1 Introduction

Digital imaging is now so commonplace that we tend to forget how complicated and exacting the process of recording and displaying a digital image is. Of course, the process is not very complicated for the average consumer, who takes pictures with a digital camera or video recorder, then views them on a computer monitor or television. It is very convenient now to obtain prints of the digital pictures at local stores or make your own with a desktop printer. Digital imaging technology can be compared to automotive technology. Most drivers do not understand the details of designing and manufacturing an automobile. They do appreciate the qualities of a good design. They understand the compromises that must be made among cost, reliability, performance, efficiency and aesthetics. This book is written for the designers of imaging systems to help them understand concepts that are needed to design and implement imaging systems that are tailored for the varying requirements of diverse technical and consumer worlds. Let us begin with a bird's eye view of the digital imaging process.

1.1 Digital imaging: overview

A digital image can be generated in many ways. The most common methods use a digital camera, video recorder or image scanner. However, digital images are also generated by image processing algorithms, by analysis of data that yields two-dimensional discrete functions and by computer graphics and animation. In most cases, the images are to be viewed and analyzed by human beings. For these applications, it is important to capture or create the image data appropriately and display the image so that it is most pleasing or best interpreted. Exceptions to human viewing are found in computer vision and automated pattern recognition applications. Even in these cases, the relevant information must be captured accurately by the imaging system. Many detection and recognition tasks are modeled on analogies to the human visual system, so recording images as the human viewer sees the scene can be important.

The most common operations in digital imaging may be illustrated by examining the capture and display of an image with a common digital camera. The camera focuses an optical image onto a sensor. The basics of the optical system of the camera are the same regardless of whether the sensor is film or a solid-state sensing array. The optics, which include lenses and apertures, must be matched to the sensitivity and resolution of the sensor. The technology has reached a point where the best digital imaging chips are comparable to high-quality consumer film. In the cases of both film and digital sensors, the characteristics of the sensors must be taken into account. In a color system, the responses of each of the color bands should be known and any interaction between them should be determined. Once the image data are recorded, the processing of film and digital images diverges. It is common to scan film, in which case the processing proceeds as in the case of a digital image.

An advantage of the digital system is that operations are performed by digital processors that have more latitude and versatility than the analog chemical-optical systems for film. Computational speed and capability have increased to the point where all the necessary processing can be done within the normal time between successive shots of the camera. At low resolution, digital cameras are capable of recording short video sequences.

The recorded data are processed to compensate for nonlinearities of the sensor and to remove any bias and nonuniform gain across the sensor array. Any defects in the sensor array can be corrected at this point by processing. Compensation for some optical defects, such as flare caused by bright light sources, is also possible. The image is then prepared for storage or output. Storage may include encoding the data to reduce the memory required. Information about the camera settings may be appended to the image data to aid the accurate interpretation and reproduction of the image. When the image is prepared for viewing on an output device, this information is combined with information about the characteristics of the output device to produce the optimal image for the user's purposes.

This text will present the material that will allow the reader to understand, analyze and evaluate each of these steps. The goal is to give the reader the analytical tools and guidance necessary to design, improve and create new digital imaging systems.

1.2 Digital imaging: short history

Electronic imaging has a longer history than most readers in this digital age would imagine. As early as 1851, the British inventor Frederick Bakewell demonstrated a device that could transmit line drawings over telegraph wires at the World's Fair in London. This device, basically the first facsimile machine, used special insulating inks at the transmitter and special paper at the receiver. It used a scanning mechanism much like a drum scanner. The drawing was wrapped around a cylinder and a stylus, attached to a lead-screw, controlled the current that was sent to a receiving unit with a synchronized scanning cylinder where the current darkened the special electro-sensitive paper.

As photography developed, methods of transforming tonal images to electronic form were considered. Just after the turn of the century, two early versions of facsimile devices were developed that used scanning but different methods for sensing tonal images. Arthur Korn, in Germany, used a selenium cell to scan a photograph directly. Edouard Belin, in France, created a relief etching from a photograph, which was scanned with a stylus. The variable resistance produced a variable current that transmitted the image. Belin's method was used to transmit the first trans-Atlantic image in 1921. Korn's methods did the same in 1923. The images could be reproduced at the receiver by modulating a light source on photographic paper or by modulating the current with electro-sensitive paper.

The first digital image was produced by the Bartlane method in 1920. This was named for the British co-inventors, Harry G. Bartholomew and Maynard D. McFarlane. This method used a series of negatives on zinc plates that were exposed for varying lengths of time, which produced varying densities. The first system used five plates, corresponding to five quantization levels. The plates were scanned simultaneously on a cylinder. A hole was punched in a paper tape to indicate that the corresponding plate was clear. The method was later increased to 15 levels. On playback, the holes could be used to modulate a light beam with the same number of intensity levels.

The first electronic television was demonstrated by Philo T. Farnsworth in 1927. This had an electronic scanning tube as well as a CRT that could be controlled to display an image. Of interest is the fact that in 1908, A. A. C. Swinton proposed, in a paper published in *Nature*, an electronic tube for recording images and sending them to a receiver. Commercial television did not appear until after World War II.

Electronic image scanners were also used in the printing industry to make color separations in the 1930s, but these were analog devices, using the electronic signals to expose film simultaneously with the scan. Thus, there was no electronic storage of the image data. The first digital image, in the sense that we know it, was produced in 1957 by Russell Kirsch at the National Bureau of Standards. His device was basically a drum scanner with a photomultiplier tube that produced digital data that could be stored in a computer.

The first designs for digital cameras were based on these scanning ideas; thus, they took a significantly long time to take a picture and were not suitable for consumer purposes. The military was very instrumental in the development of the technology and supported research that led to the first digital spy satellite, the KH-11 in 1976. Previous satellites recorded the images on film and ejected a canister that was caught in mid-air by an airplane. The limited bandwidth of the system was a great motivator for image coding and compression research in the 1970s.

The development of the charge-coupled device (CCD) in an array format made the digital camera possible. The first imaging systems to use these devices were astronomical telescopes, as early as 1973. The first black and white digital cameras were used in the 1980s but were confined to experimental and research uses. The technology made the consumer video recorder possible in the 1980s, but the low resolution of these arrays restricted their use in consumer cameras. Color could be produced by using three filters and three arrays in an arrangement similar to the common television camera, which used an electronic tube. Finally, color was added directly to the CCD array in the form of a mosaic of filters laid on top of the CCD elements. Each element recorded one band of a three-band image. A full resolution color image was obtained by spatial interpolation of the three signals. This method remains the basis of color still cameras today.

1.3 Applications

As mentioned, there are many ways to generate digital images. The emphasis of this text is on accurate input and output. Let us consider the applications that require this attention to precision. The digital still camera is the most obvious application. Since the object is to reproduce a recorded scene, accuracy on both the input and output are required. In addition to consumer photography, there are many applications where accuracy in digital imaging is important. In the medical world, color images are used to record and diagnose diseases and conditions in areas that include dermatology, ophthalmology, surgery and endoscopy. Commercial printing has a fundamental requirement for accurate color reproduction. Poor color in catalogs can lead to customer dissatisfaction and costly returns. The accuracy of commercial printing has historically been more art than science, but with innovations in technology, this application area will move to a more analytical plane. Electrophotography and copying of documents is another important application area. This combines the same basic elements of digital cameras, except the input is usually a scanner and the output device is totally under the control of the manufacturer. More exotic applications include imaging systems used on satellites and space probes. We will see that multispectral systems that record many more than the usual three color bands can be analyzed using the methods presented in this text.

There are many applications where the reproduction of the image is not the end product. In most computer vision applications, a machine interprets the recorded image. To obtain the best performance from the algorithms, the input image data should be as accurate as possible. Since many algorithms are based on human visual properties for discrimination of objects, attention to accurate input is important. Satellite imagery can be interpreted by human beings or automated, and serves as another example. The bands recorded by satellites are usually not compatible with reproduction of true color. Digital astronomy must record spatial data accurately for proper interpretation. Infrared imaging, which is common in both ground-based and satellite systems, can be accurately recorded for analysis purposes, but cannot be displayed accurately for humans, since it is beyond the range of our sensitivities.

It should be noted that several imaging modalities are not covered by this text. X-ray images from medical or industrial applications are beyond the scope of this text. The transformation from X-ray energy distributions to quantitative data is not sufficiently well modeled to determine its accuracy. X-ray computed tomography (CT) and magnetic resonance imaging (MRI) are important medical modalities, but the relationships of the physical quantities that produce the images are highly complex and still the subject of research.

1.4 Methodology

Since our goal is to present the basic methods for accurate image capture and display, it is necessary to use a mathematical approach. We have to define what we mean by accuracy and quantify errors in a meaningful way. There will be many cases where the user must make decisions about the underlying assumptions that make a mathematical algorithm optimal. We will indicate these choices and note that if the system fails to satisfy the assumed conditions, then the results may be quite useful but suboptimal.

The error measures that will be chosen are often used for mathematical convenience. For example, mean square error is often used for this reason. The use of such measures is appropriate since the methods based on them produce useful, if suboptimal, results. The analysis that is used for these methods is also important since it builds a foundation for extensions and improvements that are more accurate in the visual sense.

Errors can be measured in more than one way and in more than one space. It is not just the difference in values of the input pixel and the output pixel that is of interest. We are often interested in the difference in color values of pixels. The color values may be represented in a variety of color spaces. The transformations between the measured values and the various color spaces are important. In addition, it is not just the difference in color values of pixels that is important, but the effect of the surrounding area and the response of the human visual system. To study these effects, we need the mathematical concepts of spatial convolution and transformation to the frequency domain. Just as the eye may be more sensitive to some color ranges than others, it is more sensitive to some spatial frequency ranges than others.

The algorithms used to process images often require the setting of various parameters. The proper values for these parameters are determined by the characteristics of the images. The characteristics that are important are almost always mathematical or statistical. This text will explore the relationships between these quantitative characteristics and the qualitative visual characteristics of the images. The background that is needed to make these connections will be reviewed in the text and the appendices. We will use many examples to help the reader gain insight into these relationships.

1.5 Prerequisite knowledge

Since we are taking a mathematical approach, it is appropriate to review the mathematical and statistical concepts that are required to get the most from this text. The details of the required mathematics are reviewed in Chapter 2. The reader's background should be equivalent to an undergraduate degree in engineering or computer science. A course in linear systems is assumed. We will review the concepts of system functions, transformations and convolution from this topic. The major extension is from one dimension to two dimensions. Included in linear systems courses is an introduction to the frequency domain. The reader should have a working knowledge of the Fourier transform in both its continuous and discrete forms. Of course, knowledge of Fourier transforms requires the basic manipulation of complex numbers, along with the use of Euler's identity that relates the complex exponential to the trigonometric functions

$$e^{j\theta} = \cos(\theta) + j\sin(\theta), \tag{1.1}$$

where we will use the engineering symbol j for the imaginary number $\sqrt{-1}$. The frequency domain is important for both computational efficiency and for interpretation.

Thus, the material in Chapter 5 is the basis for much of the analysis and many methods introduced later.

The use of vectors and matrices to represent images and operations allows the use of the very powerful tools of linear algebra. The level of knowledge that is required is that covered in most undergraduate engineering programs. The reader is expected to be familiar with basic matrix-vector operations, such as addition and multiplication. The concepts of diagonalization, eigenvectors and eigenvalues should be familiar. A review of these concepts and their use in representing operations with digital images is given in Chapter 2 and Appendix B.

As mentioned previously, the optimal processing of an image depends on its characterization. The characterization is most often done statistically. The mean and variance, which is related to the signal power when computing signal-to-noise ratios, are the most common statistics. The reader should be familiar with second-order statistics, such as covariance, autocovariance and cross-covariance. These concepts are based on elementary probability, which includes knowledge of random variables, probability density functions, expected values and expected values of functions of random variables. These concepts are reviewed, along with basic probability concepts, in Appendix C. This background should be part of any undergraduate engineering degree.

1.6 Overview of the book

Because of our emphasis on the analytical methods for obtaining accurate images, we begin with a review of the basic mathematics that will be used in the rest of the book. The review in Chapter 2 uses some concepts for which the reader may need further review. For this reason, we have included additional review material on generalized functions (Dirac delta functions), matrix algebra, and probability and stochastic signals in Appendices A, B and C, respectively.

The major emphasis of the text is on accuracy in imagery. Thus, before we start using examples of images to demonstrate various concepts, we need to discuss the fundamentals of image display. Chapter 3 discusses the important points that will be used throughout the text. The rules of this chapter are used for monochrome images, which will be used to demonstrate most of the basic concepts of image capture and reproduction. The discussion of accurate reproduction of monochrome and color must await a presentation of basic photometry and colorimetry in Chapter 8.

A digital image is defined by a finite number of values for a finite number of pixels. We first consider the quantization process and its effects in Chapter 4. We use some of our statistical and probability background to derive optimal quantizers and measure the goodness of other methods. The effects of spatial sampling would naturally follow. However, the analysis of spatial sampling requires a good understanding of the two-dimensional frequency domain. We review one-dimensional Fourier transforms and extend that knowledge to two dimensions in Chapter 5. Having obtained the necessary background, we present spatial sampling in Chapter 6.

In the first six chapters, we have covered most of the basic properties of monochrome images. In Chapter 7, we put these properties in context and use them to describe image characteristics. The frequency domain is used to describe the bandwidth of an image. That concept can be extended to describe an image in terms of its relation to various subspaces. The statistical concepts can be used to characterize images by their stochastic properties, such as the signal-to-noise ratio. The statistical approach also lets us characterize an image by a stochastic model that represents the class of images to which our particular image belongs.

To capture and reproduce color images accurately, it is necessary to understand the definition and measurement of monochrome and colored objects. The imaging scientist needs to understand the relationship between the quantized pixel values and the physical quantities that they represent. While this text cannot present a complete background, Chapter 8 covers the fundamentals that are needed for practical applications. This foundation can be enhanced by further study of more complete and specialized texts. In Chapter 9, we present a topic that is missing from many color science texts. The relationship of color sampling to the concepts of sampling in the spatial domain is important in the design of digital color imaging systems and in the simulation of any color imaging system by digital computers.

Chapter 10 describes image input devices. The characterization of the images is necessary to design a device that will capture that data accurately. Likewise, in Chapter 11, we describe the various devices and methods used for image reproduction. In Chapter 12, we discuss the various methods used to characterize the input and output devices. These characterizations, together with the characterization of the images, are used to complete the cycle of accurate image capture and reproduction.

We have mentioned that characterization of images is often determined by various statistics or model parameters. These formed the basis for the image characterization of Chapter 7. In that chapter, examples are used to illustrate the effects of the parameters. However, when determining the optimal processing for a particular image, it is necessary to estimate the appropriate value of the parameters for that image. In Chapter 13, we discuss methods to estimate the important parameters that are needed for processing algorithms.

Finally, in all imaging systems, the actual components have limited accuracy. Optical systems can never produce an unblurred image. Digital systems are always subject to quantization noise, but in practical systems this is rarely the limiting noise component. A final step in producing accurate images may be the restoration of degradations caused by imperfections in a less than ideal system. The degradations of interest include noise, blurring, nonideal scanning filters, illuminant and color distortions. The basics of restoration are presented in Chapter 14. Restoration differs from enhancement in that restoration seeks to restore accurately what was degraded, whereas enhancement seeks to improve an image subjectively. While enhancement is a valid step in processing images, it requires a much different background and assumptions than does restoration. This text does not provide the background for the subjective improvement of images, and thus, we will stay within the bounds of what is quantitatively optimal.

With this motivation for the order of the topics in the text, let us begin our journey.

2 Mathematical representation

For any type of structured analysis on imaging systems, we must first have some model for the system. Most often it is a mathematical model that is most useful. Even if the model is simplified to the point that it ignores many real world physical effects, it is still very useful to give first order approximations to the behavior of the system.

In this chapter, we will review the tools needed to construct the mathematical models that are most often used in image analysis. It is assumed that the reader is already familiar with the basics of one-dimensional signals and systems. Thus, while a quick review is given for the various basic concepts, it is the relation between one and two dimensions and the two-dimensional operations that will be discussed in more depth here.

As mentioned, the mathematical models are often better or worse approximations to the real systems that are realized with optics, electronics and chemicals. As we review the concepts that are needed, we will point out where the most common problems are found and how close some of the assumptions are to reality.

2.1 Images as functions

Since images are the main topic of this book, it makes sense to discuss their representation first. Images are, by definition, defined in two spatial dimensions. Many, if not most, images represent the projection of some three-dimensional object or scene onto a two-dimensional plane. The geometrical relationship between the object and the image is left for other texts on image analysis. We will be concerned with other aspects of the physical object or scene on the recorded data. Images may be monochrome or color; they may be continuous or discrete; they may be still or moving; but always there are exactly two spatial dimensions. The value of the representation of the image may reflect the physical quantity that is measured by some imaging device, or it may reflect only the relative brightness imagined by a user.

Let us consider the common case where $f(x, y)$ is a function of two spatial variables whose value represents some physical quantity that can be measured by some instrument. Examples include:

- light intensity,
- optical or electromagnetic reflectivity,
- optical density,

- material density or attenuation,
- distance.

The cases of light intensity and reflectivity are the most common and will receive the most attention. The other cases can be modeled using the same mathematical tools. The details of the optical cases are covered in Chapter 8 on photometry and colorimetry.

For the optical cases, we are often interested in the color characteristics of the image. The model can easily be extended to include this aspect. The function $f(x, y, \lambda)$ includes the effect of wavelength or frequency of the measured radiation. Visible light has wavelengths, λ, from about 400 nm (blue) to 700 nm (red). The usual color images are represented by three bands that represent integrated power in roughly the red, green and blue regions of the spectrum. The exact representation will be discussed later. Hyperspectral images may use 100–400 bands in the visible to the near and middle infrared (IR) for satellite image applications. To reduce the problem back to the case of only two dimensions, the wavelength dimension can be eliminated by integrating the intensity over some range of wavelengths. This will produce a single-band image,

$$f_i(x, y) = \int_{\lambda_{\min}}^{\lambda_{\max}} f(x, y, \lambda) s_i(\lambda)\, \mathrm{d}\lambda, \tag{2.1}$$

where $s_i(\lambda)$ is the sensitivity of the ith sensor, which defines the ith band. We often denote the N-band image by the vector

$$\mathbf{f}(x, y) = [f_1(x, y), f_2(x, y), \ldots f_N(x, y)]^T. \tag{2.2}$$

It is often the case that each of the bands is treated as a monochrome image. We will make the case that this is inappropriate many times. The exact details of handling color bands are discussed in Chapter 8.

Light can have attributes other than intensity and wavelength. Light can be characterized as *coherent*, *partially coherent* and *noncoherent*. Coherent light is defined by a single wavelength at a single phase, $s(t) = \mathrm{e}^{\mathrm{j}(\omega t+\phi)}$. Because coherent light includes a phase parameter, it is usually represented by a complex function. Physically, coherent light is produced by a laser. Noncoherent light consists of a stochastic mixture of phases and is measured only by its power. Thus, it can be represented by a real number. Noncoherent light can be monochromatic, i.e., single wavelength, as can coherent light. The major difference is that the mixture of phases in noncoherent light prevents the narrow collimated beams possible with lasers. Almost all examples of imaging systems discussed in this course will be based on noncoherent light.

In addition, light can be polarized, that is, it has directional properties. This is most commonly observed when using polarized sunglasses. The effect has been used to display three-dimensional images in movies and on computer monitors. It is used in liquid crystal displays (see Section 11.2). Both coherent and noncoherent light can be polarized. Almost all examples of imaging systems are based on unpolarized light.

There are many applications of imaging that use radiation other than the visible band. For reference, the electromagnetic spectrum and its characteristics are given in Table 2.1.

Table 2.1. Regions of the electromagnetic spectrum

Name	Wavelength Range	Frequency	Character
Cosmic and gamma rays	10^{-7}–10^{-2} nm	300000–0.3 × 10^{20} Hz	Emissive
X-rays	10^{-2}–10^{-1} nm	300–30 × 10^{17} Hz	Heavy industrial
X-rays	10^{-1}–1 nm	30–3 × 10^{17} Hz	Medical, industrial
Ultraviolet	1–380 nm	3000–7.5 × 10^{14} Hz	Optical
Visible	400–700 nm	7.5–4.2 × 10^{14} Hz	Optical
Infrared (near)	720–1300 nm	4.2–2.3 × 10^{14} Hz	Optical
Infrared (middle)	1.3–3 μm	2.3–1 × 10^{14} Hz	Optical
Infrared (far)	7–15 μm	4.3–2 × 10^{13} Hz	Emissive
Super high (SHF)	1–10 mm	3–30 GHz	Satellite communications
Ultra high (UHF)	10–100 mm	3–0.3 GHz	UHF television, radar
Very high (VHF)	0.1–1 m	300–30 MHz	FM radio, VHF television
High (HF)	1–10 m	30–3 MHz	Amateur radio, telephone
Medium (MF)	10–100 m	3000–300 kHz	AM broadcast
Low (LF)	0.1–1 km	300–30 kHz	Marine communications
Very low (VLF)	1–10 km	30–3 kHz	Long range navigation

There are various problems for each of the various ranges that are peculiar to that band. For example, X-rays cannot be focused by ordinary means. The mathematics that is described in this text can be used for almost all bands, but should be modified according to the physical properties of each band.

The temporal aspect of the imaging can be taken into account by adding another argument to the function of spatial coordinates to represent time, $f(x, y, t)$. As with wavelength, the time variable can be discretized to produce a two-dimensional function, or series of functions,

$$f_i(x, y) = \int_{t_i}^{t_{i+1}} f(x, y, t) a_i(t)\, \mathrm{d}t, \tag{2.3}$$

where $a_i(t)$ is the aperture function during the ith time interval. Note that the image has to be integrated over some finite time interval in order to capture a finite amount of energy. The aperture function might represent the response of a CCD cell between read cycles of the array or the time integration function of an analog-to-digital converter. This will be discussed further in Chapter 6 on sampling. Examples of this actually include virtually all real still images. They all represent the image at a particular instant of time. Sequences of images are found in television, motion pictures and medical imaging.

The time and wavelength effects can be combined to produce

$$f(x, y) = \int_{t_1}^{t_2} \int_{\lambda_1}^{\lambda_2} f(x, y, \lambda, t) s(\lambda) a(t)\, \mathrm{d}\lambda\, \mathrm{d}t. \tag{2.4}$$