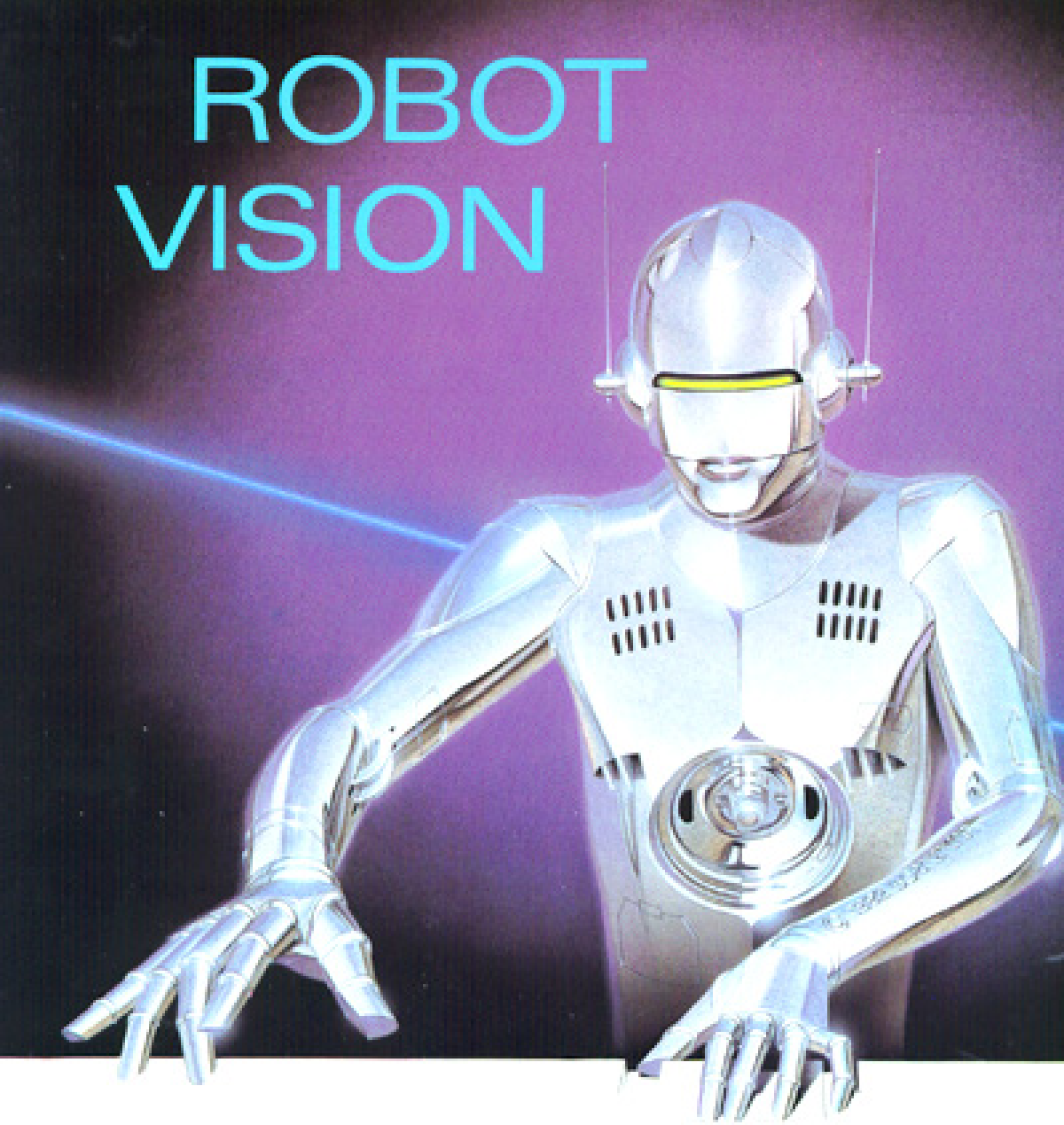
ROBOT VISION



Berthold Klaus Paul Horn

1

Introduction

In this chapter we discuss what a machine vision system is, and what tasks it is suited for. We also explore the relationship of machine vision to other fields that provide techniques for processing images or symbolic descriptions of images. Finally, we introduce the particular view of machine vision exploited in this text and outline the contents of subsequent chapters.

1.1 Machine Vision

Vision is our most powerful sense. It provides us with a remarkable amount of information about our surroundings and enables us to interact intelligently with the environment, all without direct physical contact. Through it we learn the positions and identities of objects and the relationships between them, and we are at a considerable disadvantage if we are deprived of this sense. It is no wonder that attempts have been made to give machines a sense of vision almost since the time that digital computers first became generally available.

Vision is also our most complicated sense. The knowledge we have accumulated about how biological vision systems operate is still fragmentary and confined mostly to the processing stages directly concerned with signals from the sensors. What we do know is that biological vision systems

are complex. It is not surprising, then, that many attempts to provide machines with a sense of vision have ended in failure. Significant progress has been made nevertheless, and today one can find vision systems that successfully deal with a variable environment as parts of machines.

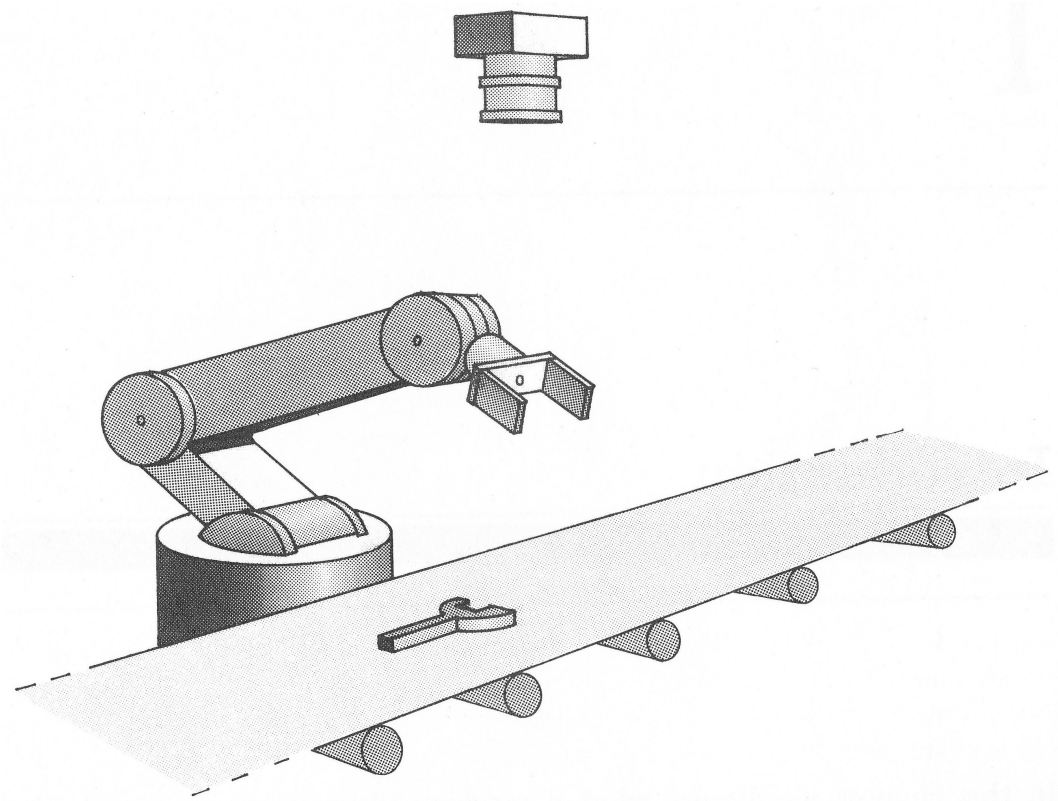


Figure 1-1. A machine vision system can make a robot manipulator much more versatile by allowing it to deal with variations in part position and orientation. In some cases simple binary image-processing systems are adequate for this purpose.

Most progress has been made in industrial applications, where the visual environment can be controlled and the task faced by the machine vision system is clear-cut. A typical example would be a vision system used to direct a robot arm to pick parts off a conveyor belt (figure 1-1).

Less progress has been made in those areas where computers have been called upon to extract ill-defined information from images that even people find hard to interpret. This applies particularly to images derived by other than the usual optical means in the visual spectrum. A typical example of such a task is the interpretation of X-rays of the human lung.

It is of the nature of research in a difficult area that some early ideas have to be abandoned and new concepts introduced as time passes. While

frustrating at times, it is part of the excitement of the search for solutions. Some believed, for example, that understanding the image-formation process was not required. Others became too enamored of specific computing methods of rather narrow utility. No doubt some of the ideas presented here will also be revised or abandoned in due course. The field is evolving too rapidly for it to be otherwise.

We cannot at this stage build a "universal" vision system. Instead, we address ourselves either to systems that perform a particular task in a controlled environment or to modules that could eventually become part of a general-purpose system. Naturally, we must also be sensitive to practical considerations of speed and cost. Because of the enormous volume of data and the nature of the computations required, it is often difficult to reach a satisfactory compromise between these factors.

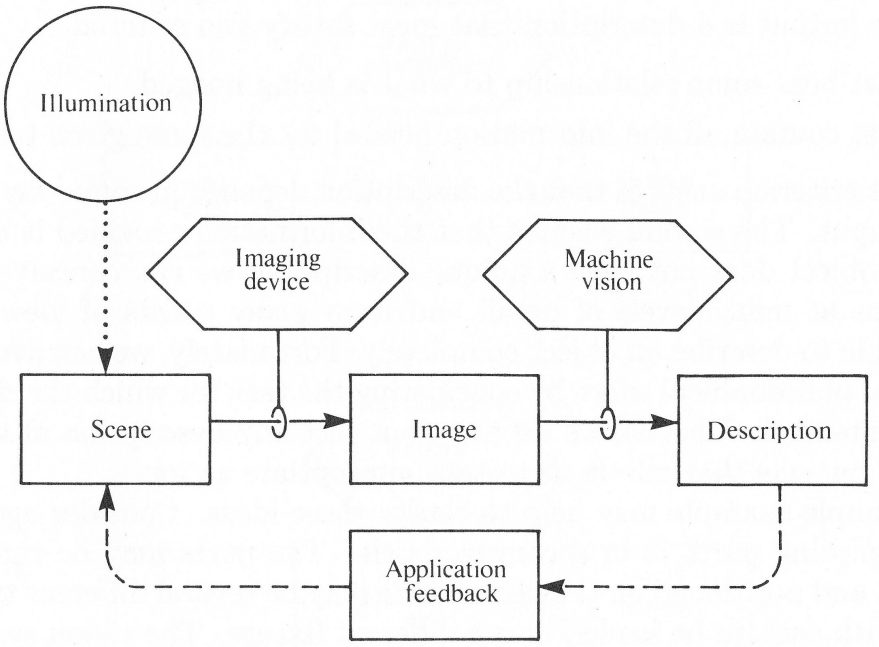


Figure 1-2. The purpose of a machine vision system is to produce a symbolic description of what is being imaged. This description may then be used to direct the interaction of a robotic system with its environment. In some sense, the vision system's task can be viewed as an inversion of the imaging process.

1.2 Tasks for a Machine Vision System

A machine vision system analyzes images and produces descriptions of what is imaged (figure 1.2). These descriptions must capture the aspects of the objects being imaged that are useful in carrying out some task. Thus we consider the machine vision system as part of a larger entity that interacts

with the environment. The vision system can be considered an element of a feedback loop that is concerned with sensing, while other elements are dedicated to decision making and the implementation of these decisions.

The input to the machine vision system is an image, or several images, while its output is a description that must satisfy two criteria:

- It must bear some relationship to what is being imaged.
- It must contain all the information needed for the some given task.

The first criterion ensures that the description depends in some way on the visual input. The second ensures that the information provided is useful.

An object does not have a unique description; we can conceive of descriptions at many levels of detail and from many points of view. It is impossible to describe an object completely. Fortunately, we can avoid this potential philosophical snare by considering the task for which the description is intended. That is, we do not want just any description of what is imaged, but one that allows us to take appropriate action.

A simple example may help to clarify these ideas. Consider again the task of picking parts from a conveyor belt. The parts may be randomly oriented and positioned on the belt. There may be several different types of parts, with each to be loaded into a different fixture. The vision system is provided with images of the objects as they are transported past a camera mounted above the belt. The descriptions that the system has to produce in this case are simple. It need only give the position, orientation, and type of each object. The description could be just a few numbers. In other situations an elaborate symbolic description may be called for.

There are cases where the feedback loop is not closed through a machine, but the description is provided as output to be interpreted by a human. The two criteria introduced above must still be satisfied, but it is harder in this case to determine whether the system was successful in solving the vision problem presented.

1.3 Relation to Other Fields

Machine vision is closely allied with three fields (figure 1-3):

- Image processing.
- Pattern classification.
- Scene analysis.

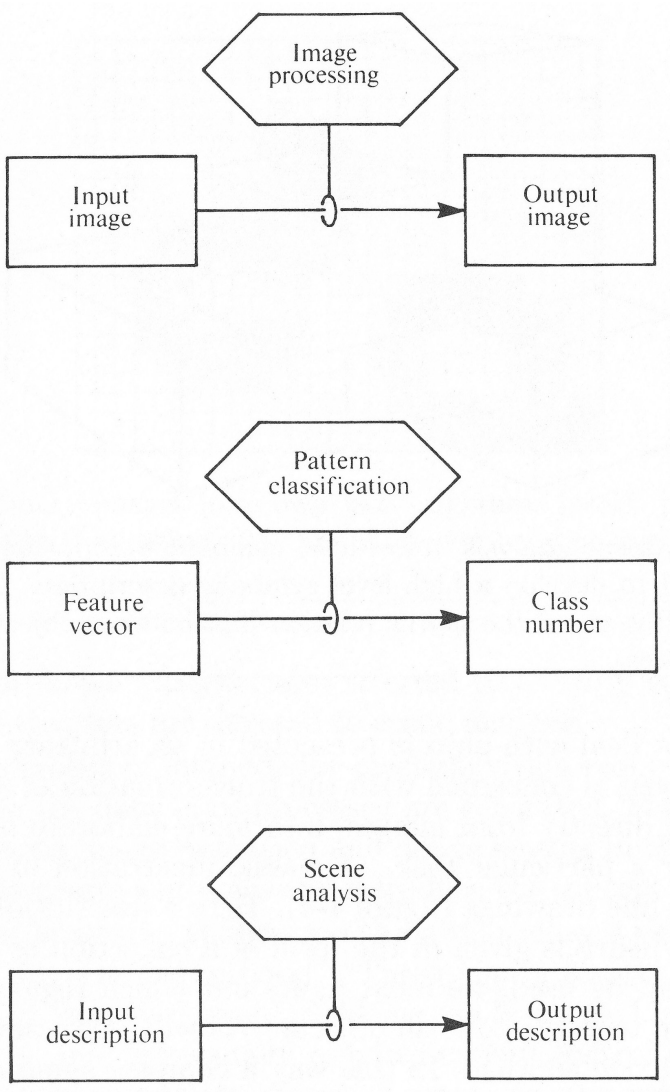


Figure 1-3. Three ancestor paradigms of machine vision are image processing, pattern classification, and scene analysis. Each contributes useful techniques, but none is central to the problem of developing symbolic descriptions from images.

Image processing is largely concerned with the generation of new images from existing images. Most of the techniques used come from linear systems theory. The new image may have noise suppressed, blurring removed, or edges accentuated. The result is, however, still an image, usually meant to be interpreted by a person. As we shall see, some of the techniques of image processing are useful for understanding the limitations of image-forming systems and for designing preprocessing modules for machine vision.

Pattern classification has as its main thrust the classification of a "pat-

tern," usually given as a set of numbers representing measurements of an object, such as height and weight. Although the input to a classifier is not an image, the techniques of pattern classification are at times useful for analyzing the results produced by a machine vision system. To recognize an object means to assign it to one of a number of known classes. Note, however, that recognition is only one of many tasks faced by the machine vision system. Researchers concerned with classification have created simple methods for obtaining measurements from images. These techniques, however, usually treat the images as a two-dimensional pattern of brightness and cannot deal with objects presented in an arbitrary attitude.

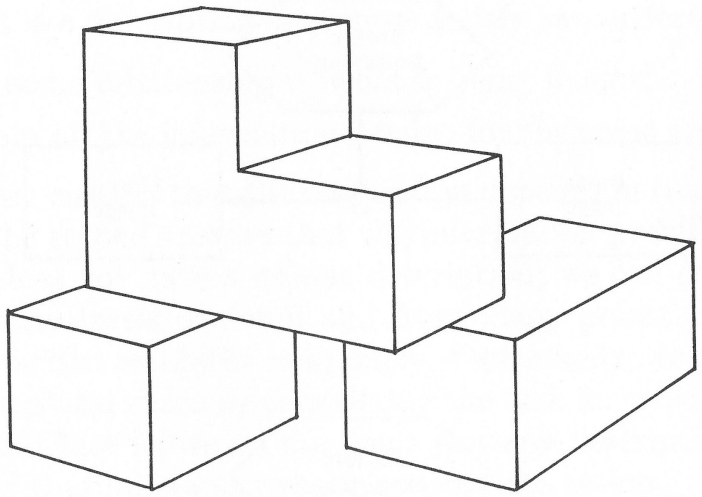


Figure 1-4. In scene analysis, a low-level symbolic description, such as a line drawing, is used to develop a high-level symbolic description. The result may contain information about the spatial relationships between objects, their shapes, and their identities.

Scene analysis is concerned with the transformation of simple descriptions, obtained directly from images, into more elaborate ones, in a form more useful for a particular task. A classic illustration of this is the interpretation of line drawings (figure 1-4). Here a description of the image of a set of polyhedra is given in the form of a collection of line segments. Before these can be used, we must figure out which regions bounded by the lines belong together to form objects. We will also want to know how objects support one another. In this way a complex symbolic description of the image can be obtained from the simple one. Note that here we do not start with an image, and thus once again do not address the central issue of machine vision:

- Generating a symbolic description from one or more images.

1.4 Outline of What Is to Come

The generation of descriptions from images can often be conveniently broken down into two stages. The first stage produces a *sketch*, a detailed but undigested description. Later stages produce more parsimonious, structured descriptions suitable for decision making. Processing in the first stage will be referred to as *image analysis*, while subsequent processing of the results will be called *scene analysis*. The division is somewhat arbitrary, except insofar as image analysis starts with an image, while scene analysis begins with a sketch. The first thirteen chapters of the book are concerned with image analysis, also referred to as *early vision*, while the remaining five chapters are devoted to scene analysis.

The development of methods for machine vision requires some understanding of how the data to be processed are generated. For this reason we start by discussing image formation and image sensing in chapter 2. There we also treat measurement noise and introduce the concept of convolution.

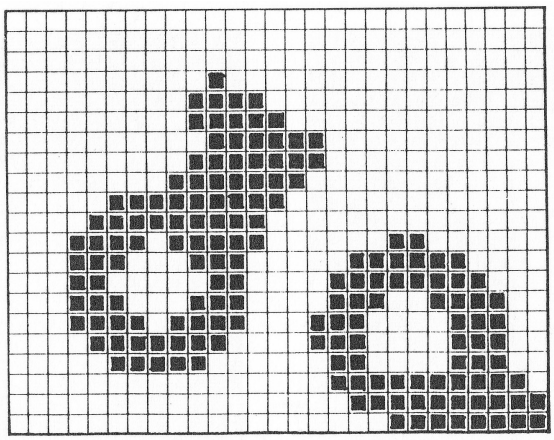


Figure 1-5. Binary images have only two brightness levels: black and white. While restricted in application, they are of interest because they are particularly easy to process.

The easiest images to analyze are those that allow a simple separation of an "object" from a "background." These *binary images* will be treated first (figure 1-5). Some industrial problems can be tackled by methods that use such images, but this usually requires careful control of the lighting. There exists a fairly complete theory of what can and cannot be accomplished with binary images. This is in contrast to the more general case of *gray-level images*. It is known, for example, that binary image techniques are useful only when possible changes in the attitude of the object are confined to rotations in a plane parallel to the image plane. Binary image

processing is covered in chapters 3 and 4.

Many image-analysis techniques are meant to be applied to regions of an image corresponding to single objects, rather than to the whole image. Because typically many surfaces in the environment are imaged together, the image must be divided up into regions corresponding to separate entities in the environment before such techniques can be applied. The required segmentation of images is discussed in chapter 5.

In chapters 6 and 7 we consider the transformation of gray-level images into new gray-level images by means of linear operations. The usual intent of such manipulations is to reduce noise, accentuate some aspect of the image, or reduce its dynamic range. Subsequent stages of the machine vision system may find the processed images easier to analyze. Such filtering methods are often exploited in edge-detection systems as preprocessing steps.

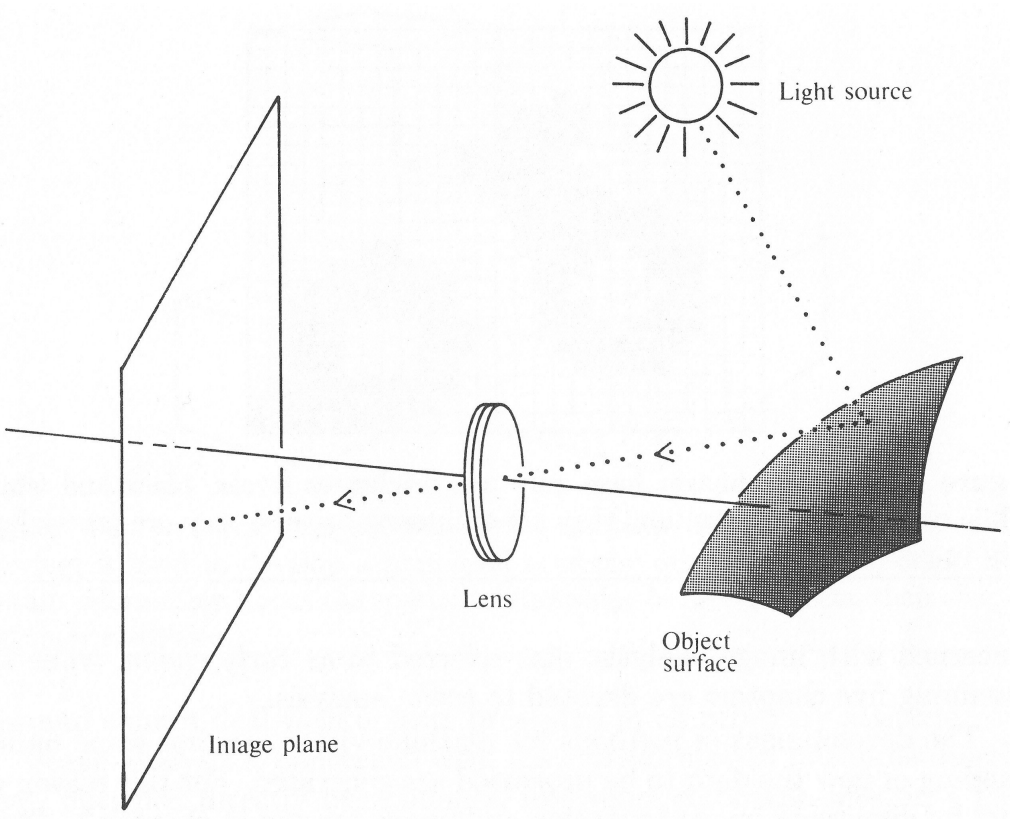


Figure 1-6. In order to use images to recover information about the world, we need to understand image formation. In some cases the image formation process can be inverted to extract estimates of the permanent properties of the surfaces of the objects being imaged.

Complementary to image segmentation is edge finding, discussed in chapter 8. Often the interesting events in a scene, such as a boundary where one object occludes another, lead to discontinuities in image brightness or in brightness gradient. Edge-finding techniques locate such features. At this point, we begin to emphasize the idea that an important aspect of machine vision is the estimation of properties of the surfaces being imaged. In chapter 9 the estimation of surface reflectance and color is addressed and found to be a surprisingly difficult task.

Finally, we confront the central issue of machine vision: the generation of a description of the world from one or more images. A point of view that one might espouse is that the purpose of the machine vision system is to invert the projection operation performed by image formation. This is not quite correct, since we want not to recover the world being imaged, but to obtain a symbolic description. Still, this notion leads us to study image formation carefully (figure 1-6). The way light is reflected from a surface becomes a central issue. The apparent brightness of a surface depends on three factors:

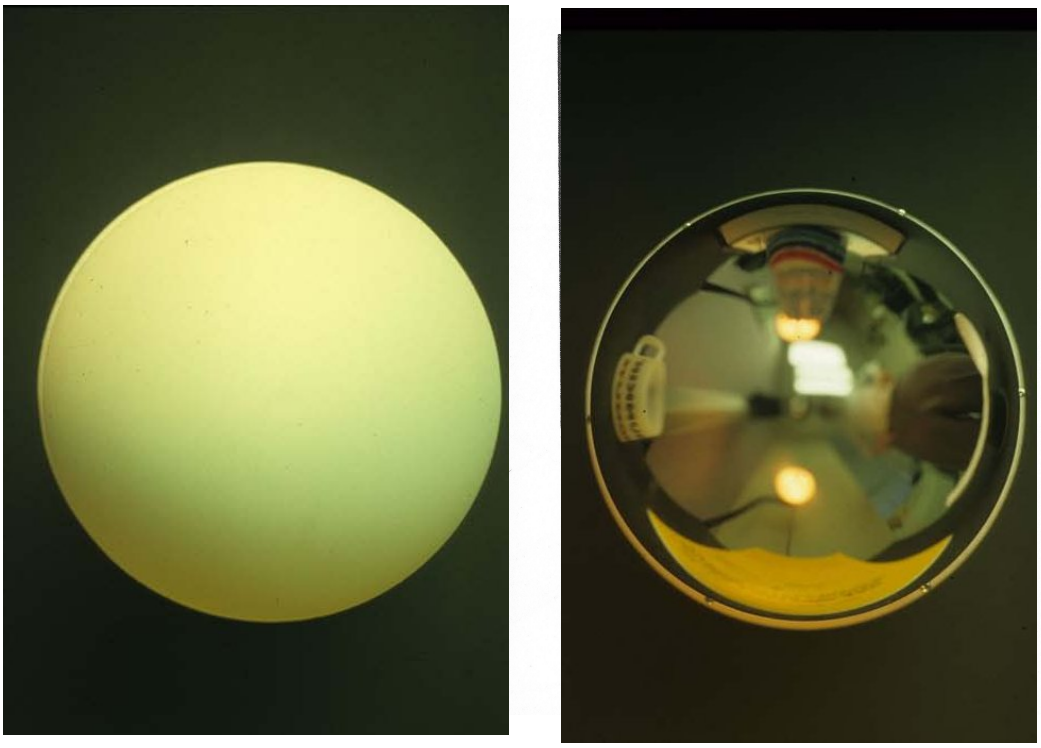


Figure 1-7. The appearance of the image of an object is greatly influenced by the reflectance properties of its surface. Perfectly matte and perfectly specular surfaces present two extreme cases.

Figure 1-8. The appearance of the image of a scene depends a lot on the lighting conditions. To recover information about the world from images we need to understand how the brightness patterns in the image are determined by the shapes of surfaces, their reflectance properties, and the distribution of light sources.

- Microstructure of the surface.
- Distribution of the incident light.
- Orientation of the surface with respect to the viewer and the light sources.

In figure 1-7 we see images of two spherical surfaces, one covered with a paint that has a matte or diffuse reflectance, the other metallic, giving rise to specular reflections. In the second case we see a virtual image of the world around the spherical object. It is clear that the microstructure of the surface is important in determining image brightness.

Figure 1-8 shows three views of Place Ville-Marie in Montreal. The three pictures were taken from the same hotel window, but under different lighting conditions. Again, we easily recognize that the same objects are depicted, but there is a tremendous difference in brightness patterns between the images taken with direct solar illumination and those obtained under a cloudy sky.

In chapters 10 and 11 we discuss these issues and apply the understanding developed to the recovery of surface shape from one or more images. Representations for the shape of a surface are also introduced there. In developing methods for recovering surface shape, we often consider the surface broken up into tiny patches, each of which can be treated as if it were planar. Light reflection from such a planar patch is governed by three angles if it is illuminated by a point source (figure 1-9).

The same systematic approach, based on an analysis of image brightness, is used in chapters 12 and 13 to recover information from time-varying images and images taken by cameras separated in space. Surface shape, object motion, and other information can be recovered from images using the methods developed in these two chapters. The relations between various coordinate systems, either viewer-centered or object-centered, are uncovered in the discussion of photogrammetry in chapter 13, along with an analysis of the binocular stereo problem. In using a machine vision system to guide a mechanical manipulator, measurements in the camera's coordinate system must be transformed into the coordinate system of the robot arm. This topic naturally fits into the discussion of this chapter also.

At this point, we turn from image analysis to scene analysis. Chapter 14 introduces methods for classifying objects based on feature measurements. Line drawings obtained from images of polyhedral objects are analyzed in chapter 15 in order to recover the spatial relationships between the objects.

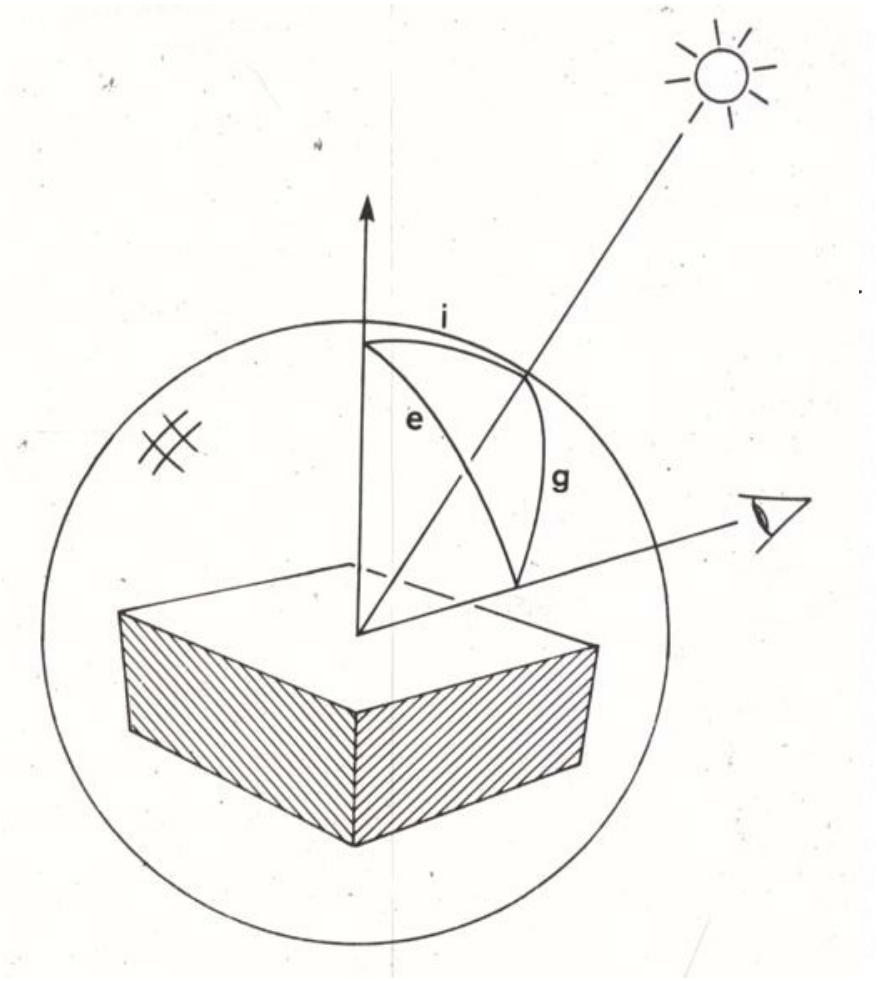


Figure 1-9. The reflection of light from a point source by a patch of an object's surface is governed by three angles: the incident angle i, the emittance angle e, and the phase angle g. Here N is the direction perpendicular, or normal, to the surface, S the direction to the light source, and V the direction toward the viewer.

The issue of how to represent visually acquired information is of great importance. In chapter 16 we develop in detail the extended Gaussian image, a representation for surface shape that is useful in recognition and allows us to determine the attitude of an object in space. Image sequences can be exploited to recover the motion of the camera. As a by-product, we obtain the shapes of the surfaces being imaged. This forms the topic of chapter 17. (The reader may wonder why this chapter does not directly follow the one on optical flow. The reason is that it does not deal with image analysis and so logically belongs in the part of the book dedicated to scene analysis.) Finally, in chapter 18 we bring together many of the concepts developed in this book to built a complete hand–eye system. A robot

arm is guided to pick up one object after another out of a pile of objects. Visual input provides the system with information about the positions of the objects and their attitudes in space. In this chapter we introduce some new topics, such as methods for representing rotations in three-dimensional space, and discuss some of the difficulties encountered in building a real-world system.

Throughout the book we start by discussing elementary issues and well-established techniques, progress to more advanced topics, and close with less certain matters and subjects of current research. In the past, machine vision may have appeared to be a collection of assorted heuristics and ad hoc tricks. To give the material coherence we maintain a particular point of view here:

- Machine vision should be based on a thorough understanding of image formation.

This emphasis allows us to derive mathematical models of the image-analysis process. Algorithms for recovering a description of the imaged world can then be based on these mathematical models.

An approach based on the analysis of image formation is, of course, not the only one possible for machine vision. One might start instead from existing biological vision systems. Artificial systems would then be based on detailed knowledge of natural systems, provided these can be adequately characterized. We shall occasionally discuss alternate approaches to given problems in machine vision, but to avoid confusion we will not dwell on them.

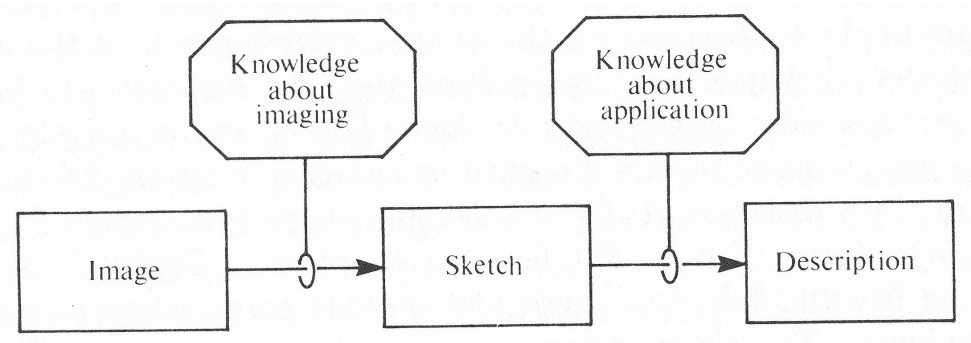


Figure 1-10. In many cases, the development of a symbolic description of a scene from one or more images can be broken down conveniently into two stages. The first stage is largely governed by our understanding of the image-formation process; the second depends more on the needs of the intended application.

The transformation from image to sketch appears to be governed mostly by what is in the image and what information we can extract directly from it (figure 1-10). The transformation from a crude sketch to a full symbolic description, on the other hand, is mostly governed by the need to generate information in a form that will be of use in the intended application.

1.5 References

Each chapter will have a section providing pointers to background reading, further explanation of the concepts introduced in that chapters, and recent results in the area. Books will be listed first, complete with authors and titles. Papers in journals, conference proceedings, and internal reports of universities and research laboratories are listed after the books, but without title. Please note that the bibliography has two sections: the first for books, the second for papers.

There are now numerous books on the subject of machine vision. Of these, *Computer Vision* by Ballard & Brown [1982] is remarkable for its broad coverage. Also notable are *Digital Picture Processing* by Rosenfeld & Kak [1982], *Computer Image Processing and Recognition* by Hall [1979], and *Machine Perception* [1982], a short book by Nevatia. A recent addition is *Vision in Man and Machine* [1985] by Levine, a book that has a biological vision point of view and emphasizes applications to biomedical problems.

Many books concentrate on the image-processing side of things, such as *Computer Techniques in Image Processing* by Andrews [1970], *Digital Image Processing* by Gonzalez & Wintz [1977], and two books dealing with the processing of images obtained by cameras in space: *Digital Image Processing* by Castleman [1979] and *Digital Image Processing: A Systems Approach* by Green [1983]. The first few chapters of *Digital Picture Processing* by Rosenfeld & Kak [1982] also provide an excellent introduction to the subject. The classic reference on image processing is still Pratt's encyclopedic *Digital Image Processing* [1978].

One of the earliest significant books in this field, *Pattern Classification and Scene Analysis* by Duda & Hart [1973], contains more on the subject of pattern classification than one typically needs to know. *Artificial Intelligence* by Winston [1984] has an easy-to-read, broad-brush chapter on machine vision that makes the connection between that subject and artificial intelligence.

A number of edited books, containing contributions from several researchers in the field, have appeared in the last ten years. Early on there

was *The Psychology of Computer Vision*, edited by Winston [1975], now out of print. Then came *Digital Picture Analysis*, edited by Rosenfeld [1976], and *Computer Vision Systems*, edited by Hanson & Riseman [1978]. Several papers on machine vision can be found in volume 2 of *Artificial Intelligence: An MIT Perspective*, edited by Winston & Brown [1979]. The collection *Structured Computer Vision: Machine Perception through Hierarchical Computation Structures*, edited by Tanimoto & Klinger, was published in 1980. Finally there appeared the fine assemblage of papers *Image Understanding 1984*, edited by Ullman & Richards [1984].

The papers presented at a number of conferences have also been collected in book form. Gardner was the editor of a book published in 1979 called *Machine-aided Image Analysis, 1978*. Applications of machine vision to robotics are explored in *Computer Vision and Sensor-Based Robots*, edited by Dodd & Rossol [1979], and in *Robot Vision*, edited by Pugh [1983]. Stucki edited *Advances in Digital Image Processing: Theory, Application, Implementation* [1979], a book containing papers presented at a meeting organized by IBM. The notes for a course organized by Faugeras appeared in *Fundamentals in Computer Vision* [1983].

Because many of the key papers in the field were not easily accessible, a number of collections have appeared, including three published by IEEE Press, namely *Computer Methods in Image Analysis*, edited by Aggarwal, Duda, & Rosenfeld [1977], *Digital Image Processing*, edited by Andrews [1978], and *Digital Image Processing for Remote Sensing*, edited by Bernstein [1978].

The IEEE Computer Society's publication *Computer* brought out a special issue on image processing in August 1977, the *Proceedings of the IEEE* devoted the May 1979 issue to pattern recognition and image processing, and *Computer* produced a special issue on machine perception for industrial applications in May 1980. A special issue (Volume 17) of the journal *Artificial Intelligence* was published in book form under the title *Computer Vision*, edited by Brady [1981]. The Institute of Electronics and Communication Engineers of Japan produced a special issue (Volume J68-D, Number 4) on machine vision work in Japan in April 1985 (in Japanese).

Not much is said in this book about biological vision systems. They provide us, on the one hand, with reassuring existence proofs and, on the other, with optical illusions. These startling effects may someday prove to be keys with which we can unlock the secrets of biological vision systems. A computational theory of their function is beginning to emerge, to a great extent due to the pioneering work of a single man, David Marr.

His approach is documented in the classic book *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information* [1982].

Human vision has, of course, always been a subject of intense curiosity, and there is a vast literature on the subject. Just a few books will be mentioned here. Gregory has provided popular accounts of the subject in *Eye and Brain* [1966] and *The Intelligent Eye* [1970]. Three books by Gibson—*The Perception of the Visual World* [1950], *The Senses Considered as Perceptual Systems* [1966], and *The Ecological Approach to Visual Perception* [1979]—are noteworthy for providing a fresh approach to the problem. Cornsweet's *Visual Perception* [1971] and *The Psychology of Visual Perception* by Haber & Hershenson [1973] are of interest also. The work of Julesz has been very influential, particularly in the area of binocular stereo, as documented in *Foundations of Cyclopean Perception* [1971]. More recently, in the wonderfully illustrated book *Seeing*, Frisby [1982] has been able to show the crosscurrents between work on machine vision and work on biological vision systems. For another point of view see *Perception* by Rock [1984].

Twenty years ago, papers on machine vision were few in number and scattered widely. Since then a number of journals have become preferred repositories for new research results. In fact, the journal *Computer Graphics and Image Processing*, published by Academic Press, had to change its name to *Computer Vision, Graphics and Image Processing* (CVGIP) when it became the standard place to send papers in this field for review. More recently, a new special-interest group of the Institute of Electrical and Electronic Engineers (IEEE) started publishing the *Transactions on Pattern Analysis and Machine Intelligence* (PAMI). Other journals, such as *Artificial Intelligence*, published by North-Holland, and *Robotics Research*, published by MIT Press, also contain articles on machine vision. There are several journals devoted to related topics, such as pattern classification.

Some research results first see the light of day at an "Image Understanding Workshop" sponsored by the Defense Advanced Research Projects Agency (DARPA). Proceedings of these workshops are published by Science Applications Incorporated, McLean, Virginia, and are available through the Defense Technical Information Center (DTIC) in Alexandria, Virginia. Many of these papers are later submitted, possibly after revision and extension, to be reviewed for publication in one of the journals mentioned above.

The Computer Society of the IEEE organizes annual conferences on

Computer Vision and Pattern Recognition (CVPR) and publishes their proceedings. Also of interest are the proceedings of the biannual International Joint Conference on Artificial Intelligence (IJCAI) and the national conferences organized by the American Association for Artificial Intelligence (AAAI), usually in the years in between.

The thorough annual surveys by Rosenfeld [1972, 1974, 1975, 1976, 1977, 1978, 1979, 1980, 1981, 1982, 1983, 1984a, 1985] in *Computer Vision, Graphics and Image Processing* are extremely valuable and make it possible to be less than complete in providing references here. The most recent survey contained 1,252 entries! There have been many analyses of the state of the field or of particular views of the field. An early survey of image processing is that of Huang, Schreiber, & Tretiak [1971]. While not really a survey, the influential paper of Barrow & Tenenbaum [1978] presents the now prevailing view that machine vision is concerned with the process of recovering information about the surfaces being imaged. More recent surveys of machine vision by Marr [1980], Barrow & Tenenbaum [1981a], Poggio [1984], and Rosenfeld [1984b] are recommended particularly. Another paper that has been influential is that by Binford [1981].

Once past the hurdles of early vision, the representation of information and the modeling of objects and the physical interaction between them become important. We touch upon these issues in the later chapters of this book. For more information see, for example, Brooks [1981] and Binford [1982].

There are many papers on the application of machine vision to industrial problems (although some of the work with the highest payoff is likely not to have been published in the open literature). Several papers in *Robotics Research: The First International Symposium*, edited by Brady & Paul [1984], deal with this topic. Chin [1982] and Chin & Harlow [1982] have surveyed the automation of visual inspection.

The inspection of printed circuit boards, both naked and stuffed, is a topic of great interest, since there are many boards to be inspected and since it is not a very pleasant job for people, nor one that they are particularly good at. For examples of work in this area, see Ejiri et al. [1973], Danielsson & Kruse [1979], Danielsson [1980], and Hara, Akiyama, & Karasaki [1983]. There is a similar demand for such techniques in the manufacture of integrated circuits. Masks are simple black-and-white patterns, and their inspection has not been too difficult to automate. The inspection of integrated circuit wafers is another matter; see, for example, Hsieh & Fu [1980].

Machine vision has been used in automated alignment. See Horn [1975b], Kashioka, Ejiri, & Sakamoto [1976], and Baird [1978] for examples in semiconductor manufacturing. Industrial robots are regularly guided using visually obtained information about the position and orientation of parts. Many such systems use binary image-processing techniques, although some are more sophisticated. See, for example, Yachida & Tsuji [1977], Gonzalez & Safabakhsh [1982], and Horn & Ikeuchi [1984]. These techniques will not find widespread application if the user has to program each application in a standard programming language. Some attempts have been made to provide tools specifically suited to the vision applications; see, for example, Lavin & Lieberman [1982].

Papers on the application of machine vision methods to the vectorization of line drawings are mentioned at the end of chapter 4; references on character recognition may be found at the end of chapter 14.

1.6 Exercises

1-1 Explain in what sense one can consider pattern classification, image processing, and scene analysis as "ancestor paradigms" to machine vision. In what way do the methods from each of these disciplines contribute to machine vision? In what way are the problems addressed by machine vision different from those to which these methods apply?

2

Image Formation & Image Sensing

In this chapter we explore how images are formed and how they are sensed by a computer. Understanding image formation is a prerequisite for full understanding of the methods for recovering information from images. In analyzing the process by which a three-dimensional world is projected onto a two-dimensional image plane, we uncover the two key questions of image formation:

- What determines where the image of some point will appear?
- What determines how bright the image of some surface will be?

The answers to these two questions require knowledge of image projection and image radiometry, topics that will be discussed in the context of simple lens systems.

A crucial notion in the study of image formation is that we live in a very special visual world. It has particular features that make it possible to recover information about the three-dimensional world from one or more two-dimensional images. We discuss this issue and point out imaging situations where these special constraint do not apply, and where it is consequently much harder to extract information from images.

We also study the basic mechanism of typical image sensors, and how information in different spectral bands may be obtained and processed. Following a brief discussion of color, the chapter closes with a discussion of noise and reviews some concepts from the fields of probability and statistics. This is a convenient point to introduce convolution in one dimension, an idea that will be exploited later in its two-dimensional generalization. Readers familiar with these concepts may omit these sections without loss of continuity. The chapter concludes with a discussion of the need for quantization of brightness measurements and for tessellations of the image plane.

2.1 Two Aspects of Image Formation

Before we can analyze an image, we must know how it is formed. An image is a two-dimensional pattern of brightness. How this pattern is produced in an optical image-forming system is best studied in two parts: first, we need to find the geometric correspondence between points in the scene and points in the image; then we must figure out what determines the brightness at a particular point in the image.

2.1.1 Perspective Projection

Consider an ideal pinhole at a fixed distance in front of an image plane (figure 2-1). Assume that an enclosure is provided so that only light coming through the pinhole can reach the image plane. Since light travels along straight lines, each point in the image corresponds to a particular direction defined by a ray from that point through the pinhole. Thus we have the familiar *perspective projection*.

We define the *optical axis*, in this simple case, to be the perpendicular from the pinhole to the image plane. Now we can introduce a convenient Cartesian coordinate system with the origin at the pinhole and z-axis aligned with the optical axis and pointing toward the image. With this choice of orientation, the z components of the coordinates of points in front of the camera are negative. We use this convention, despite the drawback, because it gives us a convenient right-hand coordinate system (with the x-axis to the right and the y-axis upward).

We would like to compute where the image P' of the point P on some object in front of the camera will appear (figure 2-1). We assume that no other object lies on the ray from P to the pinhole O. Let $\mathbf{r} = (x, y, z)^T$ be the vector connecting O to P, and $\mathbf{r}' = (x', y', f')^T$ be the vector connecting O to P'. (As explained in the appendix, vectors will be denoted by boldface

letters. We commonly deal with column vectors, and so must take the transpose, indicated by the superscript ^{T}, when we want to write them in terms of the equivalent row vectors.)

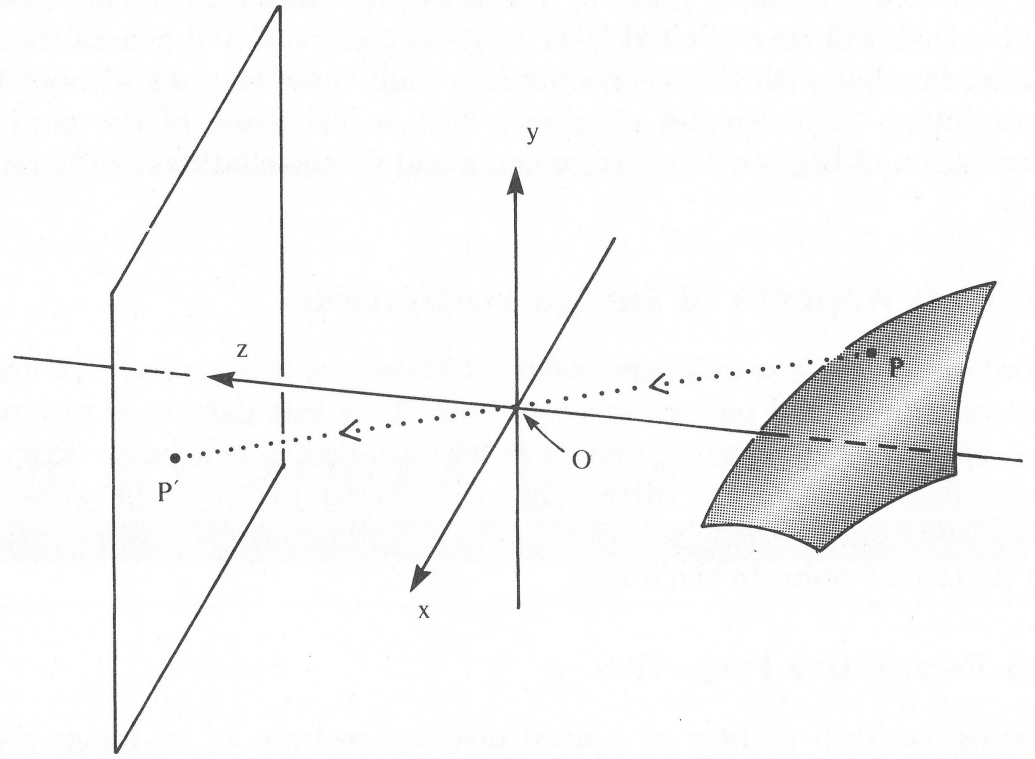


Figure 2-1. A pinhole camera produces an image that is a perspective projection of the world. It is convenient to use a coordinate system in which the xy-plane is parallel to the image plane, and the origin is at the pinhole O. The z-axis then lies along the optical axis.

Here f' is the distance of the image plane from the pinhole, while x' and y' are the coordinates of the point P' in the image plane. The two vectors $\mathbf{r}$ and $\mathbf{r}'$ are collinear and differ only by a (negative) scale factor. If the ray connecting P to P' makes an angle α with the optical axis, then the length of $\mathbf{r}$ is just

$$r = -z \sec \alpha = -(\mathbf{r} \cdot \hat{\mathbf{z}}) \sec \alpha,$$

where $\hat{\mathbf{z}}$ is the unit vector along the optical axis. (Remember that z is negative for a point in front of the camera.)

The length of $\mathbf{r}'$ is

$$r' = f' \sec \alpha,$$

and so

$$\frac{1}{f'}\mathbf{r}' = \frac{1}{\mathbf{r}\cdot\hat{\mathbf{z}}}\mathbf{r}.$$

In component form this can be written as

$$\frac{x'}{f'} = \frac{x}{z} \qquad \text{and} \qquad \frac{y'}{f'} = \frac{y}{z}.$$

Sometimes image coordinates are normalized by dividing x' and y' by f' in order to simplify the projection equations.

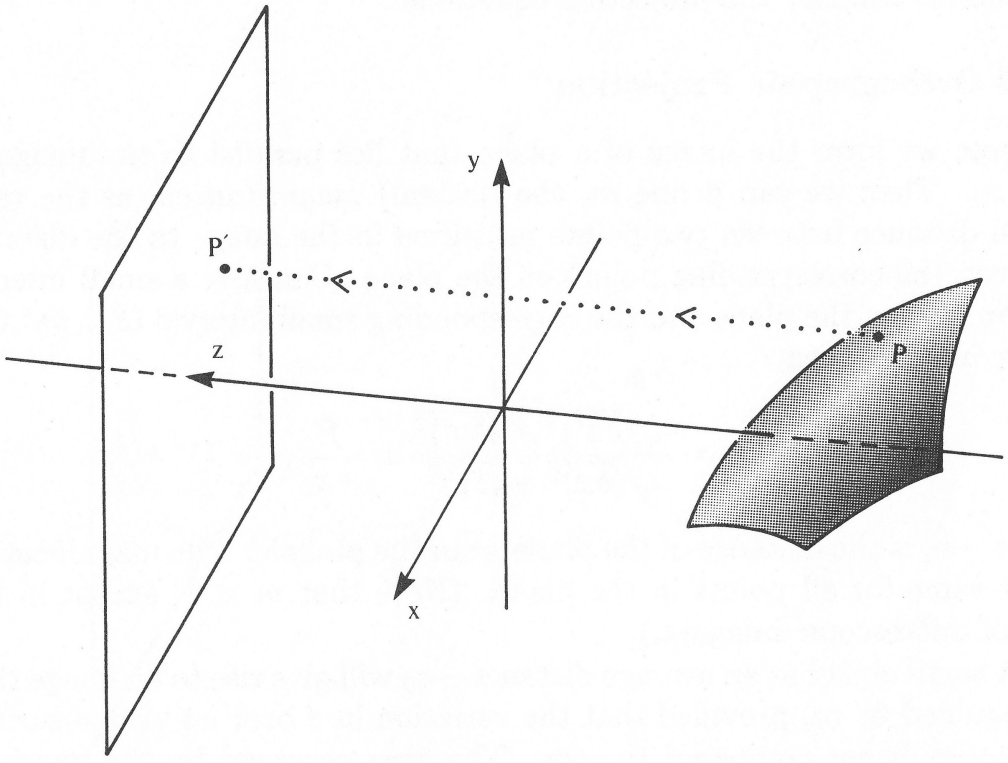


Figure 2-2. When the scene depth is small relative to the average distance from the camera, perspective projection can be approximated by orthographic projection. In orthographic projection, rays from a point in the scene are traced parallel to the projection direction until they intercept the image plane.

2.1.2 Orthographic Projection

Suppose we form the image of a plane that lies parallel to the image at $z = z_0$. Then we can define m, the (lateral) *magnification*, as the ratio of the distance between two points measured in the image to the distance between the corresponding points on the plane. Consider a small interval $(\delta x, \delta y, 0)^T$ on the plane and the corresponding small interval $(\delta x', \delta y', 0)^T$

in the image. Then

$$m = \frac{\sqrt{(\delta x')^2 + (\delta y')^2}}{\sqrt{(\delta x)^2 + (\delta y)^2}} = \frac{f'}{-z_0},$$

where $-z_0$ is the distance of the plane from the pinhole. The magnification is the same for all points in the plane. (Note that $m < 1$, except in the case of microscopic imaging.)

A small object at an average distance $-z_0$ will give rise to an image that is magnified by m, provided that the variation in z over its visible surface is not significant compared to $-z_0$. The area occupied by the image of an object is proportional to m^2. Objects at different distances from the imaging system will, of course, be imaged with different magnifications. Let the *depth range* of a scene be the range of distances of surfaces from the camera. The magnification is approximately constant when the depth range of the scene being imaged is small relative to the average distance of the surfaces from the camera. In this case we can simplify the projection equations to read

$$x' = -mx \qquad \text{and} \qquad y' = -my,$$

where $m = f'/(-z_0)$ and $-z_0$ is the average value of $-z$. Often the scaling factor m is set to 1 or -1 for convenience. Then we can further simplify the equations to become

$$x' = x \qquad \text{and} \qquad y' = y.$$

This *orthographic projection* (figure 2-2), can be modeled by rays parallel to the optical axis (rather than ones passing through the origin). The difference between perspective and orthographic projection is small when the distance to the scene is much larger than the variation in distance among objects in the scene.

The *field of view* of an imaging system is the angle of the cone of directions encompassed by the scene that is being imaged. This cone of directions clearly has the same shape and size as the cone obtained by connecting the edge of the image plane to the center of projection. A "normal" lens has a field of view of perhaps 25° by 40°. A *telephoto lens* is one that has a long focal length relative to the image size and thus a narrow field of view. Conversely, a *wide-angle lens* has a short focal length relative to the image size and thus a wide field of view. A rough rule of thumb is that perspective effects are significant when a wide-angle lens is used, while images obtained using a telephoto lenses tend to approximate

orthographic projection. We shall show in exercise 2-11 that this rule is not exact.

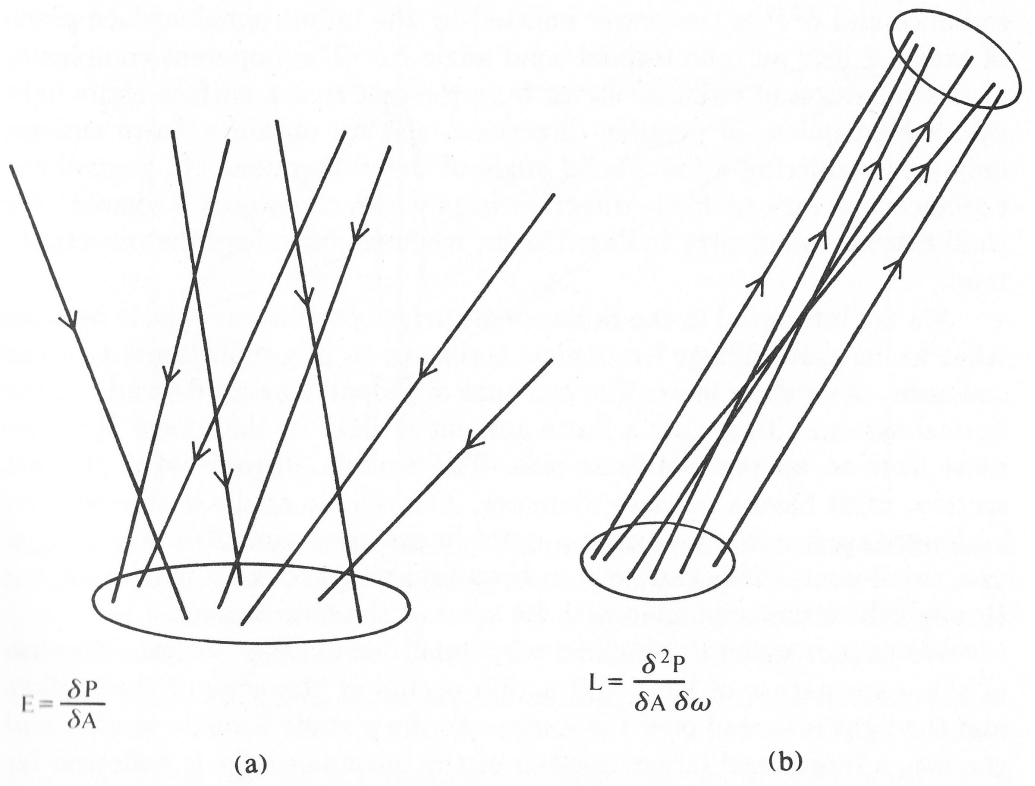


Figure 2-3. (a) Irradiance is the power per unit area falling on a surface. (b) Radiance is the power emitted per unit area into a cone of directions having unit solid angle. The term *brightness* is used informally for both concepts.

2.2 Brightness

The more difficult, and more interesting, question of image formation is what determines the brightness at a particular point in the image. *Brightness* is an informal term used to refer to at least two different concepts: image brightness and scene brightness. In the image, brightness is related to energy flux incident on the image plane and can be measured in a number of ways. Here we introduce the term *irradiance* to replace the informal term *image brightness*. Irradiance is the power per unit area ($W \cdot m^{-2}$—watts per square meter) of radiant energy falling on a surface (figure 2-3a). In the figure, E denotes the irradiance, while δP is the power of the radiant energy falling on the infinitesimal surface patch of area δA. The blackening of a film in a camera, for example, is a function of the irradiance. (As we

shall discuss a little later, the measurement of brightness in the image also depends on the spectral sensitivity of the sensor.) The irradiance at a particular point in the image will depend on how much light arrives from the corresponding object point (the point found by following the ray from the image point through the pinhole until it meets the surface of an object).

In the scene, brightness is related to the energy flux emitted from a surface. Different points on the objects in front of the imaging system will have different brightnesses, depending on how they are illuminated and how they reflect light. We now introduce the term *radiance* to substitute for the informal term *scene brightness*. Radiance is the power per unit foreshortened area emitted into a unit solid angle ($\mathrm{W \cdot m^{-2} \cdot sr^{-1}}$—watts per square meter per steradian) by a surface (figure 2-3b). In the figure, L is the radiance and $\delta^2 P$ is the power emitted by the infinitesimal surface patch of area δA into an infinitesimal solid angle $\delta\omega$. The apparent complexity of the definition of radiance stems from the fact that a surface emits light into a hemisphere of possible directions, and we obtain a finite amount only by considering a finite solid angle of these directions. In general the radiance will vary with the direction from which the object is viewed. We shall discuss radiometry in detail later, when we introduce the reflectance map.

We are interested in the radiance of surface patches on objects because what we measure, image irradiance, turns out to be proportional to scene radiance, as we show later. The constant of proportionality depends on the optical system. To gather a finite amount of light in the image plane we must have an aperture of finite size. The pinhole, introduced in the last section, must have a nonzero diameter. Our simple analysis of projection no longer applies, though, since a point in the environment is now imaged as a small circle. This can be seen by considering the cone of rays passing through the circular pinhole with its apex at the object point.

We cannot make the pinhole very small for another reason. Because of the wave nature of light, diffraction occurs at the edge of the pinhole and the light is spread over the image. As the pinhole is made smaller and smaller, a larger and larger fraction of the incoming light is deflected far from the direction of the incoming ray.

2.3 Lenses

In order to avoid the problems associated with pinhole cameras, we now consider the use of a lens in an image-forming system. An ideal lens produces the same projection as the pinhole, but also gathers a finite amount

of light (figure 2-4). The larger the lens, the larger the solid angle it subtends when seen from the object. Correspondingly it intercepts more of the light reflected from (or emitted by) the object. The ray through the center of the lens is undeflected. In a well-focused system the other rays are deflected to reach the same image point as the central ray.

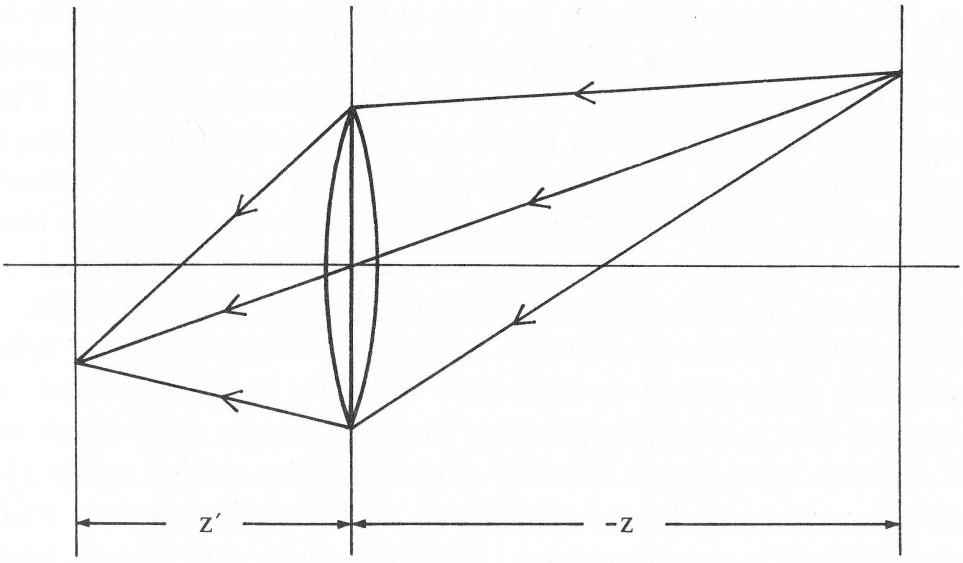


Figure 2-4. To obtain finite irradiance in the image plane, a lens is used instead of an ideal pinhole. A perfect lens generates an image that obeys the same projection equations as that generated by a pinhole, but gathers light from a finite area as well. A lens produces well-focused images of objects at a particular distance only.

An ideal lens has the disadvantage that it only brings to focus light from points at a distance $-z$ given by the familiar lens equation

$$\frac{1}{z'} + \frac{1}{-z} = \frac{1}{f},$$

where z' is the distance of the image plane from the lens and f is the *focal length* (figure 2-4). Points at other distances are imaged as little circles. This can be seen by considering the cone of light rays passing through the lens with apex at the point where they are correctly focused. The size of the blur circle can be determined as follows: A point at distance $-\overline{z}$ is imaged at a point $\overline{z}'$ from the lens, where

$$\frac{1}{\overline{z}'} + \frac{1}{-\overline{z}} = \frac{1}{f},$$

and so

$$(\overline{z}' - z') = \frac{f}{(\overline{z} + f)} \frac{f}{(z + f)} (\overline{z} - z).$$

If the image plane is situated to receive correctly focused images of objects at distance $-z$, then points at distance $-\overline{z}$ will give rise to blur circles of diameter

$$\frac{d}{z'} \left| \overline{z}' - z' \right|,$$

where d is the diameter of the lens. The *depth of field* is the range of distances over which objects are focused "sufficiently well," in the sense that the diameter of the blur circle is less than the resolution of the imaging device. The depth of field depends, of course, on what sensor is used, but in any case it is clear that the larger the lens aperture, the less the depth of field. Clearly also, errors in focusing become more serious when a large aperture is employed.

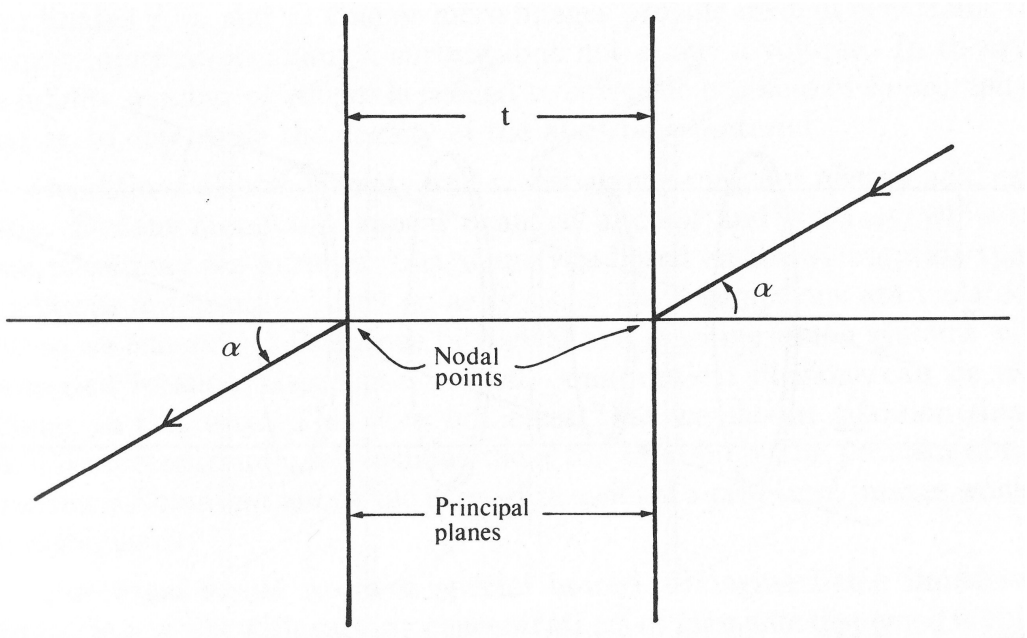


Figure 2-5. An ideal thick lens provides a reasonable model for most real lenses. It produces the same perspective projection that an ideal thin lens does, except for an additional offset, the lens thickness t, along the optical axis. It can be understood in terms of the principal planes and the nodal points at the intersections of the principal planes and the optical axis.

Simple ray-tracing rules can help in understanding simple lens combinations. As already mentioned, the ray through the center of the lens is undeflected. Rays entering the lens parallel to the optical axis converge to

a point on the optical axis at a distance equal to the focal length. This follows from the definition of focal length as the distance from the lens where the image of an object that is infinitely far away is focused. Conversely, rays emitted from a point on the optical axis at a distance equal to the focal length from the lens are deflected to emerge parallel to the optical axis on the other side of the lens. This follows from the reversibility of rays. At an interface between media of different refractive indices, the same reflection and refraction angles apply to light rays traveling in opposite directions.

A simple lens is made by grinding and polishing a glass blank so that its two surfaces have shapes that are spherical. The optical axis is the line through the centers of the two spheres. Any such simple lens will have a number of defects or aberrations. For this reason one usually combines several simple lenses, carefully lining up their individual optical axes, so as to make a compound lens with better properties.

A useful model of such a system of lenses is the *thick lens* (figure 2-5). One can define two *principal planes* perpendicular to the optical axis, and two *nodal points* where these planes intersect the optical axis. A ray arriving at the front nodal point leaves the rear nodal point without changing direction. This defines the projection performed by the lens. The distance between the two nodal points is the *thickness* of the lens. A *thin lens* is one in which the two nodal points can be considered coincident.

It is theoretically impossible to make a perfect lens. The projection will never be exactly like that of an ideal pinhole. More important, exact focusing of all rays cannot be achieved. A variety of aberrations occur. In a well-designed lens these defects are kept to a minimum, but this becomes more difficult as the aperture of the lens is increased. Thus there is a trade-off between light-gathering power and image quality.

A defect of particular interest to us here is called *vignetting*. Imagine several circular diaphragms of different diameter, stacked one behind the other, with their centers on a common line (figure 2-6). When you look along this common line, the smallest diaphragm will limit your view. As you move away from the line, some of the other diaphragms will begin to occlude more, until finally nothing can be seen. Similarly, in a simple lens, all the rays that enter the front surface of the lens end up being focused in the image. In a compound lens, some of the rays that pass through the first lens may be occluded by portions of the second lens, and so on. This will depend on the inclination of the entering ray with respect to the optical axis and its distance from the front nodal point. Thus points in the image away from the optical axis benefit less from the light-gathering

power of the lens than does the point on the optical axis. There is a falloff in sensitivity with distance from the center of the image.

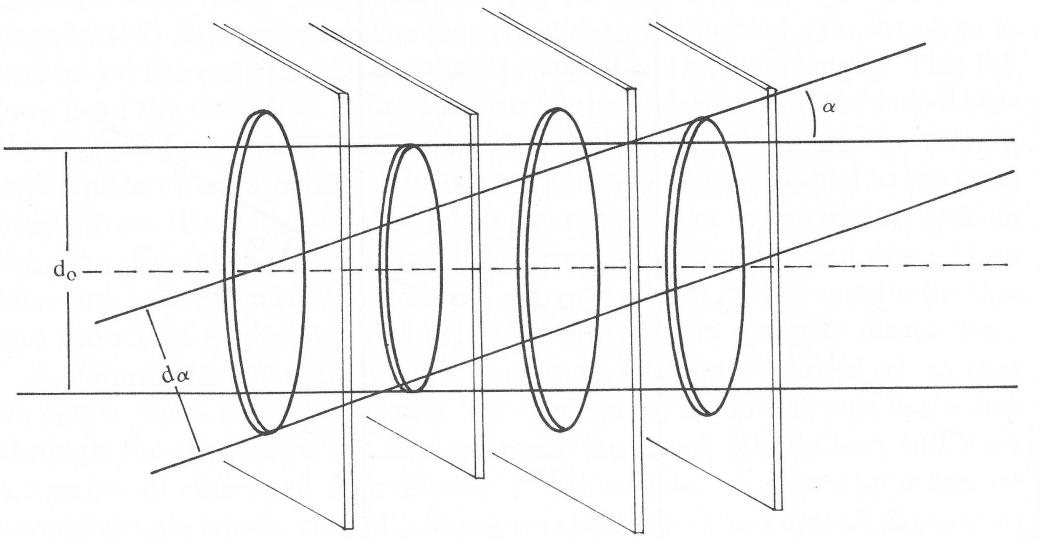


Figure 2-6. Vignetting is a reduction in light-gathering power with increasing inclination of light rays with respect to the optical axis. It is caused by apertures in the lens system occluding part of the beam of light as it passes through the lens system. Vignetting results in a smooth, but sometimes quite large, falloff in sensitivity toward the edges of the image region.

Another important consideration is that the aberrations of a lens increase in magnitude as a power of the angle between the incident ray and the optical axis. Aberrations are classified by their *order*, that is, the power of the angle that occurs in this relationship. Points on the optical axis may be quite well focused, while those in a corner of the image are smeared out. For this reason, only a limited portion of the image plane is usable. The magnitude of an aberration defect also increases as a power of the distance from the optical axis at which a ray passes through the lens. Thus the image quality can be improved by using only the central portion of a lens.

One reason for introducing diaphragms into a lens system is to improve image quality in a situation where it is not necessary to utilize fully the light-gathering power of the system. As already mentioned, fixed diaphragms ensure that rays entering at a large angle to the optical axis do not pass through the outer regions of any of the lenses. This improves image quality in the outer regions of the image, but at the same time greatly increases vignetting. In most common uses of lenses this is not an important matter, since people are astonishingly insensitive to smooth

spatial variations in image brightness. It does matter in machine vision, however, since we use the measurements of image brightness (irradiance) to determine the scene brightness (radiance).

2.4 Our Visual World

How can we hope to recover information about the three-dimensional world using a mere two-dimensional image? It may seem that the available information is not adequate, even if we take several images. Yet biological systems interact intelligently with the environment using visual information. The puzzle is solved when we consider the special nature of our usual visual world. We are immersed in a homogeneous transparent medium, and the objects we look at are typically opaque. Light rays are not refracted or absorbed in the environment, and we can follow a ray from an image point through the lens until it reaches some surface. The brightness at a point in the image depends only on the brightness of the corresponding surface patch. Surfaces are two-dimensional manifolds, and their shape can be represented by giving the distance $z(x', y')$ to the surface as a function of the image coordinates x' and y'.

This is to be contrasted with a situation in which we are looking into a volume occupied by a light-absorbing material of varying density. Here we may specify the density $\rho(x, y, z)$ of the material as a function of the coordinates x, y, and z. One or more images provide enough constraint to recover information about a surface, but not about a volume. In theory, an infinite number of images is needed to solve the problem of *tomography*, that is, to determine the density of the absorbing material.

Conditions of homogeneity and transparency may not always hold exactly. Distant mountains appear changed in color and contrast, while in deserts we may see mirages. Image analysis based on the assumption that conditions are as stated may go awry when the assumptions are violated, and so we can expect that both biological and machine vision systems will be misled in such situations. Indeed, some optical illusions can be explained in this way. This does not mean that we should abandon these additional constraints, for without them the solution of the problem of recovering information about the three-dimensional world from images would be ambiguous.

Our usual visual world is special indeed. Imagine being immersed instead in a world with varying concentrations of pigments dispersed within a gelatinous substance. It would not be possible to recover the distributions of these absorbing substances in three dimensions from one view. There

just would not be enough information. Analogously, single X-ray images are not useful unless there happens to be sharp contrast between different materials, like bone and tissue. Otherwise a very large number of views must be taken and a tomographic reconstruction attempted. It is perhaps a good thing that we do not possess Superman's X-ray vision capabilities!

By and large, we shall confine our attention to images formed by conventional optical means. We shall avoid high-magnification microscopic images, for instance, where many substances are effectively transparent, or at least translucent. Similarly, images on a very large scale often show the effects of absorption and refraction in the atmosphere. Interestingly, other modalities do sometimes provide us with images much like the ones we are used to. Examples include scanning electron microscopes (SEM) and synthetic-aperture radar systems (SAR), both of which produce images that are easy to interpret. So there is some hope of analyzing them using the methods discussed here.

In view of the importance of surfaces, we might hope that a machine vision system could be designed to recover the shapes of surfaces given one or more images. Indeed, there has been some success in this endeavor, as we shall see in chapter 10, where we discuss the recovery of shape from shading. Detailed understanding of the imaging process allows us to recover quantitative information from images. The computed shape of a surface may be used in recognition, inspection, or in planning the path of a mechanical manipulator.

2.5 Image Sensing

Almost all image sensors depend on the generation of electron–hole pairs when photons strike a suitable material. This is the basic process in biological vision as well as photography. Image sensors differ in how they measure the flux of charged particles. Some devices use an electric field in a vacuum to separate the electrons from the surface where they are liberated (figure 2-7a). In other devices the electrons are swept through a depleted zone in a semiconductor (figure 2-7b).

Not all incident photons generate an electron–hole pair. Some pass right through the sensing layer, some are reflected, and others lose energy in different ways. Further, not all electrons find their way into the detecting circuit. The ratio of the electron flux to the incident photon flux is called the *quantum efficiency*, denoted $q(\lambda)$. The quantum efficiency depends on the energy of the incident photon and hence on its wavelength λ. It also depends on the material and the method used to collect the liber-

ated electrons. Older vacuum devices tend to have coatings with relatively low quantum efficiency, while solid-state devices are near ideal for some wavelengths. Photographic film tends to have poor quantum efficiency.

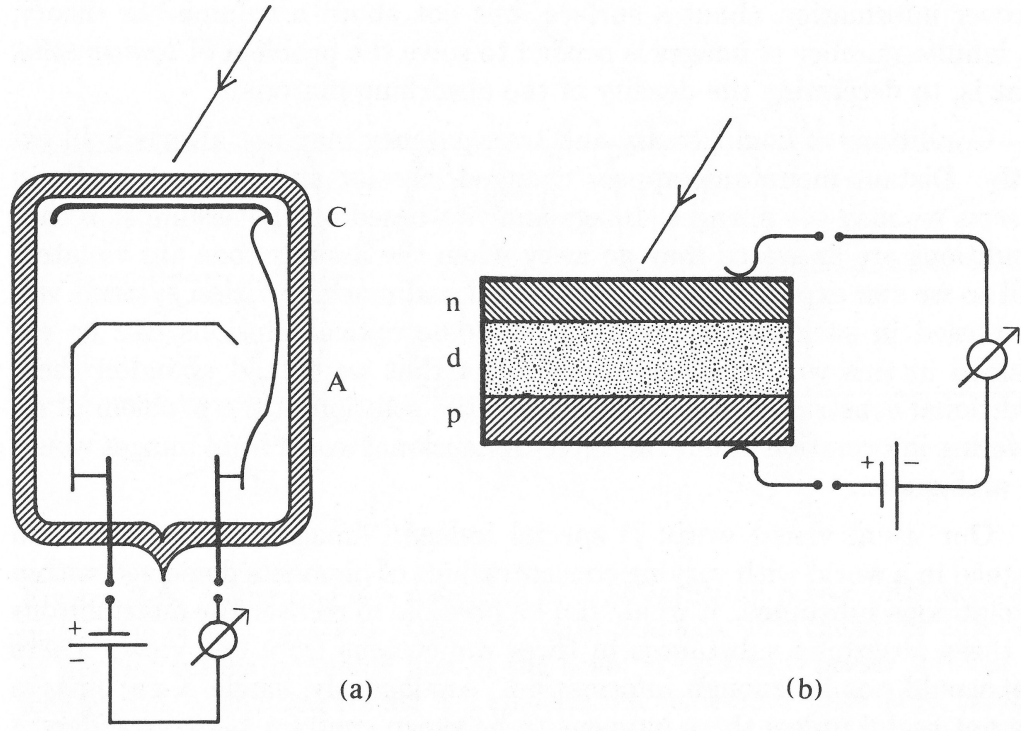


Figure 2-7. Photons striking a suitable surface generate charge carriers that are collected and measured to determine the irradiance. (a) In the case of a vacuum device, electrons are liberated from the photocathode and attracted to the positive anode. (b) In the case of a semiconductor device, electron–hole pairs are separated by the built-in field to be collected in an external circuit.

2.5.1 Sensing Color

The sensitivity of a device varies with the wavelength of the incident light. Photons with little energy tend to go right through the material, while very energetic photons may be stopped before they reach the sensitive layer. Each material has its characteristic variation of quantum efficiency with wavelength.

For a small wavelength interval $\delta\lambda$, let the flux of photons with energy equal to or greater than λ, but less than $\lambda + \delta\lambda$, be $b(\lambda)\,\delta\lambda$. Then the number of electrons liberated is

$$\int_{-\infty}^{\infty} b(\lambda)q(\lambda)\,d\lambda.$$

If we use sensors with different photosensitive materials, we obtain different images because their spectral sensitivities are different. This can be helpful in distinguishing surfaces that have similar gray-levels when imaged with one sensor, yet give rise to different gray-levels when imaged with a different sensor. Another way to achieve this effect is to use the same sensing material but place filters in front of the camera that selectively absorb different parts of the spectrum. If the transmission of the i^{th} filter is $f_i(\lambda)$, the effective quantum efficiency of the combination of that filter and the sensor is $f_i(\lambda)q(\lambda)$.

How many different filters should we use? The ability to distinguish among materials grows as more images are taken through more filters. The measurements are correlated, however, because most surfaces have a smooth variation of reflectance with wavelength. Typically, little is gained by using very many filters.

The human visual system uses three types of sensors, called *cones*, in daylight conditions. Each of these cone types has a particular spectral sensitivity, one of them peaking in the long wavelength range, one in the middle, and one in the short wavelength range of the visible spectrum, which extends from about 400 nm to about 700 nm. There is considerable overlap between the sensitivity curves. Machine vision systems often also use three images obtained through red, green, and blue filters. It should be pointed out, however, that the results have little to do with human color sensations unless the spectral response curves happen to be linear combinations of the human spectral response curves, as discussed below.

One property of a sensing system with a small number of sensor types having different spectral sensitivities is that many different spectral distributions will produce the same output. The reason is that we do not measure the spectral distributions themselves, but integrals of their product with the spectral sensitivity of particular sensor types. The same applies to biological systems, of course. Colors that appear indistinguishable to a human observer are said to be *metameric*. Useful information about the spectral sensitivities of the human visual system can be gained by systematically exploring metamers. The results of a large number of color-matching experiments performed by many observers have been averaged and used to calculate the so-called *tristimulus* or *standard observer curves*. These have been published by the *Commission Internationale de l'Eclairage* (CIE) and are shown in figure 2-8. A given spectral distribution is evaluated as follows: The spectral distribution is multiplied in turn by each of the three functions $x(\lambda)$, $y(\lambda)$, and $z(\lambda)$. The products are integrated over the visible

wavelength range. The three results $\overline{X}$, $\overline{Y}$, and $\overline{Z}$ are called the tristimulus values. Two spectral distributions that result in the same values for these three quantities appear indistinguishable when placed side by side under controlled conditions. (By the way, the spectral distributions used here are expressed in terms of energy per unit wavelength interval, not photon flux.)

The actual spectral response curves of the three types of cones cannot be determined in this way, however. There is some remaining ambiguity. It is known that the tristimulus curves are fixed linear transforms of these spectral response curves. The coefficients of the transformation are not known accurately.

We show in exercise 2-14 that a machine vision system with the same color-matching properties as the human color vision system must have sensitivities that are linear transforms of the human cone response curves. This in turn implies that the sensitivities must be linear transforms of the known standard observer curves. Unfortunately, this rule has rarely been observed when color-sensing systems were designed in the past. (Note that we are not addressing the problem of color sensations; we are only interested in having the machine confuse the same colors as the standard observer.)

2.5.2 Randomness and Noise

It is difficult to make accurate measurements of image brightness. In this section we discuss the corrupting influence of noise on image sensing. In order to do this, we need to discuss random variables and the probability density distribution. We shall also take the opportunity to introduce the concept of convolution in the one-dimensional case. Later, we shall encounter convolution again, applied to two-dimensional images. The reader familiar with these concepts may want to skip this section.

Measurements are affected by fluctuations in the signal being measured. If the measurement is repeated, somewhat differing results may be obtained. Typically, measurements will cluster around the "correct" value. We can talk of the probability that a measurement will fall within a certain interval. Roughly speaking, this is the limit of the ratio of the number of measurements that fall in that interval to the total number of trials, as the total number of trials tends to infinity. (This definition is not quite accurate, since any particular sequence of experiments may produce results that do not tend to the expected limit. It is unlikely that they are far off, however. Indeed, the probability of the limit tending to an answer that is not the desired one is zero.)

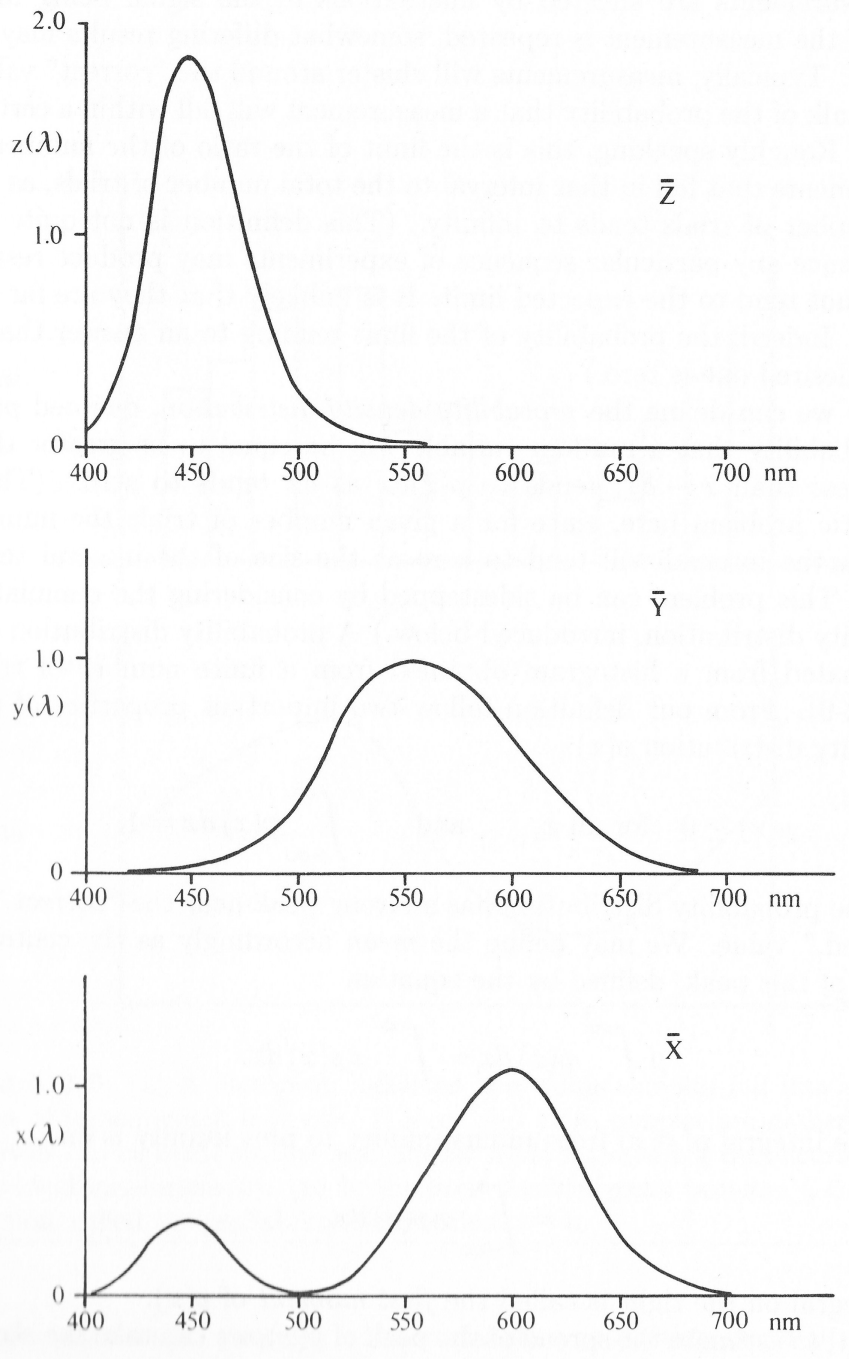


Figure 2-8. The tristimulus curves allow us to predict which spectral distributions will be indistinguishable. A given spectral distribution is multiplied by each of the functions $x(\lambda)$, $y(\lambda)$, and $z(\lambda)$, in turn, and the products integrated. In this way we obtain the tristimulus values, $\overline{X}$, $\overline{Y}$, and $\overline{Z}$, that can be used to characterize the spectral distribution. Spectral distributions that lead to the same tristimulus values appear the same when placed next to one another.

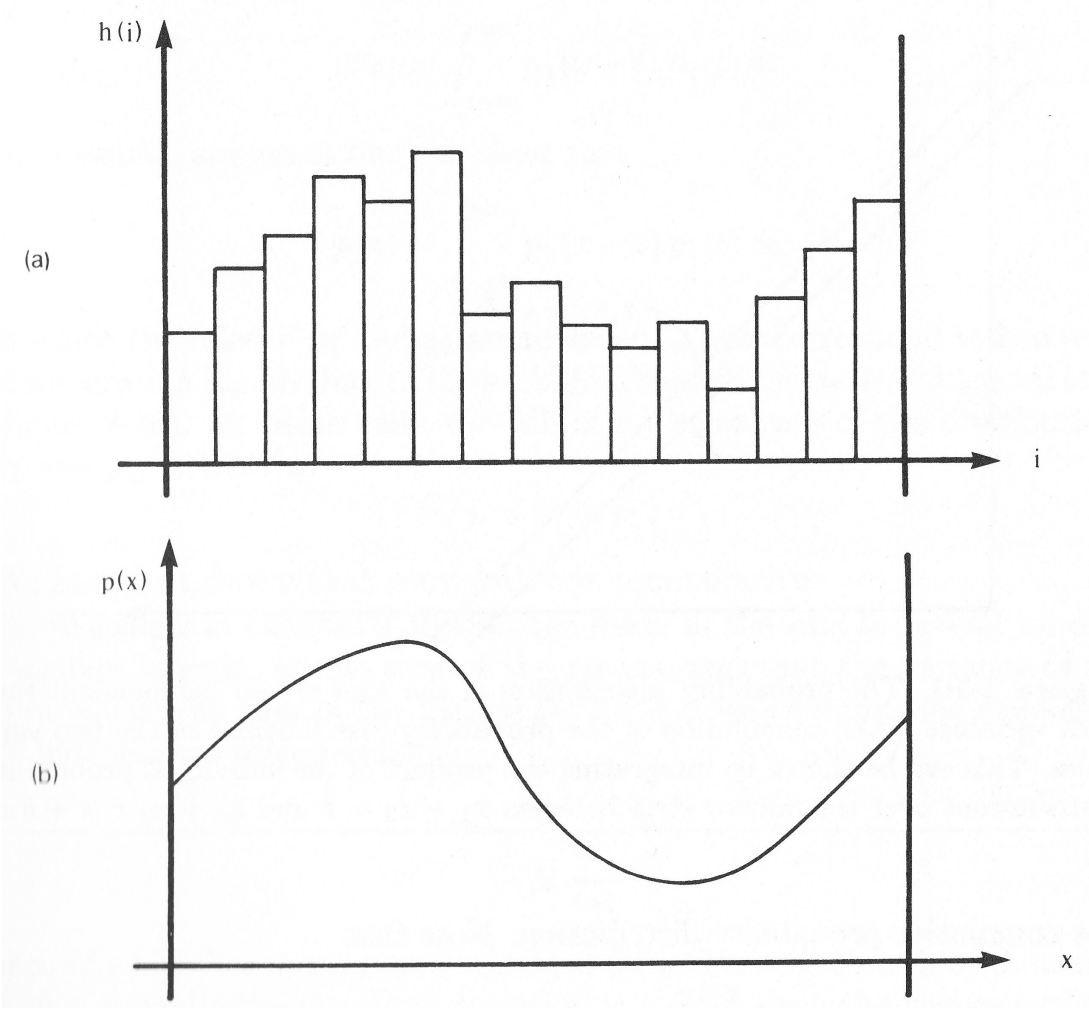


Figure 2-9. (a) A histogram indicates how many samples fall into each of a series of measurement intervals. If more and more samples are gathered, these intervals can be made smaller and smaller while maintaining the accuracy of the individual measurements. (b) In the limit the histogram becomes a continuous function, called the probability distribution.

Now we can define the *probability density distribution*, denoted $p(x)$. The probability that a random variable will be equal to or greater than x, but less than $x + \delta x$, tends to $p(x)\delta x$ as δx tends to zero. (There is a subtle problem here, since for a given number of trials the number falling in the interval will tend to zero as the size of the interval tends to zero. This problem can be sidestepped by considering the cumulative

probability distribution, introduced below.) A probability distribution can be estimated from a histogram obtained from a finite number of trials (figure 2-9). From our definition follow two important properties of any probability distribution $p(x)$:

$$p(x) \geq 0 \quad \text{for all } x, \qquad \text{and} \qquad \int_{-\infty}^{\infty} p(x)\,dx = 1.$$

Often the probability distribution has a strong peak near the "correct," or "expected," value. We may define the *mean* accordingly as the center of area, μ, of this peak, defined by the equation

$$\mu \int_{-\infty}^{\infty} p(x)\,dx = \int_{-\infty}^{\infty} x\,p(x)\,dx.$$

Since the integral of $p(x)$ from minus infinity to plus infinity is one,

$$\mu = \int_{-\infty}^{\infty} x\,p(x)\,dx.$$

The integral on the right is called the *first moment* of $p(x)$.

Next, to estimate the spread of the peak of $p(x)$, we can take the *second moment* about the mean, called the *variance*:

$$\sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2\,p(x)\,dx.$$

The square root of the variance, called the *standard deviation*, is a useful measure of the width of the distribution.

Another useful concept is the *cumulative probability distribution*,

$$P(x) = \int_{-\infty}^{x} p(t)\,dt,$$

which tells us the probability that the random variable will be less than or equal to x. The probability density distribution is just the derivative of the cumulative probability distribution. Note that

$$\lim_{x\to\infty} P(x) = 1.$$

One way to improve accuracy is to average several measurements, assuming that the "noise" in them will be independent and tend to cancel out. To understand how this works, we need to be able to compute the probability distribution of a sum of several random variables.

Suppose that x is a sum of two independent random variables x_1 and x_2 and that $p_1(x_1)$ and $p_2(x_2)$ are their probability distributions. How do we find $p(x)$, the probability distribution of $x = x_1 + x_2$? Given x_2, we know that x_1 must lie between $x - x_2$ and $x + \delta x - x_2$ in order for x to lie between x and $x + \delta x$ (figure 2-10). The probability that this will happen is $p_1(x - x_2)\,\delta x$. Now x_2 can take on a range of values, and the probability that it lies in a particular interval x_2 to $x_2 + \delta x_2$ is just $p_2(x_2)\,\delta x_2$. To find the probability that x lies between x and $x + \delta x$ we must integrate the product over all x_2. Thus

$$p(x)\,\delta x = \int_{-\infty}^{\infty} p_1(x - x_2)\,\delta x\, p_2(x_2)\,dx_2,$$

or

$$p(x) = \int_{-\infty}^{\infty} p_1(x - t)\, p_2(t)\,dt.$$

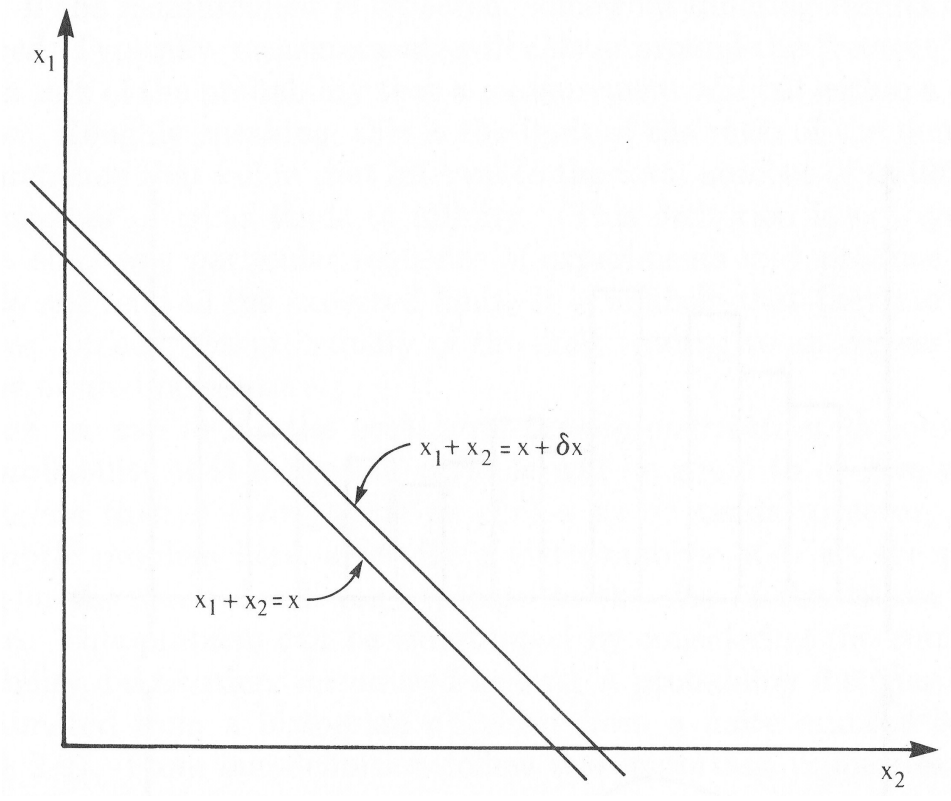


Figure 2-10. The probability distribution of the sum of two independent random variables is the convolution of the probability distributions of the two variables. This can be shown by integrating the product of the individual probability distributions over the narrow strip between $x_1 + x_2 = x$ and $x_1 + x_2 = x + \delta x$.

By a similar argument one can show that

$$p(x) = \int_{-\infty}^{\infty} p_2(x-t)\, p_1(t)\, dt,$$

in which the roles of x_1 and x_2 are reversed. These correspond to two ways of integrating the product of the probabilities over the narrow diagonal strip (figure 2-10). In either case, we talk of a *convolution* of the distributions p_1 and p_2, written as

$$p = p_1 \otimes p_2.$$

We have just shown that convolution is commutative.

We show in exercise 2-16 that the mean of the sum of several random variables is equal to the sum of the means, and that the variance of the sum equals the sum of the variances. Thus if we compute the average of N independent measurements,

$$\overline{x} = \frac{1}{N} \sum_{i=1}^{N} x_i,$$

each of which has mean μ and standard deviation σ, the mean of the result is also μ, while the standard deviation is $\sigma/\sqrt{N}$ since the variance of the sum is $N\sigma^2$. Thus we obtain a more accurate result, that is, one less affected by "noise." The relative accuracy only improves with the square root of the number of measurements, however.

A probability distribution that is of great practical interest is the *normal* or *Gaussian* distribution

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

with mean μ and standard deviation σ. The noise in many measurement processes can be modeled well using this distribution.

So far we have been dealing with random variables that can take on values in a continuous range. Analogous methods apply when the possible values are in a discrete set. Consider the electrons liberated during a fixed interval by photons falling on a suitable material. Each such event is independent of the others. It can be shown that the probability that exactly n are liberated in a time interval T is

$$P_n = e^{-m} \frac{m^n}{n!}$$

for some m. This is the *Poisson* distribution. We can calculate the average number liberated in time T as follows:

$$\sum_{n=1}^{\infty} ne^{-m}\frac{m^n}{n!} = me^{-m}\sum_{n=1}^{\infty}\frac{m^{n-1}}{(n-1)!}.$$

But

$$\sum_{n=1}^{\infty}\frac{m^{n-1}}{(n-1)!} = \sum_{n=0}^{\infty}\frac{m^n}{n!} = e^m,$$

so the average is just m. We show in exercise 2-18 that the variance is also m. The standard deviation is thus $\sqrt{m}$, so that the ratio of the standard deviation to the mean is $1/\sqrt{m}$. The measurement becomes more accurate the longer we wait, since more electrons are gathered. Again, the ratio of the "signal" to the "noise" only improves as the square root of the average number of electrons collected, however.

To obtain reasonable results, many electrons must be measured. It can be shown that a Poisson distribution with mean m is almost the same as a Gaussian distribution with mean m and variance m, provided that m is large. The Gaussian distribution is often easier to work with. In any case, to obtain a standard deviation that is one-thousandth of the mean, one must wait long enough to collect a million electrons. This is a small charge still, since one electron carries only

$$e = 1.602192\ldots \times 10^{-19} \text{ Coulomb}.$$

Even a million electrons have a charge of only about 160 fC (femto-Coulomb). (The prefix *femto-* denotes a multiplier of 10^{-15}.) It is not easy to measure such a small charge, since noise is introduced in the measurement process.

The number of electrons liberated from an area δA in time δt is

$$N = \delta A\, \delta t \int_{-\infty}^{\infty} b(\lambda)\, q(\lambda)\, d\lambda,$$

where $q(\lambda)$ is the quantum efficiency and $b(\lambda)$ is the image irradiance in photons per unit area. To obtain a usable result, then, electrons must be collected from a finite image area over a finite amount of time. There is thus a trade-off between (spatial and temporal) resolution and accuracy.

A measurement of the number of electrons liberated in a small area during a fixed time interval produces a result that is proportional to the irradiance (for fixed spectral distribution of incident photons). These measurements are quantized in order to read them into a digital computer.

This is done by analog-to-digital (A/D) conversion. The result is called a *gray-level*. Since it is difficult to measure irradiance with great accuracy, it is reasonable to use a small set of numbers to represent the irradiance levels. The range 0 to 255 is often employed—requiring just 8 bits per gray-level.

2.5.3 Quantization of the Image

Because we can only transmit a finite number of measurements to a computer, spatial quantization is also required. It is common to make measurements at the nodes of a square raster or grid of points. The image is then represented as a rectangular array of integers. To obtain a reasonable amount of detail we need many measurements. Television frames, for example, might be quantized into 450 lines of 560 *picture cells*, sometimes referred to as *pixels*.

Each number represents the average irradiance over a small area. We cannot obtain a measurement at a point, as discussed above, because the flux of light is proportional to the sensing area. At first this might appear as a shortcoming, but it turns out to be an advantage. The reason is that we are trying to use a discrete set of numbers to represent a continuous distribution of brightness, and the sampling theorem tells us that this can be done successfully only if the continuous distribution is smooth, that is, if it does not contain high-frequency components. One way to make a smooth distribution of brightness is to look at the image through a filter that averages over small areas.

What is the optimal size of the sampling areas? It turns out that reasonable results are obtained if the dimensions of the sampling areas are approximately equal to their spacing. This is fortunate because it allows us to pack the image plane efficiently with sensing elements. Thus no photons need be wasted, nor must adjacent sampling areas overlap.

We have some latitude in dividing up the image plane into sensing areas. So far we have been discussing square areas on a square grid. The picture cells could equally well be rectangular, resulting in a different resolution in the horizontal and vertical directions. Other arrangements are also possible. Suppose we want to tile the plane with regular polygons. The tiles should not overlap, yet together they should cover the whole plane. We shall show in exercise 2-21 that there are exactly three tessellations, based on triangles, squares, and hexagons (figure 2-11).

It is easy to see how a square sampling pattern is obtained simply by taking measurements at equal intervals along equally spaced lines in the

image. Hexagonal sampling is almost as easy, if odd-numbered lines are offset by half a sampling interval from even-numbered lines. In television scanning, the odd-numbered lines are read out after all the even-numbered lines because of field interlace, and so this scheme is particularly easy to implement. Hexagons on a triangular grid have certain advantages, which we shall come to later.

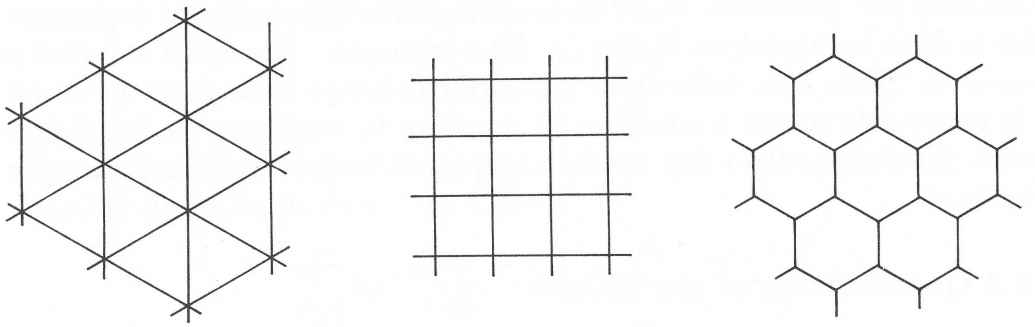


Figure 2-11. The plane can be tiled with three regular polygons: the triangle, the square, and the hexagon. Image tessellations can be based on these tilings. The gray-level of a picture cell is the quantized value of the measured power falling on the corresponding area in the image.

2.6 References

There are many standard references on basic optics, including *Principles of Optics: Electromagnetic Theory of Propagation, Interference and Diffraction of Light* by Born & Wolf [1975], *Handbook of Optics*, edited by Driscoll & Vaughan [1978], *Applied Optics: A Guide to Optical System Design* by Levi [volume 1, 1968; volume 2, 1980], and the classic *Optics* by Sears [1949]. Lens design and aberrations are covered by Kingslake in *Lens Design Fundamentals* [1978]. Norton discusses the basic workings of a large variety of sensors in *Sensor and Analyzer Handbook* [1982]. Barbe edited *Charge-Coupled Devices* [1980], a book that includes some information on the use of CCDs in image sensors.

There is no shortage of books on probability and statistics. One such is Drake's *Fundamentals of Applied Probability Theory* [1967].

Color vision is not treated in detail here, but is mentioned again in chapter 9 where we discuss the recovery of lightness. For a general discussion of color matching and tristimulus values see the first few chapters of *Color in Business, Science, and Industry* by Judd & Wyszeck [1975].

Some issues of color reproduction, including what constitutes an appropriate sensor system, are discussed by Horn [1984a]. Further references on color vision may be found at the end of chapter 9.

Straight lines in the three-dimensional world are projected as straight lines into the two-dimensional image. The projections of parallel lines intersect in a *vanishing point*. This is the point where a line parallel to the given lines passing through the center of projection intersects the image plane. In the case of rectangular objects, a great deal of information can be recovered from lines in the images and their intersections. See, for example, Barnard [1983].

When the medium between us and the scene being imaged is not perfectly transparent, the interpretation of images becomes more complicated. See, for example, Sjoberg & Horn [1983]. The reconstruction of absorbing density in a volume from measured ray attenuation is the subject of tomography; a book on this subject has been edited by Herman [1979].

2.7 Exercises

2-1 What is the shape of the image of a sphere? What is the shape of the image of a circular disk? Assume perspective projection and allow the disk to lie in a plane that can be tilted with respect to the image plane.

2-2 Show that the image of an ellipse in a plane, not necessarily one parallel to the image plane, is also an ellipse. Show that the image of a line in space is a line in the image. Assume perspective projection. Describe the brightness patterns in the image of a polyhedral object with uniform surface properties.

2-3 Suppose that an image is created by a camera in a certain world. Now imagine the same camera placed in a similar world in which everything is twice as large and all distances between objects have also doubled. Compare the new image with the one formed in the original world. Assume perspective projection.

2-4 Suppose that an image is created by a camera in a certain world. Now imagine the same camera placed in a similar world in which everything has half the reflectance and the incident light has been doubled. Compare the new image with the one formed in the original world. Hint: Ignore interflections, that is, illumination of one part of the scene by light reflected from another.

2-5 Show that in a properly focused imaging system the distance f' from the lens to the image plane equals $(1+m)f$, where f is the focal length and m is the

magnification. This distance is called the *effective focal length*. Show that the distance between the image plane and an object must be

$$\left(m + 2 + \frac{1}{m}\right) f.$$

How far must the object be from the lens for unit magnification?

2-6 What is the focal length of a compound lens obtained by placing two thin lenses of focal length f_1 and f_2 against one another? Hint: Explain why an object at a distance f_1 on one side of the compound lens will be focused at a distance f_2 on the other side.

2-7 The *f-number* of a lens is the ratio of the focal length to the diameter of the lens. The f-number of a given lens (of fixed focal length) can be increased by introducing an aperture that intercepts some of the light and thus in effect reduces the diameter of the lens. Show that image brightness will be inversely proportional to the square of the f-number. Hint: Consider how much light is

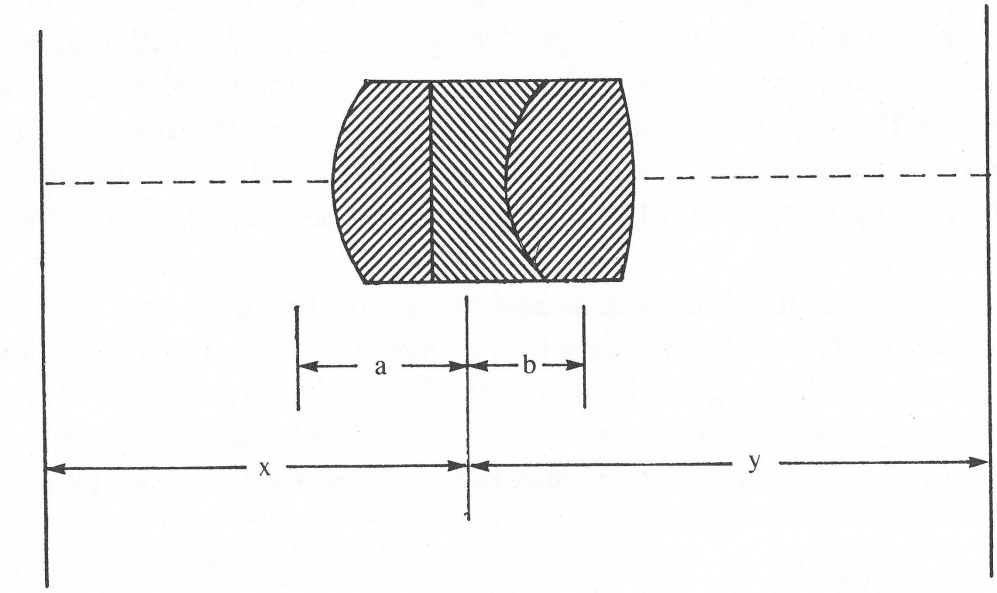


Figure 2-12. To determine the focal length and the positions of the principal planes, a number of measurements are made. Here, an object lying in a plane a distance x from an arbitrary reference point on one side of the lens is properly in focus in a plane on the other side at a distance y from the reference point. The two principal planes lie at distances a and b on either side of the reference point.

2-8 When a camera is used to obtain metric information about the world, it is important to have accurate knowledge of the parameters of the lens, including the focal length and the positions of the principal planes. Suppose that a pattern in a plane at distance x on one side of the lens is found to be focused best on a

plane at a distance y on the other side of the lens (figure 2-13). The distances x and y are measured from an arbitrary but fixed point in the lens. How many paired measurements like this are required to determine the focal length and the position of the two principal planes? (In practice, of course, more than the minimum required number of measurements would be taken, and a least-squares procedure would be adopted. Least-squares methods are discussed in the appendix.)

Suppose that the arbitrary reference point happens to lie between the two principal planes and that a and b are the distances of the principal planes from the reference point (figure 2-7). Note that $a + b$ is the thickness of the lens, as defined earlier. Show that

$$(ab + bf + fa) - \big(x_i(f + b) + y_i(f + a)\big) + x_i y_i = 0,$$

where x_i and y_i are the measurements obtained in the i^{th} experiment. Suggest a way to find the unknowns from a set of nonlinear equations like this. Can a closed-form solution be obtained for f, a, b?

2-9 Here we explore a restricted case of the problem tackled in the previous exercise. Describe a method for determining the focal length and positions of the principal planes of a lens from the following three measurements: (a) the position of a plane on which a scene at infinity on one side of the lens appears in sharp focus; (b) the position of a plane on which a scene at infinity on the other side of the lens appears in sharp focus; (c) the positions of two planes, one on each side of the lens, such that one plane is imaged at unit magnification on the other.

2-10 Here we explore what happens when the image plane is tilted slightly. Show that in a pinhole camera, tilting the image plane amounts to nothing more than changing the place where the optical axis pierces the image plane and changing the perpendicular distance of the projection center from the image plane. What happens in a camera that uses a lens? Hint: Is a camera with an (ideal) lens different from a camera with a pinhole as far as image projection is concerned?

How would you determine experimentally where the optical axis pierces the image plane? Hint: It is difficult to find this point accurately.

2-11 It has been stated that perspective effects are significant when a wide-angle lens is used, while images obtained using a telephoto lenses tend to approximate orthographic projection. Explain why these are only rough rules of thumb.

2-12 Straight lines in the three-dimensional world are projected as straight lines into the two-dimensional image. The projections of parallel lines intersect

in a *vanishing point*. Where in the image will the vanishing point of a particular family of parallel lines lie? When does the vanishing point of a family of parallel lines lie at infinity?

In the case of a rectangular object, a great deal of information can be recovered from lines in the images and their intersections. The edges of a rectangular solid fall into three sets of parallel lines, and so give rise to three vanishing points. In technical drawing one speaks of one-point, two-point, and three-point perspective. These terms apply to the cases in which two, one, or none of three vanishing points lie at infinity. What alignment between the edges of the rectangular object and the image plane applies in each case?

2-13 Typically, imaging systems are almost exactly rotationally symmetric about the optical axis. Thus distortions in the image plane are primarily radial. When very high precision is required, a lens can be calibrated to determine its radial distortion. Commonly, a polynomial of the form

$$\Delta r' = k_1(r') + k_3(r')^3 + k_5(r')^5 + \cdots$$

is fitted to the experimental data. Here $r' = \sqrt{x'^2 + y'^2}$ is the distance of a point in the image from the place where the optical axis pierces the image plane. Explain why no even powers of r' appear in the polynomial.

2-14 Suppose that a color-sensing system has three types of sensors and that the spectral sensitivity of each type is a sum of scaled versions of the human cone sensitivities. Show that two metameric colors will produce identical signals in the sensors.

Now show that a color-sensing system will have this property for all metamers only if the spectral sensitivity of each of its three sensor types is a sum of scaled versions of the human cone sensitivities. Warning: The second part of this problem is much harder than the first.

2-15 Show that the variance can be calculated as

$$\sigma^2 = \int_{-\infty}^{\infty} x^2 p(x)\, dx - \mu^2.$$

2-16 Here we consider the mean and standard deviation of the sum of two random variables.

(a) Show that the mean of $x = x_1 + x_2$ is the sum $\mu_1 + \mu_2$ of the means of the independent random variables x_1 and x_2.

(b) Show that the variance of $x = x_1 + x_2$ is the sum $\sigma_1^2 + \sigma_2^2$ of the variances of the independent random variables x_1 and x_2.

2-17 Suppose that the probability distribution of a random variable is

$$p(x) = \begin{cases} (1/2w), & \text{if } |x| \leq w; \\ 0, & \text{if } |x| > w. \end{cases}$$

What is the probability distribution of the average of two independent values from this distribution?

2-18 Here we consider some properties of the Gaussian and the Poisson distributions.

(a) Show that the mean and variance of the Gaussian distribution

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

are μ and σ^2 respectively.

(b) Show that the mean and the variance of the Poisson distribution

$$p_n = e^{-m}\frac{m^n}{n!}$$

are both equal to m.

2-19 Consider the weighted sum of independent random variables

$$\sum_{i=1}^{N} w_i x_i,$$

where x_i has mean m and standard deviation σ. Assume that the weights w_i add up to one. What are the mean and standard deviation of the weighted sum? For fixed N, what choice of weights minimizes the variance?

2-20 A television frame is scanned in 1/30 second. All the even-numbered lines in one field are followed by all the odd-numbered lines in the other field. Assume that there are about 450 lines of interest, each to be divided into 560 picture cells. At what rate must the conversion from analog to digital form occur? (Ignore time intervals between lines and between successive frames.)

2-21 Show that there are only three regular polygons with which the plane can be tiled, namely (a) the equilateral triangle, (b) the square, and (c) the hexagon. (By *tiling* we mean covering without gaps or overlap.)