicates the chemical composition of the ink. The ink must solidify on the drum while it is spinning and must completely transfer to the paper.

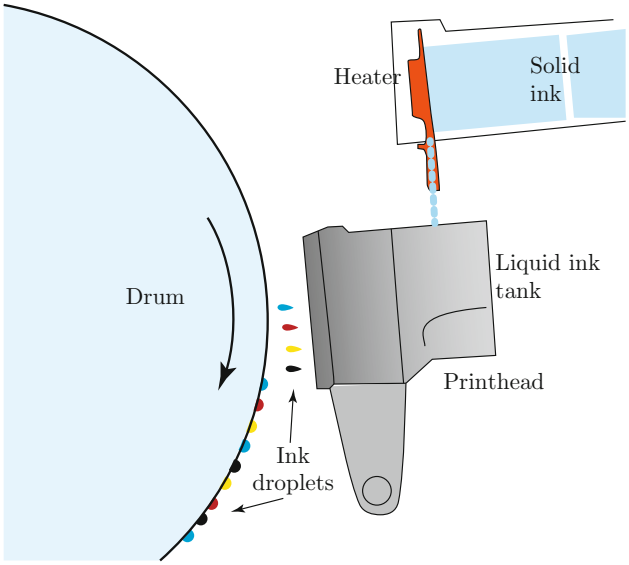


Figure 26.42: Solid Ink Printing Engine.

Imagine a printing head with 300 columns and four rows of minute apertures. Each row sprays ink droplets of one of the four CMYK colors. In one revolution, this head sprays a 300-pixel-wide stripe on the drum. The head then moves to the next empty drum area and sprays another stripe. If the printing resolution is 300 dpi, then only eight head movements (and eight drum revolutions) are needed to spray an 8-in-wide page on the drum. In the ninth revolution, the paper is pressed to the drum and the soft,

solid ink is transferred to the paper. During this revolution, the drum is also cleaned and its surface is oiled to prepare it for the next page.

This technique makes it possible to maintain the correct gap between the printing head and the drum. The gap does not depend on the paper, because the printing (offset) occurs after the entire image has been sprayed on the drum. Thus, this type of printer can print on a wide variety of papers.

The print head contains liquid ink held at 135° C, and the drum is kept at 65° C. Ink drops hit the oily surface of the drum and are instantly cooled down to its temperature to become semisoft. The paper is warmed on its way to the drum, and is pressed between the drum and a pressure roller. The ink transfers to the paper and is fully set and fused on the paper by the time the paper emerges out of the printer.

Notice that while the ink is transferred to the paper it is no longer liquid. Thus, the ink does not soak into the paper but instead fuses to it. As a result, prints produced by such a printer are water-fast, in contrast to the output of an inkjet printer.

In 2000, Xerox acquired the Tektronix color printing and imaging division and started to make its well-known Phaser line of solid-ink printers [Phaser 11].

Advantages of solid-ink printers are as follows:

- The solid ink is not absorbed into the paper like liquid ink. It sticks instead to the surface of the paper, which results in vivid colors and allows for a color gamut wider than that of inkjet printers.
- High-quality color images can be printed on many types of paper. An inkjet printer requires glossy paper to produce vivid colors, but a solid-ink printer covers the paper with glossy wax, so the paper itself can be matte.
- It takes a few minutes for a solid-ink printer to melt the solid ink and be ready, but then the first page is printed quicker than in inkjet or laser printers.
- The solid ink sticks come in different shapes depending on color. This prevents insertion of an ink block into the wrong container.
- Eco friendly. No empty ink or toner cartridges need be recycled. No ozone is produced during printing.
- Non-toxic ink. The ink blocks (or sticks) are safe to handle. In the mid 1990s, the president of Tektronix actually ate part of a stick of his company's ink, thereby demonstrating their safety (where is he now)? It has been claimed that the ink is made from food-grade processed vegetable oils.
- The printer is less sensitive to paper quality, so recycled paper can be used.
- Third-party solid-ink blocks are available and may be considerably less expensive than the original ink.

The chief disadvantages of solid-ink printers are the following:

- The wax-like ink can get scraped off the page, especially when the printer is set to the highest-print quality (which sprays more ink).
- When the printer is turned on, it may take several minutes (sometimes up to 15 minutes) for the ink to melt.

- Power consumption. When the printer is put to sleep, it maintains the ink at near melting point, which consumes 30–50 watts. In contrast, an inkjet printer may consume 2–5 watts when not actually printing.

- When power is lost, even momentarily, air enters the print-head and the printer mechanism must flush some ink from the print-head to the waste tray. This wastes ink, so in locations where electricity is erratic, an uninterrupted-power-supply (UPS) should be used with this type of printer.

- Before the printer can be moved or transported, the ink has to cool down completely. The shut-down cycle, which employs fans, may take 10–20 minutes.

- It is common for a company to print a large number of letterheads with the company logo and address. A sheet of preprinted letterhead may later be fed into a printer to print text and images. Because of the nature of the wax-ink, a solid-ink printer is unsuitable for preparing letterheads (heat in a laser printer will melt the wax of the letterhead).

- Printed documents fade over time because ultra-violet radiation from the sun interacts with the organic compounds that constitute the solid ink.

26.12 Laser Printers

Laser printers are in common use today because of two main features, they print on plain paper and they are fast. In contrast with an inkjet printer, which sprays liquid ink on the paper, a laser printer exploits electrostatic attraction to pull dry ink onto the paper. Laser printers are based on the technique of xerographic printing, which was originally developed for (analog) document copiers, with the difference that the image printed by a laser printer is digital and is produced by a laser beam that performs a raster scan. Most laser printers are monochrome, but there are color models which unfortunately produce flat, lifeless prints that cannot compare with the glossy images generated by inkjet printers.

The laser printer was invented in 1969 by a researcher at Xerox. It took less than a year to develop the concept to a fully functional prototype printer. This prototype and the models that follow it were extremely large, bulky, and expensive, but their print quality exceeded anything available at the time, so those privileged to have access to these printers loved the results. In 1981, Xerox introduced another laser printer to be used with its innovative Star 8010 computer, but it was only in 1984, when HP introduced the first laserjet (sold for $3,500), that laser printing became commonplace. IBM, Brother, and other manufacturers started competing in this field, which encouraged innovation and brought down prices to a level where many homes have such printers. The HP Laserjet P1005 printer shown in Figure 26.43 has a footprint of 8×14 in, weighs about 11 lbs and was purchased by the author in 2008 (it is now discontinued) for $60! Reference [yarin 10] has a short history of laser printer development.

When an application sends an image to be printed on an inkjet printer, it can send it pixel by pixel. The inkjet printer collects enough pixels for a complete row of the image, prints the row, and can wait for more data from the computer. In contrast, when a laser printer starts printing a page, it should not be stopped because this would

Figure 26.43: The HP Laserjet P1005 Printer.

introduce a visible gap or misalignment of the dots on the printed page. The printing mechanism (drum) must rotate continuously at least one revolution and print a sheet of paper.

Because of this requirement, images sent from a computer to a laser printer are often in PostScript format (Section 20.5) or some other page description language such as HP Printer Command Language (PCL) or Microsoft XML Page Specification (XPS). Raster image processor (RIP) software inside the printer converts the image into a complete bitmap that is stored in the printer's memory. From this memory, the image is sent at a constant rate as a stream of pixels, row by row, to the laser.

Photoconductivity is a well-understood physical phenomenon, involving both optical and electrical properties, in which a photoconductive material increases its electrical conductivity when it absorbs electromagnetic radiation.

Common examples of photoconductive materials are (1) selenium (Se, element 34, a nonmetal that is chemically related to sulfur and tellurium), which is used chiefly in photocopying (xerography) and (2) lead sulfide, used in infrared detection applications (such as heat-seeking missiles).

The heart of the printing mechanism is a drum coated with selenium. It takes a full revolution of the drum to print one sheet of paper. Printing a single sheet is done in the following stages (Figure 26.44):

1. The drum is first given a uniform positive electric charge by the corona wire (in newer models, by a charged roller). The voltage of the corona determines the amount of charge, which in turn controls the print density.

2. The laser beam is then reflected by a polygonal mirror and is swept across the drum to scan a row of image pixels (Figure 26.45). Each black pixel turns the beam on and each white pixel turns it off. As the beam scans a row on the drum, each time

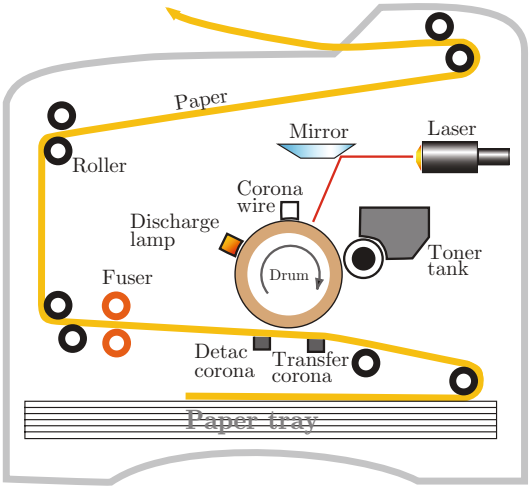


Figure 26.44: Paper Path in a Laser Printer.

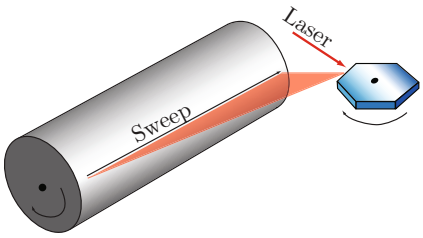


Figure 26.45: Sweeping a Laser Beam.

it is on, it reverses the electric charge on the drum to negative. Thus, the laser beam "writes" the image as negative charges on the drum, row by row.

3. Positively-charged toner (fine particles of dry plastic or wax powder mixed with carbon black or coloring agents) is released from the toner cartridge and is attracted to the negatively-charged areas of the drum. The toner is repelled from the positively-charged areas.

4. A sheet of paper is then slid from the paper tray, is strongly loaded with negative charge by the transfer corona wire, and is passed very close to the drum. Toner particles are attracted to the paper more than to the drum, so they end up on the paper. The paper then passes over the detac corona wire (or roller) that completely discharges it to prevent it from clinging to the drum.

5. The image now resides on the paper in the form of loose toner powder. It has to be bonded to the paper quickly, because any air movement or vibrations would blow it

up. Bonding is done by passing the paper through a fuser, which consists of a pair of hot rollers. The fuser temperature (up to 200° Celsius) and the paper speed are adjusted such that the toner melts and is diffused into the paper, but the paper itself does not have enough time to burn, and it emerges warm from the printer.

One roller of the fuser is often rubber and it presses the paper against the other roller which is hollow and contains an infrared lamp at its center. The lamp heats the roller from the inside to achieve uniform temperature over the entire roller. The fuser is responsible for the (relatively long) warm-up time of laser printers and also for up to 90% of their energy usage. Care must be taken to ensure that the heat is properly vented outside and does not damage sensitive printer parts. When printing stops for a few seconds, the fuser is automatically turned off, to save energy.

6. In the last stage, an electrically neutral soft plastic blade wipes any excess toner from the drum and deposits it in a waste container (this toner cannot be reused because it may be contaminated with dust and paper particles). The drum then passes under a discharge lamp whose strong light erases any residue electrical charges. If another image is ready to be printed, the drum continues its rotation and passes under the corona wire as in stage 1.

The signs (positive and negative) of electrical charges vary with printers and may be the opposite of the ones described here. Other details of printer operation may also differ from the above description.

Printer properties. A laser printer is non-impact. The image is printed by the toner particles, but no other printer parts actually come into contact with the paper. Printing resolution can be high, because the laser beam is focused and concentrated. Currently, speeds can reach about 200 pages per minute, or 3.3 pages per second. A typical 8.5×11 in sheet of paper with 1-in margins on all sides has a print area of $6.5 \times 9 = 58.5$ square inches. At 600 dpi, this translates to $58.5 \times 600^2 \approx 21$ million pixels. Thus, the laser and mirror in such a printer must be fast enough to scan about 21 million pixels each second! (Early laser printers did not have much memory, which is why they could print only text characters and not arbitrary images.)

The mechanical work in a laser printer is done by the drum and the roller that transfers toner from the reservoir to the drum. Because of this, the drum assembly tends to wear out, which is why many laser printers employ toner cartridges that also include a drum assembly. This increases the price of a cartridge, but saves many printing problems, costly repairs, and frustration in the long run.

In the early 2000s, new printer models with duplex printing were introduced. Such a printer can print on both sides of the paper, thereby saving paper. After one side is printed, the paper is almost completely ejected from the printer, but is stopped, pulled back inside, is turned over, and its other side is printed. Figure 26.46 shows one way of implementing the paper path in a duplex printer. The duplexing mechanism requires a longer paper path, which slows down the printer.

Color Laser. A laser printer can print in color by printing four images on the same sheet of paper. For each color image, software in the computer has to send the printer four bitmaps that correspond to the CMYK colors and the printer then prints the bitmaps on the same sheet of paper, each time with toner of the right color. The two main techniques for implementing this are as follows:

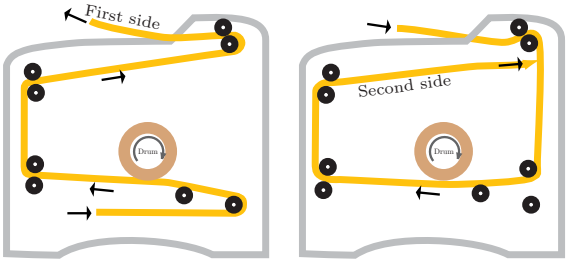


Figure 26.46: Paper Path for Duplex Printing.

- There are four toner tanks on a rotating platform. The printer rotates the platform until the Cyan tank is positioned next to the drum. The cyan image is then transferred to the paper, the platform is rotated until the magenta tank is positioned next to the drum, and that image is transferred to the paper. The complete image is ready after four such steps and it is fused. The main problem is rotating the platform precisely each time, because even the smallest misalignment is highly visible.

- The printer has four complete printing units, each with its own laser, drum, and fuser. The paper moves from one unit to the next, collecting all four colors quickly. The main problem is positioning the paper precisely at each printing station in order to avoid any misalignment. This type of color printer is fast but expensive.

Laser Steganography. The availability of high-resolution scanners and color laser printers has raised the problem of counterfeiting important documents such as money notes. One way to discourage counterfeiting is to print secret identification dots that encode important data such as the printing date, time, and printer's serial number on each sheet printed. The dots are yellow and have a diameter of about 0.1 mm, making them difficult to see. This is an example of data hiding, or steganography, and it has raised concerns of privacy (see, for example, [eff 10]).

Figure 26.47 shows 11 vertical markers with some dots. To get an idea of the size of the dots, imagine that the markers are separated by 1 mm. The dots in the image are black, for better visibility, but in a real image they are yellow.

Figure 26.47: Anti-Counterfeiting Steganographic Marks.

26.13 Plotters

A plotter is a graphics output device for printing drawings (technical, architectural, artistic, and other line art) in monochrome or color. From the previous sections in this chapter it is clear that current printers (inkjet and laser) can print any images, not just drawings, which is why plotters are currently not very popular and are used mostly for large technical drawings and also to cut patterns out of vinyl sheet and combine them to construct signs, posters, large banners, and billboards.

Calcomp (Californai Computer Corp.) was an early maker of plotters. Its model 560 drum plotter was one of the first graphics output devices ever (this was as early as 1959). The drum was 11 inches wide and both it and the pen could move in steps of 0.01 in. The company also wrote a subroutine package that made it easy to specify any drawing in a Fortran program.

For comparison, the Calcomp 563, introduced around 1973, had a 30-in-wide drum, the step size was 0.005 in, and the plotter could execute 300 steps (about 1.5 in) per second. A special plotter control card could be inserted into a PDP-8/E minicomputer, to help in controlling the plotter by software.

Notice how the speed of a plotter is measured in steps per second, instead of pages per minute as in printers.

Other technology companies, most notably HP and Tektronix, started making plotters in the 1960s and 1970s. However, with the advent of fast, inexpensive, high-resolution printers in the 1980s (and also the introduction of computers fast enough to rasterize color images), the demand for pen plotters dwindled and today older plotter models are mostly museum pieces, even though they may be in perfect working order.

Today, pen plotters are used mostly for technical and architectural drawing and CAD applications, where they excel because they can handle large paper sizes and plot large drawings in high resolution. Cutting plotters are also popular and are used for cutting complex shapes out of cloth and vinyl. Special plotters create tactile images (images that can be perceived by touch) on special thermal paper.

The business end of a plotter may be a pen, a sharp blade, or an inkjet nozzle. Early pen plotters used small, proprietary fiber-tipped or plastic nib disposable pens, but plotter makers later switched to technical pen tips. Currently, ball-point pens can be used (some after modifications) in many pen plotters.

In contrast with the limited use of pen plotters, cutting plotters have become popular. Both professional sign makers and private individuals can afford such a plotter, which makes it easy to design and cut complex patterns out of self-colored adhesive-backed vinyl sheets that have a removable paper backing material. The patterns can later be assembled into large signs and posters. Rolls of inexpensive vinyl in many colors are made in 24-in and 36-in widths. Cutting speeds are comparable to printing speeds, with the difference that the sharp blades have to be replaced periodically. The blades are normally shaped like plotter pens but are mounted on ball bearings so that the sharp edge can easily turn and always face the direction of cutting.

Sometimes, a design should only be partly cut out of the vinyl, which is why a cutting plotter often has a pressure control. The pressure exerted by the blade on the vinyl can then be adjust by software.

The material being cut by a plotter must have strong backing. It is easy to see

why, without backing, cutting a hole or a slit will cause the material to shrink, deform, droop, and generally get out of alignment. If the material does not (or cannot) have backing, it may be held in a flatbed cutting plotter by vacuum under the bed.

Recently, large-format inkjet printers started replacing cutting plotters. An inkjet printer can use special, slow fading, UV-resistant solvent-based inks, and can print directly onto fabrics, vinyls, or certain plastics. Such inkjet printers can produce smooth color transitions and raster printing, which gives them an important advantage over cutting plotters. Thus, in future, cutting plotters are expected to concentrate on applications where the material has to be both printed and cut in complex shapes.

There are two chief types of plotters, flatbed and drum.

Figure 26.48: A Flatbed Plotter.

In the former type (Figure 26.48), the paper is stationary; it lies in a bed "doing nothing" while the pen is moved over it in small steps in the x and y directions. The pen is mounted on a carriage along which it moves in the x direction, while the carriage itself moves in the y direction. In each step, the pen may be brought down (actually writing or cutting) or raised up.

In the latter type (Figure 26.49), the paper (often a very long sheet) is wrapped over a drum. Rotation of the drum moves the paper in the x direction, while the pen is moved over the paper, along the drum, in the y direction. The length of the drum limits the width of the paper, while the length of the paper is limited by the capabilities of the software.

In either type, flatbed and drum, a fast plotter requires a lightweight (i.e., low inertia) carriage and pen assembly. Figure 26.49 illustrates how small and lightweight a carriage can be. In a color plotter, the pen assembly consists of three to six pens of different colors, and the software has to specify which pen to use at every step.

For most of the history of drum plotters, the paper had perforations on both edges (clearly visible in Figure 26.49), to prevent slippage of the paper on the drum. In the 1980s, HP introduced plotters that could draw on plain, unperforated paper. The idea (dubbed "grit wheel") was to have grit wheels that pressed on the edges of the paper, thereby forming small indentations. When the paper had to be moved backward, the

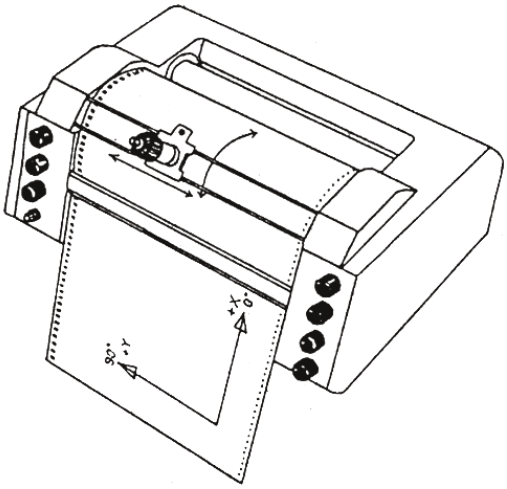


Figure 26.49: A Typical 1970s Drum Plotter.

grit particles on the wheel dropped into the perforations they made earlier and thus insured perfect movement of the paper with the drum.

Once the principles of flatbed and drum plotters are understood, it is easy to see why pen plotters are limited to line drawings. Here are the main reasons:

- Trying to paint a solid region on the paper can be done by moving the pen in parallel lines densely drawn over the region, but the results are disappointing and often cause wrinkles or tears in the paper. A solid area can be simulated by hatching (drawing several parallel lines in the area).

- In order to plot an arbitrary image, which may consist of millions of pixels, the pen would have to stop at every pixel and print a small dot. This is extremely slow and the large number of ink dots may result in bleeding over neighboring dots and noticeable visual degradation of the final image.

A plotter can draw straight lines and curves, but both are drawn as sets of short, straight segments. In a low-quality plotter, especially of the drum type, this kind of operation may lead to errors as illustrated by figure 26.50. The plotter draws the short segment from 1 to 2, then the long segment from 3 to 4. When it starts the third segment (from 5 to 6), it has to move the paper in reverse to point 5, a relatively long distance. Any slippage of the paper on the drum would cause an error and in a large, complex drawing that may consist of thousands of segments, such errors may accumulate. The figure also shows (in part 7) how a long, straight line is drawn as a set of short segments (horizontal, vertical, or at 45°), each the result of one low-level plotter command.

The plotter hardware executes basic commands of the form (x, y, pos), where x and y can be -1, 0, or 1 (for reverse, no move, and forward), and pos can be 0 (pen up) or

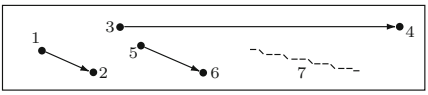


Figure 26.50: Plotting Three Lines.

1 (pen down). Thus, the triplet $(0, -1, 0)$ instructs the plotter to move one step back in the y direction with the pen up, while $(1, 1, 1)$ means to move forward in both directions (i.e., in a 45° direction) with the pen down (i.e., while drawing a short segment on the paper). The user generates the drawing in a higher-level plotter control language which generates commands such as "`move the pen up to absolute location` $(3,000,113)$" or "`move the pen down from its current location 1,240 steps in` x `and -500 steps in` y." These commands are then converted to basic triplets and are sent to the plotter for execution. Another option is to use a general-purpose programming language, such as C, combined with plotting routines taken from a software library.

Two common ASCII-based plotter control languages are Hewlett-Packard's HPGL2 and Houston Instruments DMPL. Examples of special software packages for plotting are the Calcomp library, Hewlett-Packard's AGL libraries, and the high-end routines that constitute the DISSPLA package (Display Integrated Software System and Plotting Language [dissplа 11]).

26.14 Interactive Devices

Many computer graphics applications, such as drawing, illustrating, and virtual reality, are interactive, which is why pioneers in the CG field have come up with specialized, interactive graphics I/O devices that transmit data to and from the graphics software. This data is not generated by a keyboard or a mouse but arrives in a form that is natural to the user. Such devices can convert hand or head movements and gestures to signals that are input by the software and are interpreted in different ways depending on the application. Most interactive devices are input, but certain outputs can also be sent by the graphics software to some interactive devices.

Two such devices, the wired glove and the head-mounted display, are described in this section.

26.14.1 The Wired Glove

A wired glove, also known as a data glove, is an input device worn on the hand like a glove. Sensors built into the glove capture data such as the bending of fingers and rotation of the wrist. The simpler types of data glove use potentiometers (or optical flex sensors) located at the knuckles, fingertips, and finger joints. Such devices receive an input voltage and vary it according to the amount of flexing at each joint or knuckle. The output voltage of a potentiometer is digitized and it tells the software how much flexing happened at the knuckle. More expensive gloves employ a magnetic or inertial tracking device that can output the absolute location of each knuckle and fingertip. High-end data gloves may contain actuators that apply forces, vibrations, movements, and other

types of haptic feedback to the user's hand, thereby acting also as an output device, adding the sense of touch to previously visual-only outputs, and in general doing for the sense of touch what computer graphics does for vision. Typical haptic actuators include vibratory motors, electroactive polymers, piezoelectric crystals, and electrostatic surface actuation.

The word haptic, from the Greek $\alpha\pi\tau\iota\kappa o\varsigma$ (haptikos), means pertaining to the sense of touch. It comes from the Greek verb $\alpha\pi\tau\epsilon\sigma\theta\alpha\iota$ (haptesthai), meaning to contact or touch.

Data gloves have a long history and go back to the Sayre glove of 1977. Most of the development of these gloves was done in the 1980s and 1990s. Recently, several models of wireless gloves have been introduced. Such a glove must have a power source (battery) and it sends its signals to a special receiver that is plugged into a USB port in the computer.

The cost of a typical wired glove is high, so [Pamplona et al. 08] proposed a simple, inexpensive input device that they dub the image-based data glove. Markers are attached to four of the five fingertips (the thumb is left bare) with a different dot pattern printed on each marker. A camera is attached to the arm and there is no actual glove. When the fingers are bent, the camera follows the movements of the individual fingertips and uses this information to estimate the positions of the joints and knuckles. When a motion-tracker device is added to this configuration, it can also map pitch, yaw, roll (Figure 4.26), and XYZ-translations of the user's hand, and recreate almost perfectly all the gestures and postures performed by the hand.

The new wireless CyberGlove II [vrealities 10] is an example of a wireless modern data glove. This is a fully instrumented glove that provides up to 22 high-accuracy joint-angle measurements. It uses proprietary resistive bend-sensing technology to accurately transform hand and finger motions into real-time digital joint-angle data.

The 18-sensor model features two bend sensors on each finger, four abduction sensors, plus sensors measuring thumb crossover, palm arch, wrist flexion, and wrist abduction.

The 22-sensor model has three flexion sensors per finger, four abduction sensors, a palm-arch sensor, and sensors to measure flexion and abduction. Each sensor is extremely thin and flexible being virtually undetectable in the lightweight elastic glove.

The CyberGlove II Wireless Glove transforms hand and finger motion into real-time digital joint-angle data—and works without cumbersome wires that can impede movement and slow your project.

Reference [Sturman and Zeltzer 94] is a survey of sensor technologies used in wired gloves.

26.14.2 Head-Mounted Display

A head-mounted display (HMD, also known as a helmet-mounted display) is often an output device, but can also be an I/O device. The device is sometimes attached to several belts that are mounted on the head, but it can also be part of a helmet. The output part of an HMD is a small display optic placed in front of one eye or both eyes (monocular or binocular HMD, respectively). The input part, if it exists, consists of sensors, similar to the ones inside a wired glove, that sense the user's head movements

and generate data that can be input by the software, interpreted, and acted upon.

The idea of an HMD was first proposed by Ivan Sutherland, an early pioneer of computer graphics. As early as 1968, he demonstrated an HMD that followed the user's head movements and could, with the help of software, vary the display accordingly. This may have been the first implementation of virtual reality. In addition to being binocular and having sensors in the helmet, this HMD also included a hand-held wand whose position was also tracked (by means of inertial sensors) by the software.

A typical HMD includes one or two small displays with lenses and semi-transparent mirrors. These are either embedded in a helmet, attached to normal eye-glasses, or clipped to a visor. The resolution of the display units is one of the chief factors in determining the price of the HMD.

There are three main types of HMDs according to what they display as follows:

- Only computer-generated images (CGI) can be displayed. This is the most-common type.
- Only real-world views are displayed. This may be useful for night vision or in cases where telescopic vision is needed.
- A CGI is superimposed on a real-world image. This is the most useful (and also most expensive) type of HMD and is known as augmented reality. Such superimposing can be done either optically or electronically. In the former case, the CGI is projected through a partially reflective mirror and the real image is viewed directly. In the latter type, video from a camera is mixed electronically with the CGI.

Following are a few common examples of HMD applications.

- The helmet worn by a soldier, firefighter, or a pilot may include a rugged, waterproof HMD that provides night vision and superimposes on this real scene a CGI such as flight data, maps, thermal images, or instant commands from an officer.
- Architects and civil engineers may benefit from an HMD that displays plans in three dimensions. Surgeons would love to wear an HMD that combines what they see with an X-ray image of deeper parts that they cannot see.
- Those addicted to computer games would like to "graduate" from flat scenes to three-dimensional virtual worlds.

The input part of an HMD tracks the user's head movements and varies the CGI that is being displayed according to the position and angle of the head. Thus, the user may see parts of a panoramic image as they pan their heads horizontally or tilt it vertically. This generates the illusion of looking around a virtual world and moving in any direction. When an HMD is combined with a wired glove, the user can extend his hand, point in any direction, and even touch and "grab" a virtual object.

It is possible to create stereoscopic images in an HMD (and thus provide the viewer with depth perception) with the following techniques:

- Have two video signals sent to the two screens inside the HMD.
- Send a single video signal with the left and right images multiplexed. Block the left-eye display when the right image is sent, and block the right-eye display when the

left image is sent. This technique is termed time-based multiplexing (see page-flipped techniques on Page 345).

- Have a single, large display, partitioned such that each eye can see only half the display. This is side-by-side multiplexing.

Reference [wearcam 10] is a review (with pictures) of many types of HMDs, and [sensics 08] is a user survey of HMD requirements.

The world has arrived at an age of cheap complex devices of great reliability; and something is bound to come of it.

—Vannevar Bush

Plate S.1. A Simple Panorama. Notice the Large Overlap Between Individual Images.

Plate S.2. Variations on a Theme (http://www.photofunia.com/).

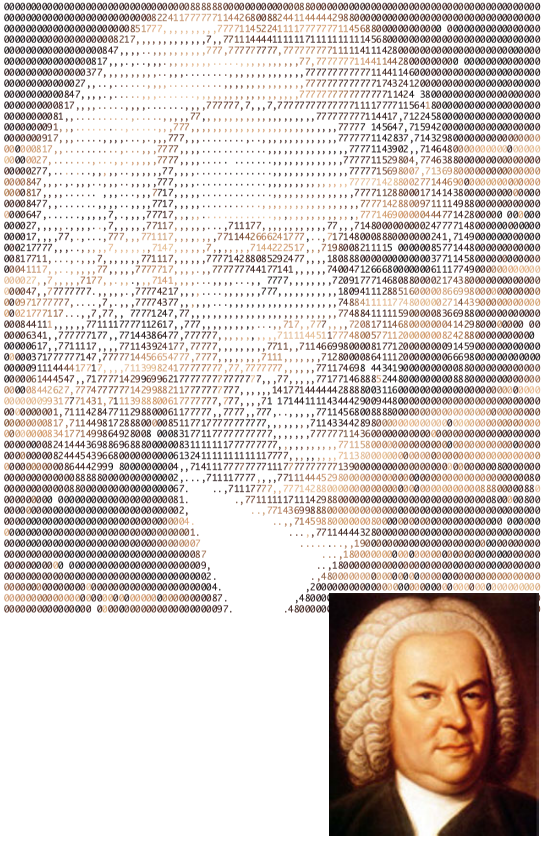


Plate T.1. Bach and Beethoven in ASCII Art.

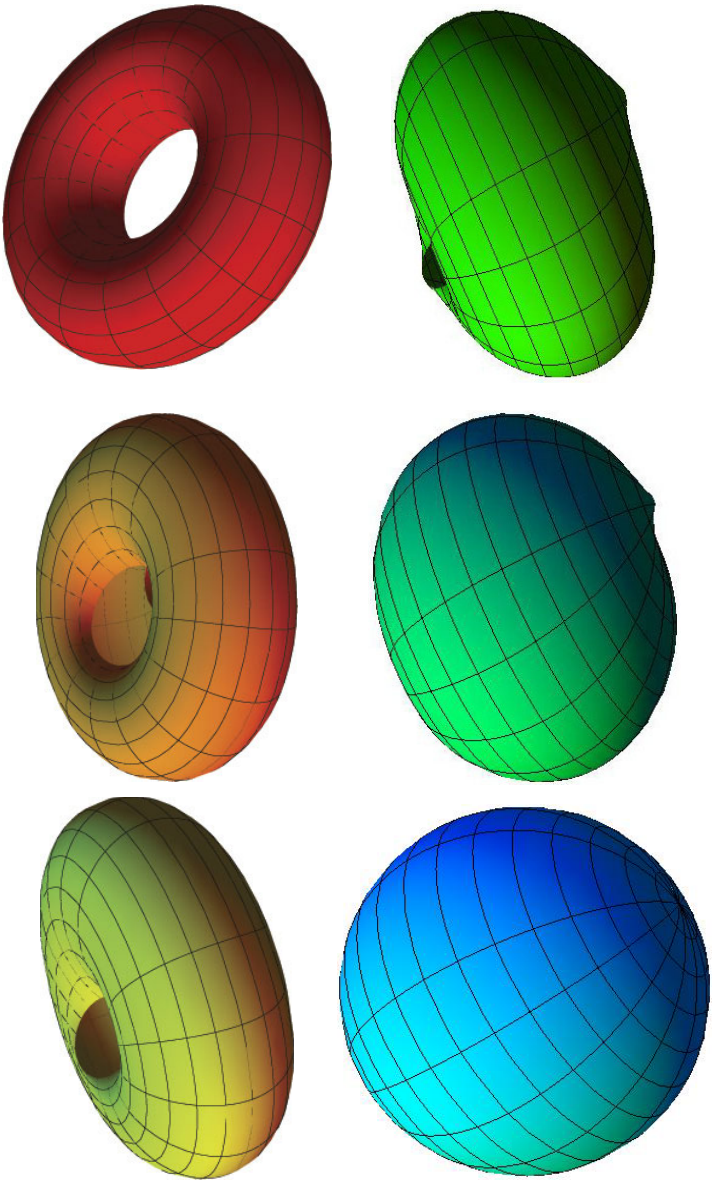


Plate T.2. Morphing a Torus to a Sphere (Mathematica).

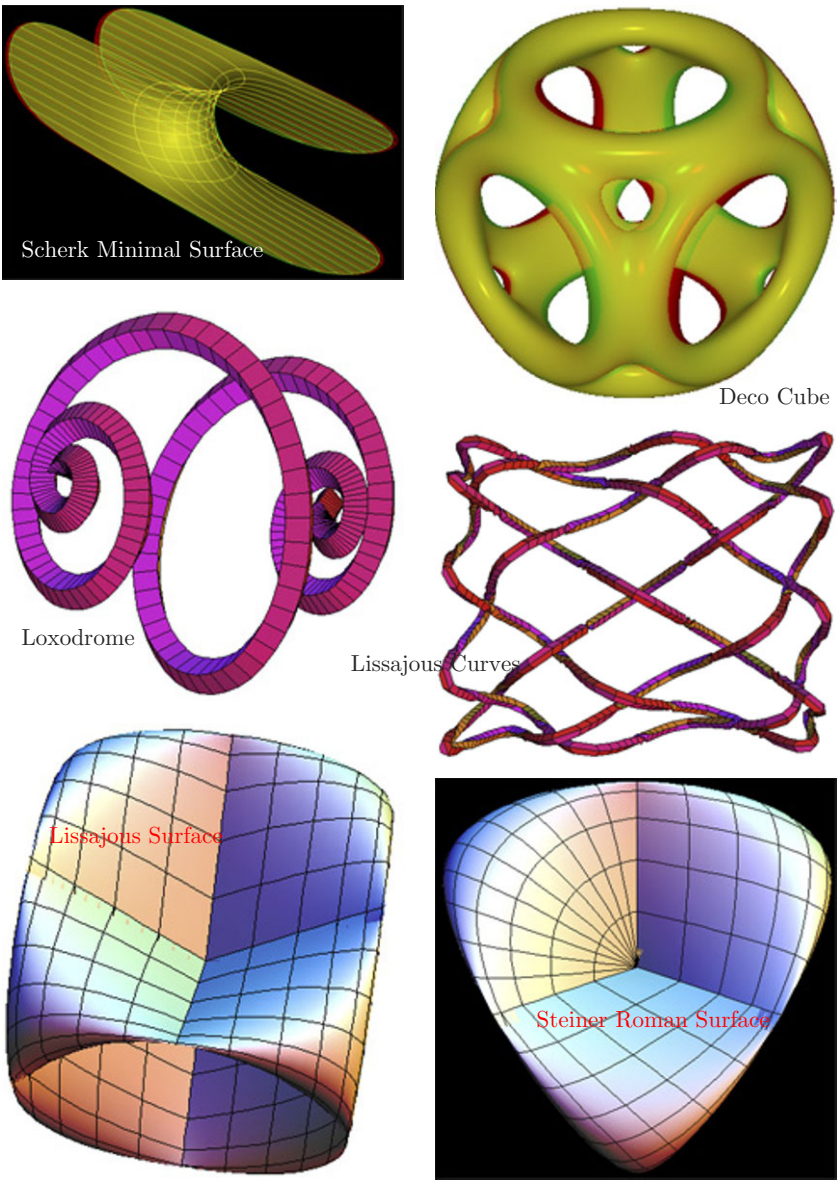


Plate U.1. Various Mathematical Objects (3D-ExplorMath).

Plate U.2. Photoshop Effects, Original, Halftone, Painting on Canvas, and Pixel Dots.