Chapter 3
Image Formation

And since geometry is the right foundation of all painting, I have decided to teach its rudiments and principles to all youngsters eager for art...

– Albrecht Dürer, *The Art of Measurement, 1525*

This chapter introduces simple mathematical models of the image formation process. In a broad figurative sense, vision is the inverse problem of image formation: the latter studies how objects give rise to images, while the former attempts to use images to recover a description of objects in space. Therefore, designing vision algorithms requires first developing a suitable model of image formation. Suitable, in this context, does not necessarily mean physically accurate: the level of abstraction and complexity in modeling image formation must trade off physical constraints and mathematical simplicity in order to result in a manageable model (i.e. one that can be inverted with reasonable effort). Physical models of image formation easily exceed the level of complexity necessary and appropriate for this book, and determining the right model for the problem at hand is a form of engineering art.

It comes as no surprise, then, that the study of image formation has for centuries been in the domain of artistic reproduction and composition, more so than of mathematics and engineering. Rudimentary understanding of the geometry of image formation, which includes various models for projecting the three-dimensional world onto a plane (e.g., a canvas), is implicit in various forms of visual arts. The roots of formulating the geometry of image formation can be traced back to the work of Euclid in the fourth century B.C. Examples of partially
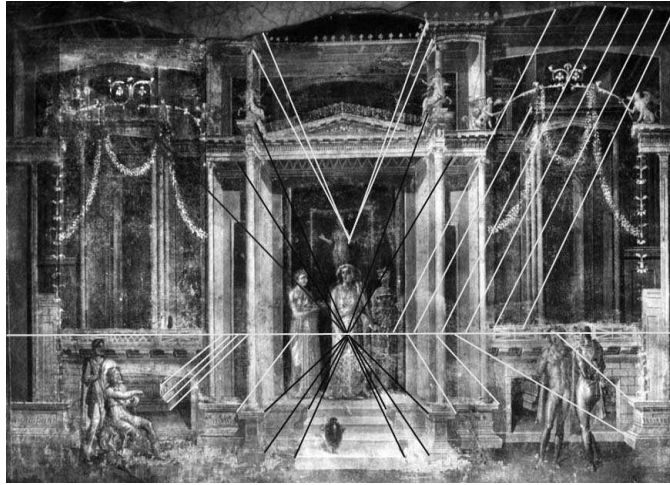
Figure 3.1. Frescoes from the first century B.C. in Pompeii. Partially correct perspective projection is visible in the paintings, although not all parallel lines converge to the vanishing point. The skill was lost during the middle ages, and it did not reappear in paintings until the Renaissance (image courtesy of C. Taylor).

correct perspective projection are visible in the frescoes and mosaics of Pompeii (Figure 3.1) from the first century B.C. Unfortunately, these skills seem to have been lost with the fall of the Roman empire, and it took over a thousand years for correct perspective projection to emerge in paintings again in the late fourteenth century. It was the early Renaissance painters who developed systematic methods for determining the perspective projection of three-dimensional landscapes. The first treatise on perspective, *Della Pictura*, was published by Leon Battista Alberti, who emphasized the "eye's view" of the world capturing correctly the geometry of the projection process. It is no coincidence that early attempts to formalize the rules of perspective came from artists proficient in architecture and engineering, such as Alberti and Brunelleschi. Geometry, however, is only a part of the image formation process: in order to obtain an image, we need to decide not only where to draw a point, but also what brightness value to assign to it. The interaction of light with matter is at the core of the studies of Leonardo Da Vinci in the 1500s, and his insights on perspective, shading, color, and even stereopsis are vibrantly expressed in his notes. Renaissance painters such as Caravaggio and Raphael exhibited rather sophisticated skills in rendering light and color that remain compelling to this day.[1]

In this book, we restrict our attention to the geometry of the scene, and therefore, we need a simple geometric model of image formation, which we derive

---

[1]There is some evidence that suggests that some Renaissance artists secretly used camera-like devices (*camera obscura*) [Hockney, 2001].

in this chapter. More complex photometric models are beyond the scope of this book; in the next two sections as well as in Appendix 3.A at the end of this chapter, we will review some of the basic notions of radiometry so that the reader can better evaluate the assumptions by which we are able to reduce image formation to a purely geometric process.

3.1   Representation of images

An *image*, as far as this book is concerned, is a two-dimensional brightness array.[2] In other words, it is a map I, defined on a compact region Ω of a two-dimensional surface, taking values in the positive real numbers. For instance, in the case of a camera, Ω is a planar, rectangular region occupied by the photographic medium or by the CCD sensor. So I is a function

$$I : \Omega \subset \mathbb{R}^2 \to \mathbb{R}_+; \quad (x, y) \mapsto I(x, y). \tag{3.1}$$

Such an image (function) can be represented, for instance, using the graph of I as in the example in Figure 3.2. In the case of a digital image, both the domain Ω and the range $\mathbb{R}_+$ are discretized. For instance, $\Omega = [1, 640] \times [1, 480] \subset \mathbb{Z}^2$, and $\mathbb{R}_+$ is approximated by an interval of integers $[0, 255] \subset \mathbb{Z}_+$. Such an image can be represented by an array of numbers as in Table 3.1.
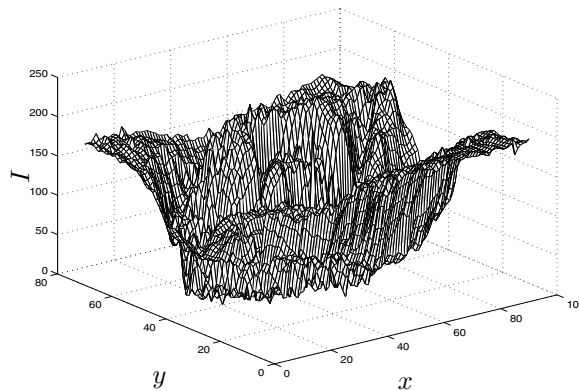

PSfrag replacements

Figure 3.2. An image I represented as a two-dimensional surface, the graph of I.

The values of the image I depend upon physical properties of the scene being viewed, such as its shape, its material reflectance properties, and the distribution of the light sources. Despite the fact that Figure 3.2 and Table 3.1 do not seem very indicative of the properties of the scene they portray, this is how they are represented in a computer. A different representation of the same image that is

---

[2]If it is a color image, its RGB (red, green, blue) values represent three such arrays.

```
188 186 188 187 168 130 101  99 110 113 112 107 117 140 153 153 156 158 156 153
189 189 188 181 163 135 109 104 113 113 110 109 117 134 147 152 156 163 160 156
190 190 188 176 159 139 115 106 114 123 114 111 119 130 141 154 165 160 156 151
190 188 188 175 158 139 114 103 113 126 112 113 127 133 137 151 165 156 152 145
191 185 189 177 158 138 110  99 112 119 107 115 137 140 135 144 157 163 158 150
193 183 178 164 148 134 118 112 119 117 118 106 122 139 140 152 154 160 155 147
185 181 178 165 149 135 121 116 124 120 122 109 123 139 141 154 156 159 154 147
175 176 176 163 145 131 120 118 125 123 125 112 124 139 142 155 158 158 155 148
170 170 172 159 137 123 116 114 119 122 126 113 123 137 141 156 158 159 157 150
171 171 173 157 131 119 116 113 114 118 125 113 122 135 140 155 156 160 160 152
174 175 176 156 128 120 121 118 113 112 123 114 122 135 141 155 155 158 159 152
176 174 174 151 123 119 126 121 112 108 122 115 123 137 143 156 155 152 155 150
175 169 168 144 117 117 127 122 109 106 122 116 125 139 145 158 156 147 152 148
179 179 180 155 127 121 118 109 107 113 125 133 130 129 139 153 161 148 155 157
176 183 181 153 122 115 113 106 105 109 123 132 131 131 142 157 157 149 156 159
180 181 177 147 115 110 111 107 107 105 120 132 133 133 141 150 154 148 155 157
181 174 170 141 113 111 115 112 113 105 119 130 132 134 144 153 156 148 152 151
180 172 168 140 114 114 118 113 112 107 119 128 130 134 146 157 162 153 153 148
186 176 171 142 114 114 116 110 108 104 116 125 128 134 148 161 165 159 157 149
185 178 171 138 109 110 114 110 109  97 110 121 127 136 150 160 163 158 156 150
```

Table 3.1. The image I represented as a two-dimensional matrix of integers (subsampled).

better suited for interpretation by the human visual system is obtained by generating a *picture*. A picture can be thought of as a scene different from the true one that produces on the imaging sensor (the eye in this case) the same image as the true one. In this sense *pictures* are "controlled illusions": they are scenes different from the true ones (they are *flat*) that produce in the eye the same image as the original scenes. A picture of the same image I described in Figure 3.2 and Table 3.1 is shown in Figure 3.3. Although the latter seems more *informative* as to the content of the scene, it is merely a different representation and contains exactly the same information.
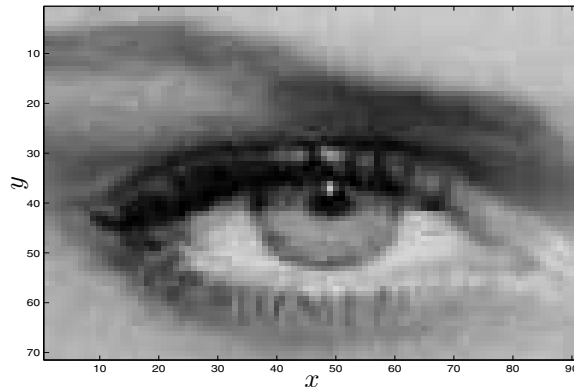

PSfrag replacements

Figure 3.3. A "picture" of the image I (compare with Figure 3.2 and Table 3.1).

3.2   Lenses, light, and basic photometry

In order to describe the image formation process, we must specify the value of $I(x, y)$ at each point (x, y) in Ω. Such a value $I(x, y)$ is typically called *image*

intensity or *brightness*, or more formally *irradiance*. It has the units of power per unit area (W/m^2) and describes the energy falling onto a small patch of the imaging sensor. The irradiance at a point of coordinates (x, y) is obtained by integrating energy both in time (e.g., the shutter interval in a camera, or the integration time in a CCD array) and in a region of space. The region of space that contributes to the irradiance at (x, y) depends upon the shape of the object (surface) of interest, the optics of the imaging device, and it is by no means trivial to determine. In Appendix 3.A at the end of this chapter, we discuss some common simplifying assumptions to approximate it.

3.2.1   *Imaging through lenses*

A camera (or in general an optical system) is composed of a set of lenses used to "direct" light. By directing light we mean a controlled change in the direction of propagation, which can be performed by means of diffraction, refraction, and reflection. For the sake of simplicity, we neglect the effects of diffraction and reflection in a lens system, and we consider only refraction. Even so, a complete description of the functioning of a (purely refractive) lens is well beyond the scope of this book. Therefore, we will consider only the simplest possible model, that of a *thin lens*. For a more germane model of light propagation, the interested reader is referred to the classic textbook [Born and Wolf, 1999].

A *thin lens* (Figure 3.4) is a mathematical model defined by an axis, called the *optical axis*, and a plane perpendicular to the axis, called the *focal plane*, with a circular aperture centered at the *optical center*, i.e. the intersection of the focal plane with the optical axis. The thin lens has two parameters: its *focal length f* and its *diameter d*. Its function is characterized by two properties. The first property is that all rays entering the aperture parallel to the optical axis intersect on the optical axis at a distance f from the optical center. The point of intersection is called the *focus* of the lens (Figure 3.4). The second property is that all rays through the optical center are undeflected. Consider a point $p \in \mathbb{E}^3$ not too far from the optical axis at a distance Z along the optical axis from the optical center. Now draw two rays from the point p: one parallel to the optical axis, and one through the optical center (Figure 3.4). The first one intersects the optical axis at the focus; the second remains undeflected (by the defining properties of the thin lens). Call x the point where the two rays intersect, and let z be its distance from the optical center. By decomposing any other ray from p into a component ray parallel to the optical axis and one through the optical center, we can argue that all rays from p intersect at x on the opposite side of the lens. In particular, a ray from x parallel to the optical axis, must go through p. Using similar triangles, from Figure 3.4, we obtain the following *fundamental equation of the thin lens*:

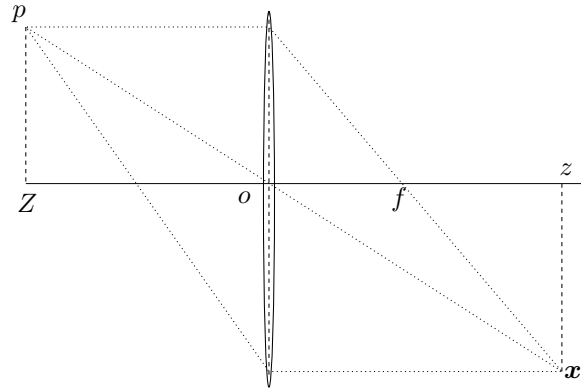$$\boxed{\frac{1}{Z} + \frac{1}{z} = \frac{1}{f}.}$$

PSfrag replacements

Figure 3.4. The image of the point p is the point $\boldsymbol{x}$ at the intersection of rays going parallel to the optical axis and the ray through the optical center.

The point $\boldsymbol{x}$ will be called the *image*[3] of the point p. Therefore, under the assumption of a thin lens, the irradiance $I(\boldsymbol{x})$ at the point $\boldsymbol{x}$ with coordinates (x, y) on the image plane is obtained by integrating all the energy emitted from the region of space contained in the cone determined by the geometry of the lens, as we describe in Appendix 3.A.

3.2.2   *Imaging through a pinhole*

If we let the aperture of a thin lens decrease to zero, all rays are forced to go through the optical center o, and therefore they remain undeflected. Consequently, the aperture of the cone decreases to zero, and the only points that contribute to the irradiance at the image point $\boldsymbol{x} = [x, y]^T$ are on a line through the center o of the lens. If a point p has coordinates $\boldsymbol{X} = [X, Y, Z]^T$ relative to a reference frame centered at the optical center o, with its z-axis being the optical axis (of the lens), then it is immediate to see from similar triangles in Figure 3.5 that the coordinates of p and its image $\boldsymbol{x}$ are related by the so-called ideal *perspective projection*

$$x = -f\frac{X}{Z}, \quad y = -f\frac{Y}{Z}, \tag{3.2}$$

where f is referred to as the *focal length*. Sometimes, we simply write the projection as a map π:

$$\pi : \mathbb{R}^3 \to \mathbb{R}^2; \quad \boldsymbol{X} \mapsto \boldsymbol{x}. \tag{3.3}$$

We also often write $\boldsymbol{x} = \pi(\boldsymbol{X})$. Note that any other point on the line through o and p projects onto the same coordinates $\boldsymbol{x} = [x, y]^T$. This imaging model is called an *ideal pinhole camera* model. It is an idealization of the thin lens model, since

---

[3]Here the word "image" is to be distinguished from the irradiance image $I(\boldsymbol{x})$ introduced before. Whether "image" indicates $\boldsymbol{x}$ or $I(\boldsymbol{x})$ will be made clear by the context.
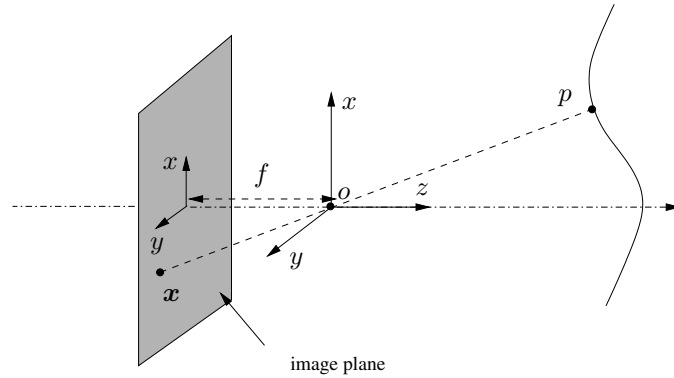
Figure 3.5. Pinhole imaging model: The image of the point p is the point $\boldsymbol{x}$ at the intersection of the ray going through the optical center o and an image plane at a distance f away from the optical center.

when the aperture decreases, diffraction effects become dominant, and therefore the (purely refractive) thin lens model does not hold [Born and Wolf, 1999]. Furthermore, as the aperture decreases to zero, the energy going through the lens also becomes zero. Although it is possible to actually build devices that approximate pinhole cameras, from our perspective the pinhole model will be just a good geometric approximation of a well-focused imaging system.

Notice that there is a negative sign in each of the formulae (3.2). This makes the image of an object appear to be upside down on the image plane (or the retina). To eliminate this effect, we can simply flip the image: $(x, y) \mapsto (-x, -y)$. This corresponds to placing the image plane $\{z = -f\}$ in front of the optical center instead $\{z = +f\}$. In this book we will adopt this more convenient "frontal" pinhole camera model, illustrated in Figure 3.6. In this case, the image $\boldsymbol{x} = [x, y]^T$ of the point p is given by

$$x = f\frac{X}{Z}, \quad y = f\frac{Y}{Z}. \tag{3.4}$$

We often use the same symbol, $\boldsymbol{x}$, to denote the homogeneous representation $[fX/Z, fY/Z, 1]^T \in \mathbb{R}^3$, as long as the dimension is clear from the context.[4]

In practice, the size of the image plane is usually limited, hence not every point p in space will generate an image $\boldsymbol{x}$ inside the image plane. We define the *field of view* (FOV) to be the angle subtended by the spatial extent of the sensor as seen from the optical center. If $2r$ is the largest spatial extension of the sensor (e.g., the

---

[4]In the homogeneous representation, it is only the direction of the vector $\boldsymbol{x}$ that is important. It is not crucial to normalize the last entry to 1 (see Appendix 3.B). In fact, $\boldsymbol{x}$ can be represented by $\lambda \boldsymbol{X}$ for any nonzero $\lambda \in \mathbb{R}$ as long as we remember that any such vector uniquely determines the intersection of the image ray and the actual image plane, in this case $\{Z = f\}$.
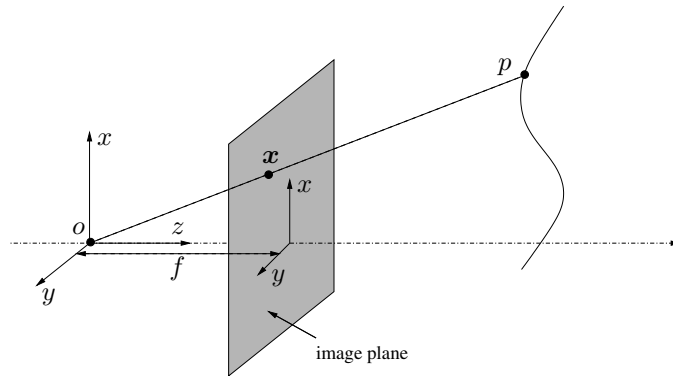
Figure 3.6. Frontal pinhole imaging model: the image of a 3-D point p is the point $\boldsymbol{x}$ at the intersection of the ray going through the optical center o and the image plane at a distance f in front of the optical center.

side of the CCD), then the field of view is $\theta = 2\arctan(r/f)$. Notice that if a flat plane is used as the image plane, the angle θ is always less than $180°$.[5]

In Appendix 3.A we give a concise description of a simplified model to determine the intensity value of the image at the position $\boldsymbol{x}$, $I(\boldsymbol{x})$. This depends upon the ambient light distribution, the material properties of the visible surfaces and, their geometry. There we also show under what conditions this model can be reduced to a purely geometric one, where the intensity measured at a pixel is identical to the amount of energy radiated at the corresponding point in space, independent of the vantage point, e.g., a Lambertian surface. Under these conditions, the image formation process can be reduced to tracing rays from surfaces in space to points on the image plane. How to do so is explained in the next section.

3.3   A geometric model of image formation

As we have mentioned in the previous section and we elaborate further in Appendix 3.A, under the assumptions of a pinhole camera model and Lambertian surfaces, one can essentially reduce the process of image formation to tracing rays from points on objects to pixels. That is, knowing which point in space projects onto which point on the image plane allows one to directly associate the radiance at the point to the irradiance of its image; see equation (3.36) in Appendix 3.A. In order to establish a precise correspondence between points in 3-D space (with respect to a fixed global reference frame) and their projected images in a 2-D image plane (with respect to a local coordinate frame), a mathematical model for this process must account for three types of transformations:

---

[5]In case of a spherical or ellipsoidal imaging surface, common in omnidirectional cameras, the field of view can often exceed $180°$.

1. coordinate transformations between the camera frame and the world frame;

2. projection of 3-D coordinates onto 2-D image coordinates;

3. coordinate transformation between possible choices of image coordinate frame.

In this section we will describe such a (simplified) image formation process as a series of transformations of coordinates. Inverting such a chain of transformations is generally referred to as "camera calibration," which is the subject of Chapter 6 and also a key step to 3-D reconstruction.

3.3.1 An ideal perspective camera

Let us consider a generic point p, with coordinates $\boldsymbol{X}_0 = [X_0, Y_0, Z_0]^T \in \mathbb{R}^3$ relative to the world reference frame.[6] As we know from Chapter 2, the coordinates $\boldsymbol{X} = [X, Y, Z]^T$ of the same point p relative to the camera frame are given by a rigid-body transformation $g = (R, T)$ of $\boldsymbol{X}_0$:

$$\boldsymbol{X} = R\boldsymbol{X}_0 + T \quad \in \mathbb{R}^3.$$

Adopting the frontal pinhole camera model introduced in the previous section (Figure 3.6), we see that the point $\boldsymbol{X}$ is projected onto the image plane at the point

$$\boldsymbol{x} = \begin{bmatrix} x \\ y \end{bmatrix} = \frac{f}{Z} \begin{bmatrix} X \\ Y \end{bmatrix}.$$

In homogeneous coordinates, this relationship can be written as

$$Z \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \begin{bmatrix} f & 0 & 0 & 0 \\ 0 & f & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix}. \tag{3.5}$$

We can rewrite the above equation equivalently as

$$Z\boldsymbol{x} = \begin{bmatrix} f & 0 & 0 & 0 \\ 0 & f & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \boldsymbol{X}, \tag{3.6}$$

where $\boldsymbol{X} \doteq [X, Y, Z, 1]^T$ and $\boldsymbol{x} \doteq [x, y, 1]^T$ are now in homogeneous representation. Since the coordinate Z (or the depth of the point p) is usually unknown, we may simply write it as an arbitrary positive scalar $\lambda \in \mathbb{R}_+$. Notice that in the

---

[6]We often indicate with $\boldsymbol{X}_0$ the coordinates of the point relative to the initial position of a moving camera frame.

above equation we can decompose the matrix into

$$\begin{bmatrix} f & 0 & 0 & 0 \\ 0 & f & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} = \begin{bmatrix} f & 0 & 0 \\ 0 & f & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}.$$

Define two matrices

$$K_f \doteq \begin{bmatrix} f & 0 & 0 \\ 0 & f & 0 \\ 0 & 0 & 1 \end{bmatrix} \in \mathbb{R}^{3\times 3}, \quad \Pi_0 \doteq \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \in \mathbb{R}^{3\times 4}. \qquad (3.7)$$

The matrix Π_0 is often referred to as the *standard (or "canonical") projection matrix*. From the coordinate transformation we have for $\boldsymbol{X} = [X, Y, Z, 1]^T$,

$$\begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix} = \begin{bmatrix} R & T \\ 0 & 1 \end{bmatrix} \begin{bmatrix} X_0 \\ Y_0 \\ Z_0 \\ 1 \end{bmatrix}. \qquad (3.8)$$

To summarize, using the above notation, the overall geometric model for *an ideal camera* can be described as

$$\lambda \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \begin{bmatrix} f & 0 & 0 \\ 0 & f & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} R & T \\ 0 & 1 \end{bmatrix} \begin{bmatrix} X_0 \\ Y_0 \\ Z_0 \\ 1 \end{bmatrix},$$

or in matrix form,

$$\lambda \boldsymbol{x} = K_f \Pi_0 \boldsymbol{X} = K_f \Pi_0 g \boldsymbol{X}_0. \qquad (3.9)$$

If the focal length f is known and hence can be normalized to 1, this model reduces to a Euclidean transformation g followed by a standard projection Π_0, i.e.

$$\boxed{\lambda \boldsymbol{x} = \Pi_0 \boldsymbol{X} = \Pi_0 g \boldsymbol{X}_0.} \qquad (3.10)$$

3.3.2 Camera with intrinsic parameters

The ideal model of equation (3.9) is specified relative to a very particular choice of reference frame, the "canonical retinal frame," centered at the optical center with one axis aligned with the optical axis. In practice, when one captures images with a digital camera the measurements are obtained in terms of pixels (i, j), with the origin of the image coordinate frame typically in the upper-left corner of the image. In order to render the model (3.9) usable, we need to specify the relationship between the retinal plane coordinate frame and the pixel array.

The first step consists in specifying the units along the x- and y-axes: if (x, y) are specified in terms of metric units (e.g., millimeters), and (x_s, y_s) are scaled
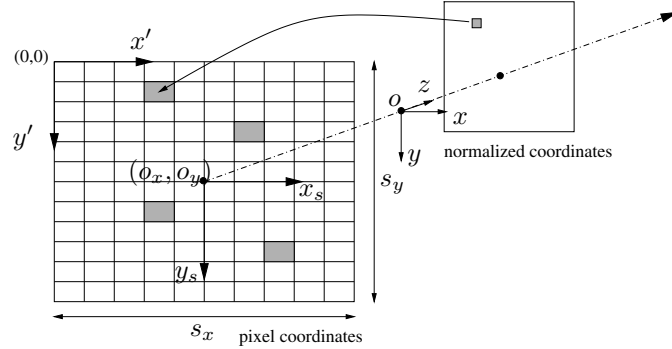
Figure 3.7. Transformation from normalized coordinates to coordinates in pixels.

versions that correspond to coordinates of the pixel, then the transformation can be described by a scaling matrix

$$\begin{bmatrix} x_s \\ y_s \end{bmatrix} = \begin{bmatrix} s_x & 0 \\ 0 & s_y \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \tag{3.11}$$

that depends on the size of the pixel (in metric units) along the x and y directions (Figure 3.7). When $s_x = s_y$, each pixel is square. In general, they can be different, and then the pixel is rectangular. However, here x_s and y_s are still specified relative to the *principal point* (where the z-axis intersects the image plane), whereas the pixel index (i, j) is conventionally specified relative to the upper-left corner, and is indicated by positive numbers. Therefore, we need to translate the origin of the reference frame to this corner (as shown in Figure 3.7),

$$x' = x_s + o_x,$$
$$y' = y_s + o_y,$$

where (o_x, o_y) are the coordinates (in pixels) of the principal point relative to the image reference frame. So the actual image coordinates are given by the vector $\boldsymbol{x}' = [x', y', 1]^T$ instead of the ideal image coordinates $\boldsymbol{x} = [x, y, 1]^T$. The above steps of coordinate transformation can be written in the homogeneous representation as

$$\boldsymbol{x}' \doteq \begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix} = \begin{bmatrix} s_x & 0 & o_x \\ 0 & s_y & o_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix}, \tag{3.12}$$

where x' and y' are actual image coordinates in pixels. This is illustrated in Figure 3.7. In case the pixels are not rectangular, a more general form of the scaling matrix can be considered,

$$\begin{bmatrix} s_x & s_\theta \\ 0 & s_y \end{bmatrix} \quad \in \mathbb{R}^{2 \times 2},$$

where s_θ is called a *skew factor* and is proportional to $\cot(\theta)$, where θ is the angle between the image axes x_s and y_s.[7] The transformation matrix in (3.12) then takes the general form

$$K_s \doteq \begin{bmatrix} s_x & s_\theta & o_x \\ 0 & s_y & o_y \\ 0 & 0 & 1 \end{bmatrix} \in \mathbb{R}^{3\times 3}. \tag{3.13}$$

In many practical applications it is common to assume that $s_\theta = 0$.

Now, combining the projection model from the previous section with the scaling and translation yields a more realistic model of a transformation between homogeneous coordinates of a 3-D point relative to the camera frame and homogeneous coordinates of its image expressed in terms of pixels,

$$\lambda \begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix} = \begin{bmatrix} s_x & s_\theta & o_x \\ 0 & s_y & o_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} f & 0 & 0 \\ 0 & f & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix}.$$

Notice that in the above equation, the effect of a real camera is in fact carried through two stages:

- The first stage is a standard perspective projection with respect to a *normalized coordinate system* (as if the focal length were $f = 1$). This is characterized by the standard projection matrix $\Pi_0 = [I, 0]$.

- The second stage is an additional transformation (on the obtained image x) that depends on parameters of the camera such as the focal length f, the scaling factors s_x, s_y, and s_θ, and the center offsets o_x, o_y.

The second transformation is obviously characterized by the combination of the two matrices K_s and K_f:

$$K \doteq K_s K_f \doteq \begin{bmatrix} s_x & s_\theta & o_x \\ 0 & s_y & o_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} f & 0 & 0 \\ 0 & f & 0 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} fs_x & fs_\theta & o_x \\ 0 & fs_y & o_y \\ 0 & 0 & 1 \end{bmatrix}. \tag{3.14}$$

The coupling of K_s and K_f allows us to write the projection equation in the following way:

$$\lambda x' = K\Pi_0 X = \begin{bmatrix} fs_x & fs_\theta & o_x \\ 0 & fs_y & o_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix}. \tag{3.15}$$

The constant 3×4 matrix Π_0 represents the perspective projection. The upper triangular 3×3 matrix K collects all parameters that are "intrinsic" to a particular camera, and is therefore called the *intrinsic parameter matrix*, or the *calibration*

---

[7]Typically, the angle θ is very close to $90°$, and hence s_θ is very close to zero.

matrix of the camera. The entries of the matrix K have the following geometric interpretation:

- o_x: x-coordinate of the principal point in pixels,

- o_y: y-coordinate of the principal point in pixels,

- $f s_x = \alpha_x$: size of unit length in horizontal pixels,

- $f s_y = \alpha_y$: size of unit length in vertical pixels,

- α_x / α_y: aspect ratio σ,

- $f s_\theta$: skew of the pixel, often close to zero.

Note that the height of the pixel is not necessarily identical to its width unless the aspect ratio σ is equal to 1.

When the calibration matrix K is known, the *calibrated* coordinates $\boldsymbol{x}$ can be obtained from the pixel coordinates $\boldsymbol{x}'$ by a simple inversion of K:

$$\lambda \boldsymbol{x} = \lambda K^{-1} \boldsymbol{x}' = \Pi_0 \boldsymbol{X} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix}. \tag{3.16}$$

The information about the matrix K can be obtained through the process of camera calibration to be described in Chapter 6. With the effect of K compensated for, equation (3.16), expressed in the normalized coordinate system, corresponds to the ideal pinhole camera model with the image plane located in front of the center of projection and the focal length f equal to 1.

To summarize, the geometric relationship between a point of coordinates $\boldsymbol{X}_0 = [X_0, Y_0, Z_0, 1]^T$ relative to the world frame and its corresponding image coordinates $\boldsymbol{x}' = [x', y', 1]^T$ (in pixels) depends on the rigid-body motion (R, T) between the world frame and the camera frame (sometimes referred to as the *extrinsic calibration parameters*), an ideal projection Π_0, and the camera intrinsic parameters K. The overall model for image formation is therefore captured by the following equation:

$$\lambda \begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix} = \begin{bmatrix} f s_x & f s_\theta & o_x \\ 0 & f s_y & o_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} R & T \\ 0 & 1 \end{bmatrix} \begin{bmatrix} X_0 \\ Y_0 \\ Z_0 \\ 1 \end{bmatrix}.$$

In matrix form, we write

$$\lambda \boldsymbol{x}' = K \Pi_0 \boldsymbol{X} = K \Pi_0 g \boldsymbol{X}_0, \tag{3.17}$$

or equivalently,

$$\lambda \boldsymbol{x}' = K \Pi_0 \boldsymbol{X} = [KR, KT] \boldsymbol{X}_0. \tag{3.18}$$

Often, for convenience, we call the 3×4 matrix $K \Pi_0 g = [KR, KT]$ a (general) *projection matrix* Π, to be distinguished from the standard projection matrix Π_0.

Hence, the above equation can be simply written as

$$\boxed{\lambda x' = \Pi X_0 = K\Pi_0 g X_0.}$$ (3.19)

Compared to the ideal camera model (3.10), the only change here is the standard projection matrix Π_0 being replaced by a general one Π.

At this stage, in order to explicitly see the nonlinear nature of the perspective projection equation, we can divide equation (3.19) by the scale λ and obtain the following expressions for the image coordinates (x', y', z'),

$$x' = \frac{\pi_1^T X_0}{\pi_3^T X_0}, \quad y' = \frac{\pi_2^T X_0}{\pi_3^T X_0}, \quad z' = 1,$$ (3.20)

where $\pi_1^T, \pi_2^T, \pi_3^T \in \mathbb{R}^4$ are the three rows of the projection matrix Π.

Example 3.1 (Spherical perspective projection). The perspective pinhole camera model outlined above considers planar imaging surfaces. An alternative imaging surface that is also commonly used is that of a sphere, shown in Figure 3.8.
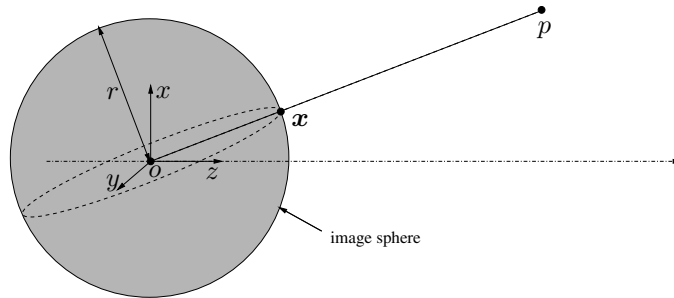
Figure 3.8. Spherical perspective projection model: the image of a 3-D point p is the point x at the intersection of the ray going through the optical center o and a sphere of radius r around the optical center. Typically r is chosen to be 1.

This choice is partly motivated by retina shapes often encountered in biological systems. For spherical projection, we simply choose the imaging surface to be the unit sphere $\mathbb{S}^2 = \{p \in \mathbb{R}^3 \mid \|X(p)\| = 1\}$. Then, the spherical projection is defined by the map π_s from $\mathbb{R}^3$ to $\mathbb{S}^2$:

$$\pi_s : \mathbb{R}^3 \to \mathbb{S}^2; \quad X \mapsto x = \frac{X}{\|X\|}.$$

As in the case of planar perspective projection, the relationship between pixel coordinates of a point and their 3-D metric counterpart can be expressed as

$$\lambda x' = K\Pi_0 X = K\Pi_0 g X_0,$$ (3.21)

where the scale is given by $\lambda = \sqrt{X^2 + Y^2 + Z^2}$ in the case of spherical projection while $\lambda = Z$ in the case of planar projection. Therefore, mathematically, spherical projection and planar projection can be described by the same set of equations. The only difference is that the unknown (depth) scale λ takes different values. ∎

For convenience, we often write $\boldsymbol{x} \sim \boldsymbol{y}$ for two (homogeneous) vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ equal up to a scalar factor (see Appendix 3.B for more detail). From the above example, we see that for any perspective projection we have

$$\boldsymbol{x}' \sim \Pi\boldsymbol{X}_0 = K\Pi_0 g\boldsymbol{X}_0, \tag{3.22}$$

and the shape of the imaging surface chosen does not matter. The imaging surface can be any (regular) surface as long as any ray $\overrightarrow{op}$ intersects with the surface at one point at most. For example, an entire class of ellipsoidal surfaces can be used, which leads to the so-called *catadioptric model* popular in many omnidirectional cameras. In principle, all images thus obtained contain exactly the same information.

3.3.3   Radial distortion

In addition to linear distortions described by the parameters in K, if a camera with a wide field of view is used, one can often observe significant distortion along radial directions. The simplest effective model for such a distortion is:

$$x = x_d(1 + a_1 r^2 + a_2 r^4),$$
$$y = y_d(1 + a_1 r^2 + a_2 r^4),$$

where (x_d, y_d) are coordinates of the distorted points, $r^2 = x_d^2 + y_d^2$ and a_1, a_2 are additional camera parameters that model the amount of distortion. Several algorithms and software packages are available for compensating radial distortion via calibration procedures. In particular, a commonly used approach is that of [Tsai, 1986a], if a calibration rig is available (see Chapter 6 for more details).

In case the calibration rig is not available, the radial distortion parameters can be estimated directly from images. A simple method suggested by [Devernay and Faugeras, 1995] assumes a more general model of radial distortion:

$$\boldsymbol{x} = c + f(r)(\boldsymbol{x}_d - c),$$
$$f(r) = 1 + a_1 r + a_2 r^2 + a_3 r^3 + a_4 r^4,$$

where $\boldsymbol{x}_d = [x_d, y_d]^T$ are the distorted image coordinates, $r^2 = \|\boldsymbol{x}_d - c\|^2$, $c = [c_x, c_y]^T$ is the center of the distortion, not necessarily coincident with the center of the image, and $f(r)$ is the distortion correction factor. The method assumes a set of straight lines in the world and computes the best parameters of the radial distortion model which would transform the curved images of the lines into straight segments. One can use this model to transform Figure 3.9 (left) into 3.9 (right) via preprocessing algorithms described in [Devernay and Faugeras, 1995]. Therefore, in the rest of this book we assume that radial distortion has been compensated for, and a camera is described simply by the parameter matrix K. The interested reader may consult classical references such as [Tsai, 1986a, Tsai, 1987, Tsai, 1989, Zhang, 1998b], which are available as software packages. Some authors have shown that radial distortion

Figure 3.9. Left: image taken by a camera with a short focal length; note that the straight lines in the scene become curved on the image. Right: image with radial distortion compensated for.

can be recovered from multiple corresponding images: a simultaneous estimation of 3-D geometry and radial distortion can be found in the more recent work of [Zhang, 1996, Stein, 1997, Fitzgibbon, 2001]. For more sophisticated lens aberration models, the reader can refer to classical references in geometric optics given at the end of this chapter.

3.3.4 Image, preimage, and coimage of points and lines

The preceding sections have formally established the notion of a perspective image of a point. In principle, this allows us to define an image of any other geometric entity in 3-D that can be defined as a set of points (e.g., a line or a plane). Nevertheless, as we have seen from the example of spherical projection, even for a point, there exist seemingly different representations for its image: two vectors $x \in \mathbb{R}^3$ and $y \in \mathbb{R}^3$ may represent the same image point as long as they are related by a nonzero scalar factor; i.e. $x \sim y$ (as a result of different choices in the imaging surface). To avoid possible confusion that can be caused by such different representations for the same geometric entity, we introduce a few abstract notions related to the image of a point or a line.

Consider the perspective projection of a straight line L in 3-D onto the 2-D image plane (Figure 3.10). To specify a line in 3-D, we can typically specify a point p_o, called the base point, on the line and specify a vector v that indicates the direction of the line. Suppose that $\boldsymbol{X}_o = [X_o, Y_o, Z_o, 1]^T$ are the homogeneous coordinates of the base point p_o and $\boldsymbol{V} = [V_1, V_2, V_3, 0]^T \in \mathbb{R}^4$ is the homogeneous representation of v, relative to the camera coordinate frame. Then the (homogeneous) coordinates of any point on the line L can be expressed as

$$\boldsymbol{X} = \boldsymbol{X}_o + \mu \boldsymbol{V}, \quad \mu \in \mathbb{R}.$$
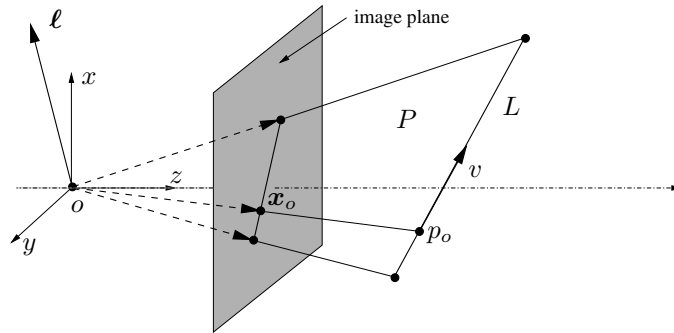
PSfrag replacements

Figure 3.10. Perspective image of a line L in 3-D. The collection of images of points on the line forms a plane P. Intersection of this plane and the image plane gives a straight line ℓ which is the image of the line.

Then, the image of the line L is given by the collection of image points with homogeneous coordinates given by

$$\boldsymbol{x} \; \sim \; \Pi_0 \boldsymbol{X} = \Pi_0 (\boldsymbol{X}_o + \mu \boldsymbol{V}) = \Pi_0 \boldsymbol{X}_o + \mu \Pi_0 \boldsymbol{V}.$$

It is easy to see that this collection of points $\{\boldsymbol{x}\}$, treated as vectors with origin at o, span a 2-D subspace P, shown in Figure 3.10. The intersection of this subspace with the image plane gives rise to a straight line in the 2-D image plane, also shown in Figure 3.10. This line is then the (physical) image of the line L.

Now the question is how to efficiently represent the image of the line. For this purpose, we first introduce the notion of preimage:

Definition 3.2 (Preimage). *A preimage of a point or a line in the image plane is the set of 3-D points that give rise to an image equal to the given point or line.*

Note that the given image is constrained to lie in the image plane, whereas the preimage lies in 3-D space. In the case of a point $\boldsymbol{x}$ on the image plane, its preimage is a one-dimensional subspace, spanned by the vector joining the point $\boldsymbol{x}$ to the camera center o. In the case of a line, the preimage is a plane P through o (hence a subspace) as shown in Figure 3.10, whose intersection with the image plane is exactly the given image line. Such a plane can be represented as the span of any two linearly independent vectors in the same subspace. Thus the preimage is really the largest set of 3-D points or lines that gives rise to the same image. The definition of a preimage can be given not only for points or lines in the image plane but also for curves or other more complicated geometric entities in the image plane as well. However, when the image is a point or a line, the preimage is a subspace, and we may also represent this subspace by its (unique) orthogonal complement in $\mathbb{R}^3$. For instance, a plane can be represented by its normal vector. This leads to the following notion of coimage:

Definition 3.3 (Coimage). *The coimage of a point or a line is defined to be the subspace in $\mathbb{R}^3$ that is the (unique) orthogonal complement of its preimage.*

The reader must be aware that the image, preimage, and coimage are *equivalent* representations, since they uniquely determine one another:

$$\text{image } = \text{ preimage } \cap \text{ image plane}, \quad \text{preimage } = \text{ span(image)},$$
$$\text{preimage } = \text{ coimage}^{\perp}, \quad \text{coimage } = \text{ preimage}^{\perp}.$$

Since the preimage of a line L is a two-dimensional subspace, its coimage is represented as the span of the normal vector to the subspace. The notation we use for this is $\boldsymbol{\ell} = [a, b, c]^T \in \mathbb{R}^3$ (Figure 3.10). If $\boldsymbol{x}$ is the image of a point p on this line, then it satisfies the orthogonality equation

$$\boldsymbol{\ell}^T \boldsymbol{x} = 0. \tag{3.23}$$

Recall that we use $\widehat{u} \in \mathbb{R}^{3\times 3}$ to denote the skew-symmetric matrix associated with a vector $u \in \mathbb{R}^3$. Its column vectors span the subspace orthogonal to the vector u. Thus the column vectors of the matrix $\widehat{\boldsymbol{\ell}}$ span the plane that is *orthogonal* to $\boldsymbol{\ell}$; i.e. they span the preimage of the line L. In Figure 3.10, this means that $P = \text{span}(\widehat{\boldsymbol{\ell}})$. Similarly, if $\boldsymbol{x}$ is the image of a point p, its coimage is the plane orthogonal to $\boldsymbol{x}$ given by the span of the column vectors of the matrix $\widehat{\boldsymbol{x}}$. Thus, in principle, we should use the notation in Table 3.2 to represent the image, preimage, or coimage of a point and a line.

Notation	Image	Preimage	Coimage
Point	$\text{span}(\boldsymbol{x}) \cap$ image plane	$\text{span}(\boldsymbol{x}) \subset \mathbb{R}^3$	$\text{span}(\widehat{\boldsymbol{x}}) \subset \mathbb{R}^3$
Line	$\text{span}(\widehat{\boldsymbol{\ell}}) \cap$ image plane	$\text{span}(\widehat{\boldsymbol{\ell}}) \subset \mathbb{R}^3$	$\text{span}(\boldsymbol{\ell}) \subset \mathbb{R}^3$

Table 3.2. The image, preimage, and coimage of a point and a line.

Although the (physical) image of a point or a line, strictly speaking, is a notion that depends on a particular choice of imaging surface, mathematically it is more convenient to use its preimage or coimage to represent it. For instance, we will use the vector $\boldsymbol{x}$, defined up to a scalar factor, to represent the preimage (hence the image) of a point; and the vector $\boldsymbol{\ell}$, defined up to a scalar factor, to represent the coimage (hence the image) of a line. The relationships between preimage and coimage of points and lines can be expressed in terms of the vectors $\boldsymbol{x}, \boldsymbol{\ell} \in \mathbb{R}^3$ as

$$\widehat{\boldsymbol{x}}\boldsymbol{x} = 0, \quad \widehat{\boldsymbol{\ell}}\boldsymbol{\ell} = 0.$$

Often, for a simpler language, we may refer to either the preimage or coimage of points and lines as the "image" if its actual meaning is clear from the context. For instance, in Figure 3.10, we will, in future chapters, often mark in the image plane the image of the line L by the same symbol $\boldsymbol{\ell}$ as the vector typically used to denote its coimage.

3.4   Summary

In this chapter, perspective projection is introduced as a model of the image formation for a pinhole camera. In the ideal case (e.g., when the calibration matrix K is the identity), homogeneous coordinates of an image point are related to their 3-D counterparts by an unknown (depth) scale λ,

$$\lambda \boldsymbol{x} = \Pi_0 \boldsymbol{X} = \Pi_0 g \boldsymbol{X}_0.$$

If K is not the identity, the standard perspective projection is augmented by an additional linear transformation K on the image plane

$$\boldsymbol{x}' = K\boldsymbol{x}.$$

This yields the following relationship between coordinates of an (uncalibrated) image and their 3-D counterparts:

$$\lambda \boldsymbol{x}' = K\Pi_0 \boldsymbol{X} = K\Pi_0 g \boldsymbol{X}_0.$$

As equivalent representations for an image of a point or a line, we introduced the notions of image, preimage, and coimage, whose relationships were summarized in Table 3.2.

3.5   Exercises

Exercise 3.1 Show that any point on the line through o and p projects onto the same image coordinates as p.

Exercise 3.2 Consider a thin lens imaging a plane parallel to the lens at a distance z from the focal plane. Determine the region of this plane that contributes to the image I at the point $\boldsymbol{x}$. (Hint: consider first a one-dimensional imaging model, then extend to a two-dimensional image.)

Exercise 3.3 (Field of view). An important parameter of the imaging system is the *field of view* (FOV). The field of view is twice the angle between the optical axis (z-axis) and the end of the retinal plane (CCD array). Imagine having a camera system with focal length 24 mm, and retinal plane (CCD array) (16 mm $\times$ 12 mm) and that your digitizer samples your imaging surface at 500×500 pixels in the horizontal and vertical directions.

1. Compute the FOV.

2. Write down the relationship between the image coordinate and a point in 3-D space expressed in the camera coordinate system.

3. Describe how the size of the FOV is related to the focal length and how it affects the resolution in the image.

4. Write a software program (in Matlab) that simulates the geometry of the projection process; given the coordinates of an object with respect to the calibrated camera frame, create an image of that object. Experiment with changing the parameters of the imaging system.

Exercise 3.4 Under the standard perspective projection (i.e. $K = I$):

1. What is the image of a sphere?

2. Characterize the objects for which the image of the centroid is the centroid of the image.

Exercise 3.5 (Calibration matrix). Compute the calibration matrix K that represents the transformation from image I to I' as shown in Figure 3.11. Note that from the definition of the calibration matrix, you need to use homogeneous coordinates to represent image points. Suppose that the resulting image I' is further digitized into an array of 640×480 pixels and the intensity value of each pixel is quantized to an integer in $[0, 255]$. Then how many *different* digitized images can one possibly get from such a process?
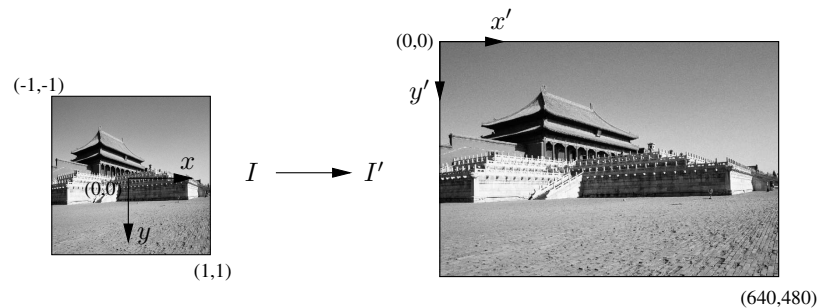

Figure 3.11. Transformation of a normalized image into pixel coordinates.

Exercise 3.6 (Image cropping). In this exercise, we examine the effect of cropping an image from a change of coordinate viewpoint. Compute the coordinate transformation between pixels (of same points) between the two images in Figure 3.12. Represent this transformation in homogeneous coordinates.
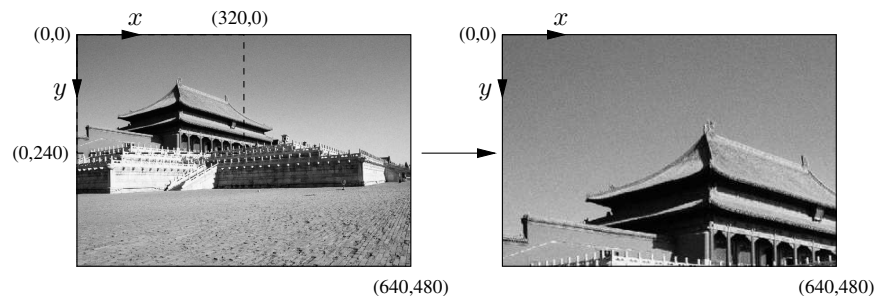

Figure 3.12. An image of size 640×480 pixels is cropped by half and then the resulting image is up-sampled and restored as a 640×480-pixel image.

Exercise 3.7 (Approximate camera models). The most commonly used approximation to the perspective projection model is *orthographic projection*. The light rays in the orthographic model travel along lines parallel to the optical axis. The relationship between

image points and 3-D points in this case is particularly simple: $x = X; y = Y$. So, the geometric model for orthographic projection can be expressed as

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \end{bmatrix}, \tag{3.24}$$

or simply in matrix form

$$\boldsymbol{x} = \Pi_o \boldsymbol{X}, \tag{3.25}$$

where $\Pi_o \doteq [I_{2\times2}, 0] \in \mathbb{R}^{2\times3}$. A scaled version of the orthographic model leads to the so-called *weak-perspective* model

$$\boldsymbol{x} = s\Pi_o \boldsymbol{X}, \tag{3.26}$$

where s is a constant scalar independent of the point $\boldsymbol{x}$. Show how the (scaled) orthographic projection approximates perspective projection when the scene occupies a volume whose diameter (or depth variation of the scene) is small compared to its distance from the camera. Characterize at least one more condition under which the two projection models produce similar results (equal in the limit).

Exercise 3.8 (Scale ambiguity). It is common sense that with a perspective camera, one cannot tell an object from another object that is exactly *twice as big but twice as far*. This is a classic ambiguity introduced by the perspective projection. Use the ideal camera model to explain why this is true. Is the same also true for the orthographic projection? Explain.

Exercise 3.9 (Image of lines and their intersection). Consider the image of a line L (Figure 3.10).

1. Show that there exists a vector in $\mathbb{R}^3$, call it $\boldsymbol{\ell}$, such that

$$\boldsymbol{\ell}^T \boldsymbol{x} = 0$$

   for the image $\boldsymbol{x}$ of every point on the line L. What is the geometric meaning of the vector $\boldsymbol{\ell}$? (Note that the vector $\boldsymbol{\ell}$ is defined only up to an arbitrary scalar factor.)

2. If the images of two points on the line L are given, say $\boldsymbol{x}^1, \boldsymbol{x}^2$, express the vector $\boldsymbol{\ell}$ in terms of $\boldsymbol{x}^1$ and $\boldsymbol{x}^2$.

3. Now suppose you are given two images of two lines, in the above vector form $\boldsymbol{\ell}^1, \boldsymbol{\ell}^2$. If $\boldsymbol{x}$ is the intersection of these two image lines, express $\boldsymbol{x}$ in terms of $\boldsymbol{\ell}^1, \boldsymbol{\ell}^2$.

Exercise 3.10 (Vanishing points). A straight line on the 3-D world is projected onto a straight line in the image plane. The projections of two parallel lines intersect in the image plane at the *vanishing point*.

1. Show that projections of parallel lines in 3-D space intersect at a point on the image.

2. Compute, for a given family of parallel lines, where in the image the vanishing point will be.

3. When does the vanishing point of the lines in the image plane lie at infinity (i.e. they do not intersect)?

The reader may refer to Appendix 3.B for a more formal treatment of vanishing points as well as their mathematical interpretation.

3.A Basic photometry with light sources and surfaces

In this section we give a concise description of a basic radiometric image formation model, and show that some simplifications are necessary in order to reduce the model to a purely geometric one, as described in this chapter. The idea is to describe how the intensity at a pixel on the image is generated. Under suitable assumptions, we show that such intensity depends only on the amount of energy radiated from visible surfaces in space and not on the vantage point.

Let S be a smooth visible surface in space; we denote the tangent plane to the surface at a point p by $T_p S$ and its outward unit normal vector by ν_p. At each point $p \in S$ we can construct a local coordinate frame with its origin at p, its z-axis parallel to the normal vector ν_p, and its xy-plane parallel to $T_p S$ (see Figure 3.13). Let L be a smooth surface that is irradiating light, which we call the *light source*. For simplicity, we may assume that L is the only source of light in space. At a point $q \in L$, we denote with $T_q S$ and ν_q the tangent plane and the outward unit normal of L, respectively, as shown in Figure 3.13.
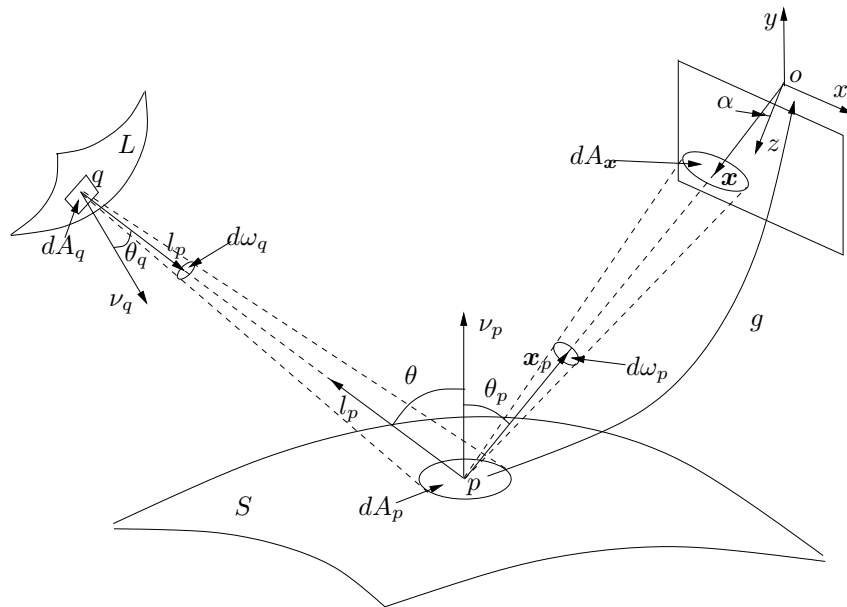

Figure 3.13. Generative model.

The change of coordinates between the local coordinate frame at p and the camera frame, which we assume coincides with the world frame, is indicated by a rigid-body transformation g; then g maps coordinates in the local coordinate

frame at p into those in the camera frame, and any vector u in the local coordinate frame to a vector $v = g_*(u)$ in the camera frame.[8]

Foreshortening and solid angle

When considering interactions between a light source and a surface, we need to introduce the notion of *foreshortening* and that of *solid angle*. Foreshortening encodes how the light distribution on a surface changes as we change the surface orientation with respect to the source of illumination. In formulas, if dA_p is the area element in T_pS, and l_p is the unit vector that indicates the direction from p to q (see Figure 3.13), then the corresponding foreshortened area as seen from q is

$$\cos(\theta)dA_p,$$

where θ is the angle between the direction l_p and the normal vector ν_p; i.e. $\cos(\theta) = \langle \nu_p, l_p \rangle$. A solid angle is defined to be the area of a cone cut out on a unit sphere. Then, the infinitesimal solid angle $d\omega_q$ seen from a point q of the infinitesimal area dA_p is

$$d\omega_q \doteq \frac{\cos(\theta)dA_p}{d(p,q)^2}, \tag{3.27}$$

where $d(p,q)$ is the distance between p and q.

Radiance and irradiance

In radiometry, *radiance* is defined to be the amount of energy emitted along a certain direction, per unit area perpendicular to the direction of emission (the foreshortening effect), per unit of solid angle, and per unit of time, following the definition in [Sillion, 1994]. According to our notation, if we denote the radiance at the point q in the direction of p by $\mathcal{R}(q, l_p)$, the energy emitted by the light L at a point q toward p on S is

$$dE(p, l_p) \doteq \mathcal{R}(q, l_p) \cos(\theta_q) \, dA_q \, d\omega_q \, dt, \tag{3.28}$$

where $\cos(\theta_q) \, dA_q$ is the foreshortened area of dA_q seen from the direction of p, and $d\omega_q$ is the solid angle given in equation (3.27), as shown in Figure 3.13. Notice that the point p on the left hand side of the equation above and the point q on the right hand side are related by the direction l_p of the vector connecting p to q.

While the radiance is used for energy that is emitted, the quantity that describes incoming energy is called *irradiance*. The irradiance is defined as the amount of energy received along a certain direction, per unit area and per unit time. Notice that in the case of the irradiance, we *do not* foreshorten the surface area as in the case of the radiance. Denote the irradiance at p received in the direction l_p by

---

[8]We recall from the previous chapter that if we represent the change of coordinates g with a rotation matrix $R \in SO(3)$ and a translation vector T, then the action of g on a point p of coordinates $\boldsymbol{X} \in \mathbb{R}^3$ is given by $g(\boldsymbol{X}) \doteq R\boldsymbol{X} + T$, while the action of g on a *vector* of coordinates u is given by $g_*(u) \doteq Ru$.

$dI(p, l_p)$. By energy preservation, we have $dI(p, l_p)\, dA_p\, dt = dE(p, l_p)$. Then the radiance $\mathcal{R}$ at a point q that illuminates the surface dA_p along the direction l_p with a solid angle $d\omega$ and the irradiance dI measured at the same surface dA_p received from this direction are related by

$$dI(p, l_p) = \mathcal{R}(q, l_p) \cos(\theta)\, d\omega, \tag{3.29}$$

where $d\omega = \frac{\cos(\theta_q)}{d(p,q)^2} dA_q$ is the solid angle of dA_q seen from p.

Bidirectional reflectance distribution function

For many common materials, the portion of energy coming from a direction l_p that is reflected onto a direction $\boldsymbol{x}_p$ (i.e. the direction of the vantage point) by the surface S, is described by $\beta(\boldsymbol{x}_p, l_p)$, the *bidirectional reflectance distribution function* (BRDF). Here both $\boldsymbol{x}_p$ and l_p are vectors expressed in local coordinates at p. More precisely, if $d\mathcal{R}(p, \boldsymbol{x}_p, l_p)$ is the radiance emitted in the direction $\boldsymbol{x}_p$ from the irradiance $dI(p, l_p)$, the BRDF is given by the ratio

$$\beta(\boldsymbol{x}_p, l_p) \doteq \frac{d\mathcal{R}(p, \boldsymbol{x}_p, l_p)}{dI(p, l_p)} = \frac{d\mathcal{R}(p, \boldsymbol{x}_p, l_p)}{\mathcal{R}(q, l_p) \cos(\theta)\, d\omega}. \tag{3.30}$$

To obtain the total radiance at a point p in the outgoing direction $\boldsymbol{x}_p$, we need to integrate the BRDF against all the incoming irradiance directions l_p in the hemisphere Ω at p:

$$\mathcal{R}(p, \boldsymbol{x}_p) = \int_\Omega d\mathcal{R}(p, \boldsymbol{x}_p, l_p) = \int_\Omega \beta(\boldsymbol{x}_p, l_p)\, \mathcal{R}(q, l_p) \cos(\theta)\, d\omega. \tag{3.31}$$

Lambertian surfaces

The above model can be considerably simplified if we restrict our attention to a class of materials, called *Lambertian*, that do not change appearance depending on the viewing direction. For example, matte surfaces are to a large extent well approximated by the Lambertian model, since they diffuse light almost uniformly in all directions. Metal, mirrors, and other shiny surfaces, however, do not. Figure 3.14 illustrates a few common surface properties.

For a perfect Lambertian surface, its radiance $\mathcal{R}(p, \boldsymbol{x}_p)$ only depends on how the surface faces the light source, but not on the direction $\boldsymbol{x}_p$ from which it is viewed. Therefore, $\beta(\boldsymbol{x}_p, l_p)$ is actually independent of $\boldsymbol{x}_p$, and we can think of the radiance function as being "glued," or "painted" on the surface S, so that at each point p the radiance $\mathcal{R}$ depends only on the surface. Hence, the perceived irradiance will depend only on which point on the surface is seen, not on in which direction it is seen. More precisely, for Lambertian surfaces, we have

$$\beta(\boldsymbol{x}_p, l_p) = \rho(p),$$

where $\rho(p) : \mathbb{R}^3 \mapsto \mathbb{R}_+$ is a scalar function. In this case, we can easily compute the *surface albedo* ρ_a, which is the percentage of incident irradiance reflected in
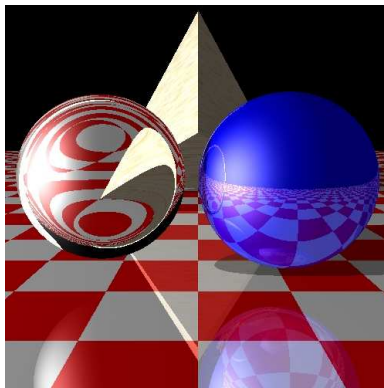
Figure 3.14. This figure demonstrates different surface properties widely used in computer graphics to model surfaces of natural objects: Lambertian, diffuse, reflective, specular (highlight), transparent with refraction, and textured. Only the (wood textured) pyramid exhibits Lambertian reflection. The ball on the right is partly ambient, diffuse, reflective and specular. The checkerboard floor is partly ambient, diffuse and reflective. The glass ball on the left is both reflective and refractive.

any direction, as

$$
\begin{aligned}
\rho_a(p) &= \int_\Omega \beta(\boldsymbol{x}_p, l_p) \cos(\theta_p)\, d\omega_p = \rho(p) \int_0^{2\pi} \int_0^{\frac{\pi}{2}} \cos(\theta_p) \sin(\theta_p)\, d\theta_p\, d\phi_p \\
&= \pi \rho(p),
\end{aligned}
$$

where $d\omega_p$, as shown in Figure 3.13, is the infinitesimal solid angle in the outgoing direction, which can be parameterized by the space angles (θ_p, ϕ_p) as $d\omega_p = \sin(\theta_p) d\theta_p d\phi_p$. Hence the radiance from the point p on a Lambertian surface S is

$$
\mathcal{R}(p) = \int_\Omega \frac{1}{\pi} \rho_a(p)\, \mathcal{R}(q, l_p) \cos(\theta)\, d\omega. \tag{3.32}
$$

This equation is known as *Lambertian cosine law*. Therefore, for a Lambertian surface, the radiance $\mathcal{R}$ depends only on the surface S, described by its generic point p, and on the light source L, described by its radiance $\mathcal{R}(q, l_p)$.

Image intensity for a Lambertian surface

In order to express the direction $\boldsymbol{x}_p$ in the camera frame, we consider the change of coordinates from the local coordinate frame at the point p to the camera frame: $\boldsymbol{X}(p) \doteq g(0)$ and $\boldsymbol{x} \sim g_*(\boldsymbol{x}_p)$, where we note that g_* is a rotation.[9] The reader should be aware that the transformation g itself depends on local shape of the

---

[9]The symbol $\sim$ indicates equivalence up to a scalar factor. Strictly speaking, $\boldsymbol{x}$ and $g_*(\boldsymbol{x}_p)$ do not represent the same vector, but only the same direction (they have opposite sign and different lengths). To obtain a rigorous expression, we would have to write $\boldsymbol{x} = \pi(-g_*(\boldsymbol{x}_p))$. However, these

surface at p, in particular its tangent plane T_pS and its normal ν_p at the point p. We now can rewrite the expression (3.31) for the radiance in terms of the camera coordinates and obtain

$$\mathcal{R}(\boldsymbol{X}) \doteq \mathcal{R}(p, g_*^{-1}(\boldsymbol{x})), \quad \text{where} \quad \boldsymbol{x} = \pi(\boldsymbol{X}). \tag{3.33}$$

If the surface is Lambertian, the above expression simplifies to

$$\mathcal{R}(\boldsymbol{X}) = \mathcal{R}(p). \tag{3.34}$$

Suppose that our imaging sensor is well modeled by a thin lens. Then, by measuring the amount of energy received along the direction $\boldsymbol{x}$, the irradiance (or image intensity) I at $\boldsymbol{x}$ can be expressed as a function of the radiance from the point p:

$$I(\boldsymbol{x}) = \mathcal{R}(\boldsymbol{X})\frac{\pi}{4}\left(\frac{d}{f}\right)^2 \cos^4(\alpha), \tag{3.35}$$

where d is the lens diameter, f is the focal length, and α is the angle between the optical axis (i.e. the z-axis) and the image point $\boldsymbol{x}$, as shown in Figure 3.13. The quantity $\frac{d}{f}$ is called the *F-number* of the lens. A detailed derivation of the above formula can be found in [Horn, 1986] (page 208). For a Lambertian surface, we have

$$\begin{aligned} I(\boldsymbol{x}) &= \mathcal{R}(\boldsymbol{X})\frac{\pi}{4}\left(\frac{d}{f}\right)^2\cos^4(\alpha) = \mathcal{R}(p)\frac{\pi}{4}\left(\frac{d}{f}\right)^2\cos^4(\alpha) \\ &= \frac{1}{4}\left(\frac{d}{f}\right)^2\cos^4(\alpha)\int_\Omega \rho_a(p)\,\mathcal{R}(q, l_p)\cos(\theta)\,d\omega, \end{aligned}$$

where $\boldsymbol{x}$ is the image of the point p taken at the vantage point g. Notice that in the above expression, only the angle α depends on the vantage point. In general, for a thin lens with a small field of view, α is approximately constant. Therefore, in our ideal pin-hole model, we may assume that the image intensity (i.e. irradiance) is related to the surface radiance by the *irradiance equation*:

$$\boxed{I(\boldsymbol{x}) = \gamma\mathcal{R}(p),} \tag{3.36}$$

where $\gamma \doteq \frac{\pi}{4}\left(\frac{d}{f}\right)^2\cos^4(\alpha)$ is a constant factor that is independent of the vantage point.

In all subsequent chapters we will adopt this simple model. The fact that the irradiance I does not change with the vantage point for Lambertian surfaces constitutes a fundamental condition that allows to establish correspondence across multiple images of the same object. This condition and its implications will be studied in more detail in the next chapter.

---

two vectors do represent the same ray through the camera center, and therefore we will regard them as the same.

3.B   Image formation in the language of projective geometry

The perspective pinhole camera model described by (3.18) or (3.19) has retained the physical meaning of all parameters involved. In particular, the last entry of both x' and X is normalized to 1 so that the other entries may correspond to actual 2-D and 3-D coordinates (with respect to the metric unit chosen for respective coordinate frames). However, such a normalization is not always necessary as long as we know that it is the direction of those homogeneous vectors that matters. For instance, the two vectors

$$[X, Y, Z, 1]^T, \quad [XW, YW, ZW, W]^T \quad \in \mathbb{R}^4 \tag{3.37}$$

can be used to represent the same point in $\mathbb{R}^3$. Similarly, we can use $[x', y', z']^T$ to represent a point $[x, y, 1]^T$ on the 2-D image plane as long as $x'/z' = x$ and $y'/z' = y$. However, we may run into trouble if the last entry W or z' happens to be 0. To resolve this problem, we need to generalize the interpretation of homogeneous coordinates introduced in the previous chapter.

Definition 3.4 (Projective space and its homogeneous coordinates). *An n-dimensional* projective space $\mathbb{P}^n$ *is the set of one-dimensional subspaces (i.e. lines through the origin) of the vector space $\mathbb{R}^{n+1}$. A point p in $\mathbb{P}^n$ can then be assigned* homogeneous coordinates $X = [x_1, x_2, \ldots, x_{n+1}]^T$ *among which at least one x_i is nonzero. For any nonzero $\lambda \in \mathbb{R}$ the coordinates $Y = [\lambda x_1, \lambda x_2, \ldots, \lambda x_{n+1}]^T$ represent the same point p in $\mathbb{P}^n$. We say that X and Y are equivalent, denoted by $X \sim Y$.*

Example 3.5 (Topological models for the projective space $\mathbb{P}^2$). Figure 3.15 demonstrates two equivalent geometric interpretations of the 2-D projective space $\mathbb{P}^2$. According
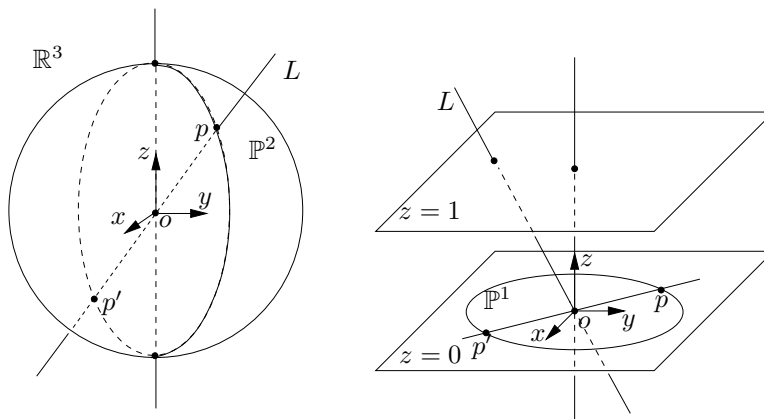
PSfrag replacements

Figure 3.15. Topological models for $\mathbb{P}^2$.

to the definition, it is simply a family of 1-D lines $\{L\}$ in $\mathbb{R}^3$ through a point o (typically chosen to be the origin of the coordinate frame). Hence, $\mathbb{P}^2$ can be viewed as a 2-D sphere $\mathbb{S}^2$ with any pair of antipodal points (e.g., p and p' in the figure) identified as one point in $\mathbb{P}^2$. On the right-hand side of Figure 3.15, lines through the center o in general intersect with the plane $\{z = 1\}$ at a unique point except when they lie on the plane $\{z = 0\}$. Lines in the plane $\{z = 0\}$ simply form the 1-D projective space $\mathbb{P}^1$ (which is in fact a circle). Hence, $\mathbb{P}^2$ can be viewed as a 2-D plane $\mathbb{R}^2$ (i.e. $\{z = 1\}$) with a circle $\mathbb{P}^1$ attached. If we adopt the view that lines in the plane $\{z = 0\}$ intersect the plane $\{z = 1\}$ infinitely far, this circle $\mathbb{P}^1$ then represents a *line at infinity*. Homogeneous coordinates for a point on this circle then take the form $[x, y, 0]^T$; on the other hand, all regular points in $\mathbb{R}^2$ have coordinates $[x, y, 1]^T$. In general, any projective space $\mathbb{P}^n$ can be visualized in a similar way: $\mathbb{P}^3$ is then $\mathbb{R}^3$ with a plane $\mathbb{P}^2$ attached at infinity; and $\mathbb{P}^n$ is $\mathbb{R}^n$ with $\mathbb{P}^{n-1}$ attached at infinity, which is, however, harder to illustrate on a piece of paper. ∎

Using this definition, $\mathbb{R}^n$ with its homogeneous representation can then be identified as a subset of $\mathbb{P}^n$ that includes exactly those points with coordinates $\boldsymbol{X} = [x_1, x_2, \ldots, x_{n+1}]^T$ where $x_{n+1} \neq 0$. Therefore, we can always normalize the last entry to 1 by dividing $\boldsymbol{X}$ by x_{n+1} if we so wish. Then, in the pinhole camera model described by (3.18) or (3.19), $\lambda \boldsymbol{x}'$ and $\boldsymbol{x}'$ now represent the same projective point in $\mathbb{P}^2$ and therefore the same 2-D point in the image plane. Suppose that the projection matrix is

$$\Pi = K\Pi_0 g = [KR, KT] \quad \in \mathbb{R}^{3 \times 4}. \tag{3.38}$$

Then the camera model simply reduces to a projection from a three-dimensional projective space $\mathbb{P}^3$ to a two-dimensional projective space $\mathbb{P}^2$,

$$\pi : \mathbb{P}^3 \to \mathbb{P}^2; \quad \boldsymbol{X}_0 \mapsto \boldsymbol{x}' \sim \Pi \boldsymbol{X}_0, \tag{3.39}$$

where λ is omitted here, since the equivalence "$\sim$" is defined in the homogeneous sense, i.e. up to a nonzero scalar factor.

Intuitively, the remaining points in $\mathbb{P}^3$ with the fourth coordinate $x_4 = 0$ can be interpreted as points that are "infinitely far away from the origin." This is because for a very small value ϵ, if we normalize the last entry of $\boldsymbol{X} = [X, Y, Z, \epsilon]^T$ to 1, it gives rise to a point in $\mathbb{R}^3$ with 3-D coordinates $\boldsymbol{X} = [X/\epsilon, Y/\epsilon, Z/\epsilon]^T$. The smaller $|\epsilon|$ is, the farther away is the point from the origin. In fact, all points with coordinates $[X, Y, Z, 0]^T$ form a two-dimensional plane described by the equation $[0, 0, 0, 1]^T \boldsymbol{X} = 0$.[10] This plane is called *plane at infinity*. We usually denote this plane by P_∞. That is,

$$P_\infty \doteq \mathbb{P}^3 \setminus \mathbb{R}^3 \; (= \mathbb{P}^2).$$

Then the above imaging model (3.39) is well-defined on the entire projective space $\mathbb{P}^3$ including points in this plane at infinity. This slight generalization allows us to talk about images of points that are infinitely far away from the camera.

---

[10] It is two-dimensional because X, Y, Z are not totally free: the coordinates are determined only up to a scalar factor.

Example 3.6 (Image of points at infinity and "vanishing points"). Two parallel lines in $\mathbb{R}^3$ do not intersect. However, we can view them as intersecting at infinity. Let $\boldsymbol{V} = [V_1, V_2, V_3, 0]^T \in \mathbb{R}^4$ be a (homogeneous) vector indicating the direction of two parallel lines L^1, L^2. Let $\boldsymbol{X}_o^1 = [X_o^1, Y_o^1, Z_o^1, 1]^T$ and $\boldsymbol{X}_o^2 = [X_o^2, Y_o^2, Z_o^2, 1]^T$ be two base points on the two lines, respectively. Then (homogeneous) coordinates of points on L^1 can be expressed as

$$\boldsymbol{X}^1 = \boldsymbol{X}_o^1 + \mu \boldsymbol{V}, \quad \mu \in \mathbb{R},$$

and similarly for points on L^2. Then the two lines can be viewed as intersecting at a point at infinity with coordinates $\boldsymbol{V}$. The "image" of this intersection, traditionally called a *vanishing point*, is simply given by

$$\boldsymbol{x}' \sim \Pi \boldsymbol{V}.$$

This can be shown by considering images of points on the lines and letting $\mu \to \infty$ asymptotically. If the images of these two lines are given, the image of this intersection can be easily computed or measured. Figure 3.16 shows the intersection of images of parallel lines at the vanishing point, a concept well known to Renaissance artists.   ∎
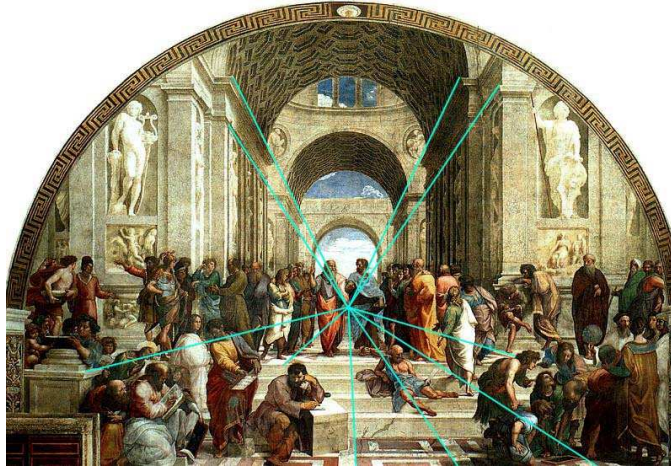

Figure 3.16. "The School of Athens" by Raphael (1518), a fine example of architectural perspective with a central vanishing point, marking the end of the classical Renaissance (courtesy of C. Taylor).

Example 3.7 (Image "outside" the image plane). Consider the standard perspective projection of a pair of parallel lines as in the previous example. We further assume that they are also parallel to the image plane, i.e. the xy-plane. In this case, we have

$$\Pi = \Pi_0 = [I, 0] \quad \text{and} \quad \boldsymbol{V} = [V_1, V_2, 0, 0]^T.$$

Hence, the "image" of the intersection is given in homogeneous coordinates as

$$\boldsymbol{x}' = [V_1, V_2, 0]^T.$$

This does not correspond to any physical point on the 2-D image plane (whose points supposedly have homogeneous coordinates of the form $[x, y, 1]^T$). It is, in fact, a vanishing point at infinity. Nevertheless, we can still treat it as a valid image point. One way is to view it as the image of a point with zero depth (i.e. with the z-coordinate zero). Such a problem will automatically go away if we choose the imaging surface to be an entire sphere rather than a flat plane. This is illustrated in Figure 3.17. ∎
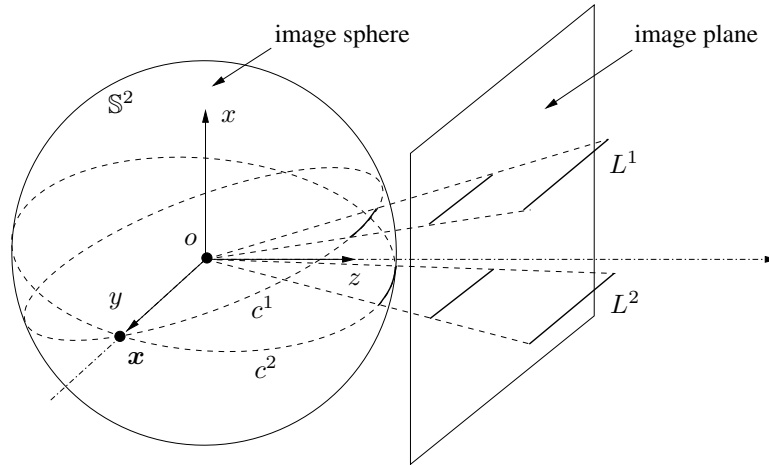
PSfrag replacements

Figure 3.17. Perspective images of two parallel lines that are also parallel to the 2-D image plane. In this case they are parallel to the y-axis. The two image lines on the image plane are also parallel, and hence they do not intersect. On an image sphere, however, the two image circles c^1 and c^2 do intersect at the point $\boldsymbol{x}$. Clearly, $\boldsymbol{x}$ is the direction of the two image lines.

Further readings

Deviations from the pinhole model

As we mentioned earlier in this chapter, the analytical study of pinhole perspective imaging dates back to the Renaissance. Nevertheless, the pinhole perspective model is a rather ideal approximation to actual CCD photosensors or film-based cameras. Before the pinhole model can be applied to such cameras, a correction is typically needed to convert them to an exact perspective device; see [Brank et al., 1993] and references therein.

In general, the pinhole perspective model is not adequate for modeling complex optical systems that involve a zoom lens or multiple lenses. For a systematic introduction to photographic optics and lens systems, we recommend the classic books [Stroebel, 1999, Born and Wolf, 1999]. For a more detailed account of models for a zoom lens, the reader may refer to [Horn, 1986, Lavest et al., 1993]

and references therein. Other approaches such as using a two-plane model [Wei and Ma, 1991] have also been proposed to overcome the limitations of the pinhole model.

Other simple camera models

In the computer vision literature, besides the pinhole perspective model, there exist many other types of simple camera models that are often used for modeling various imaging systems under different practical conditions. This book will *not* cover these cases. The interested reader may refer to [Tomasi and Kanade, 1992] for the study of the orthographic projection, to [Ohta et al., 1981, Aloimonos, 1990, Poelman and Kanade, 1997, Basri, 1996] for the study of the paraperspective projection, to [Konderink and van Doorn, 1991, Mundy and Zisserman, 1992], and [Quan and Kanade, 1996, Quan, 1996] for the study of the affine camera model, and to [Geyer and Daniilidis, 2001] and references therein for catadioptric models often used for omnidirectional cameras.