1.1

Introduction to Digital Image and Video Processing

Alan C. Bovik
The University of Texas at Austin

In this new millenium, scarcely a week passes where we do not hear an announcement of some new technological breakthrough in the areas of digital computation and telecommunication. Particularly exciting has been the participation of the general public in these developments, as affordable computers and the incredible explosion of the World Wide Web have brought a flood of instant information into a large and increasing percentage of homes and businesses. Indeed, the advent of broadband wireless devices is bringing these technologies into the pocket and purse. Most of this information is designed for *visual* consumption in the form of text, graphics, and pictures, or integrated *multimedia* presentations. *Digital images* and *digital video* are, respectively, pictures and movies that have been converted into a computer-readable binary format consisting of logical 0s and 1s. Usually, by an image we mean a still picture that does not change with time, whereas a video evolves with time and generally contains moving and/or changing objects. Digital images/video are usually obtained by converting continuous signals into digital format, although "direct digital" systems are becoming more prevalent. Likewise, digital visual signals are viewed using diverse display media, included digital printers, computer monitors, and digital projection devices. The frequency with which information is transmitted, stored, processed, and displayed in a digital visual format is increasing rapidly, and as such, the design of engineering methods for efficiently transmitting, maintaining, and even improving the visual integrity of this information is of heightened interest.

One aspect of image processing that makes it such an interesting topic of study is the amazing diversity of applications that use image processing or analysis techniques. Virtually every branch of science has subdisciplines that use recording devices or sensors to collect image data from the universe around us, as depicted in Fig. 1. This data is often multidimensional and can be arranged in a format that is suitable for human viewing. Viewable datasets like this can be regarded as images, and processed using established techniques for image processing, even if the information has not been derived from visible-light sources. Moreover, the data may be recorded as it changes over time, and with faster sensors and recording devices, it is becoming easier to acquire and analyze digital video data sets. By mining the rich spatio-temporal information that is available in video, it is often possible to analyze the growth or evolutionary properties of dynamic physical phenomena or of living specimens.

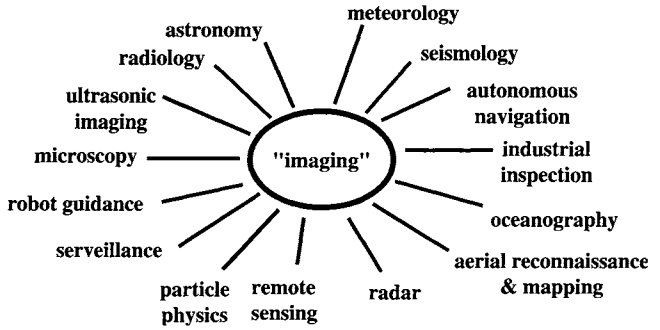

FIGURE 1 Part of the universe of image processing applications.

Types of Images

Another rich aspect of digital imaging is the diversity of image types that arise, and which can derive from nearly every type of radiation. Indeed, some of the most exciting developments in medical imaging have arisen from new sensors that record image data from previously little-used sources of radiation, such as PET (positron emission tomography) and MRI (magnetic resonance imaging), or that sense radiation in new ways, as in CAT (computer-aided tomography), where x-ray data is collected from multiple angles to form a rich aggregate image.

There is an amazing availability of radiation to be sensed, recorded as images or video, and viewed, analyzed, transmitted, or stored. In our daily experience we think of "what we see" as being "what is there," but in truth, our eyes record very little of the information that is available at any given moment. As with any sensor, the human eye has a limited bandwidth. The band of electromagnetic (EM) radiation that we are able to see, or "visible light," is quite small, as can be seen from the plot of the EM band in Fig. 2. Note that the horizontal axis is logarithmic! At any given moment, we see very little of the available radiation that is going on around us, although certainly enough to get around. From an evolutionary perspective, the band of EM wavelengths that the human eye perceives is perhaps optimal, since the volume is data is reduced, and the data that is used is highly reliable and abundantly available (the sun emits strongly in the visible bands, and the earth's atmosphere is also largely transparent in the visible wavelengths). Nevertheless, radiation from other bands can be quite useful as we attempt to glean the fullest possible amount of information from the world around us. Indeed, certain branches of science sense and record images from nearly all of the EM spectrum, and use the information to give a better picture of physical reality. For example, astronomers are often identified according to the type of data that they specialize in, e.g., radio astronomers, x-ray astronomers, etc. Non-EM radiation is also useful for imaging. A good example are the high-frequency sound waves (ultrasound) that are used to create images of the human body, and the low-frequency sound waves that are used by prospecting companies to create images of the earth's subsurface.

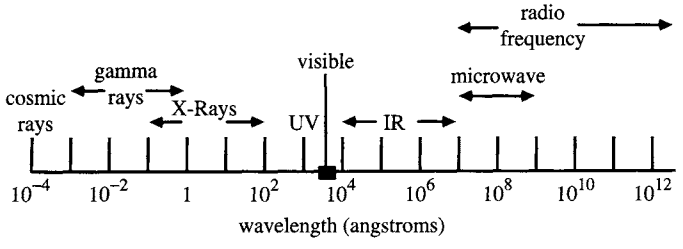

FIGURE 2 The electromagnetic spectrum.

One commonality that can be made regarding nearly all images is that radiation is emitted from some source, then interacts with some material, then is sensed and ultimately transduced into an electrical signal which may then be digitized. The resulting images can then be used to extract information about the radiation source, and/or about the objects with which the radiation interacts.

We may loosely classify images according to the way in which the interaction occurs, understanding that the division is sometimes unclear, and that images may be of multiple types. Figure 3 depicts these various image types.

Reflection images sense radiation that has been reflected from the surfaces of objects. The radiation itself may be ambient or artificial, and it may be from a localized source, or from multiple or extended sources. Most of our daily experience of optical imaging through the eye is of reflection images. Common nonvisible examples include radar images, sonar images, laser images, and some types of electron microscope images. The type of information that can be extracted from reflection images is primarily about object surfaces, viz., their shapes, texture, color, reflectivity, and so on.

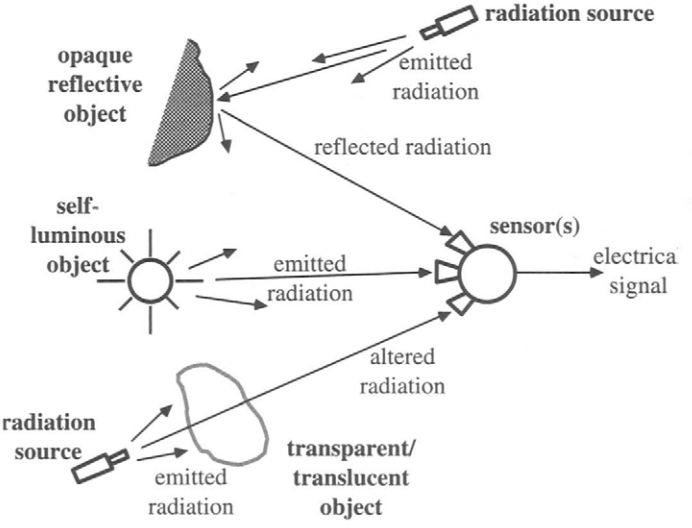

FIGURE 3 Recording the various types of interaction of radiation with matter.

Emission images are even simpler, since in this case the objects being imaged are self-luminous. Examples include thermal or infrared images, which are commonly encountered in medical, astronomical, and military applications, self-luminous visible-light objects, such as light bulbs and stars, and MRI images, which sense particle emissions. In images of this type, the information to be had is often primarily internal to the object; the image may reveal how the object creates radiation, and thence something of the internal structure of the object being imaged. However, it may also be external; for example, a thermal camera can be used in low-light situations to produce useful images of a scene containing warm objects, such as people.

Finally, *absorption images* yield information about the internal structure of objects. In this case, the radiation passes through objects, and is partially absorbed or attenuated by the material composing them. The degree of absorption dictates the level of the sensed radiation in the recorded image. Examples include x-ray images, transmission microscopic images, and certain types of sonic images.

Of course, the above classification into types is informal, and a given image may contain objects which interacted with radiation in different ways. More important is to realize that images come from many different radiation sources and objects, and that the purpose of imaging is usually to extract information about either the source and/or the objects, by sensing the reflected/transmitted radiation, and examining the way in which it has interacted with the objects, which can reveal physical information about both source and objects.

Figure 4 depicts some representative examples of each of the above categories of images. Figures 4(a) and 4(b) depict reflection images arising in the visible-light band and in the microwave band, respectively. The former is quite recognizable; the latter is a synthetic aperture radar image of DFW airport. Figures 4(c) and 4(d) are emission images, and depict, respectively, a forward-looking infrared (FLIR) image, and a visible-light image of the globular star cluster Omega Centauri. Perhaps the reader can probably guess the type of object that is of interest in 4(c). The object in 4(d), which consists of over a million stars, is visible with the unaided eye at lower northern latitudes. Lastly, Figs. 4(e) and 4(f), which are absorption images, are of a digital (radiographic) mammogram and a conventional light micrograph, respectively.

Scale of Images

Examining the figures in Fig. 4 reveals another image diversity: *scale.* In our daily experience we ordinarily encounter and visualize objects that are within 3 or 4 orders of magnitude of 1 meter. However, devices for image magnification and amplification have made it possible to extend the realm of "vision" into the cosmos, where it has become possible to image extended structures extending over as much as 10^{30} meters, and into the microcosmos, where it has become possible to acquire images of objects as small as 10^{-10} meters. Hence we are able to image from the grandest scale to the minutest scales, over a range of 40 orders of magnitude, and as we will find, the techniques of image and video processing are generally applicable to images taken at any of these scales.

Scale has another important interpretation, in the sense that any given image can contain objects that exist at scales different from other objects in the same image, or that even exist at multiple scales simultaneously. In fact, this is the rule rather than the exception. For example, in Fig. 4(a), at a small scale of observation, the image contains the bas-relief patterns cast onto the coins. At a slightly larger scale, strong circular structures arose. However, at a yet larger scale, the coins can be seen to be organized into a highly coherent spiral pattern. Similarly, examination of Fig. 4(d) at a small scale reveals small bright objects corresponding to stars; at a larger scale, it is found that the stars are nonuniformly distributed over the image, with a tight cluster having a density that sharply increases towards the center of the image. This concept of multi-scale is a powerful one, and is the basis for many of the algorithms that will be described in the chapters of this *Handbook.*

Dimension of Images

An important feature of digital images and video is that they are *multidimensional signals,* meaning that they are functions of more than a single variable. In the classic study of *digital signal processing,* the signals are usually one-dimensional functions of time. Images, however, are functions of two, and perhaps three space dimensions, whereas digital video as a function includes a third (or fourth) time dimension as well. The dimension of a signal is the number of coordinates that are required to index a given point in the image, as depicted in Fig. 5. A consequence of this is that digital image processing, and especially digital video processing, is quite data-intensive, meaning that significant computational and storage resources are often required.

Digitization of Images

The environment around us exists, at any reasonable scale of observation, in a space/time continuum. Likewise, the signals and images that are abundantly available in the environment (before being sensed) are naturally *analog.* By analog, we mean two things: that the signal exists on a continuous (space/time) domain, and that also takes values that comes from a continuum of possibilities. However, this *Handbook* is about processing *digital* image and video signals, which means that

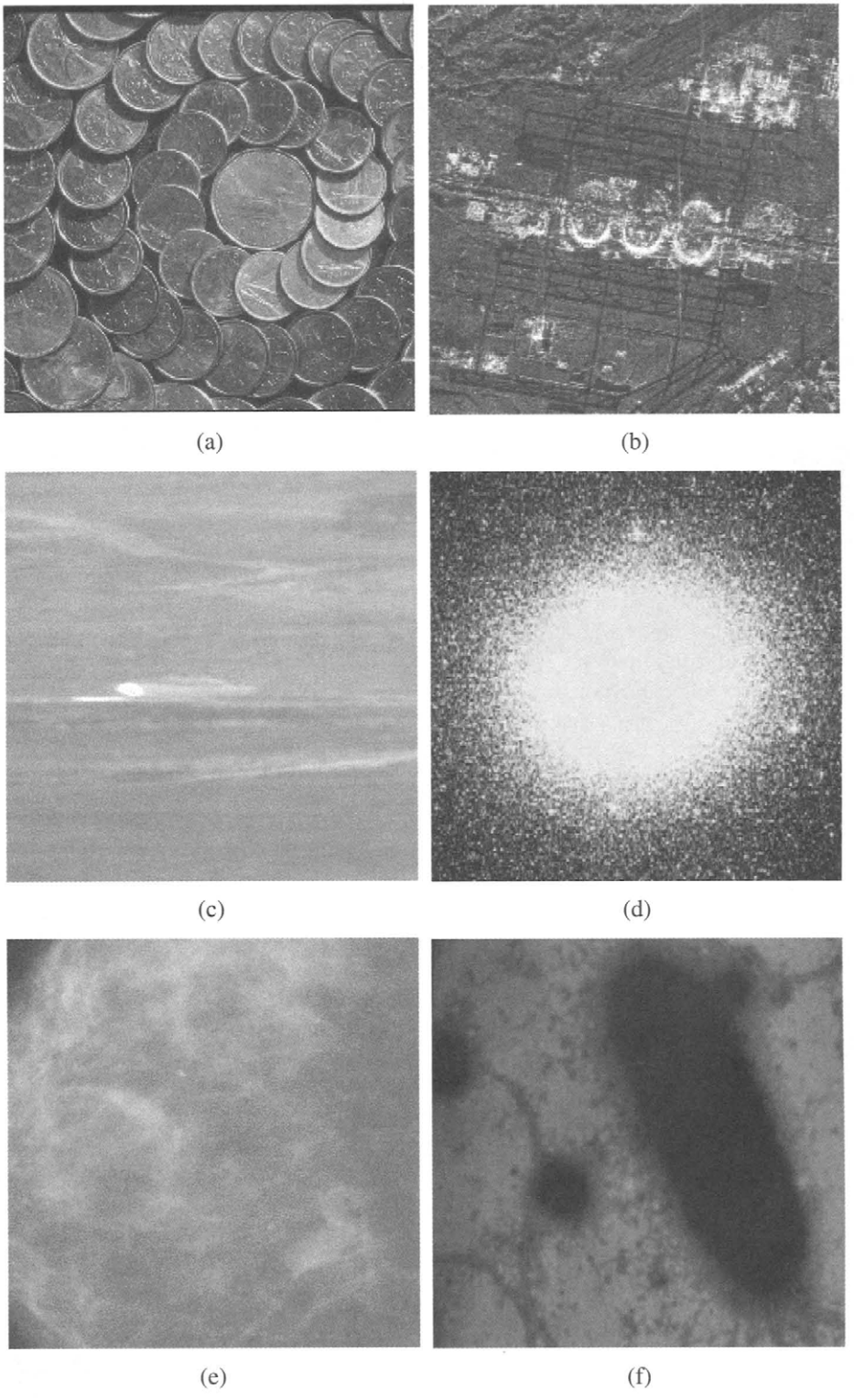

FIGURE 4 Examples of reflection (a), (b); emission (c), (d); and absorption (e), (f) image types.

once the image/video signal is sensed, it must be converted into a computer-readable, digital format. By digital, we also mean two things: that the signal is defined on a discrete (space/time) domain, and that it takes values from a discrete set of possibilities. Before digital processing can commence, a process of *analog-to-digital conversion* (A/D conversion) must occur. A/D conversion consists of two distinct subprocesses: *sampling* and *quantization.*

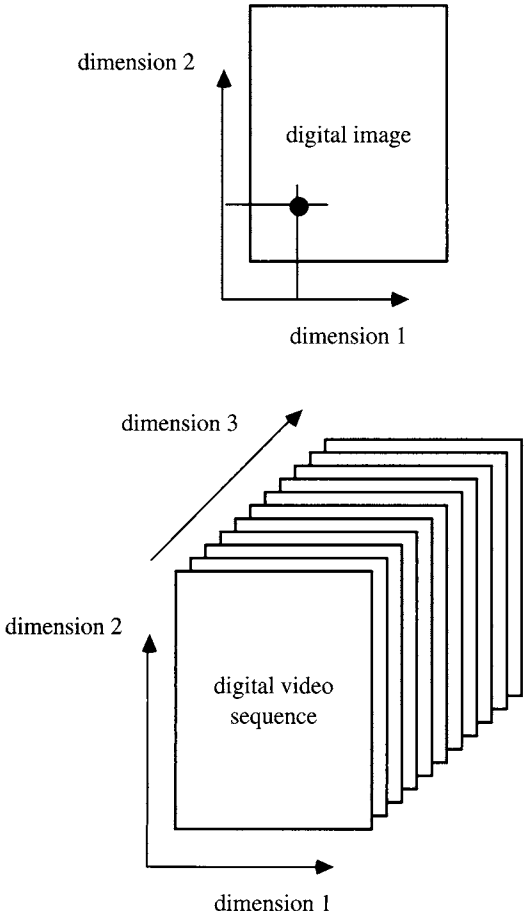

FIGURE 5 The dimensionality of images and video.

Sampled Images

Sampling is the process of converting a continuous-space (or continuous-space/time) signal into a discrete-space (or discrete-space/time) signal. The sampling of continuous signals is a rich topic that is effectively approached using the tools of linear systems theory. The mathematics of sampling, along with practical implementations are addressed elsewhere in this *Handbook*. In this introductory chapter, however, it is worth giving the reader a feel for the process of sampling and the need to sample a signal sufficiently densely. For a continuous signal of given space/time dimensions, there are mathematic reasons why there is a lower bound on the space/time sampling frequency (which determines the minimum possible number of samples) required to retain the information in the signal. However, image processing is a visual discipline, and it is more fundamental to realize that what is usually important is that the process of sampling does not lose *visual information*. Simply stated, the sampled image/video signal must "look good," meaning that it does not suffer too much from a loss of visual resolution, or from artifacts that can arise from the process of sampling.

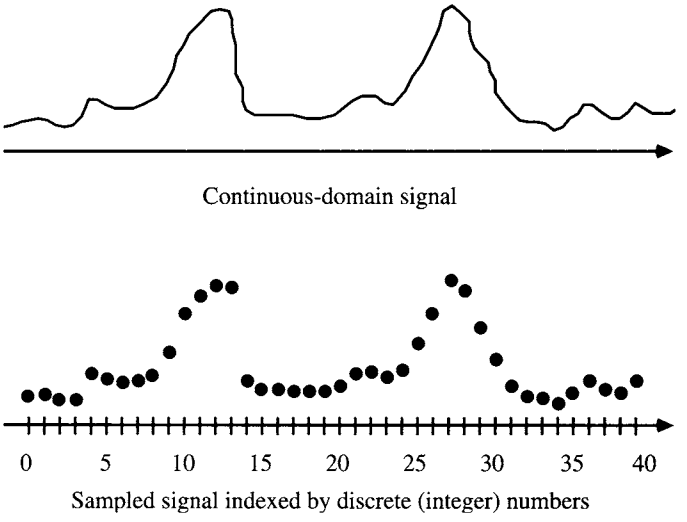

FIGURE 6 Sampling a continuous-domain one-dimensional signal.

Figure 6 illustrates the result of sampling a one-dimensional continuous-domain signal. It is easy to see that the samples collectively describe the gross shape of the original signal very nicely, but that smaller variations and structures are harder to discern or may be lost. Mathematically, information may have been lost, meaning that it might not be possible to reconstruct the original continuous signal from the samples (as determined by the Sampling Theorem, see Chapters 2.3 and 7.1). Supposing that the signal is part of an image, e.g., is a single scan-line of an image displayed on a monitor, then the visual quality may or may not be reduced in the sampled version. Of course, the concept of visual quality varies from person-to-person, and it also depends on the conditions under which the image is viewed, such as the viewing distance.

Note that in Fig. 6, the samples are indexed by integer numbers. In fact, the sampled signal can be viewed as a vector of numbers. If the signal is finite in extent, then the signal vector can be stored and digitally processed as an array, hence the integer indexing becomes quite natural and useful. Likewise, image and video signals that are space/time sampled are generally indexed by integers along each sampled dimension, allowing them to be easily processed as multi-dimensional arrays of numbers. As shown in Fig. 7, a sampled

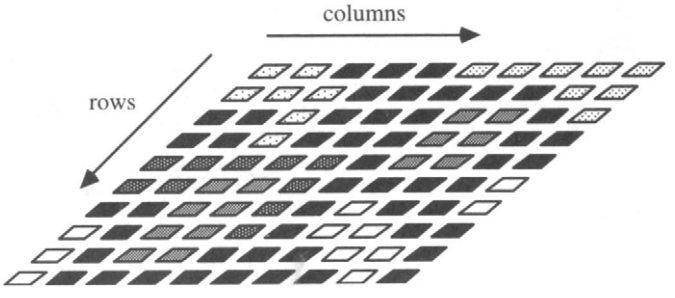

FIGURE 7 Depiction of a very small (10 × 10) piece of an image array.

image is an array of sampled image values that are usually arranged in a row–column format. Each of the indexed array elements is often called a *picture element,* or *pixel* for short. The term *pel* has also been used, but has faded in usage probably since it is less descriptive and not as catchy. The number of rows and columns in a sampled image is also often selected to be a power of 2, since it simplifies computer addressing of the samples, and also since certain algorithms, such as discrete Fourier transforms, are particularly efficient when operating on signals that have dimensions that are powers of 2. Images are nearly always rectangular (hence indexed on a Cartesian grid), and are often square, although the horizontal dimensional is often longer, especially in video signals, where an aspect ratio of 4:3 is common.

As mentioned above, the effects of insufficient sampling ("undersampling") can be visually obvious. Figure 8 shows two very illustrative examples of image sampling. The two images, which we'll call "mandrill" and "fingerprint," both contain a significant amount of interesting visual detail that substantially defines the content of the images. Each image is shown at three different sampling densities: 256×256 (or $2^8 \times 2^8 = 65{,}536$ samples), 128×128 (or $2^7 \times 2^7 = 16{,}384$ samples), and 64×64 (or $2^6 \times 2^6 = 4{,}096$ samples). Of course, in both cases, all three scales of images are digital, and so there is potential loss of information relative to the original analog image. However, the perceptual quality of the images can easily be seen to degrade rather rapidly; note the whiskers on the mandrill's face, which lose all coherency in the 64×64 image. The 64×64 fingerprint is very interesting, since the pattern has completely changed! It almost appears as a different fingerprint. This results from an undersampling effect know as *aliasing,* where image frequencies appear that have no physical meaning (in this case, creating a false pattern). Aliasing, and its mathematic interpretation, will be discussed further in Chapter 2.3 in the context of the Sampling Theorem.

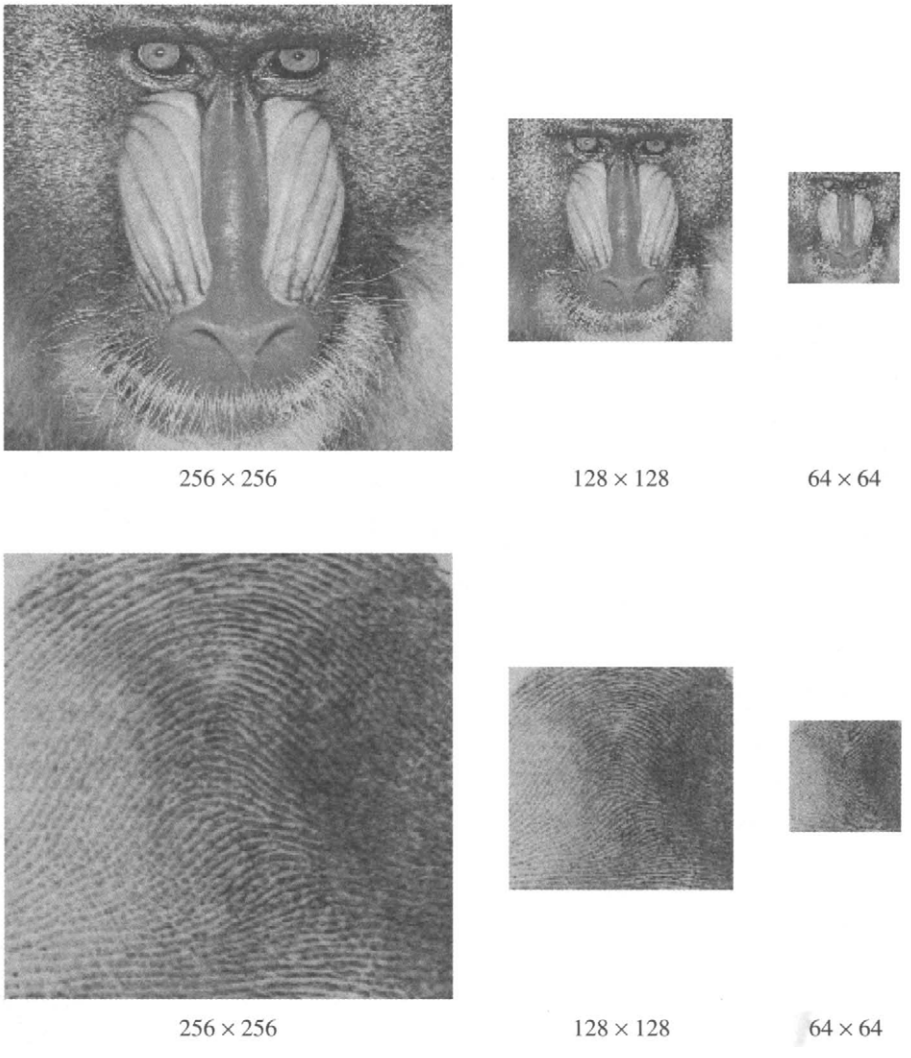

FIGURE 8 Examples of the visual effect of different image sampling densities.

Quantized Images

The other part of image digitization is *quantization.* The values that a (single-valued) image takes are usually *intensities,* since they are a record of the intensity of the signal incident on the sensor, e.g., the photon count or the amplitude of a measured wave function. Intensity is a positive quantity. If the image is represented visually using shades of gray (like a black-and-white photograph), then the pixel values are referred to as *gray levels.* Of course, broadly speaking, an image may be multi-valued at each pixel (such as a color image), or an image may have negative pixel values, in which case, it is not an intensity function. In any case, the image values must be quantized for digital processing.

Quantization is the process of converting a *continuous-valued image,* that has a continuous range (set of values that it can take), into a *discrete-valued image,* that has a discrete range. This is ordinarily done by a process of rounding, truncation, or some other irreversible, nonlinear process of information destruction. Quantization is a necessary precursor to digital processing, since the image intensities must be represented with a finite precision (limited by wordlength) in any digital processor.

When the gray level of an image pixel is quantized, it is assigned to be one of a finite set of numbers which is the *gray-level range.* Once the discrete set of values defining the gray-level range is known or decided, then a simple and efficient method of quantization is simply to round the image pixel values to the respective nearest members of the intensity range. These rounded values can be any numbers, but for conceptual convenience and ease of digital formatting, they are then usually mapped by a linear transformation into a finite set of non-negative integers $\{0, \ldots, K-1\}$, where K is a power of two: $K = 2^B$. Hence the number of allowable gray levels is K, and the number of bits allocated to each pixel's gray level is B. Usually $1 \leq B \leq 8$ with $B = 1$ (for binary images) and $B = 8$ (where each gray level conveniently occupies a byte) being the most common bit depths (see Fig. 9). Multi-valued images, such as color images, require quantization of the components either individually or collectively ("vector quantization"); for example, a three-component color image is frequently represented with 24 bits per pixel of color precision.

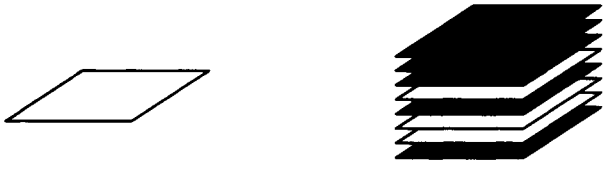

FIGURE 9 Illustration of 8-bit representation of a quantized pixel.

Unlike sampling, quantization is a difficult topic to analyze since it is nonlinear. Moreover, most theoretical treatments of signal processing assume that the signals under study are *not* quantized, since it tends to greatly complicate the analysis. On the other hand, quantization is an essential ingredient of any (lossy) signal compression algorithm, where the goal can be thought of as finding an optimal quantization strategy that simultaneously minimizes the volume of data contained in the signal, while disturbing the fidelity of the signal as little as possible. With simple quantization, such as gray-level rounding, the main concern is that the pixel intensities or gray levels must be quantized with sufficient precision that excessive information is not lost. Unlike sampling, there is no simple mathematic measurement of information loss from quantization. However, while the effects of quantization are difficult to express mathematically, the effects are visually obvious.

Each of the images depicted in Figs. 4 and 8 is represented with 8 bits of gray-level resolution — meaning that bits less significant than the 8th bit have been rounded or truncated. This number of bits is quite common for two reasons: first, using more bits will generally *not* improve the visual appearance of the image — the adapted human eye usually is unable to see improvements beyond 6 bits (although the total range that can be seen under different conditions can exceed 10 bits) — hence using more bits would be wasteful. Secondly, each pixel is then conveniently represented by a byte. There are exceptions: in certain scientific or medical applications, 12, 16, or even more bits may be retained for more exhaustive examination by human or by machine.

Figures 10 and 11 depict two images at various levels of gray-level resolution. Reduced resolution (from 8 bits) was obtained by simply truncating the appropriate number of less-significant bits from each pixel's gray level. Figure 10 depicts the 256×256 digital image "fingerprint" represented at 4, 2, and 1 bits of gray-level resolution. At 4 bits, the fingerprint is nearly indistinguishable from the 8-bit representation of Fig. 8. At 2 bits, the image has lost a significant amount of information, making the print difficult to read. At one bit, the *binary* image that results is likewise hard to read. In practice, binarization of fingerprints is often used to make the print more distinctive. Using simple truncation–quantization, most of the print is lost since it was inked insufficiently on the left, and to excess on the right. Generally, bit truncation is a poor method for creating a binary image from a gray-level image. See Chapter 2.2 for better methods of image binarization.

Figure 11 shows another example of gray-level quantization. The image "eggs" is quantized at 8, 4, 2, and 1 bit of gray level resolution. At 8 bits, the image is very agreeable. At four bits, the eggs take on the appearance of being striped or painted like Easter eggs. This effect is known as "false contouring," and results when inadequate gray-scale resolution is used to represent smoothly-varying regions of an image. In such places, the effects of a (quantized) gray level can be visually

FIGURE 10 Quantization of the 256 × 256 image "fingerprint." Left to right: 4, 2, and 1 bits per pixel.

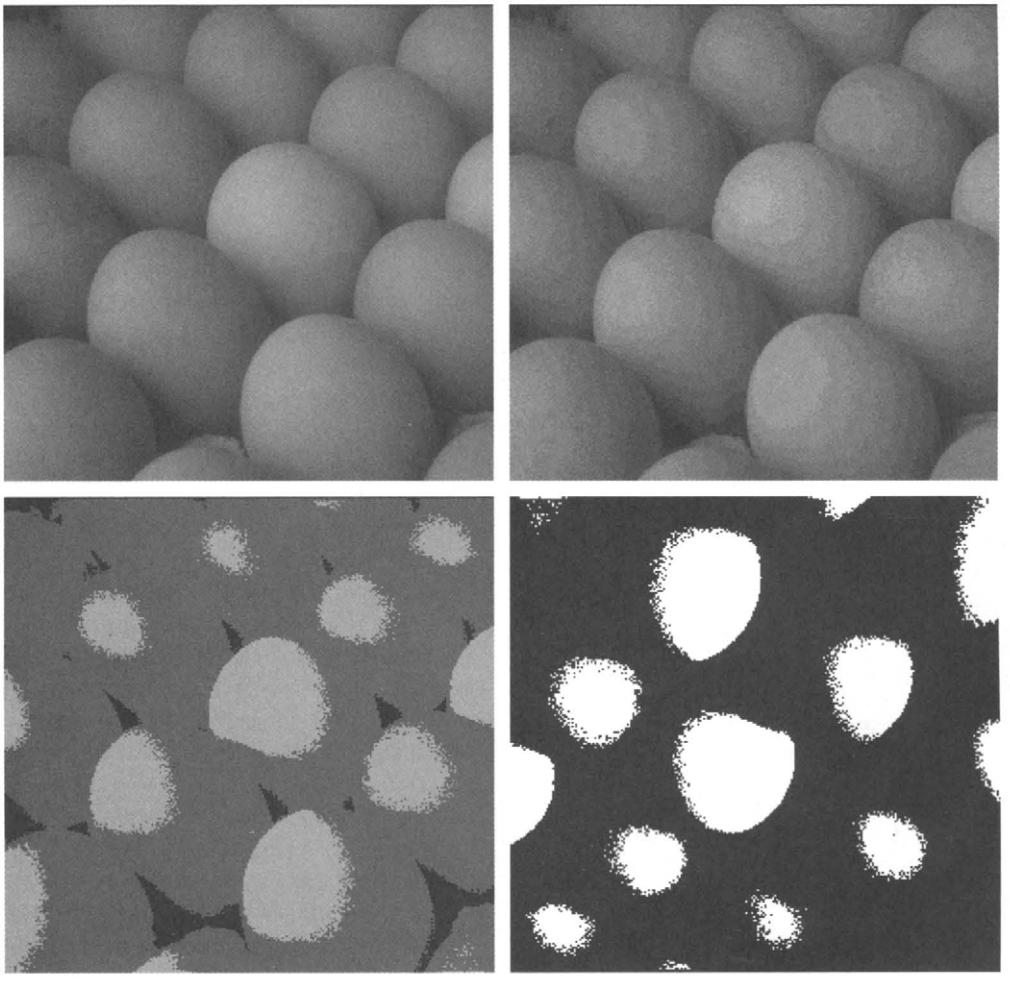

FIGURE 11 Quantization of the 256 × 256 image "eggs." Clockwise from upper left: 8, 4, 2, and 1 bits per pixel.

exaggerated, leading to an appearance of false structures. At 2 bits and 1 bit, significant information has been lost from the image, making it difficult to recognize.

A quantized image can be thought of as a stacked set of single-bit images (known as "bit planes") corresponding to the gray-level resolution depths. The most significant bits of every pixel comprise the top bit plane, and so on. Figure 12 depicts a 10×10 digital image as a stack of B bit planes. Special-purpose image processing algorithms are occasionally applied to the individual bit planes.

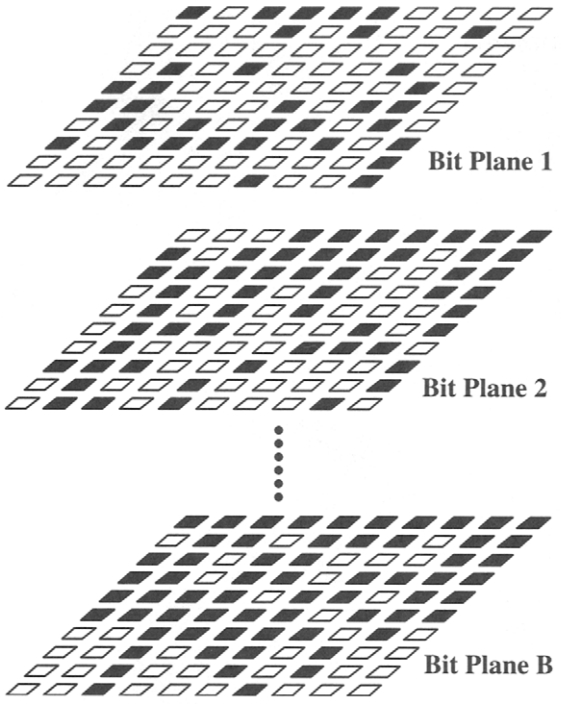

FIGURE 12 Depiction of a small (10×10) digital image as a stack of bit planes ranging from most significant (top) to least significant (bottom).

Color Images

Of course, the visual experience of the normal human eye is not limited to gray scales — *color* is an extremely important aspect of images. It is also an important aspect of digital images. In a very general sense, color conveys a variety of rich information that describes the *quality* of objects, and as such, it has much to do with visual *impression.* For example, it is known that different colors have the potential to evoke different emotional responses. The perception of color is allowed by the color-sensitive neurons known as *cones* that are located in the retina of the eye. The cones are responsive to normal light levels and are distributed with greatest density near the center of the retina, known as *fovea* (along the direct line of sight). The *rods* are neurons that are sensitive at low-light levels, and are not capable of distinguishing color wavelengths. They are distributed with greatest density around the periphery of the fovea, with very low density near the line-of-sight. Indeed, this may be observed by observing a dim point target (such as a star) under dark conditions. If the gaze is shifted slightly off-center, then the dim object suddenly becomes easier to see.

In the normal human eye, colors are sensed as near-linear combinations of long, medium, and short wavelengths, which roughly correspond to the three *primary colors* that are used in standard video camera systems: red (R), green (G), and blue (B). The way in which visible light wavelengths map to RGB camera color coordinates is a complicated topic, although standard tables have been devised based on extensive experiments. A number of other color coordinate systems are also used in image processing, printing, and display systems, such as the YIQ (luminance, in-phase chromatic, quadratic chromatic) color coordinate system. Loosely speaking, the YIQ coordinate system attempts to separate the perceived image *brightness* (luminance) from the chromatic components of the image via an invertible linear transformation:

$$\begin{bmatrix} Y \\ I \\ Q \end{bmatrix} = \begin{bmatrix} 0.299 & 0.587 & 0.114 \\ 0.596 & -0.275 & -0.321 \\ 0.212 & -0.523 & 0.311 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix}. \quad (1)$$

The RGB system is used by color cameras and video display systems, while the YIQ is the standard color representation used in broadcast television. Both representations are used in practical image and video processing systems along with several other representations.

Most of the theory and algorithms for digital image and video processing has been developed for single-valued, monochromatic (gray level), or intensity-only images, whereas color images are vector-valued signals. Indeed, many of the approaches described in this *Handbook* are developed for single-valued images. However, these techniques are often applied (suboptimally) to color image data by regarding each color component as a separate image to be processed, and recombining the results afterwards. As seen in Fig. 13, the R, G, and B components contain a considerable amount of overlapping information. Each of them is a valid image in the same sense is as the image seen through colored spectacles, and can be processed as such. Conversely, however, if the color components are collectively available, then vector image processing algorithms can often be designed that achieve optimal results by taking this information into account. For example, a vector-based image enhancement algorithm applied to the "cherries" image in Fig. 13 might adapt by giving less importance to enhancing the Blue component, since the image signal is weaker in that band.

Chromanance is usually associated with slower amplitude variations than is luminance, since it usually is associated with fewer image details or rapid changes in value. The human eye has a greater spatial bandwidth allocated for luminance perception than for chromatic perception. This is exploited by compression algorithms that use alternate color representations such as YIQ, and store, transmit, or process the chromatic components using a lower bandwidth (fewer bits) than the luminance component. Image and video compression algorithms achieve increased efficiencies through this strategy.

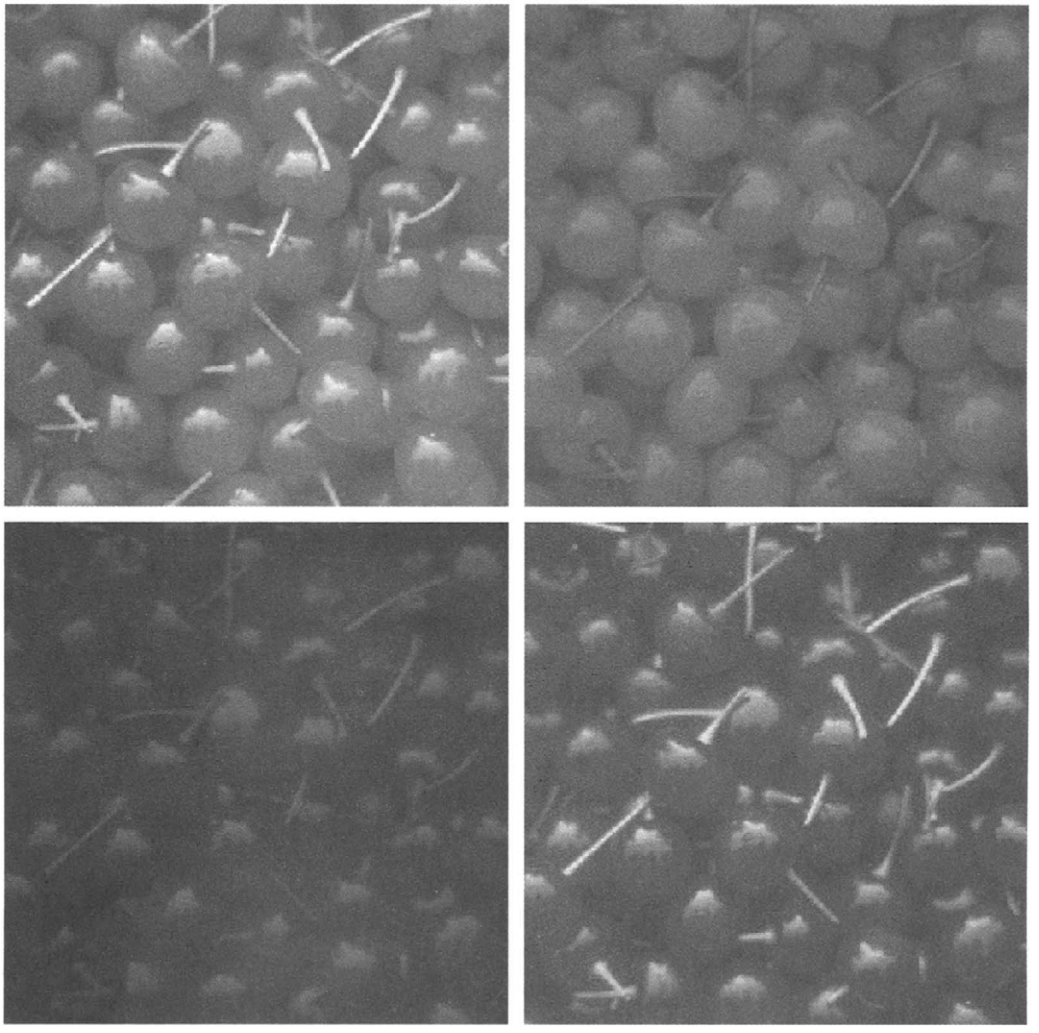

FIGURE 13 Color image "cherries" (top left) and (clockwise) its Red, Green, and Blue components. (See color insert.)

Size of Image Data

The amount of data in visual signals is usually quite large, and increases geometrically with the dimensionality of the data. This impacts nearly every aspect of image and video processing; data volume is a major issue in the processing, storage, transmission, and display of image and video information. The storage required for a single monochromatic digital still image that has (row × column) dimensions $N \times M$ and B bits of gray-level resolution is NMB bits. For the purpose of discussion we will assume that the image is square ($N = M$), although images of any aspect ratio are common. Most commonly, $B = 8$ (1 byte/pixel) unless the image is binary or is special-purpose. If the image is vector-valued, e.g., color, then the data volume is multiplied by the vector dimension. Digital images that are delivered by commercially-available image digitizers are typically of approximate size 512×512 pixels, which is large enough to fill much of a monitor screen. Images both larger (ranging up to $4{,}096 \times 4{,}096$ or more) and smaller (as small as 16×16) are commonly encountered. Table 1 depicts the required storage for a variety of image resolution parameters, assuming that there has been no compression of the data. Of course, the spatial extent (area) of the image exerts the greatest effect on the data volume. A single $512 \times 512 \times 8$ color image requires nearly a megabyte of digital storage space, which only a few years ago, was a lot. More recently, even large images are suitable for viewing and manipulation on home personal computers (PCs), although somewhat inconvenient for transmission over existing telephone networks.

However, when the additional time dimension is introduced, the picture changes completely. Digital video is extremely storage-intensive. Standard video systems display visual information at a rate of 30 images/s for reasons related to human visual latency (at slower rates, there is a perceivable "flicker"). A $512 \times 512 \times 24$ color video sequence thus occupies 23.6 megabytes for *each* second of viewing. A two-hour digital film at the same resolution levels would thus

TABLE 1 Data-volume requirements for digital still images of various sizes, bit depths, and vector dimension

Spatial Dimensions	Pixel Resolution (bits)	Image Type	Data Volume (bytes)
128 × 128	1	Monochromatic	2,048
256 × 256	1	Monochromatic	8,192
512 × 512	1	Monochromatic	32,768
1,024 × 1,024	1	Monochromatic	131,072
128 × 128	8	Monochromatic	16,384
256 × 256	8	Monochromatic	65,536
512 × 512	8	Monochromatic	262,144
1,024 × 1,024	8	Monochromatic	1,048,576
128 × 128	3	Trichromatic	6,144
256 × 256	3	Trichromatic	24,576
512 × 512	3	Trichromatic	98,304
1,024 × 1,024	3	Trichromatic	393,216
128 × 128	24	Trichromatic	49,152
256 × 256	24	Trichromatic	196,608
512 × 512	24	Trichromatic	786,432
1,024 × 1,024	24	Trichromatic	3,145,728

require about *85 gigabytes* of storage at nowhere near theatre quality. That's a lot of data, even for today's computer systems. Fortunately, images and video generally contain a significant degree of redundancy along each dimension. Taking this into account along with measurements of human visual response, it is possible to significantly compress digital images and video streams to acceptable levels. Sections 5 and 6 of this *Handbook* contain a number of chapters devoted to these topics. Moreover, the pace of information delivery is expected to significantly increase in the future, as significant additional bandwidth becomes available in the form of gigabit and terabit Ethernet networks, digital subscriber lines that use existing telephone networks, and public cable systems. These developments in telecommunications technology, along with improved algorithms for digital image and video transmission, promise a future that will be rich in visual information content in nearly every medium.

Digital Video

A significant portion of this *Handbook* is devoted to the topic of *digital video processing.* In recent years, hardware technologies and standards activities have matured to the point that it is becoming feasible to transmit, store, process, and view video signals that are stored in digital formats, and to share video signals between different platforms and application areas. This is a natural evolution, since temporal change, which is usually associated with motion of some type, is often the most important property of a visual signal.

Beyond this, there is a wealth of applications which stand to benefit from digital video technologies, and it is no exaggeration to say that the blossoming digital video industry represents many billions of dollars in research investments. The payoff from this research will be new advances in digital video processing theory, algorithms, and hardware that are expected to result in many billions more in revenues and profits. It is safe to say that digital video is very much the current frontier and the future of image processing research and development. The existing and expected applications of digital video are either growing rapidly or are expected to explode once the requisite technologies become available.

Some of the notable emerging digital video applications are:

- Video teleconferencing
- Video telephony
- Digital TV, including High-Definition Television (HDTV)
- Internet video
- Medical video
- Dynamic scientific visualization
- Multimedia video
- Video instruction
- Digital cinema

Sampled Video

Of course, digital processing of video requires that the video stream be in a digital format, meaning that it must be sampled and quantized. Video quantization is essentially the same as image quantization. However, video sampling involves taking samples along a new and different (time) dimension. As such, it involves some different concepts and techniques.

First and foremost, the time dimension has a direction associated with it, unlike the space dimensions, which are ordinarily regarded as directionless until a coordinate system is artificially imposed upon it. Time proceeds from the past towards the future, with an origin that exists only in the current moment. Video is often processed in "real time," which (loosely) means that the result of processing appears effectively "instantaneously" (usually in a perceptual sense) once the input becomes available. Such a processing system cannot depend more than a few future video samples. Moreover, it must process the video data quickly enough that the result appears instantaneous. Because of the vast data volume involved, the design of fast algorithms and hardware devices is a major priority.

In principle, an analog video signal $I(x, y, t)$, where (x, y) denote continuous space coordinates and t denotes continuous time, is continuous in both the space and time dimensions, since the radiation flux that is incident on a video sensor is continuous at normal scales of observation. However, the analog video that is viewed on display monitors is *not* truly analog, since it is sampled along one space dimension and

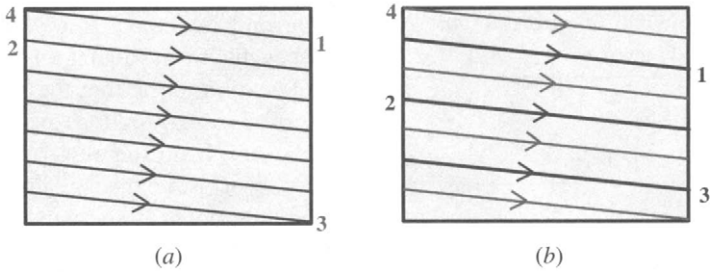

FIGURE 14 Video scanning. (*a*) Progressive video scanning. At the end of a scan (1), the electron gun spot snaps back to (2). A blank signal is sent in the interim. After reaching the end of a frame (3), the spot snaps back to (4). A synchronization pulse then signals the start of another frame. (*b*) Interlaced video scanning. Red and blue fields are alternately scanned left-to-right and top-to-bottom. At the end of scan (1), the spot snaps to (2). At the end of the blue field (3), the spot snaps to (4) (new field).

along the time dimension. Practical so-called analog video systems, such as television and monitors, represent video as a one-dimensional electrical signal $V(t)$. Prior to display, a one-dimensional signal is obtained by sampling $I(x, y, t)$ along the vertical (y) space direction and along the time (t) direction. This is called scanning and the result is a series of time samples, which are complete pictures or *frames*, each of which is composed as space samples, or *scan lines*.

Two types of video scanning are commonly used: *progressive scanning* and *interlaced scanning*. A progressive scan traces a complete frame, line-by-line from top-to-bottom, at a scan rate of Δt s/frame. High-resolution computer monitors are a good example, with a scan rate of $\Delta t = 1/72$ s. Figure 14(*a*) depicts progressive scanning on a standard monitor.

A description of interlaced scanning requires that some other definitions be made. For both types of scanning, the *refresh rate* is the frame rate at which information is displayed on a monitor. It is important that the frame rate be high enough, since otherwise the displayed video will appear to "flicker." The human eye detects flicker if the refresh rate is less than about 50 frames/s. Clearly, computer monitors (72 frames/s) exceed this rate by almost 50%. However, in many other systems, notably television, such fast refresh rates are not possible unless spatial resolution is severely compromised because of bandwidth limitations. Interlaced scanning is a solution to this. In P:1 interlacing, every Pth line is refreshed at each frame refresh. The sub-frames in interlaced video are called *fields*, hence P fields constitute a frame. The most common is 2:1 interlacing which is used in standard television systems, as depicted in Fig. 14(*b*). In 2:1 interlacing, the two fields are usually referred to as the top and bottom fields. In this way, flicker is effectively eliminated provided that the field refresh rate is above the visual limit of about 50 Hertz (Hz). Broadcast television in the U.S. uses a frame rate of 30 Hz, hence the field rate is 60 Hz, which is well above the limit. The reader may wonder if there is a loss of visual information, since the video is being effectively sub-sampled by a factor of two in the vertical space dimension in order to increase the apparent frame rate. In fact there is, since image motion may change the picture between fields. However, the effect is ameliorated to a significant degree by standard monitors and TV screens, which have screen phosphors with a *persistence* (glow time) that just matches the frame rate, hence each field persists until the matching field is sent.

Digital video is obtained either by sampling an analog video signal $V(t)$, or by directly sampling the three-dimensional space–time intensity distribution that is incident on a sensor. In either case, what results is a time sequence of two-dimensional spatial intensity arrays, or equivalently, a three-dimensional space–time array. If progressive analog video is sampled, then the sampling is rectangular and properly indexed in an obvious manner, as illustrated in Fig. 15. If interlaced analog video is sampled, then the digital video is interlaced also as shown in Fig. 16. Of course, if an interlaced

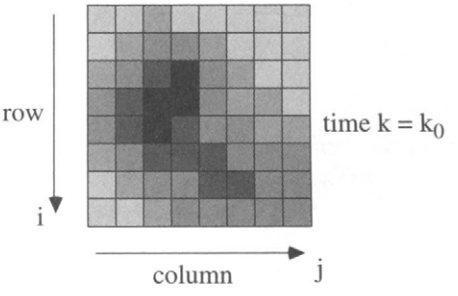

FIGURE 15 A single frame from a sampled progressive video sequence.

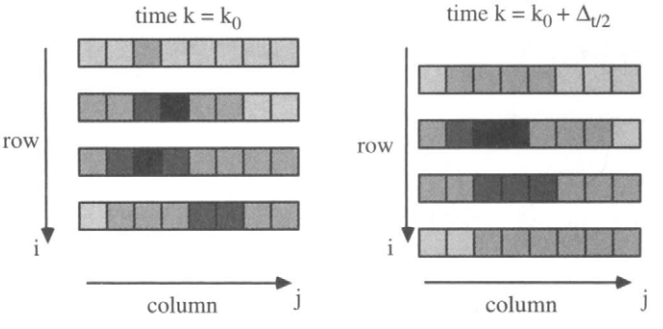

FIGURE 16 A single frame (two fields) from a sampled 2:1 interlaced video sequence.

video stream is sent to a system that processes or displays non-interlaced video, then the video data must first be converted or *de-interlaced* to obtain a standard progressive video stream before the accepting system will be able to handle it.

Video Transmission

The data volume of digital video is usually described in terms of bandwidth or bitrate. As described in Chapter 6.1, the bandwidth of digital video streams (without compression) that match the current visual resolution of current television systems exceeds 100 megabits/s (mbps). Modern television formats such as HDTV can multiply this by a factor of four or more. By contrast, the networks that are currently available to handle digital data are limited. Conventional telephone lines (POTS) delivers only 56 kilobits/s 9kbps), although digital subscriber lines (DSL) multiply this by a factor of 30 or more. ISDN (Integrated Services Digital Network) lines allow for data bandwidths equal to $64p$ kbps, where $1 \leq p \leq 30$, which falls short of the necessary datarate to handle full digital video. Dedicated T1 lines (1.5 mbps) also handle only a small fraction of the necessary bandwidth. Ethernet and cable systems, which deliver data in the gigabit/s (gbps) range are capable of handling raw digital video, but have problems delivering multiple video streams over the same network. The problem is similar to that of delivering large amounts of water to through small pipelines. Either the data rate (water pressure) must be increased, or the data volume must be reduced.

Fortunately, unlike water, digital video can be compressed very effectively because of the redundancy inherent in the data, and because of an increased understanding of what components in the video stream are actually visible. Because of many years of research into image and video compression, it is now possible to transmit digital video data over a broad spectrum of networks, and we may expect that digital video will arrive in a majority of homes in the near future. Based on research developments along these lines, a number of world standards have recently emerged, or are under discussion, for video compression, video syntax, and video formatting. The use of standards allows for a common protocol for video and ensures that the consumer will be able to accept the same video inputs using products from different manufacturers. The current and emerging video standards broadly extend standards for still images that have been in use for a number of years. Several chapters are devoted to describing these standards, while others deal with emerging techniques which may effect future standards. It is certain, in any case, that we have entered a new era where digital visual data will play an important role in education, entertainment, personal communications, broadcast, the internet, and many other aspects of daily life.

Objectives of this *Handbook*

The goals of this *Handbook* are ambitious, since it is intended to reach a broad audience that is interested in a wide variety of image and video processing applications. Moreover, it is intended to be accessible to readers that have a diverse background, and that represent a wide spectrum of levels of preparation and engineering/computer education. However, a *Handbook* format is ideally suited for this multi-user purpose, since it allows for a presentation that adapts to the readers needs. In the early part of the *Handbook* we present very basic material that is easily accessible even for novices to the image processing field. These chapters are also useful for review, for basic reference, and as support for later chapters. In every major section of the *Handbook*, basic introductory material is presented as well as more advanced chapters that take the reader deeper into the subject.

Unlike textbooks on image processing, the *Handbook* is therefore not geared towards a specified level of presentation, nor does it uniformly assume a specific educational background. There is material that is available for the beginning image processing user, as well as for the expert. The *Handbook* is also unlike a textbook in that it is not limited to a specific point of view given by a single author. Instead, leaders from image and video processing education, industry, and research have been called upon to explain the topical material from their own daily experience. By calling upon most of the leading experts in the field, we have been able to provide a complete coverage of the image and video processing area without sacrificing any level of understanding of any particular area.

Because of its broad spectrum of coverage, we expect that the *Handbook of Image and Video Processing* will serve as an excellent textbook as well as reference. It has been our objective to keep the students needs in mind, and we feel that the material contained herein is appropriate to be used for classroom presentations ranging from the introductory undergraduate level, to the upper-division undergraduate, to the graduate level. Although the *Handbook* does not include "problems in the back," this is not a drawback since the many examples provided in every chapter are sufficient to give the student a deep understanding of the function of the various image and video processing algorithms. This field is very much a visual science, and the principles underlying it are best taught via visual examples. Of course, we also foresee the *Handbook* as providing easy reference, background, and guidance for image and video processing professionals working in industry and research.

Our specific objectives are to:

- Provide the practicing engineer and the student with a highly accessible resource for learning and using image/video processing algorithms and theory
- Provide the essential understanding of the various image and video processing standards that exist or

are emerging, and that are driving today's explosive industry
- Provide an understanding of what images are, how they are modeled, and give an introduction to how they are perceived
- Provide the necessary practical background to allow the engineer student to acquire and process his/her own digital image or video data
- Provide a diverse set of example applications, as separate complete Chapters, that are explained in sufficient depth to serve as extensible models to the readers own potential applications

The *Handbook* succeeds in achieving these goals, primarily because of the many years of broad educational and practical experience that the many contributing authors bring to bear in explaining the topics contained herein.

Organization of the *Handbook*

Since this *Handbook* is emphatically about *processing* images and video, the next section is immediately devoted to basic algorithms for image processing, instead of surveying methods and devices for image acquisition at the outset, as many textbooks do. Section II is divided into three Chapters, which respectively, introduce the reader to the most fundamental two-dimensional image processing techniques. Chapter 2.1 lays out basic methods for gray-level image processing, which includes point operations, the image histogram, and simple image algebra. The methods described there stand alone as algorithms that can be applied to most images, but also they set the stage and the notation for the more involved methods discussed in later Chapters. Chapter 2.2 describes basic methods for image binarization and for binary image processing, with emphasis on morphological binary image processing. The algorithms described there are among the most widely used in applications, especially in the biomedical area. Chapter 2.3 explains the basics of the Fourier transform and frequency-domain analysis, including discretization of the Fourier transform and discrete convolution. Special emphasis is laid on explaining frequency-domain concepts through visual examples. Fourier image analysis provides a unique opportunity for visualizing the meaning of frequencies as components of signals. This approach reveals insights which are difficult to capture in one-dimensional, graphical discussions.

Section III of the *Handbook* deals with methods for correcting distortions or uncertainties in images and for improving image information by combining images taken from multiple views. Quite frequently the visual data that is acquired has been in some way corrupted. Acknowledging this and developing algorithms for dealing with it is especially critical since the human capacity for detecting errors, degradations, and delays in digitally-delivered visual data is quite high. Image and video signals are derived from imperfect sensors, and the processes of digitally converting and transmitting these signals are subject to errors. There are many types of errors that can occur in image/video data, including, for example, blur from motion or defocus; noise that is added as part of a sensing or transmission process; bit, pixel, or frame loss as the data is copied or read; or artifacts that are introduced by an image or video compression algorithm. As such, it is important to be able to model these errors, so that numerical algorithms can be developed to ameliorate them in such a way as to improve the data for visual consumption. Section III contains three broad categories of topics. The first is Image/Video Enhancement, where the goal is to remove noise from an image while retaining the perceptual fidelity of the visual information; these are seen to be conflicting goals. Chapters are included that include very basic linear methods; highly efficient nonlinear methods; and recently developed and very powerful wavelet methods; and also extensions to video enhancement. The second broad category is Image/Video Restoration, where it is assumed that the visual information has been degraded by a distortion function, since as defocus, motion blur, or atmospheric distortion, and more than likely, by noise as well. The goal is to remove the distortion and attenuate the noise, while again preserving the perceptual fidelity of the information contained within. And again, it is found that a balanced attack on conflicting requirements is required in solving these difficult, ill-posed problems. The treatment again begins with a basic, introductory chapter; ensuing chapters build on this basis and discuss methods for restoring multi-channel images (such as color images); multi-frame images (meaning, using information from multiple images taken of the same scene); iterative methods for restoration; and extensions to video restoration. Related topics that are considered are motion detection and estimation, which is essential for handling many problems in video processing, and a general framework for regularizing ill-posed restoration problems. Finally, the third category involves the extraction of enriched information about the environment by combining images taken from multiple views of the same scene. This includes chapters on methods for computed stereopsis and for image stabilization and mosaicking.

Section IV of the *Handbook* deals with methods for image and video analysis. Not all images or videos are intended for direct human visual consumption. Instead, in many situations it is of interest to automate the process of repetitively interpreting the content of multiple images or video data through the use of an *image* or *video analysis algorithm.* For example, it may be desired to *classify* parts of images or videos as being of some type, or it may be desired to *detect* or *recognize* objects contained in the data sets. If one is able to develop a reliable computer algorithm that consistently achieves success in the desired task, and if one has access to a computer that is fast enough, then a tremendous savings in man-hours can be attained. The advantage of such a system

increases with the number of times that the task must be done and with the speed with which it can be automatically accomplished. Of course, problems of this type are typically quite difficult, and in many situations it is not possible to approach, or even come close to, the efficiency of the human visual system. However, if the application is specific enough, and if the process of image acquisition can be sufficiently controlled (to limit the variability of the image data), then tremendous efficiencies can be achieved. With some exceptions, image/video analysis systems are quite complex, but are often composed at least in part of subalgorithms that are common to other image/video analysis applications. Section IV of this *Handbook* outlines some of the basic models and algorithms that are encountered in practical systems. The first set of chapters deals with image models and representations that are commonly used in every aspect of image/video processing. This starts with a chapter on models of the human visual system. Much progress has been made in recent years in modeling the brain and the functions of the optics and the neurons along the visual pathway (although much remains to be learned as well). Since images and videos that are processed are nearly always intended for eventual visual consumption by humans, then in the design of these algorithms, it is imperative the receiver be taken into account, as with any communication system. After all, vision is very much a form of dense communication, and images are the medium of information. The human eye–brain system is the receiver. This is followed by chapters on wavelet image representations, random field image models, image modulation models, image noise models and image color models that are referred to in many other places in the *Handbook*. These chapters may be thought of as a core reference section of the *Handbook* that supports the entire presentation. Methods for image/video classification and segmentation are described next; these basic tools are used in a wide diversity of analysis applications. Complementary to these are two chapters on edge and boundary detection, where the goal is to find the boundaries of regions, *viz.*, sudden changes in image intensities, rather than finding (segmenting out) and classifying regions directly. The approach taken depends on the application. Finally, a chapter is given that reviews currently available software for image and video processing.

As described earlier in this introductory chapter, image and video information is highly data-intensive. Sections V and VI of the *Handbook* deal with methods for compressing this data. Section V deals with still image compression, beginning with several basic chapters of lossless compression, and on several useful general approaches for image compression. In some realms, these approaches compete, but each has its advantages and subsequent appropriate applications. The existing JPEG standards for both lossy and lossless compression are described next. Although these standards are quite complex, they are described in sufficient detail to allow for the practical design of systems that accept and transmit JPEG data sets.

Section VI extends these ideas to video compression, beginning with an introductory chapter that discusses the basic ideas and that uses the H.261 standard as an example. The H.261 standard, which is used for video teleconferencing systems, is the starting point for later video compression standards, such as MPEG. The following two chapters are on especially promising methods for future and emerging video compression systems: wavelet-based methods, where the video data is decomposed into multiple subimages (scales or subbands), and object-based methods, where objects in the video stream are identified and coded separately across frames, even (or especially) in the presence of motion. Finally, chapters on the existing MPEG-1 and MPEG-2 and emerging MPEG-4 and MPEG-7 standards for video compression are given, again in sufficient detail to enable the practicing engineer to put the concepts to use.

Section VII deals with image and video scanning, sampling and interpolation. These important topics give the basics for understanding image acquisition, converting images and video into digital format, and for resizing or spatially manipulating images. Section VIII deals with the visualization of image and video information. One chapter focuses on the halftoning and display of images, and another on methods for assessing the quality of images, especially compressed images.

With the recent significant activity in *multimedia*, of which image and video is the most significant component, methods for databasing, access/retrieval, archiving, indexing, networking, and securing image and video information are of high interest. These topics are dealt with in detail in Section IX of the *Handbook*.

Finally, Section X includes eight chapters on a diverse set of image processing applications that are quite representative of the universe of applications that exist. Many of the chapters in this section have *analysis, classification* or *recognition* as a main goal, but reaching these goals inevitably requires the use of a broad spectrum of image/video processing subalgorithms for enhancement, restoration, detection, motion, and so on. The work that is reported in these chapters is likely to have significant impact on science, industry, and even on daily life. It is hoped that the reader is able to translate the lessons learned in these chapters, and in the preceding material, into their own research or product development work in image and/or video processing. For the student, it is hoped that s/he now possesses the required reference material that will allow her/him to acquire the basic knowledge to be able to begin a research or development career in this fast-moving and rapidly growing field.

Acknowledgment

Many thanks to Professor Joel Trussell for carefully reading and commenting on this introductory chapter.

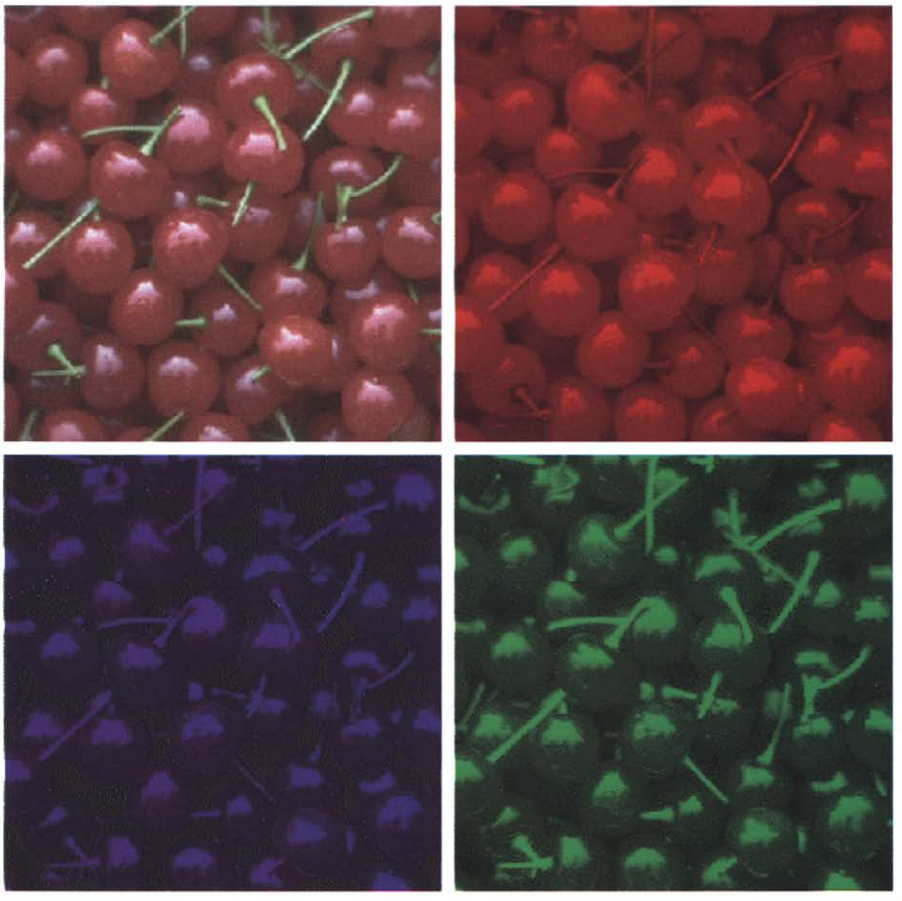

FIGURE 1.1.13 Color image "cherries" (top left) and (clockwise) its red, green, and blue components.

FIGURE 3.2.8 Impulse noise cleaning with a 5×5 CWM smoother: (top left) original "portrait" image, (top right) image with "salt and pepper" noise, (bottom left) CWM smoother with $W_c = 16$, (bottom right) CWM smoother with $W_c = 5$.