2

Digital Snapshots

Verweile doch! Du bist so schön![1]
Goethe, *Faust*

This chapter deals with digital images and their relation to the physical world. We learn the principles of image formation, define the two main types of images in this book (*intensity* and *range images*), and discuss how to acquire and store them in a computer.

Chapter Overview

Section 2.2 considers the basic optical, radiometric, and geometric principles underlying the formation of intensity images.

Section 2.3 brings the computer into the picture, laying out the special nature of digital images, their acquisition, and some mathematical models of intensity cameras.

Section 2.4 discusses the fundamental mathematical models of intensity cameras and their parameters.

Section 2.5 introduces range images and describes a class of range sensors based on intensity cameras, so that we can use what we learn about intensity imaging.

What You Need to Know to Understand this Chapter

- Sampling theorem (Appendix, section A.3).
- Rotation matrices (Appendix, section A.9).

[1] Stop! You are so beautiful!

2.1 Introduction

This chapter deals with the main ingredients of computer vision: *digital images*. We concentrate on two types of images frequently used in computer vision:

intensity images, the familiar, photographlike images encoding light intensities, acquired by television cameras;

range images, encoding shape and distance, acquired by special sensors like sonars or laser scanners.

Intensity images measure the amount of light impinging on a photosensitive device; range images estimate directly the 3-D structure of the viewed scene through a variety of techniques. Throughout the book, we will develop algorithms for both types of images.[2]

It is important to stress immediately that *any digital image, irrespective of its type, is a 2-D array (matrix) of numbers*. Figure 2.1 illustrates this fact for the case of intensity images. Depending on the nature of the image, the numbers may represent light intensities, distances, or other physical quantities. This fact has two fundamental consequences.

- The exact relationship of a digital image to the physical world (i.e., its nature of range or intensity image) is determined by the acquisition process, which depends on the sensor used.
- Any information contained in images (e.g., shape, measurements, or object identity) must ultimately be extracted (computed) from 2-D numerical arrays, in which it is encoded.

In this chapter, we investigate the origin of the numbers forming a digital image; the rest of the book is devoted to computational techniques that make *explicit* some of the information contained *implicitly* in these numbers.

2.2 Intensity Images

We start by introducing the main concepts behind intensity image formation.

2.2.1 Main Concepts

In the visual systems of many animals, including man, the process of image formation begins with the light rays coming from the outside world and impinging on the photoreceptors in the retina. A simple look at any ordinary photograph suggests the variety of physical parameters playing a role in image formation. Here is an incomplete list:

Optical parameters of the lens characterize the sensor's optics. They include:

- lens type,
- focal length,

[2] Although, as we shall see, some algorithms make sense for intensity images only.

117	125	133	127	130	130	133	121	116	115	100	91	93	94	99	103	112	105	109	106
134	133	138	138	132	134	130	133	128	123	121	113	106	102	99	106	113	109	109	113
146	147	138	140	125	134	124	115	102	96	93	94	99	96	99	100	103	110	109	110
144	141	136	130	120	108	88	74	53	37	31	37	35	39	53	79	93	100	109	116
139	136	129	119	102	85	58	31	41	77	51	53	53	33	37	41	69	94	105	108
132	127	117	102	87	57	49	77	42	28	17	15	13	13	17	41	53	69	88	100
124	120	108	94	72	74	72	31	35	31	15	13	15	11	15	13	46	75	83	96
125	115	102	93	88	82	42	79	113	41	19	100	82	11	11	17	31	91	99	100
124	116	109	99	91	113	99	140	144	57	20	20	15	11	15	17	63	87	119	124
136	133	133	135	138	133	132	144	150	120	24	17	15	15	17	20	115	113	88	150
158	157	157	154	149	145	133	127	146	150	116	35	20	19	28	105	124	128	141	171
155	154	156	155	146	155	154	154	147	139	148	150	138	120	128	129	130	151	156	165
150	151	154	162	166	167	169	174	172	167	177	166	164	140	134	120	121	120	127	172
145	149	151	157	165	169	173	179	176	166	166	157	145	136	129	124	120	136	163	168
144	148	153	160	159	158	165	172	165	169	157	151	149	141	130	140	151	162	169	167
144	141	147	155	154	149	156	151	157	157	151	144	147	147	149	159	158	159	166	165
139	140	140	150	153	151	150	146	140	139	138	140	145	151	149	156	156	162	162	161
136	134	138	146	156	164	153	146	145	136	139	139	140	141	149	157	159	161	169	166
136	133	136	135	144	159	168	159	151	142	141	145	139	146	153	156	164	167	172	168
133	129	140	142	146	159	167	165	154	151	146	141	147	154	156	160	161	157	153	154

Figure 2.1 Digital images are 2-D arrays of numbers: a 20 × 20 grey-level image of an eye (pixels have been enlarged for display) and the corresponding 2-D array.

- field of view,
- angular apertures.

Photometric parameters appear in models of the light energy reaching the sensor after being reflected from the objects in the scene. They include:

- type, intensity, and direction of illumination,
- reflectance properties of the viewed surfaces,
- effects of the sensor's structure on the amount of light reaching the photoreceptors.

Geometric parameters determine the image position on which a 3-D point is projected. They include:

- type of projections,
- position and orientation of camera in space,
- perspective distortions introduced by the imaging process.

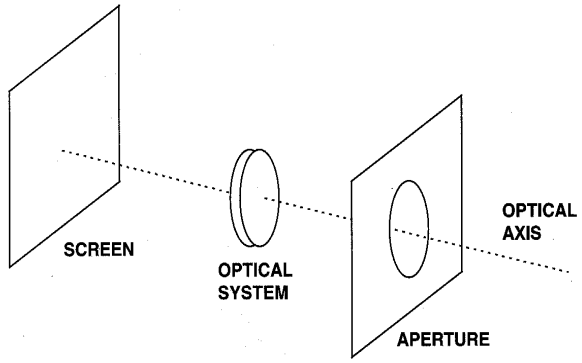

Figure 2.2 The basic elements of an imaging device.

All the above plays a role in *any* intensity imaging device, be it a photographic camera, camcorder, or computer-based system. However, further parameters are needed to characterize digital images and their acquisition systems. These include:

- the physical properties of the photosensitive matrix of the viewing camera,
- the discrete nature of the photoreceptors,
- the quantization of the intensity scale.

We will now review the optical, radiometric, and geometric aspects of image formation.

2.2.2 Basic Optics

We first need to establish a few fundamental notions of optics. As for many natural visual systems, the process of image formation in computer vision begins with the light rays which enter the camera through an *angular aperture* (or *pupil*), and hit a screen or *image plane* (Figure 2.2), the camera's photosensitive device which registers light intensities. Notice that most of these rays are the result of the reflections of the rays emitted by the light sources and hitting object surfaces.

Image Focusing. Any *single* point of a scene reflects light coming from possibly many directions, so that many rays reflected by the same point may enter the camera. In order to obtain sharp images, all rays coming from a single scene point, P, must converge onto a single point on the image plane, p, the *image of* P. If this happen, we say that the image of P is *in focus*; if not, the image is spread over a circle. *Focusing* all rays from a scene point onto a single image point can be achieved in two ways:

1. Reducing the camera's aperture to a point, called a *pinhole*. This means that only one ray from any given point can enter the camera, and creates a one-to-one correspondence between visible points, rays, and image points. This results in very sharp, undistorted images of objects at different distances from the camera (see Project 2.1).
2. Introducing an *optical system* composed of lenses, apertures, and other elements, explicitly designed to make all rays coming from the same 3-D point converge onto a single image point.

An obvious disadvantage of a pinhole aperture is its *exposure time*; that is, how long the image plane is allowed to receive light. Any photosensitive device (camera film, electronic sensors) needs a minimum amount of light to register a legible image. As a pinhole allows very little light into the camera per time unit, the exposure time necessary to form the image is too long (typically several seconds) to be of practical use.[3] Optical systems, instead, can be adjusted to work under a wide range of illumination conditions and exposure times (the exposure time being controlled by a *shutter*).

☞ Intuitively, an optical system can be regarded as a device that aims at producing the same image obtained by a pinhole aperture, but by means of a much larger aperture and a shorter exposure time. Moreover, an optical system enhances the light gathering power.

Thin Lenses. Standard optical systems are quite sophisticated, but we can learn the basic ideas from the simplest optical system, the *thin lens*. The optical behavior of a thin lens (Figure 2.3) is characterized by two elements: an axis, called *optical axis*, going through the lens center, O, and perpendicular to the plane; and two special points, F_l and F_r, called *left* and *right focus*, placed on the optical axis, on the opposite sides of the lens, and at the same distance from O. This distance, called the *focal length* of the lens, is usually indicated by f.

By construction, a thin lens deflects all rays parallel to the optical axis and coming from one side onto the focus on the other side, as described by two *basic properties*.

Thin Lens: Basic Properties

1. Any ray entering the lens parallel to the axis on one side goes through the focus on the other side.
2. Any ray entering the lens from the focus on one side emerges parallel to the axis on the other side.

The Fundamental Equation of Thin Lenses. Our next task is to derive the *fundamental equation of thin lenses* from the basic properties 1 and 2. Consider a point P, not too far from the optical axis, and let $Z + f$ be the distance of P from the lens along the optical axis (Figure 2.4). By assumption, a thin lens focuses all the rays from P onto the same point, the image point p. Therefore, we can locate p by intersecting only two known rays, and we do not have to worry about tracing the path of any other.

[3] The exposure time is, roughly, inversely proportional to the square of the aperture diameter, which in turn is proportional to the amount of light that enters the imaging system.

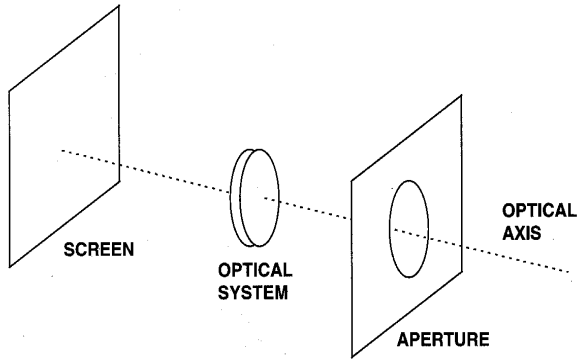

Figure 2.2 The basic elements of an imaging device.

All the above plays a role in *any* intensity imaging device, be it a photographic camera, camcorder, or computer-based system. However, further parameters are needed to characterize digital images and their acquisition systems. These include:

- the physical properties of the photosensitive matrix of the viewing camera,
- the discrete nature of the photoreceptors,
- the quantization of the intensity scale.

We will now review the optical, radiometric, and geometric aspects of image formation.

2.2.2 Basic Optics

We first need to establish a few fundamental notions of optics. As for many natural visual systems, the process of image formation in computer vision begins with the light rays which enter the camera through an *angular aperture* (or *pupil*), and hit a screen or *image plane* (Figure 2.2), the camera's photosensitive device which registers light intensities. Notice that most of these rays are the result of the reflections of the rays emitted by the light sources and hitting object surfaces.

Image Focusing. Any *single* point of a scene reflects light coming from possibly many directions, so that many rays reflected by the same point may enter the camera. In order to obtain sharp images, all rays coming from a single scene point, P, must converge onto a single point on the image plane, p, the *image of* P. If this happen, we say that the image of P is *in focus*; if not, the image is spread over a circle. *Focusing* all rays from a scene point onto a single image point can be achieved in two ways:

1. Reducing the camera's aperture to a point, called a *pinhole*. This means that only one ray from any given point can enter the camera, and creates a one-to-one correspondence between visible points, rays, and image points. This results in very sharp, undistorted images of objects at different distances from the camera (see Project 2.1).
2. Introducing an *optical system* composed of lenses, apertures, and other elements, explicitly designed to make all rays coming from the same 3-D point converge onto a single image point.

An obvious disadvantage of a pinhole aperture is its *exposure time*; that is, how long the image plane is allowed to receive light. Any photosensitive device (camera film, electronic sensors) needs a minimum amount of light to register a legible image. As a pinhole allows very little light into the camera per time unit, the exposure time necessary to form the image is too long (typically several seconds) to be of practical use.[3] Optical systems, instead, can be adjusted to work under a wide range of illumination conditions and exposure times (the exposure time being controlled by a *shutter*).

☞ Intuitively, an optical system can be regarded as a device that aims at producing the same image obtained by a pinhole aperture, but by means of a much larger aperture and a shorter exposure time. Moreover, an optical system enhances the light gathering power.

Thin Lenses. Standard optical systems are quite sophisticated, but we can learn the basic ideas from the simplest optical system, the *thin lens*. The optical behavior of a thin lens (Figure 2.3) is characterized by two elements: an axis, called *optical axis*, going through the lens center, O, and perpendicular to the plane; and two special points, F_l and F_r, called *left* and *right focus*, placed on the optical axis, on the opposite sides of the lens, and at the same distance from O. This distance, called the *focal length* of the lens, is usually indicated by f.

By construction, a thin lens deflects all rays parallel to the optical axis and coming from one side onto the focus on the other side, as described by two *basic properties*.

Thin Lens: Basic Properties

1. Any ray entering the lens parallel to the axis on one side goes through the focus on the other side.
2. Any ray entering the lens from the focus on one side emerges parallel to the axis on the other side.

The Fundamental Equation of Thin Lenses. Our next task is to derive the *fundamental equation of thin lenses* from the basic properties 1 and 2. Consider a point P, not too far from the optical axis, and let $Z + f$ be the distance of P from the lens along the optical axis (Figure 2.4). By assumption, a thin lens focuses all the rays from P onto the same point, the image point p. Therefore, we can locate p by intersecting only two known rays, and we do not have to worry about tracing the path of any other.

[3] The exposure time is, roughly, inversely proportional to the square of the aperture diameter, which in turn is proportional to the amount of light that enters the imaging system.

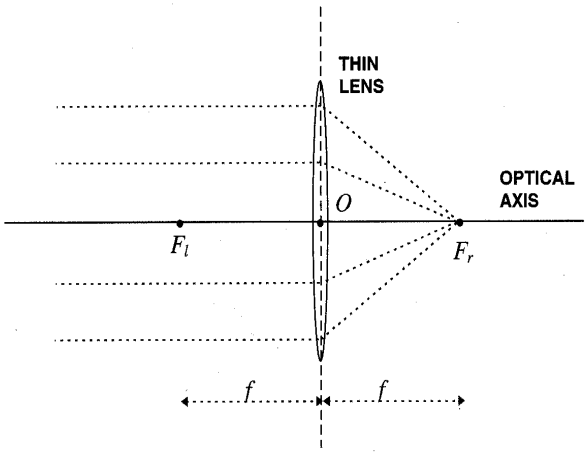

Figure 2.3 Geometric optics of a thin lens (a perpendicular view to the plane approximating the lens).

Note that by applying property 1 to the ray PQ and property 2 to the ray PR, PQ and PR are deflected to intersect at a certain point on the other side of the thin lens. But since the lens focuses all rays coming from P onto the same point, PQ and PR must intersect at p! From Figure 2.4 and using the two pairs of similar triangles $< PF_lS >$ and $< ROF_l >$ and $< psF_r >$ and $< QOF_r >$, we obtain immediately

$$Zz = f^2. \tag{2.1}$$

Setting $\hat{Z} = Z + f$ and $\hat{z} = z + f$, (2.1) reduces to our target equation.

The Fundamental Equation of Thin Lenses

$$\frac{1}{\hat{Z}} + \frac{1}{\hat{z}} = \frac{1}{f}. \tag{2.2}$$

☞ The ray going through the lens center, O, named the *principal ray*, goes through p undeflected.

Field of View. One last observation about optics. Let d be the *effective diameter of the lens*, identifying the portion of the lens actually reachable by light rays.

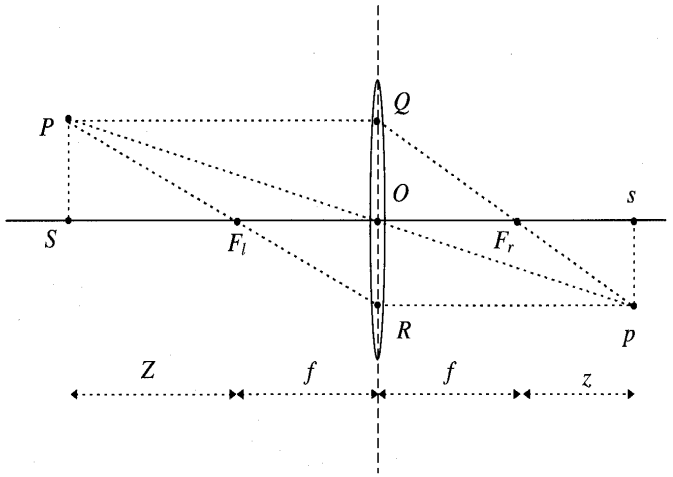

Figure 2.4 Imaging by a thin lens. Notice that, in general, a real lens has two different focal lengths, because the curvatures of its two surfaces may be different. The situation depicted here is a special case, but it is sufficient for our purposes. See the Further Readings at the end of this chapter for more on optics.

☞ We call d the *effective* diameter to emphasize the difference between d and the *physical* diameter of the lens. The aperture may prevent light rays from reaching the peripheral points of the lens, so that d is usually smaller than the physical diameter of the lens.

The effective lens diameter and the focal length determine the *field of view* of the lens, which is *an angular measure of the portion of 3-D space actually seen by the camera.* It is customary to define the field of view, w, as half of the angle subtended by the lens diameter as seen from the focus:

$$\tan w = \frac{d}{2f}. \tag{2.3}$$

This is the minimum amount of optics needed for our purposes. Optical models of real imaging devices are a great deal more complicated than our treatment of thin (and ideal) lenses; problems and phenomena not considered here include *spherical aberration* (defocusing of nonparaxial rays), *chromatic aberration* (different defocusing of rays of different colors), and focusing objects at different distances from the camera.[4]

[4] The fundamental equation of thin lenses implies that scene points at different distances from the lens come in focus at different image distances. The optical lens systems of real cameras are designed so that all points within a given range of distances are imaged on or close to the image plane, and therefore acceptably in focus. This range is called the *depth of field* of the camera.

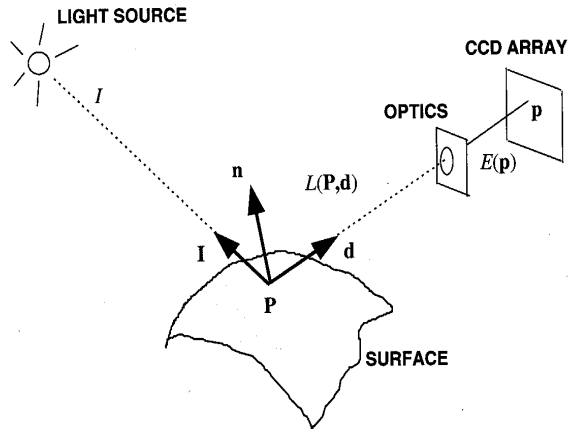

Figure 2.5 Illustration of the basic radiometric concepts.

The Further Readings section at the end of this chapter tell where to find more about optics.

2.2.3 Basic Radiometry

Radiometry is the essential part of image formation concerned with the relation among the amounts of light energy emitted from light sources, reflected from surfaces, and registered by sensors. We shall use radiometric concepts to pursue two objectives:

1. modelling how much of the illuminating light is reflected by object surfaces;
2. modelling how much of the reflected light actually reaches the image plane of the camera.

Definitions. We begin with some definitions, illustrated in Figure 2.5 and summarized as follows:

Image Irradiance and Scene Radiance

The *image irradiance* is the power of the light, per unit area and at each point **p** of the image plane.

The *scene radiance* is the power of the light, per unit area, ideally emitted by each point **P** of a surface in 3-D space in a given direction **d**.

☞ *Ideally* refers to the fact that the surface in the definition of scene radiance might be the illuminated surface of an object, the radiating surface of a light source, or even a fictitious surface. The term *scene radiance* denotes the total radiance emitted by a point; sometimes *radiance* refers to the energy radiated from a surface (emitted or reflected), whereas *irradiance* refers to the energy incident on a surface.

Surface Reflectance and Lambertian Model. A model of the way in which a surface reflects incident light is called a *surface reflectance model*. A well-known one is the *Lambertian model*, which assumes that each surface point appears equally bright from all viewing directions. This approximates well the behavior of rough, nonspecular surfaces, as well as various materials like matte paint and paper. If we represent the direction and amount of incident light by a vector **I**, the scene radiance of an ideal Lambertian surface, L, is simply proportional to the dot product between **I** and the unit normal to the surface, **n**:

$$L = \rho \mathbf{I}^\top \mathbf{n} \tag{2.4}$$

with $\rho > 0$ a constant called the surface's *albedo*, which is typical of the surface's material. We also assume that $\mathbf{I}^\top \mathbf{n}$ is *positive*; that is, the surface faces the light source. This is a necessary condition for the ray of light to reach **P**. If this condition is not met, the scene radiance should be set equal to 0.

We will use the Lambertian model in several parts of this book; for example, while analyzing image sequences (Chapter 8) and computing shape from shading (Chapter 9). Intuitively, the Lambertian model is based on the exact cancellation of two factors. Neglecting constant terms, the amount of light reaching *any* surface is always proportional to the cosine of the angle between the illuminant and the surface normal **n** (that is, the effective area of the surface as seen from the illuminant direction). According to the model, a Lambertian surface reflects light in a given direction **d** proportionally to the cosine of θ, the angle between **d** and **n**. But since the surface's area seen from the direction **d** is inversely proportional to $\cos\theta$, the two $\cos\theta$ factors cancel out and do not appear in (2.4).

Linking Surface Radiance and Image Irradiance. Our next task is to link the amounts of light reflected by the surfaces, L, and registered by the imaging sensor, E.

Assumptions and Problem Statement

Given a thin lens of diameter d and focal length f, an object at distance Z from the lens, and an image plane at distance Z' from the lens, with f, Z, and Z' as in (2.1), find the relation between image irradiance and scene radiance.

In order to derive this fundamental relation, we need to recall the geometric notion of *solid angle*. The solid angle of a cone of directions is the area cut out by the cone on the unit sphere centered in the cone's vertex. Therefore, the solid angle $\delta\omega$ subtended by a small, planar patch of area δA at distance r from the origin (Figure 2.6) is

$$\delta\omega = \frac{\delta A \cos\psi}{r^2} \tag{2.5}$$

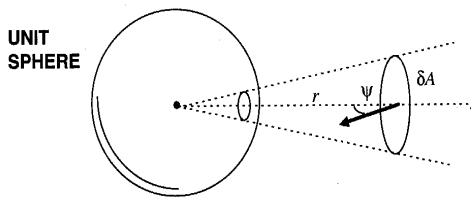

Figure 2.6 The definition of solid angle.

with ψ the angle between the normal to δA and the ray that points from the origin to δA. The factor $\cos\psi$ ensures the proper foreshortening of the area δA as seen from the origin.

We now write the image irradiance at an image point, **p**, as the ratio between δP, the power of light over a small image patch, and δI, the area of the small image patch:

$$E = \frac{\delta P}{\delta I}. \tag{2.6}$$

If δO is the area of a small surface patch around **P**, L the scene radiance at **P** in the direction toward the lens, $\Delta\Omega$ the solid angle subtended by the lens, and θ the angle between the normal to the viewed surface at **P** and the principal ray (Figure 2.7), the power δP is given by $\delta O L \Delta\Omega$ (the total power emitted in the direction of the lens) multiplied by $\cos\theta$ (the foreshortening of the area δO as seen from the lens):

$$\delta P = \delta O L \Delta\Omega \cos\theta. \tag{2.7}$$

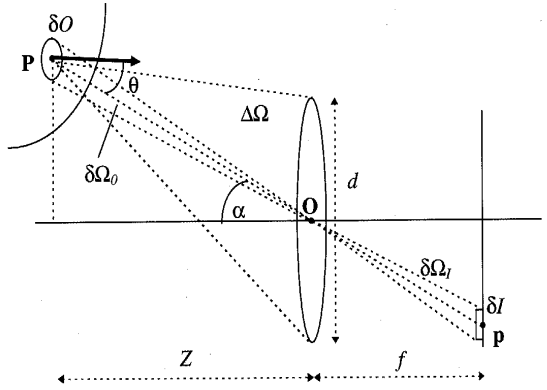

Figure 2.7 Radiometry of the image formation process.

Combining (2.6) and (2.7), we find

$$E = L\Delta\Omega\cos\theta\frac{\delta O}{\delta I}. \tag{2.8}$$

We still need to evaluate $\Delta\Omega$ and $\delta O/\delta I$. For the solid angle $\Delta\Omega$ (Figure 2.7), (2.5) with $\delta A = \pi d^2/4$ (lens area), $\psi = \alpha$ (angle between the principal ray and the optical axis), and $r = Z/\cos\alpha$ (distance of P from the lens center) becomes

$$\Delta\Omega = \frac{\pi}{4}d^2\frac{\cos^3\alpha}{Z^2}. \tag{2.9}$$

For the solid angle $\delta\Omega_I$, subtended by a small image patch of area δI (see Figure 2.7), (2.5) with $\delta A = \delta I$, $\psi = \alpha$, and $r = f/\cos\alpha$ gives

$$\delta\Omega_I = \frac{\delta I\cos\alpha}{(f/\cos\alpha)^2}. \tag{2.10}$$

Similarly, for the solid angle $\delta\Omega_O$ subtended by the patch δO on the object side, we have

$$\delta\Omega_O = \frac{\delta O\cos\theta}{(Z/\cos\alpha)^2}. \tag{2.11}$$

It is clear from Figure 2.7 that $\delta\Omega_I = \delta\Omega_O$; hence, their ratio is 1, so that dividing (2.11) by (2.10) we obtain

$$\frac{\delta O}{\delta I} = \frac{\cos\alpha}{\cos\theta}\left(\frac{Z}{f}\right)^2 \tag{2.12}$$

Ignoring energy losses within the system, and plugging (2.9) and (2.12) into (2.8), we finally obtain the desired relation between E and L.

The Fundamental Equation of Radiometric Image Formation

$$E(\mathbf{p}) = L(\mathbf{P})\frac{\pi}{4}\left(\frac{d}{f}\right)^2\cos^4\alpha. \tag{2.13}$$

Equation (2.13) says that *the illumination of the image at* **p** *decreases as the fourth power of the cosine of the angle formed by the principal ray through* **p** *with the optical axis*. In the case of small angular aperture, this effect can be neglected; therefore, the image irradiance can be regarded as uniformly proportional to the scene radiance over the *whole* image plane.

The nonuniform illumination predicted by (2.13) is hard to notice in ordinary images, because the major component of brightness changes is usually due to the spatial gradient of the image irradiance. You can try a simple experiment to verify the effect predicted by (2.13) by acquiring an image of a Lambertian surface illuminated by diffuse light (see Exercise 2.2).

☞ We make no substantial distinction among *image irradiance, intensity,* and *brightness*. Be warned, however, that a distinction does exist; although it is not relevant for our purposes. The *intensity* is the grey level recorded by an image and is linked to *irradiance* by a monotonic relation depending on the sensor. *Brightness* often indicates the subjective human perception of the image intensity.

The fundamental equation of radiometric image formation also shows that the quantity f/d, called the *F-number*, influences how much light is collected by the camera: the smaller the F-number, the larger the fraction of L which reaches the image plane. The F-number is one of the characteristics of the optics. As shown by (2.13), image irradiance is inversely proportional to the square of the F-number (see footnote 3 in this chapter).

2.2.4 Geometric Image Formation

We now turn to the geometric aspect of image formation. The aim is to link the position of scene points with that of their corresponding image points. To do this, we need to model the *geometric projection* performed by the sensor.

The Perspective Camera. The most common geometric model of an intensity camera is the *perspective* or *pinhole* model (Figure 2.8). The model consists of a plane π, the *image plane*, and a 3-D point $\mathbf{O}$, the *center* or *focus of projection*. The distance between π and $\mathbf{O}$ is the *focal length*. The line through $\mathbf{O}$ and perpendicular to π is the *optical axis*,[5] and $\mathbf{o}$, the intersection between π and the optical axis, is named *principal point* or *image center*. As shown in Figure 2.8, $\mathbf{p}$, the image of $\mathbf{P}$, is the point at which the straight line through $\mathbf{P}$ and $\mathbf{O}$ intersects the image plane π. Consider the 3-D reference frame in which $\mathbf{O}$ is the origin and the plane π is orthogonal to the Z axis, and let $\mathbf{P} = [X, Y, Z]^\top$ and $\mathbf{p} = [x, y, z]^\top$. This reference frame, called the *camera frame*, has fundamental importance in computer vision. We now will write the basic equations of perspective projections in the camera frame.

Perspective Camera: Fundamental Equations

In the camera frame, we have

$$x = f\frac{X}{Z}$$
$$y = f\frac{Y}{Z} \tag{2.14}$$

☞ In the camera frame, the third component of an image point is always equal to the focal length (as the equation of the plane π is $z = f$). For this reason, we will often write $\mathbf{p} = [x, y]^\top$ instead of $\mathbf{p} = [x, y, f]^\top$.

[5] You should link these definitions of focal length and optical axis with those in section 2.2.2.

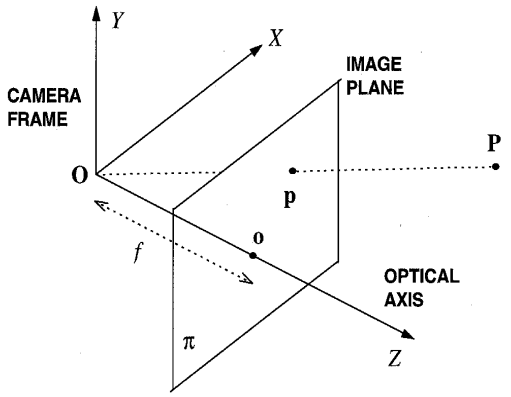

Figure 2.8 The perspective camera model.

Note that (2.14) are nonlinear because of the factor $1/Z$, and do not preserve either distances between points (not even up to a common scaling factor), or angles between lines. However, they do map lines into lines (see Exercise 2.3).

The Weak-Perspective Camera. A classical approximation that turns (2.14) into linear equations is the *weak-perspective* camera model. This model requires that the relative distance along the optical axis, δz, of any two scene points (that is, the scene's depth) is much smaller than the average distance, $\bar{Z}$, of the points from the viewing camera. In this case, for each scene point, $\mathbf{P}$, we can write

$$x = f\frac{X}{Z} \approx \frac{f}{\bar{Z}}X$$
$$y = f\frac{Y}{Z} \approx \frac{f}{\bar{Z}}Y \tag{2.15}$$

☞ Indicatively, the weak-perspective approximation becomes viable for $\delta z < \bar{Z}/20$ approximately.

These equations, (2.15), describe a sequence of two transformations: an *orthographic projection*, in which world points are projected along rays parallel to the optical axis,[6] that is,

$$x = X$$
$$y = Y,$$

[6] Analytically, the orthographic projection is the limit of the perspective projection for $f \to \infty$. For $f \to \infty$, we have $Z \to \infty$ and thus $f)Z \to 1$.

followed by *isotropic scaling* by the factor $f/\bar{Z}$. Section 2.4 shows that this and other camera models can also be derived in a compact matrix notation. Meanwhile, it is time for a summary.

The Perspective Camera Model

In the *perspective camera model* (nonlinear), the coordinates (x, y) of a point $\mathbf{p}$, image of the 3-D point $\mathbf{P} = [X, Y, Z]^\top$, are given by

$$x = f\frac{X}{Z}$$
$$y = f\frac{Y}{Z}$$

The Weak-Perspective Camera Model

If the average depth of the scene, $\bar{Z}$, is much larger than the relative distance between any two scene points along the optical axis, the *weak-perspective camera model* (linear) holds:

$$x = \frac{f}{\bar{Z}}X$$
$$y = \frac{f}{\bar{Z}}Y.$$

All equations are written in the camera reference frame.

2.3 Acquiring Digital Images

In this section, we now discuss the aspects of image acquisition that are special to *digital* images, namely:

- the essential structure of a typical image acquisition system
- the representation of digital images in a computer
- practical information on spatial sampling and camera noise

2.3.1 Basic Facts

How do we acquire a digital image into a computer? A digital image acquisition system consists of three hardware components: a *viewing camera*, typically a *CCD* (*Charged Coupled Device*) camera, a *frame grabber*, and a *host computer*, on which processing takes place (Figure 2.9).[7]

[7] This is the standard configuration, but not the only possibility. For instance, several manufacturers commercialize *smart cameras*, which can acquire images *and* perform a certain amount of image processing.

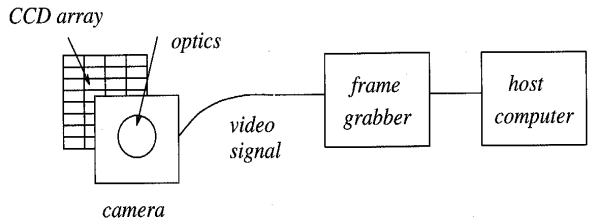

Figure 2.9 Essential components of a digital image acquisition system.

The input to the camera is, as we know, the incoming light, which enters the camera's lens and hits the image plane. In a CCD camera, the physical image plane is the *CCD array*, a $n \times m$ rectangular grid of photosensors, each sensitive to light intensity. Each photosensor can be regarded as a tiny, rectangular black box which converts light energy into a voltage. The output of the CCD array is usually a continuous electric signal, the *video signal*, which we can regard as generated by scanning the photosensors in the CCD array in a given order (e.g., line by line) and reading out their voltages. The video signal is sent to an electronic device called a *frame grabber*, where it is digitized into a 2-D, rectangular array of $N \times M$ integer values and stored in a memory buffer. At this point, the image can be conveniently represented by a $N \times M$ matrix, E, whose entries are called *pixels* (an acronym for *pic*ture *el*ements), with N and M being two fixed integers expressing the image size, in pixels, along each direction. Finally, the matrix E is transferred to a host computer for processing.

For the purposes of the following chapters, *the starting point of computer vision is the digitized image*, E. Here are the main assumptions we make about E.

Digital Images: Representation

A digital image is represented by a numerical matrix, E, with N rows and M columns.

$E(i, j)$ denotes the image value (image brightness) at pixel (i, j) (i-th row and j-th column), and encodes the intensity recorded by the photosensors of the CCD array contributing to that pixel.

$E(i, j)$ is an integer in the range $[0, 255]$.

This last statement about the range of $E(i, j)$ means that the brightness of an image point can be represented by one byte, or 256 grey levels (typically 0 is black, 255 white). This is an adequate resolution for ordinary, *monochromatic* (or *grey-level*) images and is suitable for many vision tasks. *Color images* require three monochromatic *component images* (red, green, blue) and therefore three numbers. Throughout this book, we shall always refer to grey-level images.

If we assume that the chain of sampling and filtering procedures performed by camera and frame buffer does not distort the video signal, the image stored in the

frame buffer is a faithful digitization of the image captured by the CCD array. However, *the number of elements along each side of the CCD arrays is usually different from the dimensions, in pixels, of the frame buffer.* Therefore, the position of the same point on the image plane will be different if measured in CCD elements or image pixels; more precisely, measuring positions from the upper left corner, the relation between the position (x_{im}, y_{im}) (in pixels) in the frame buffer image and the position (x_{CCD}, y_{CCD}) (in CCD elements) on the CCD array is given by

$$x_{im} = \frac{n}{N} x_{CCD}$$
$$y_{im} = \frac{m}{M} y_{CCD}. \tag{2.16}$$

Note that n/N and m/M in (2.16) are not the only parameters responsible for a different scaling of the image with respect to the CCD array along the horizontal and vertical direction; the different ratio of horizontal and vertical sizes of the CCD array has exactly the same effect. This is illustrated in Figure 2.10. The image stored in the computer memory consists of a squared grid of $N \times N$ pixels (Figure 2.10(*a*)). By inspection, it is easy to see that a grid of $n \times n$ CCD elements with an aspect ratio of n/m between the horizontal and vertical CCD element size (Figure 2.10(*b*)) produces exactly the same distorsion of an $m \times n$ CCD array with squared elements (Figure 2.10(*c*)).

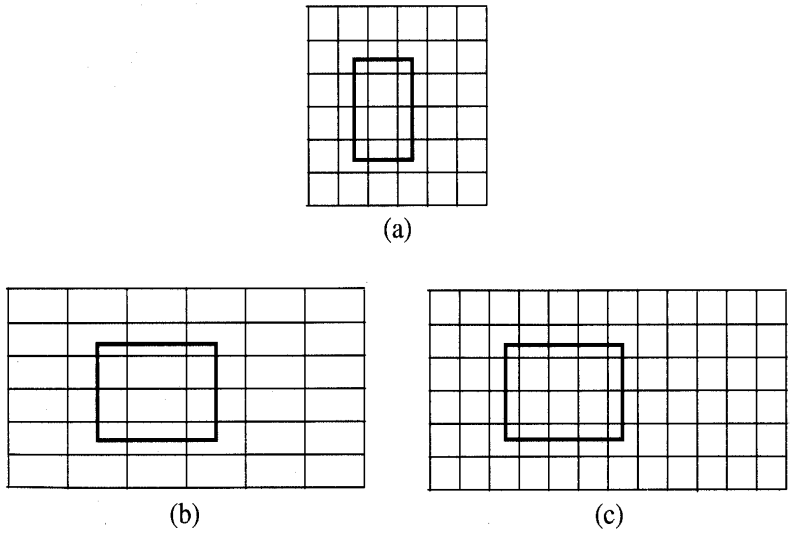

Figure 2.10 The same distortion of a given pattern on the CCD array (a) is produced by a $n \times n$ grid of rectangular elements of aspect ratio n/m (b), and by a $m \times n$ grid of squared elements (c).

In summary, it is convenient to assume that the CCD elements are always in one-to-one correspondence with the image pixels and to introduce *effective horizontal and vertical sizes* to account for the possible different scaling along the horizontal and vertical direction. The effective sizes of the CCD elements are our first examples of *camera parameters*, which are the subject of section 2.4.

2.3.2 Spatial Sampling

The spatial quantization of images originates at the very early stage of the image formation process, as the photoreceptors of a CCD sensor are organized in a rectangular array of photosensitive elements packed closely together. For simplicity, we assume that the distance d between adjacent CCD elements (specified by the camera manufacturer) is the same in the horizontal and vertical directions. We know from the *sampling theorem* that d determines the highest spatial frequency, ν_c, that can be captured by the system, according to the relation

$$\nu_c = \frac{1}{2d}$$

How does this characteristic frequency compare with the spatial frequency spectrum of images? A classical result of the diffraction theory of aberrations states that *the imaging process can be expressed in terms of a linear low-pass filtering of the spatial frequencies of the visual signal.* (For more information about the diffraction theory of aberrations, see the Further Readings.) In particular, if a is the linear size of the angular aperture of the optics (e.g., the diameter of a circular aperture), λ the wavelength of light, and f the focal length, spatial frequencies larger than

$$\nu_c' = \frac{a}{\lambda f}$$

do not contribute to the spatial spectrum of the image (that is, they are filtered out).

In a typical image acquisition system, the spatial frequency ν_c is nearly one order of magnitude smaller than ν_c'. Therefore, since the viewed pattern may well contain spatial frequencies larger than ν_c, *we expect aliasing*. You can convince yourself of the reality of spatial aliasing by taking images of a pattern of equally spaced thin black lines on a white background (see Exercise 2.6) at increasing distance from the camera. As predicted by the sampling theorem, if n is the number of the CCD elements in the horizontal direction, the camera cannot *see* more than n' vertical lines (with n' somewhat less than $n/2$, say $n' \sim n/3$). Until the number of lines within the field of view remains smaller than n', all the lines are correctly imaged and resolved. Once the limit is reached, if the distance of the pattern is increased further, but *before* blurring effects take over, the number of imaged lines *decreases* as the distance of the pattern *increases*!

☞ The main reason why spatial aliasing is often neglected is that the amplitude (that is, the information content) of high-frequency components of ordinary images is usually, though by no means always, very small.

2.3.3 Acquisition Noise and How to Estimate It

Let us briefly touch upon the problem of *noise* introduced by the imaging system and how it is estimated. The effect of noise is, essentially, that image values are not those expected, as these are corrupted during the various stages of image acquisition. As a consequence, the pixel values of two images of the same scene taken by the same camera and in the same light conditions are never *exactly* the same (try it!). Such fluctuations will introduce errors in the results of calculations based on pixel values; it is therefore important to estimate the magnitude of the noise.

The main objective of this section is *to suggest a simple characterization of image noise*, which can be used by the algorithms of following chapters. Noise attenuation, in particular, is the subject of Chapter 3.

An obvious way to proceed is to regard noisy variations as random variables, and try to characterize their statistical behavior. To do this, we acquire a sequence of images of the same scene, in the same acquisition conditions, and compute the pointwise average of the image brightness over all the images. The same sequence can also be used to estimate the *signal-to-noise ratio* of the acquisition system, as follows.[8]

Algorithm EST_NOISE

We are given n images of the same scene, $E_0, E_1, \ldots, E_{n-1}$, which we assume square ($N \times N$) for simplicity.

For each $i, j = 0, \ldots, N-1$, let

$$\overline{E(i,j)} = \frac{1}{n}\sum_{k=0}^{n-1} E_k(i,j)$$

$$\sigma(i,j) = \left(\frac{1}{n-1}\sum_{k=0}^{n-1}(\overline{E(i,j)} - E_k(i,j))^2\right)^{\frac{1}{2}} \qquad (2.17)$$

The quantity $\sigma(i,j)$ is an estimate of the standard deviation of the acquisition noise at each pixel. The average of $\sigma(i,j)$ over the image is an estimate of the average noise, while $\max_{i,j\in[0,\ldots,N-1]}\{\sigma(i,j)\}$ an estimate of the worst case acquisition noise.

☞ Notice that the beat frequency of some fluorescent room lights may skew the results of EST_NOISE.

Figure 2.11 shows the noise estimates relative to a particular acquisition system. A static camera was pointed at a picture posted on the wall. A sequence of $n = 100$ images was then acquired. The graphs in Figure 2.11 reproduce the average plus and minus

[8] The signal-to-noise ratio is usually expressed in *decibel* (dB), and is defined as 10 times the logarithm in base 10 of the ratio of two powers (in our case, of signal and noise). For example, a signal-to-noise ratio of 100 corresponds to $10\log_{10}100 = 20dB$.

Figure 2.11 Estimated acquisition noise. Graphs of the average image brightness, plus (solid line) and minus (dotted line) the estimated standard deviation, over a sequence of images of the same scene along the same horizontal scan line. The image brightness ranges from 73 to 211 grey levels.

the standard deviation of the image brightness (pixel values) over the entire sequence, along an horizontal scanline (image row). Notice that the standard deviation is almost independent of the average, typically less than 2 and never larger than 2.5 grey values. This corresponds to an average signal-to-noise ratio of nearly one hundred.

Another cause of noise, which is important when a vision system is used for fine measurements, is that *pixel values are not completely independent of each other:* some *cross-talking* occurs between adjacent photosensors in each row of the CCD array, due to the way the content of each CCD row is read in order to be sent to the frame buffer. This can be verified by computing the autocovariance $C_{EE}(i,j)$ of the image of a spatially uniform pattern parallel to the image plane and illuminated by diffuse light.

Algorithm AUTO_COVARIANCE

Let $c = 1/N^2$, $N_{i'} = N - i' - 1$, and $N_{j'} = N - j' - 1$. Given an image E, for each $i', j' = 0, \ldots, N-1$ compute

$$C_{EE}(i',j') = c\sum_{i=0}^{N_{i'}}\sum_{j=0}^{N_{j'}}(E(i,j) - \overline{E(i,j)})(E(i+i',j+j') - \overline{E(i+i',j+j')}) \qquad (2.18)$$

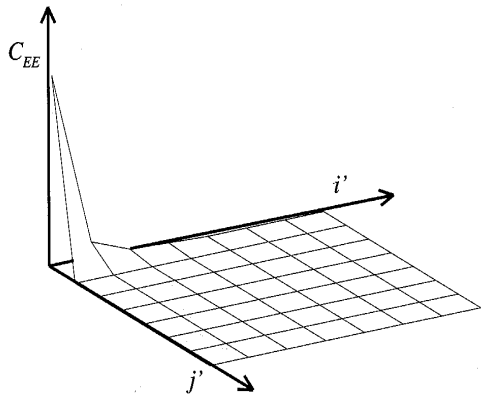

Figure 2.12 Autocovariance of the image of a uniform pattern for a typical image acquisition system, showing cross-talking between adjacent pixels along i'.

☞ The autocovariance should actually be estimated as the average of the autocovariance computed on many images of the same pattern. To minimize the effect of radiometric nonlinearities (see (2.13)), C_{EE} should be computed on a patch in the central portion of the image.

Figure 2.12 displays the graph of the average of the autocovariance computed on many images acquired by the same acquisition system used to generate Figure 2.11. The autocovariance was computed by means of (2.18) on a patch of 16×16 pixels centered in the image center. Notice the small but visible covariance along the horizontal direction: consistently with the physical properties of many CCD cameras, this indicates that the grey value of each pixel is not completely independent of that of its neighbors.

2.4 Camera Parameters

We now come back to discuss the geometry of a vision system in greater detail. In particular, we want to characterize the parameters underlying camera models.

2.4.1 Definitions

Computer vision algorithms reconstructing the 3-D structure of a scene or computing the position of objects in space need equations linking the coordinates of points in 3-D space with the coordinates of their corresponding image points. These equations are written in the camera reference frame (see (2.14) and section 2.2.4), but it is often assumed that

- the camera reference frame can be located with respect to some other, known, reference frame (the *world reference frame*), and

- the coordinates of the image points in the camera reference frame can be obtained from *pixel coordinates*, the only ones directly available from the image.

This is equivalent to assume knowledge of some camera's characteristics, known in vision as the camera's *extrinsic* and *intrinsic* parameters. Our next task is to understand the exact nature of the intrinsic and extrinsic parameters and why the equivalence holds.

Definition: Camera Parameters

The *extrinsic parameters* are the parameters that define the location and orientation of the camera reference frame with respect to a known world reference frame.

The *intrinsic parameters* are the parameters necessary to link the pixel coordinates of an image point with the corresponding coordinates in the camera reference frame.

In the next two sections, we write the basic equations that allow us to define the extrinsic and intrinsic parameters in practical terms. The problem of estimating the value of these parameters is called *camera calibration*. We shall solve this problem in Chapter 6, since calibration methods need algorithms which we discuss in Chapters 4 and 5.

2.4.2 Extrinsic Parameters

The camera reference frame has been introduced for the purpose of writing the fundamental equations of the perspective projection (2.14) in a simple form. However, *the camera reference frame is often unknown*, and a common problem is determining the location and orientation of the camera frame with respect to some known reference frame, *using only image information*. The extrinsic parameters are defined as *any set of geometric parameters that identify uniquely the transformation between the unknown camera reference frame and a known reference frame*, named the *world reference frame*.

A typical choice for describing the transformation between camera and world frame is to use

- a 3-D translation vector, $\mathbf{T}$, describing the relative positions of the origins of the two reference frames, and
- a 3×3 rotation matrix, R, an orthogonal matrix ($R^\top R = RR^\top = I$) that brings the corresponding axes of the two frames onto each other.

The orthogonality relations reduce the number of degrees of freedom of R to three (see section A.9 in the Appendix).

In an obvious notation (see Figure 2.13), the relation between the coordinates of a point $\mathbf{P}$ in world and camera frame, $\mathbf{P}_w$ and $\mathbf{P}_c$ respectively, is

$$\mathbf{P}_c = R(\mathbf{P}_w - \mathbf{T}), \tag{2.19}$$

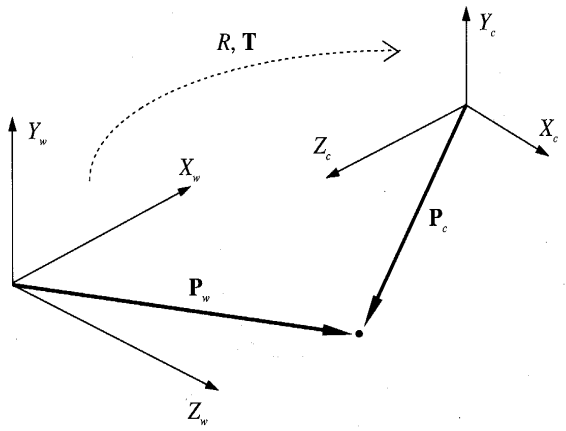

Figure 2.13 The relation between camera and world coordinate frames.

with

$$R = \begin{pmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \end{pmatrix}.$$

Definition: Extrinsic Parameters

The camera extrinsic parameters are the translation vector, $\mathbf{T}$, and the rotation matrix, R (or, better, its free parameters), which specify the transformation between the camera and the world reference frame.

2.4.3 Intrinsic Parameters

The intrinsic parameters can be defined as the set of parameters needed to characterize the optical, geometric, and digital characteristics of the viewing camera. For a pinhole camera, we need three sets of intrinsic parameters, specifying respectively

- the perspective projection, for which the only parameter is the focal length, f;
- the transformation between camera frame coordinates and pixel coordinates;
- the geometric distortion introduced by the optics.

From Camera to Pixel Coordinates. To find the second set of intrinsic parameters, we must link the coordinates (x_{im}, y_{im}) of an image point in pixel units with the coordinates (x, y) of the same point in the camera reference frame. The coordinates

(x_{im}, y_{im}) can be thought of as coordinates of a new reference frame, sometimes called *image reference frame.*

The Transformation between Camera and Image Frame Coordinates

Neglecting any geometric distorsions possibly introduced by the optics and in the assumption that the CCD array is made of a rectangular grid of photosensitive elements, we have

$$\begin{aligned} x &= -(x_{im} - o_x)s_x \\ y &= -(y_{im} - o_y)s_y \end{aligned} \tag{2.20}$$

with (o_x, o_y) the coordinates in pixel of the image center (the principal point), and (s_x, s_y) the effective size of the pixel (in millimeters) in the horizontal and vertical direction respectively.

Therefore, the current set of intrinsic parameters is f, o_x, o_y, s_x, s_y.

☞ The sign change in (2.20) is due to the fact that the horizontal and vertical axes of the image and camera reference frames have opposite orientation.

In several cases, the optics introduces image distortions that become evident at the periphery of the image, or even elsewhere using optics with large fields of view. Fortunately, these distortions can be modelled rather accurately as simple *radial distortions*, according to the relations

$$\begin{aligned} x &= x_d(1 + k_1 r^2 + k_2 r^4) \\ y &= y_d(1 + k_1 r^2 + k_2 r^4) \end{aligned}$$

with (x_d, y_d) the coordinates of the distorted points, and $r^2 = x_d^2 + y_d^2$. As shown by the equations above, this distortion is a radial displacement of the image points. The displacement is null at the image center, and increases with the distance of the point from the image center. k_1 and k_2 are further intrinsic parameters. Since they are usually very small, radial distortion is ignored whenever high accuracy is not required in all regions of the image, or when the peripheral pixels can be discarded. If not, as $k_2 << k_1$, k_2 is often set equal to 0, and k_1 is the only intrinsic parameter to be estimated in the radial distortion model.

☞ The magnitude of geometric distortion depends on the quality of the lens used. As a rule of thumb, with optics of average quality and CCD size around 500 × 500, expect distortions of several pixels (say around 5) in the outer cornice of the image. Under these circumstances, a model with $k_2 = 0$ is still accurate.

It is now time for a summary.

Intrinsic Parameters

The camera intrinsic parameters are defined as the focal length, f, the location of the image center in pixel coordinates, (o_x, o_y), the effective pixel size in the horizontal and vertical direction (s_x, s_y), and, if required, the radial distortion coefficient, k_1.

2.4.4 Camera Models Revisited

We are now fully equipped to write relations linking directly the pixel coordinates of an image point with the world coordinates of the corresponding 3-D point, *without explicit reference to the camera reference frame* needed by (2.14).

Linear Version of the Perspective Projection Equations. Plugging (2.19) and (2.20) into (2.14) we obtain

$$-(x_{im} - o_x)s_x = f\frac{\mathbf{R}_1^\top(\mathbf{P}_w - \mathbf{T})}{\mathbf{R}_3^\top(\mathbf{P}_w - \mathbf{T})}$$
$$-(y_{im} - o_y)s_y = f\frac{\mathbf{R}_2^\top(\mathbf{P}_w - \mathbf{T})}{\mathbf{R}_3^\top(\mathbf{P}_w - \mathbf{T})} \qquad (2.21)$$

where $\mathbf{R}_i, i = 1, 2, 3$, is a 3-D vector formed by the i-th row of the matrix R. Indeed, (2.21) relates the 3-D coordinates of a point in the world frame to the image coordinates of the corresponding image point, via the camera extrinsic and intrinsic parameters.

☞ Notice that, due to the particular form of (2.21), not all the intrinsic parameters are independent. In particular, the focal length could be absorbed into the effective sizes of the CCD elements.

Neglecting radial distortion, we can rewrite (2.21) as a simple matrix product. To this purpose, we define two matrices, M_{int} and M_{ext}, as

$$M_{int} = \begin{pmatrix} -f/s_x & 0 & o_x \\ 0 & -f/s_y & o_y \\ 0 & 0 & 1 \end{pmatrix}$$

and

$$M_{ext} = \begin{pmatrix} r_{11} & r_{12} & r_{13} & -\mathbf{R}_1^\top\mathbf{T} \\ r_{21} & r_{22} & r_{23} & -\mathbf{R}_2^\top\mathbf{T} \\ r_{31} & r_{32} & r_{33} & -\mathbf{R}_3^\top\mathbf{T} \end{pmatrix},$$

so that *the* 3×3 *matrix* M_{int} *depends only on the intrinsic parameters, while the* 3×4 *matrix* M_{ext} *only on the extrinsic parameters.* If we now add a "1" as a fourth coordinate of $\mathbf{P}_w$ (that is, express $\mathbf{P}_w$ in homogeneous coordinates), and form the product $M_{int}M_{ext}\mathbf{P}_w$, we obtain a linear matrix equation describing perspective projections.

The Linear Matrix Equation of Perspective Projections

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = M_{int}M_{ext}\begin{pmatrix} X_w \\ Y_w \\ Z_w \\ 1 \end{pmatrix}.$$

What is interesting about vector $[x_1, x_2, x_3]^\top$ is that the ratios (x_1/x_3) and (x_2/x_3) are nothing but the image coordinates:

$$x_1/x_3 = x_{im}$$
$$x_2/x_3 = y_{im}.$$

Moreover, we have separated nicely the two steps of the world-image projection:

- M_{ext} performs the transformation between the world and the camera reference frame;
- M_{int} performs the transformation between the camera reference frame and the image reference frame.

☞ In more formal terms, the relation between a 3-D point and its perspective projection on the image plane can be seen as a linear transformation from the *projective space*, the space of vectors $[X_w, Y_w, Z_w, 1]^\top$, to the *projective plane*, the space of vectors $[x_1, x_2, x_3]^\top$. This transformation is defined *up to an arbitrary scale factor* and so that the matrix M has only 11 independent entries (see review questions). This fact will be discussed in Chapter 6.

The Perspective Camera Model. Various camera models, including the perspective and weak-perspective ones, can be derived by setting appropriate constraints on the matrix $M = M_{int}M_{ext}$. Assuming, for simplicity, $o_x = o_y = 0$ and $s_x = s_y = 1$, M can then be rewritten as

$$M = \begin{pmatrix} -fr_{11} & -fr_{12} & -fr_{13} & f\mathbf{R}_1^\top\mathbf{T} \\ -fr_{21} & -fr_{22} & -fr_{23} & f\mathbf{R}_2^\top\mathbf{T} \\ r_{31} & r_{32} & r_{33} & -\mathbf{R}_3^\top\mathbf{T} \end{pmatrix}.$$

When unconstrained, M describes the full-perspective camera model and is called *projection matrix*.

The Weak-Perspective Camera Model. To derive the form of M for the weak-perspective camera model, we observe that the image $\mathbf{p}$ of a point $\mathbf{P}$ is given by

$$\mathbf{p} = M\begin{pmatrix} X_w \\ Y_w \\ Z_w \\ 1 \end{pmatrix} = \begin{pmatrix} f\mathbf{R}_1^\top(\mathbf{T} - \mathbf{P}) \\ f\mathbf{R}_2^\top(\mathbf{T} - \mathbf{P}) \\ \mathbf{R}_3^\top(\mathbf{P} - \mathbf{T}) \end{pmatrix}. \qquad (2.22)$$

But $\|\mathbf{R}_3^\top(\mathbf{P}-\mathbf{T})\|$ is simply the distance of $\mathbf{P}$ from the projection center along the optical axis; therefore, the basic constraint for the weak-perspective approximation can be written as

$$\left|\frac{\mathbf{R}_3^\top(\mathbf{P}_i-\bar{\mathbf{P}})}{\mathbf{R}_3^\top(\bar{\mathbf{P}}-\mathbf{T})}\right| << 1, \tag{2.23}$$

where $\mathbf{P}_1, \mathbf{P}_2$ are two points in 3-D space, and $\bar{\mathbf{P}}$ the centroid of $\mathbf{P}_1$ and $\mathbf{P}_2$. Using (2.23), (2.22) can be written for $\mathbf{P}=\mathbf{P}_i$, $i=1,2$, as

$$\mathbf{p}_i \approx \begin{pmatrix} f\mathbf{R}_1^\top(\mathbf{T}-\mathbf{P}_i) \\ f\mathbf{R}_2^\top(\mathbf{T}-\mathbf{P}_i) \\ \mathbf{R}_3^\top(\bar{\mathbf{P}}-\mathbf{T}) \end{pmatrix}.$$

Therefore, the projection matrix M becomes

$$M_{wp} = \begin{pmatrix} -fr_{11} & -fr_{12} & -fr_{13} & f\mathbf{R}_1^\top\mathbf{T} \\ -fr_{21} & -fr_{22} & -fr_{23} & f\mathbf{R}_2^\top\mathbf{T} \\ 0 & 0 & 0 & \mathbf{R}_3^\top(\bar{\mathbf{P}}-\mathbf{T}) \end{pmatrix}.$$

The Affine Camera Model. Another interesting camera model, widely used in the literature for its simplicity, is the so-called *affine model*, a mathematical generalization of the weak-perspective model. In the affine model, the first three entries in the last row of the matrix M are equal to zero. All other entries are unconstrained. The affine model does not appear to correspond to any physical camera, but leads to simple equations and has appealing geometric properties. The affine projection does not preserve angles but does preserve parallelism.

The main difference with the weak-perspective model is the fact that, in the affine model, only the ratio of distances measured along parallel directions is preserved. We now move on to consider range images.

2.5 Range Data and Range Sensors

In many applications, one wants to use vision to measure distances; for example, to steer vehicles away from obstacles, estimate the shape of surfaces, or inspect manufactured objects. A single intensity image proves of limited use, as pixel values are related to surface geometry only indirectly; that is, through the optical and geometrical properties of the surfaces as well as the illumination conditions. All these are usually complex to model and often unknown. As we shall see in Chapter 9, reconstructing 3-D shape from a single intensity image is difficult and often inaccurate. Can we acquire images encoding shape *directly*? Yes: this is exactly what range sensors do.

Range Images

Range images are a special class of digital images. Each pixel of a range image expresses the *distance between a known reference frame and a visible point in the scene.* Therefore, a range image reproduces the 3-D *structure* of a scene, and is best thought of as a *sampled surface.*

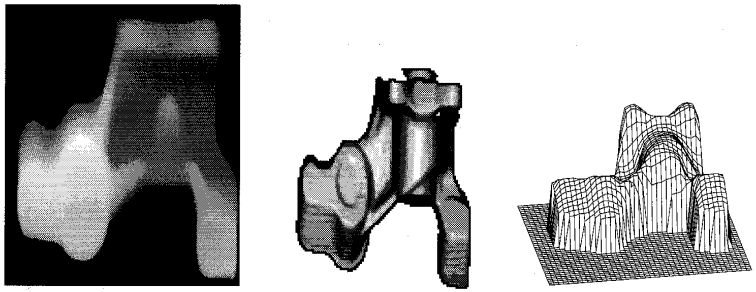

Figure 2.14 Range views of a mechanical component displayed as intensity image (left, the lighter the closer), cosine shaded (middle), and 3-D surface (right). Courtesy of R. B. Fisher, Department of Artificial Intelligence, University of Edinburgh.

2.5.1 Representing Range Images

Range images can be represented in two basic forms. One is a list of 3-D coordinates in a given reference frame, called *xyz form* or *cloud of points*, for which no specific order is required. The other is a matrix of depth values of points along the directions of the x, y image axes, called the r_{ij} *form*, which makes spatial information explicit. Notice that xyz data can be more difficult to process than r_{ij} data, as no spatial order is assumed. Range images are also referred to as *depth images, depth maps, xyz maps, surface profiles*, and *2.5-D images*.

☞ Obviously, r_{ij} data can always be visualized as a normal intensity image; the term "range image" refers indeed to the r_{ij} form. We will assume this in the following, unless otherwise specified. One can also display a range image as *cosine shaded*, whereby the grey level of each pixel is proportional to the norm of the gradient to the range surface. Figure 2.14 illustrates the three main methods of displaying range images.

☞ Dense range images prove useful for estimating the differential properties (local shape) of object surfaces.

2.5.2 Range Sensors

An *optical range sensor* is a device using optical phenomena to acquire range images. We concentrate on optical sensors as we are concerned with vision. Range sensors may measure depth at one point only, or the distance and shape of surface profiles, or of full surfaces. It is useful to distinguish between *active* and *passive* range sensors.

Definition: Active and Passive Range Sensors

Active range sensors project energy (e.g., a pattern of light, sonar pulses) on the scene and detect its position to perform the measure; or exploit the effect of controlled changes of some sensor parameters (e.g., focus).

Passive range sensors rely only on intensity images to reconstruct depth (e.g., stereopsis, discussed in Chapter 7).

Passive range sensors are the subject of Chapters 7, 8, and 9, and are not discussed further here. Active range sensors exploit a variety of physical principles; examples are radars and sonars, Moiré interferometry, focusing, and triangulation. Here, we sketch the first three, and concentrate on the latter in greater detail.

Radars and Sonars. The basic principle of these sensors is to emit a short electromagnetic or acoustic wave, or *pulse*, and detect the return (echo) reflected from surrounding surfaces. Distance is obtained as a function of the time taken by the wave to hit a surface and come back, called *time of flight*, which is measured directly. By sweeping such a sensor across the target scene, a full range image can be acquired. Different principles are used in imaging laser radars; for instance, such sensors can emit an amplitude-modulated laser beam and measure the phase difference between the transmitted and received signals.

Moiré Interferometry. A Moiré interference pattern is created when two gratings with regularly spaced patterns (e.g., lines) are superimposed on each other. Moiré sensors project such gratings onto surfaces, and measure the phase differences of the observed interference pattern. Distance is a function of such phase difference. Notice that such sensors can recover absolute distance only if the distance of one reference point is known; otherwise, only relative distances between scene points are obtained (which is desirable for inspection).

Active Focusing/Defocusing. These methods infer range from two or more images of the same scene, acquired under varying focus settings. For instance, *shape-from-focus* sensors vary the focus of a motorized lens continuously, and measure the amount of blur for each focus value. Once determined the best focused image, a model linking focus values and distance yields the distance. In *shape-from-defocus*, the blur-focus model is fitted to two images only to estimate distance.

In the following section, we concentrate on *triangulation-based* range sensors. The main reason for this choice is that they are based on intensity cameras, so we can exploit everything we know on intensity imaging. Moreover, such sensors can give accurate and dense 3-D coordinate maps, are easy to understand and build (as long as limited accuracy is acceptable), and are commonly found in applications.

2.5.3 Active Triangulation

We start by discussing the basic principle of active triangulation. Then, we discuss a simple sensor, and how to evaluate its performance. As we do not know yet how to calibrate intensity cameras, nor how to detect image features, you will be able to implement the algorithms in this section only after reading Chapters 4 and 5.

The basic geometry for an active triangulation system is shown in Figure 2.15. A light projector is placed at a distance b (called *baseline*) from the center of projection of

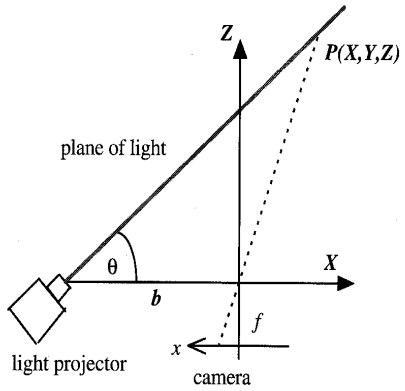

Figure 2.15 The basic geometry of active, optical triangulation (planar XZ view). The Y and y axes are perpendicular to the plane of the figure.

a pinhole camera.[9] The center of projection is the origin of the reference frame XYZ, in which all the sensor's measurements are expressed. The Z axis and the camera's optical axis coincide. The y and Y, and x and X axes are respectively parallel but point in opposite directions. Let f be the focal length. The projector emits a plane of light perpendicular to the plane XZ and forming a controlled angle, θ, with the XY plane. The Y axis is parallel to the plane of light and perpendicular to the page, so that only the profile of the plane of light is shown. The intersection of the plane of light with the scene surfaces is a planar curve called the *stripe*, which is observed by the camera. In this setup, the coordinates of a stripe point $\mathbf{P} = [X, Y, Z]^\top$ are given by

$$\begin{pmatrix} X \\ Y \\ Z \end{pmatrix} = \frac{b}{f \cot\theta - x} \begin{pmatrix} x \\ y \\ f \end{pmatrix} \tag{2.24}$$

☞ The focal length and the other intrinsic parameters of the intensity camera can be calibrated with the same procedures to be used for intensity cameras (Chapter 6).

Applying this equation to all the visible stripe points, we obtain the 3-D profile of the surface points under the stripe (a cross-section of the surface). We can acquire multiple, adjacent profiles by advancing the object under the stripe, or sweeping the stripe across the object, and repeat the computation for each relative position of stripe and object. The sequence of all profiles is a full range image of the scene.

[9] Notice that, in Figure 2.15, the center of projection is *in front* of the screen, not behind, and is not the origin of the camera frame. This does not alter the geometry of image formation. Why?.

In order to measure (x, y), we must identify the stripe points in the image. To facilitate this task, we try to make the stripe stand out in the image. We can do this by projecting laser light, which makes the stripe brighter than the rest of the image; or we can project a black line onto a matte white or light grey object, so that the only really dark image points are the stripe's. Both solutions are popular but have drawbacks. In the former case, concavities on shiny surfaces may create reflections that confuse the stripe detection; in the latter, stripe location may be confused by shadows, marks and dark patches. In both cases, no range data can be obtained where the stripe is invisible to the camera because of occlusions. Sensors based on laser light are called *3-D laser scanners*, and are found very frequently in applications. A real sensor, modelled closely after the basic geometry in Figure 2.15, is shown in Figure 2.16.

☞ To limit occlusions one often uses two or more cameras, so that the stripe is nearly always visible from at least one camera.

2.5.4 A Simple Sensor

In order to use (2.24) we must calibrate f, b and θ. Although it is not difficult to devise a complete calibration procedure based on the projection equations and the geometry of Figure 2.17, we present here a simple and efficient method, called *direct calibration*, which does not require any equations at all. Altogether we shall describe a small but complete range sensor, how to calibrate it, and how to use it for measuring range profiles of 3-D objects. The algorithms require knowledge of some simple image processing operations, that you will be able to implement after going through the next three chapters.

The direct calibration procedure builds a lookup table (LUT) linking image and 3-D coordinates. Notice that this is possible because a one-to-one correspondence exists between image and 3-D coordinates, thanks to the fact that the stripe points are constrained to lie in the plane of light. The LUT is built by measuring the image coordinates of a grid of known 3-D points, and recording both image and world coordinates for each point; the depth values of all other visible points are obtained by interpolation.

The procedure uses a few rectangular blocks of known heights δ (Figure 2.17). One block (call it G) must have a number (say n) of parallel, rectangular grooves. We assume the image size (in pixels) is $x_{max} \times y_{max}$.

Algorithm RANGE_CAL

Set up the system and reference frame as in Figure 2.17. With no object in the scene, the vertical stripe falls on $Z = 0$ (background plane) and should be imaged near $y = y_{max} - 1$.

1. Place block G under the stripe, with grooves perpendicular to the stripe plane. Ensure the stripe appears parallel to x (constant y).
2. Acquire an image of the stripe falling on G. Find the y coordinates of the stripe points falling on G's higher surface (i.e., not in the groove) by scanning the image columns.
3. Compute the coordinates $[x_i, y_Z]^\top$, $i = 1, \ldots, n$, of the centers of the stripe segments on G's top surface, by taking the centers of the segments in the scanline $y = y_Z$. Enter each image point $[x_i, y_Z]^\top$ and its corresponding 3-D points $[X, Z]^\top$ (known) into a table T.

4. Put another block under G, raising G's top surface by δ. Ensure that the conditions of step 1 still apply. Be careful not to move the XYZ reference frame.
5. Repeat steps 2, 3, 4 until G's top surface is imaged near $y = 0$.
6. Convert T into a 2-D lookup table L, indexed by image coordinates $[x, y]^\top$, with x between 0 and $x_{max} - 1$, and y between 0 and $y_{max} - 1$, and returning $[X, Z]^\top$. To associate values to the pixels not measured directly, interpolate linearly using the four nearest neighbors.

The output is a LUT linking coordinates of image points and coordinates of scene points.

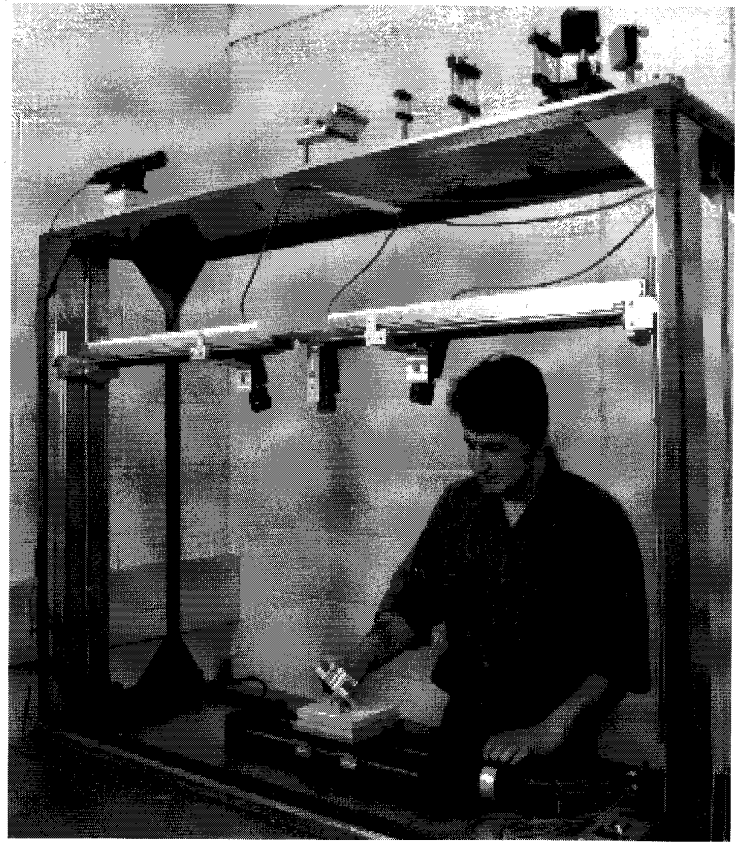

Figure 2.16 A real 3-D triangulation system, developed at Heriot-Watt University by A. M. Wallace and coworkers. Notice the laser source (top left), which generates a laser beam; the optical components forming the plane of laser light (top middle and left); the cameras; and the motorized platform (bottom middle) supporting the object and sweeping it through the stationary plane of light.

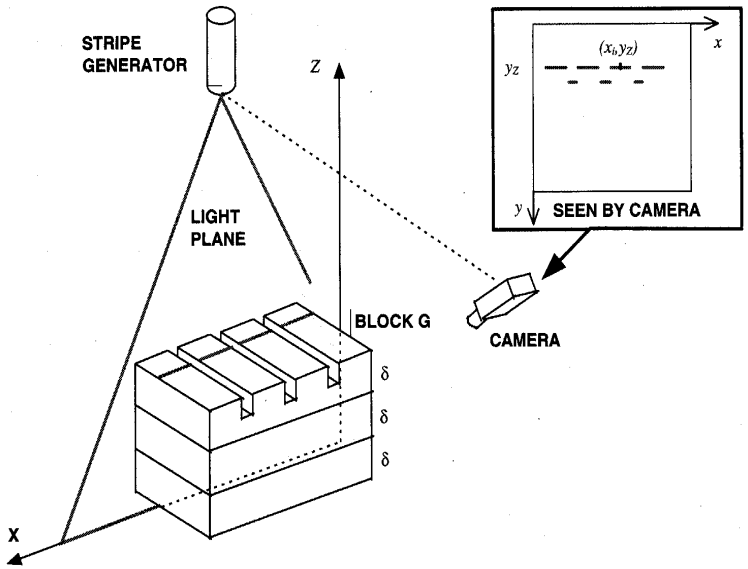

Figure 2.17 Setup for direct calibration of a simple profile range sensor.

And here is how to use L to acquire a range profile.

Algorithm RANGE_ACQ

The input is the LUT, L, built by RANGE_CAL.

1. Put an object under the stripe and acquire an image of the stripe falling on G.
2. Compute the image coordinates $[x, y]^\top$ of the stripe points by scanning each image column.
3. Index L using the image coordinates (x, y) of the stripe point, to obtain range points $[X, Z]^\top$.

The output is the set of 3-D coordinates corresponding to the stripe points imaged.

Notice that the numbers computed by such a sensor grow from the background plane ($Z = 0$), not from the camera.

☞ When a new block is added to the calibration scene, the stripe should move up by at least one or two pixels; if not, the calibration will not discriminate between Z levels. Be sure to use the *same* code for peak location in RANGE_ACQ *and* RANGE_CAL! The more sparse the calibration grid, the less accurate the range values obtained by interpolation in L.

When is a range sensor better than another for a given application? The following list of parameters is a basis for characterizing and comparing range sensors. Most parameters apply for non-triangulation sensors too.

Basic Parameters of Range Sensors

Workspace: the volume of space in which range data can be collected.

Stand-off distance: the approximate distance between the sensor and the workspace.

Depth of field: the depth of the workspace (along Z).

Accuracy: statistical variations of repeated measurements of a known true value (ground truth). Accuracy specifications should include at least the mean absolute error, the RMS error, and the maximum absolute error over N measures of a same object, with $N \gg 1$.

Resolution or precision: the smallest change in range that the sensor can measure or represent.

Speed: the number of range points measured per second.

Size and weight: important in some applications (e.g., only small sensors can be fitted on a robot arm).

☞ It is often difficult to know the *actual* accuracy of a sensor without carrying out your own measurements. Accuracy figures are sometimes reported without specifying to which error they refer to (e.g., RMS, absolute mean, maximum), and often omitting the experimental conditions and the optical properties of the surfaces used.

2.6 Summary

After working through this chapter you should be able to:

- explain how digital images are formed, represented and acquired
- estimate experimentally the noise introduced in an image by an acquisition system
- explain the concept of intrinsic and extrinsic parameters, the most common models of intensity cameras, and their applicability
- design (but not yet implement) an algorithm for calibrating and using a complete range sensor based on direct calibration

2.7 Further Readings

It is hard to find more on the content of this chapter on just one book. As a result if you want to know more you must be willing to do some bibliographic search. A readable account of basic optics can be found in the Feynman's Lecture on Physics [4]. A classic on the subject and beyond is the Born and Wolf [3] . The Born and Wolf also covers topics like image formation and spatial frequency filtering (though it is not always simple to go through). Our derivation of (2.13) is based on Horn and Sjoberg [6]. Horn [5] gives an extensive treatment of surface reflectance models. Of the many, very good textbooks on signal theory, our favorite is the Oppenheim, Willsky, and Young [11]. The discussion on camera models via the projection matrix is based on the appendix of Mundy and Zisserman's book *Geometric Invariants in Computer Vision* [9].

Our discussion of range sensors is largely based on Besl [1], which is a very good introduction to the principles, types and evaluation of range sensors. A recent,

detailed review of commercial laser scanners can be found in [14]. Two laser-based, active triangulation range sensors are described in [12, 13]; the latter is based on direct calibration, the former uses a geometric camera model. References [8] and [2] are examples of triangulation sensors projecting patterns of lines generated using incoherent light (as opposed to laser light) onto the scene. Krotkow [7] and Nayar and Nakagawa [10] make good introductions to focus- based ranging.

2.8 Review

Questions

- ❒ *2.1* How does an image change if the focal length is varied?
- ❒ *2.2* Give an intuitive explanation of the reason why a pinhole camera has an infinite depth of field.
- ❒ *2.3* Use the definition of F-number to explain geometrically why this quantity measures the fraction of the light entering the camera which reaches the image plane.
- ❒ *2.4* Explain why the beat frequency of fluorescent room light (e.g., 60 Hz) can skew the results of EST_NOISE.
- ❒ *2.5* *Intensity thresholding* is probably the simplest way to locate interesting objects in an image (a problem called *image segmentation*). The idea is that only the pixels whose value is above a threshold belong to interesting objects. Comment on the shortcomings of this technique, particularly in terms of the relation between scene radiance and image irradiance. Assuming that scene and illumination can be controlled, what would you do to guarantee successful segmentation by thresholding?
- ❒ *2.6* The projection matrix M is a 3×4 matrix defined up to an arbitrary scale factor. This leaves only 11 of the 12 entries of T independent. On the other hand, we have seen that the matrix T can be written in terms of 10 parameters (4 intrinsic and 6 extrinsic independent parameters). Can you guess the independent intrinsic parameter that has been left out? If you cannot guess now, you have to wait for Chapter 6.
- ❒ *2.7* Explain the problem of camera calibration, and why calibration is necessary at all.
- ❒ *2.8* Explain why the length *in millimeters* of an image line of endpoints $[x_1, y_1]$ and $[x_2, y_2]$ is not simply $\sqrt{(x_2 - x_1)^2 - (y_2 - y_1)^2}$. What does this formula miss?
- ❒ *2.9* Explain the difference between a range and an intensity image. Could range images be acquired using intensity cameras only (i.e., no laser light or the like)?
- ❒ *2.10* Explain the reason for the word "shaded" in "cosine-shaded rendering of a range image". What assumptions on the illumination does a cosine-shaded image imply? How is the surface gradient linked to shading?
- ❒ *2.11* What is the reason for step 1 in RANGE_CAL?

- ❒ *2.12* Consider a triangulation sensor which scans a whole surface profile by translating an object through a plane of laser light. Now imagine the surface is scanned by making the laser light sweep the object. In both cases the camera is stationary. What parts of the triangulation algorithm change? Why?
- ❒ *2.13* The performance of a range sensor based on (2.24) depend on the values of f, b, θ. How would you define and determine "optimal" values of f, b, θ for such a sensor?

Exercises

- ○ *2.1* Show that (2.1) and (2.2) are equivalent.
- ○ *2.2* Devise an experiment that checks the prediction of (2.13) on your own system. *Hint:* Use a spatially uniform object (like a flat sheet of matte gray paper) illuminated by perfectly diffuse light. Use optics with a wide field of view. Repeat the experiment by averaging the acquired image over time. What difference does this averaging step make?
- ○ *2.3* Show that, in the pinhole camera model, three collinear points in 3-D space are imaged into three collinear points on the image plane.
- ○ *2.4* Use the perspective projection equations to explain why, in a picture of a face taken frontally and from a very small distance, the nose appears much larger than the rest of the face. Can this effect be reduced by acting on the focal length?
- ○ *2.5* Estimate the noise of your acquisition system using procedures EST_NOISE and AUTO_COVARIANCE.
- ○ *2.6* Use the equations of section 2.3.2 to estimate the spatial aliasing of your acquisition system, and devise a procedure to estimate, roughly, the number of CCD elements of your camera.
- ○ *2.7* Write a program which displays a range image as a normal image (grey levels encode distance) or as a cosine shaded image.
- ○ *2.8* Derive (2.24) from the geometry shown in Figure 2.15. *Hint:* Use the law of sines and the pinhole projection equation. Why have we chosen to position the reference frame as in Figure 2.15?
- ○ *2.9* We can predict the sensitivity of measurements obtained through (2.24) by taking partial derivatives with respect to the formula's parameters. Compare such predictions with respect to b and f.

Projects

- ● *2.1* You can build your own pinhole camera, and join the adepts of *pinhole photography*. Pierce a hole about 5 mm in diameter on one side of an old tin box, 10 to 30 cm in depth. Spray the inside of box and lid with black paint. Pierce a pinhole in a piece of thick aluminium foil (e.g., the one used for milk tops), and fix the foil to the hole in the box with black tape. In a dark room, fix a piece of black and white photographic film on the hole in the box, and seal the box with black tape. The nearer the pinhole to the film, the wider the field of view. Cover the pinhole with

a piece of black paper to be used as shutter. Your camera is ready. Indicatively, a 125-ASA film may require an exposure of about 5 seconds. Make sure that the camera does not move as you open and close the shutter. Some experimentation will be necessary, but results can be striking!

- *2.2* Although you will learn how to locate image features and extract straight lines automatically in the next chapter, you can get ready for an implementation of the profile scanner described in section 2.5.4, and set up the equipment necessary. All you need (in addition to camera, frame buffer and computer) is a projector creating a black stripe (easily done with a slide projector and an appropriate slide, or even with a flashlight) and a few, accurately cut blocks. You must also work out the best arrangement for projector, stripe and camera.

References

[1] P.J. Besl, Active, Optical Imaging Sensors, *Machine Vision and Applications*, Vol. 1, pp. 127-152 (1988).

[2] A. Blake, H.R. Lo, D. McCowen and P. Lindsey, Trinocular Active Range Sensing, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 15, pp. 477-483 (1993).

[3] M. Born and E. Wolf, *Principles of Optics*, Pergamon Press, New York (1959).

[4] R.P. Feynman, R.B. Leighton, and M. Sands, *The Feynman Lectures on Physics*, Addison-Wesley, Reading, Mass (1965).

[5] B.K P. Horn, *Robot Vision*, MIT Press, Cambridge, MA (1986).

[6] B.K.P. Horn and R.W. Sjoberg, Calculating the Reflectance Map, *Applied Optics*, Vol. 18, pp 1770-1779 (1979).

[7] E. Krotkow, Focusing, *International Journal of Computer Vision*, Vol. 1, pp. 223–237 (1987).

[8] M. Maruyama and S. Abe, Range Sensing by Projecting Multiple Slits with Random Cuts, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 15, no. 6, pp. 647-651 (1988).

[9] J.L. Mundy and A. Zisserman, Appendix - Projective Geometry for Machine Vision. In *Geometric Invariants in Computer Vision*, Mundy, J.L. and Zisserman, A., eds., MIT Press, Cambridge, MA (1992).

[10] S.K. Nayar and Y. Nakagawa, Shape from Focus, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 16, pp. 824-831 (1994).

[11] A.V. Oppenheim, A.S. Willsky and I.T. Young, *Signals and Systems*, Prentice-Hall International Editions (1983).

[12] P. Saint-Marc, J.-C. Jezouin and G. Medioni, A Versatile PC-Based Range Finding System, *IEEE Transactions on Robotics and Automation*, Vol. RA-7, no. 2, pp. 250–256 (1991).

[13] E. Trucco and R.B. Fisher, Acquisition of Consistent Range Data Using Direct Calibration, *Proc. IEEE Int. Conf. on Robotics and Automation*, San Diego, pp. 3410–3415 (1994).

[14] T. Wohlers, 3-D Digitizers, *Computer Graphics World*, July, pp. 73-77 (1992).

3 Dealing with Image Noise

The mariachis would serenade
And they would not shut up till they were paid.
Tom Lehrer, *In Old Mexico*

Attenuating or, ideally, suppressing image noise is important because any computer vision system begins by processing intensity values. This chapter introduces a few, basic noise models and filtering methods, which constitute an initial but useful toolkit for many practical situations.

Chapter Overview

Section 3.1 discusses the concept of noise and how to quantify it. It also introduces *Gaussian* and *impulsive noise*, and their effects on images.

Section 3.2 discusses some essential linear and a nonlinear filtering methods, aimed to attenuate random and impulsive noise.

What You Need to Know to Understand this Chapter

- The basics of signal theory: sampling theorem (Appendix, section A.3), Fourier transforms, and linear filtering.

3.1 Image Noise

Chapter 2 introduced the concept of acquisition noise, and suggested a method to estimate it. But, in general, the term *noise* covers much more.