1 Image Formation

The images we process in computer vision are formed by light bouncing off surfaces in the world and into the lens of the system. The light then hits an array of sensors inside the camera. Each sensor produces electric charges that are read by an electronic circuit and converted to voltages. These are in turn sampled by a device called a digitizer (or analog-to-digital converter) to produce the numbers that computers eventually process, called pixel values. Thus, the pixel values are a rather indirect encoding of the physical properties of visible surfaces.

Is it not amazing that all those numbers in an image file carry information on how the properties of a packet of photons were changed by bouncing off a surface in the world? Even more amazing is that from this information we can perceive shapes and colors. Although we are used to these notions nowadays, the discovery of how images form, say, on our retinas, is rather recent. In ancient Greece, Euclid, in 300 B.C., attributed sight to the action of rectilinear rays issuing from the observer's eye, a theory that remained prevalent until the sixteenth Century when Johannes Kepler explained image formation as we understand it now. In Euclid's view, then, the eye is an active participant in the visual process. Not a receptor, but an agent that reaches out to apprehend its object. One of Euclid's postulates on vision maintained that any given object can be removed to a distance from which it will no longer be visible because it falls between adjacent visual rays. This is ray tracing in a very concrete, physical sense!

Studying image formation amounts to formulating models of the process that encodes the properties of light off a surface in the world into brightness values in the image array. We start from what happens once light leaves a visible surface. What happens thereafter is in fact a function only of the imaging device, if we assume that the medium in-between is transparent. In contrast, what happens at the visible surface, although definitely of great interest in computer vision, is so to speak out of our control, because it depends on the reflectance properties of the surface. In other words, reflectance is about the world, not about the imaging process.

The study of image formation can be further divided into what happens up to the point when light hits the sensor, and what happens thereafter. The first part occurs in the realm of optics, the second is a matter of electronics. We will look at the optics first and at what is called sensing (the electronic part) later.

1.1 Optics

A camera projects light from surfaces onto a two-dimensional sensor. Two aspects of this projection are of interest here: *where* light goes is the geometric aspect, and *how much* of it lands on the sensor is the photometric, or radiometric, aspect.

1.1.1 Projection Geometry

Our idealized model for the optics of a camera is the so-called *pinhole* camera model, for which we define the geometry of *perspective* projection. All rays in this model, as we will see, go through a small hole, and therefore form a star of lines.

For ever more distant scenes of fixed size, the rays of the star become more and more parallel to each other, and the *perspective* projection transformation performed by a pinhole camera tends to a limit called *orthographic* projection, where all rays are exactly parallel. Because orthographic projection is mathematically simpler than perspective, it is sometimes a more convenient and more reliable model to use. We will look at both the perspective projection of the pinhole camera and the orthographic projection model. Finally, we briefly sketch how real lenses behave differently from these idealized models.

The Pinhole Camera. A pinhole camera is a box with five opaque faces and a translucent one (Figure 1.1(a)). A very small hole is punched in the face of the box opposite to the translucent face. If you consider a single point in the world, such as the tip of the candle flame in the figure, only a thin beam from that point enters the pinhole and hits the translucent screen. Thus, the pinhole acts as a selector of light rays: without the pinhole and the box, any point on the screen would be illuminated from a whole hemisphere of directions, yielding a uniform coloring. With the pinhole, on the other hand, an inverted image of the visible world is formed on the screen. When the pinhole is reduced to a single point, this image is formed by the star of rays through the pinhole, intersected by the plane of the screen. Of course, a pinhole reduced to a point is an idealization: no power would pass through such a pinhole, and the image would be infinitely dim (black).

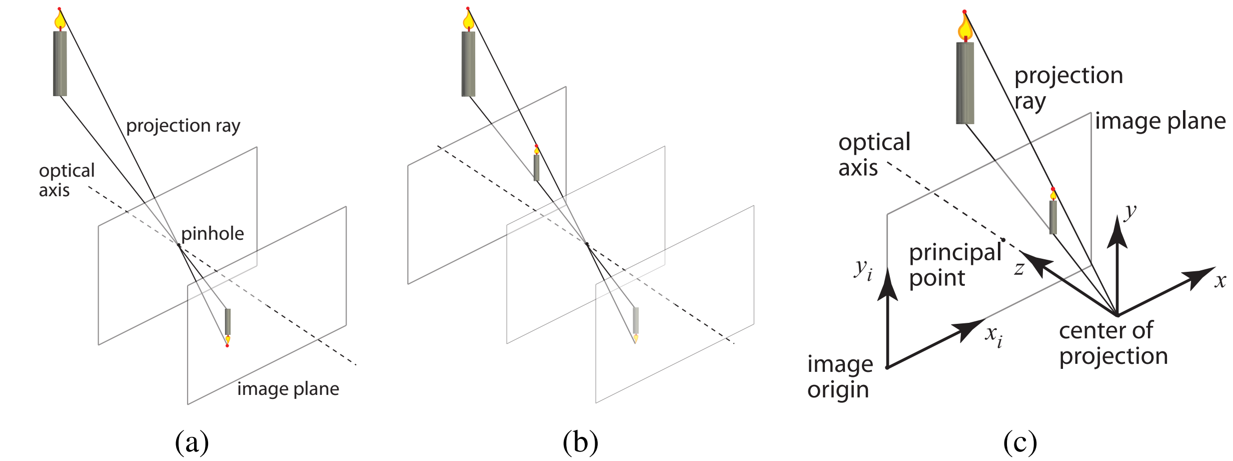


Figure 1.1: (a) Projection geometry for a pinhole camera. (b) If a screen could be placed in front of the pinhole, rather than behind, without blocking the projection rays, then the image would be upside-up. (c) What is left is the so-called *pinhole camera model*. The camera coordinate frame (x, y, z) is left-handed.

The fact that the image on the screen is inverted is mathematically inconvenient. It is therefore customary to consider instead the intersection of the star of rays through the pinhole with a plane parallel to the screen and *in front* of the pinhole as shown in Figure 1.1(b). This is of course an even greater idealization, since a screen in this position would block the light rays. The new image is isomorphic to the old one, but upside-up.

In this model, the pinhole is called more appropriately the *center of projection*. The front screen is the *image plane*. The distance between center of projection and image plane is the *focal distance*, and is denoted with f. The *optical axis* is the line through the center of projection that is perpendicular to the image plane. The point where the optical axis pierces the sensor plane is the *principal point*.

In keeping with standard conventions in computer graphics, the origin of the *image coordinate system* (x_i, y_i) is placed in the bottom left corner of the image. The *camera reference system* (x, y, z) axes are respectively parallel to x_i, y_i, and the optical axis, and the z axis points towards the scene. With the choice in Figure 1.1(c), the camera reference system is left-handed. The z coordinate of a point in the world is called the point's *depth*.

The units used to measure point coordinates in the camera reference system (x, y, z) are often different from those used in the image reference system (x_i, y_i). Typically, metric units (meters, centimeters, millimeters) are used in the camera system and pixels in the image system. As we will see in the Section on sensing below, pixels are the individual, rectangular elements on a digital camera's sensing array. Since pixels are not necessarily square, there may be a different number of pixels in a millimeter measured horizontally on the array than in a millimeter measured vertically, so two separate conversion units are needed to convert pixels to millimeters in the two directions.

Every point on the image plane has a z coordinate of f in the camera reference system. The image reference system, on the other hand, is two-dimensional, so the third coordinate is undefined. The other two coordinates differ by a translation and two separate unit conversions:

Let x_0 and y_0 be the coordinates in pixels of the principal point of the image in the image reference system (x_i, y_i). Then an image point with coordinates (x, y, f) in millimeters in the camera reference frame has image coordinates (in pixels)

$$x_i = s_x x + x_0 \quad \text{and} \quad y_i = s_y y + y_0 \tag{1}$$

where s_x and s_y are scaling constants expressed in pixels per milllimeter.

The *projection equations* relate the camera-system coordinates **P** $= (X, Y, Z)$ of a point in space to the camera-system coordinates **p** $= (x, y)$ of the projection of **P** onto the image plane and then, in turn, to the image-system coordinates $\mathbf{p}_i = (x_i, y_i)$ of the projection. These equations can be easily derived for the x coordinate from the top view of Figure 1.2. From this Figure we see that the triangle with orthogonal sides of length X and Z is similar to that with orthogonal sides of length x and f (the focal distance), so that $X/Z = x/f$. Similarly, for the Y coordinate, one gets $Y/Z = y/f$. In conclusion, Under perspective projection, the world point with coordinates (X, Y, Z) projects to the image point with coordinates

$$\begin{aligned} x &= f\frac{X}{Z} \\ y &= f\frac{Y}{Z} \,. \end{aligned} \tag{2}$$

One way to make units of measure consistent in these projection equations is to measure all quantities in the same unit, say, millimeters. In this case, the two constants s_x and s_y in equation (1) have the dimension of pixels per millimeter. However, it is sometimes more convenient to express x, y, and f in pixels (image dimensions) and X, Y, Z in millimeters (world dimensions). The ratios x/f, y/f, X/Z, and Y/Z are dimensionless, so the equations (2) are dimensionally consistent with this choice as well. In this case, the two constants s_x and s_y in equation (1) are dimensionless as well.

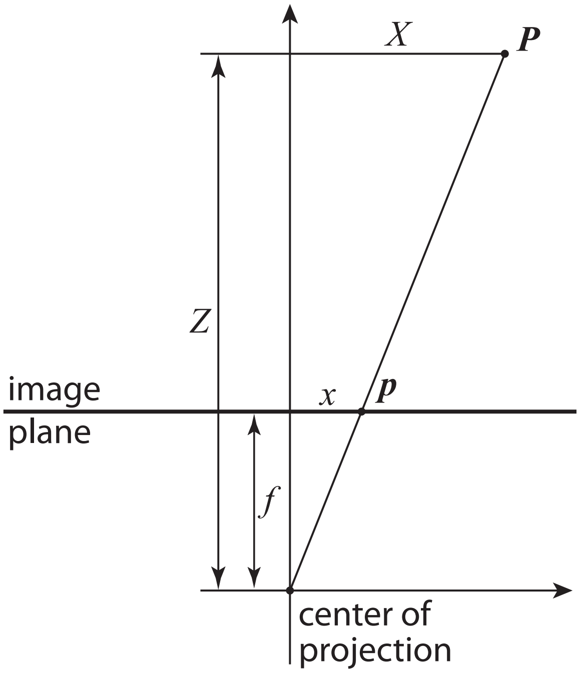


Figure 1.2: A top view of figure 1.1 (c).

Orthographic Projection. As the camera recedes and gets farther away from a scene of constant size, the projection rays become more parallel to each other. At the same time, the image becomes smaller, and eventually reduces to a point. To avoid image shrinking, one can magnify the image by Z_0/f, where Z_0 is the depth of, say, the centroid of all visible points, or that of an arbitrary point in the world. For the magnified coordinates x and y one then obtains

$$\begin{aligned} x &= X\frac{Z_0}{Z} \\ y &= Y\frac{Z_0}{Z} \; . \end{aligned}$$

As the camera recedes to infinity, Z and Z_0 grow at the same rate, and their ratio tends to 1. This situation, in which the projection rays are parallel to each other and orthogonal to the image plane, is called *orthographic* projection:

> Under orthographic projection, the world point with coordinates (X, Y, Z) projects to the image point with coordinates
>
> $$\begin{aligned} x &= X \\ y &= Y \; . \end{aligned} \tag{3}$$

The linearity of these projection equations makes orthographic projection an appealing assumption whenever warranted, that is, whenever a telephoto lens is used.

1.1.2 Lenses and Discrepancies from the Pinhole Model

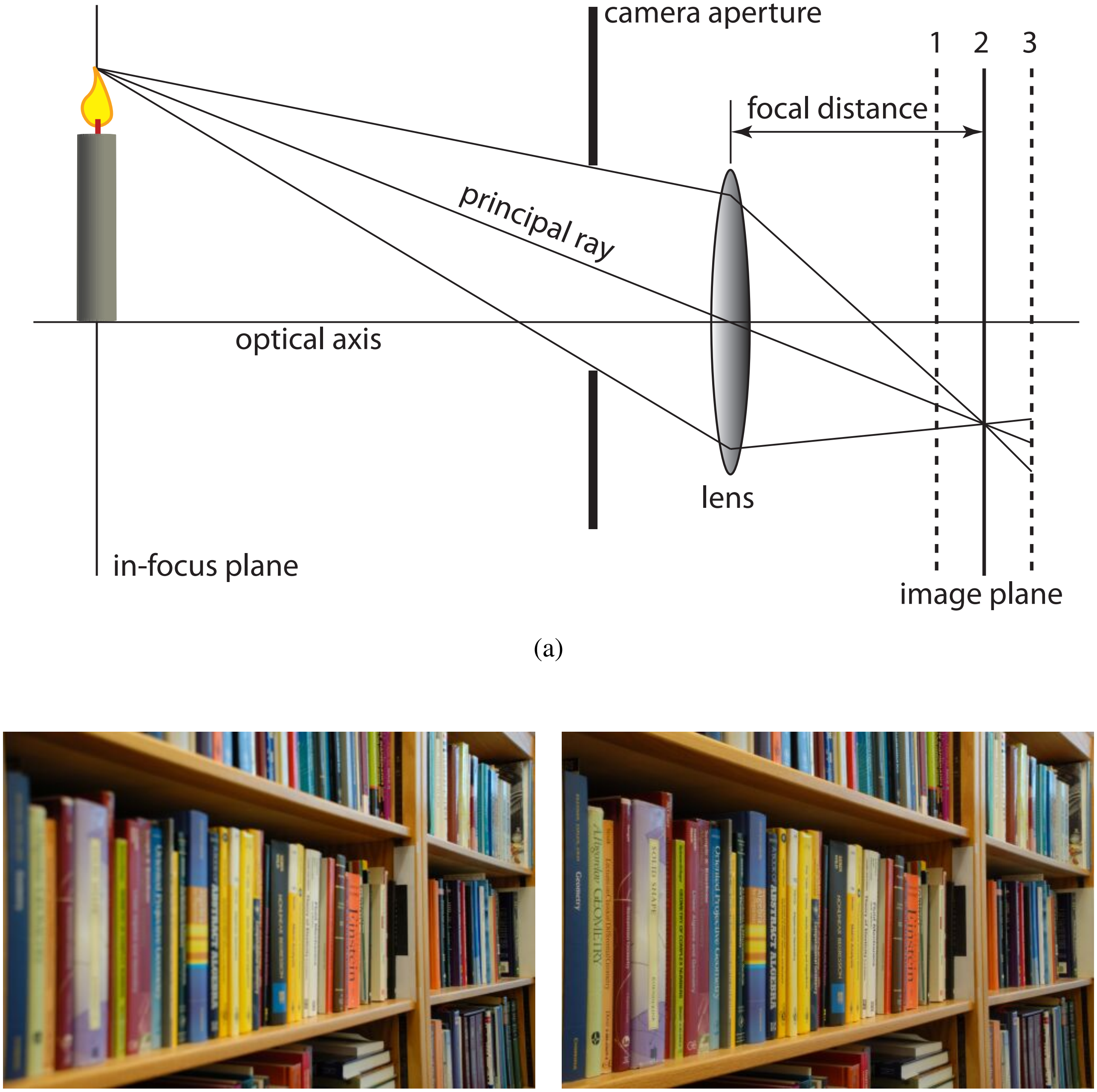
As pointed out above, the pinhole camera has a fundamental problem: if the pinhole is large, the image is blurred, and if it is small, the image is dim. When the diameter of the pinhole tends to zero, the image vanishes.[1] For this reason, lenses are used instead. Ideally, a lens gathers a whole cone of light from every point of a visible surface, and refocuses this cone onto a single point on the sensor. Unfortunately, lenses only approximate the geometry of a pinhole camera. The most obvious discrepancies concern focusing and distortion.

Focusing Figure 1.3 (a) illustrate the geometry of image focus. In front of the camera lens[2] there is a circular diaphragm of adjustable diameter called the *aperture*. This aperture determines the width of the cone of rays that hits the lens from any given point in the world.

Consider for instance the tip of the candle flame in the Figure. If the image plane is at the wrong distance (cases 1 and 3 in the Figure), the cone of rays from the candle tip intersects the image plane in an ellipse, which for usual imaging geometries is very close to a circle. This is called the *circle of confusion* for that point. When every point in the world projects onto a circle of confusion, the image appears to be blurred.

[1] In fact, blurring cannot be reduced at will, because of diffraction limits.

[2] Or inside the block of lenses, depending on various issues.



(a)

(b)

(c)

Figure 1.3: (a) If the image plane is at the correct focal distance (2), the lens focuses the entire cone of rays that the aperture allows through the lens onto a single point on the image plane. If the image plane is either too close (1) or too far (3) from the lens, the cone of rays from the candle tip intersects the image in a small ellipse (approximately a circle), producing a blurred image of the candle tip. (b) Image taken with a large aperture. Only a shallow range of depths is in focus. (c) Image taken with a small aperture. Everything is in focus.

For the image of the candle tip to be sharply focused, it is necessary for the lens to funnel all of the rays that the aperture allows through from that point onto a single point in the image. This condition is achieved by changing the focal distance, that is, the distance between the lens and the image plane. By studying the optics of light diffraction through the lens, it can be shown that the further the point in the world, the shorter the focal distance must be for sharp focusing. All distances are measured along the optical axis of the lens.

Since the correct focal distance depends on the distance of the world point from the lens, for any fixed focal distance, only the points on a single plane in the world are in focus. An image plane in position 1 in the Figure would focus points that are farther away than the candle, and an image plane in position 3 would focus points that are closer by. The dependence of focus on distance is visible in Figure 1.3(b): the lens was focused on the vertical, black and white stripe visible in the image, and the books that are closer are out of focus. The books that are farther away are out of focus as well, but by a lesser amount, since the effect of depth is not symmetric around the optimal focusing distance. Photographers say that the lens with the settings in Figure 1.3(b) has a *shallow* (or *narrow*) *depth of field*.

The depth of field can be increased, that is, the effects of poor focusing can be reduced, by making the lens aperture smaller. As a result, the cone of rays that hit the lens from any given point in the world becomes narrower, the circle of confusion becomes smaller, and the image becomes more sharply focused everywhere. This can be seen by comparing Figures 1.3 (b) and (c). Image (b) was taken with the lens aperture opened at its greatest diameter, resulting in a shallow depth of field. Image (c), on the other hand, was taken with the aperture closed down as much as possible for the given lens, resulting in a much greater depth of field: all books are in focus to the human eye. The price paid for a sharper image was exposure time: a small aperture lets little light through, so the imaging sensor had to be exposed longer to the incoming light: 1/8 of a second for image (b) and 5 seconds, forty times as long, for image (c).

Quantitative Aspects of Focusing. The focal distance at which a given lens focuses objects at infinite distance from the camera is called the *rear focal length* of the lens, or *focal length* for short.[3] All distances are measured from the center of the lens and along the optical axis. Note that the focal length is a lens property, which is usually printed on the barrel of the lens. In contrast, the focal distance is the distance between lens and image plane that a photographer selects to place a certain plane of the world in focus. So the focal distance varies even for the same lens.[4]

For a lens that is sufficiently thin, if f is the focal length, d the focal distance, and D the distance to a frontal[5] plane in the world, then the plane is in focus if the following *thin lens equation* is satisfied:

$$\frac{1}{D} + \frac{1}{d} = \frac{1}{f} \ . \tag{4}$$

[3] The *front focal length* is the converse: the distance to a world object that would be focused on an image plane at infinite distance from the lens.

[4] This has nothing to do with zooming. A zoom lens lets you change the focal length as well, that is, modify the optical properties of the lens.

[5] Orthogonal to the optical axis.

For instance, an object that is 2 meters (2000 millimeters) away from a lens with a focal length of 50 millimeters is in focus when the image plane is moved to a distance from the lens equal to the following:

$$d = \frac{1}{\frac{1}{D} + \frac{1}{f}} = \frac{1}{\frac{1}{2000} + \frac{1}{50}} \approx 48.78 \text{ mm} .$$

Consistently with the definition of focal length, when the distance D to the object goes to infinity we have

$$\lim_{D \to \infty} \frac{1}{D} + \frac{1}{d} = \frac{1}{d}$$

so that the thin lens equation yields

$$\frac{1}{d} = \frac{1}{f} \quad \text{that is,} \quad d = f .$$

[Make sure you understand why this makes the thin lens equation consistent with the definition of focal length.]

In photography, the aperture is usually measured in *stops*, or f*-numbers*. For a focal length f, an aperture of diameter a is said to have an f-number

$$n = \frac{f}{a} ,$$

so a large aperture has a small f-number. To remind one of this fact, apertures are often denoted with the notation f/n. For instance, the shallow depth of view image in Figure 1.3 (b) was obtained with a relatively wide aperture $f/4.2$, while the greater depth of field of the image in Figure 1.3 (c) was achieved with a much narrower aperture $f/29$.

Why use a wide aperture at all, if images can be made sharp with a small aperture? As was mentioned earlier, sharper images are darker, or require longer exposure times. In the example above, the ratio between the *areas* of the apertures is $(29/4.2)^2 \approx 48$. This is more or less consistent with the fact that the sharper image required forty times the exposure of the blurrier one: 48 times the area means that the lens focuses 48 times as much light on any given small patch on the image, and the exposure time can be decreased accordingly by a factor of 48. So, wide apertures are required for subjects that move very fast (for instance, in sports photography). In these cases, long exposure times are not possible, as they would lead to *motion blur*, a blur of a different origin (motion in the world) than poor focusing. Wide apertures are often aesthetically desirable also for static subjects, as they attract attention to what is in focus, at the expense of what is not. This is illustrated in Figure 1.4, from `http://www.hp.com/united-states/consumer/digital_photography/take_better_photos/tips/depth.html` .

In computer vision, image blurring has also been used as an asset, in systems that determine depth by measuring the amount of blur in different parts of the image. See for instance [4, 5].

Distortion. Even the high quality lens[6] used for the images in Figure 1.3 exhibits distortion. For instance, if you place a ruler along the vertical edge of the blue book on the far left of the

[6] Nikkor AF-S 18-135 zoom lens, used for both images (b) and (c).

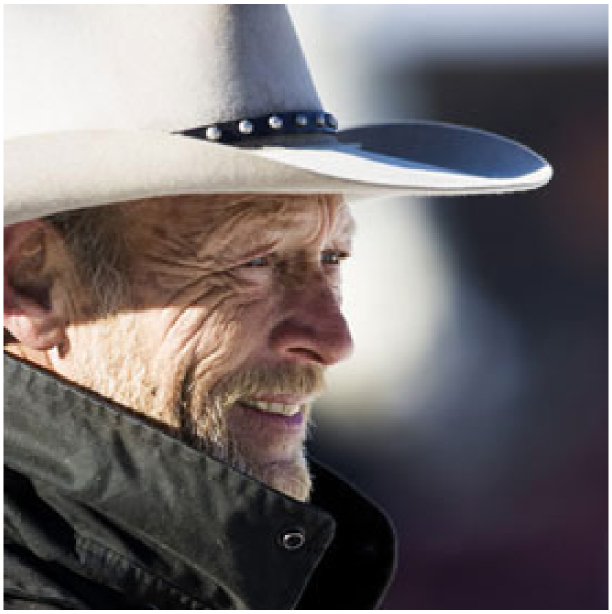


Figure 1.4: A shallow depth of field draw attention to what is in focus, at the expense of what is not.

Figure, you will notice that the edge is not straight. Curvature is visible also in the top shelf. This is geometric *pincushion distortion*. This type of distortion, illustrated in Figure 1.5(b), moves every point in the image away from the principal point, by an amount that is proportional to the square of the distance of the point from the principal point. The reverse type of distortion is called *barrel distortion*, and draws image points closer to the principal point by an amount proportional to the square of their distance from it. Because it moves image points towards or away from the principal point, both types of distortion are called *radial*. While non-radial distortion does occur, it is typically negligible in common lenses, and is henceforth ignored.

Distortion can be quite substantial, either by design (such as in non-perspective lenses like fisheye lenses) or to keep the lens inexpensive and with a wide field of view. Accounting for distortion is crucial in computer vision algorithms that use cameras as measuring devices, for instance, to reconstruct the three-dimensional shape of objects from two or more images of them.

Quantitative Aspects of Distortion. An excellent treatment of the mathematical theory that relates distortions to properties of light and lenses can be found in [1], but is beyond the scope of these notes. Lens designers must understand this theory. For vision, it suffices to note that distortion is necessarily a circularly symmetric function around the principal point of the image. This is because lenses are built by grinding glass that spins precisely around what becomes the lens' optical axis. This symmetry constrains the form that a mathematical description of distortion can take.

To understand this, let x and y be the first two camera-system coordinates of the image that an ideal, pinhole camera would form of some point in the world. If the pinhole is replaced by a lens

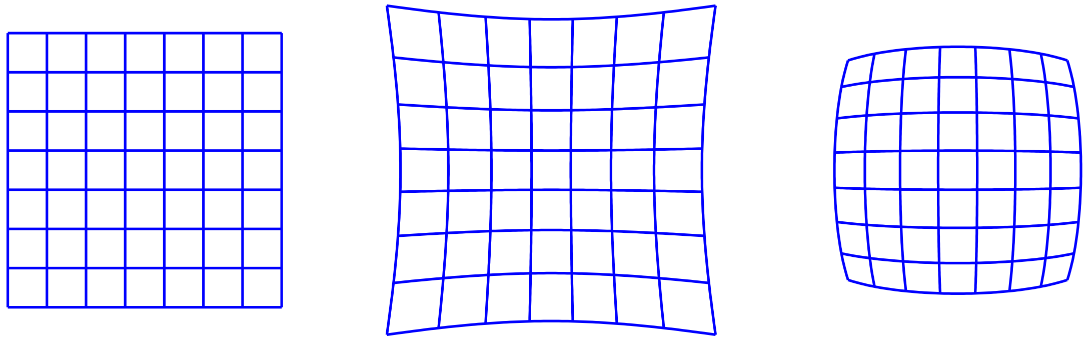


Figure 1.5: (a) An undistorted grid. (b) The grid in (a) with pincushion distortion. (c) The grid in (a) with barrel distortion.

with the same focal distance,[7] the same point in the world generally projects to a different point in the image, because of lens distortion. Let x_d and y_d be the coordinates of this new point.

Because of symmetry, distortion can only be a function of the distance

$$r = \sqrt{x^2 + y^2}$$

of the ideal image point from the principal point, and act in the same way on x and y:

$$x_d = x d(r) \quad \text{and} \quad y_d = y d(r)$$

where $d(r)$ is called the *distortion function*. In addition, and again because symmetry, the function $d(r)$ must be an even function of r, and can therefore be approximated with a polynomial whose odd-degree terms vanish. For most purposes in computer vision, a second- or fourth-order polynomial suffices[8]:

$$d(r) = 1 + k_2 r^2 + k_4 r^4 \ .$$

The images in Figure 1.5 all have $k_4 = 0$ (second-order distortion). The value of k_2 is 0 for (a), 0.1 for (b) (pincushion) and -0.1 for (c) (barrel). The values of k_2 and k_4 are determined for a particular lens through a procedure called *lens calibration*, or for a lens/camera combination through what is called *interior camera calibration*. This is the topic of a later Section.

The constant (zero-th order term) in the polynomial approximation for $d(r)$ must be 1, because distortion can be shown to vanish at the principal point for any symmetric lens. This is important: since distortion is continuous, if it is zero at the principal point it must be small in a sufficiently small neighborhood of it.

[7] More precisely, the *front nodal point* of the lens must be placed where the pinhole used to be. The front nodal point of a lens is a point on the optical axis and in front of the lens, defined by the property that any light ray that traverses it as it enters the lens leaves the lens in the same direction in space. The point where these rays leave the lens is called the *rear nodal point*. When the pinhole is replaced with a lens, the image plane needs to be moved away from the lens by the distance between the two nodal points, because the focal distance is measured from the rear focal point.

[8] Trucco and Verri in their book [6] approximate $1/d(r)$ instead. As we will see in the Section on calibration, this is less convenient.

Practical Aspects: Achieving Low Distortion. Since the linear term in $d(r)$ is zero, this neighborhood is typically fairly large. As a consequence, very low distortion can be obtained by mounting a lens designed for a large sensor onto a camera with a smaller sensor. The latter only sees the central portion of the field of view of the lens, where distortion is usually small.

For instance, lenses for the Nikon D200 used for Figure 1.3 are designed for a 23.6 by 15.8 millimeter sensor. Distortion is small but not negligible (see Figure 1.3 (c)) at the boundaries of the image when a sensor of this size is used. Distortion would be much smaller if the same lens were mounted onto a camera with what is called a "1/2 inch" sensor, which is really 6.4 by 4.8 millimeters in size, because the periphery of the lens would not be used. Lens manufacturers sell relatively inexpensive adaptors for this purpose. The real price paid for this reduction of distortion is a concomitant reduction of the camera's field of view (more on this in the Section on sensing below).

1.2 Radiometry

The other aspect of image formation, besides geometry, is radiometry, which describes how light is attenuated in different parts of the field of view and for world surfaces with different geometry and optical properties.

Radiometry became important in computer vision mainly through the seminal work of B. K. P. Horn [3, 2], who developed algorithms that reconstruct the shape of world surfaces from measurements of the variations in their brightness in the image ("shape from shading"). The qualitative aspects of these studies are of fundamental conceptual importance in understanding image formation. However, their quantitative aspects are of limited practical usefulness in computer vision, because their exploitation often implies knowing *a priori* unrealistically many facts about the sensing system and, more importantly, about the surfaces being viewed. Because of this, we only touch on radiometry here.

If we place a light source of fixed intensity at different points on a frontal plane and view it through an ideal lens, the apparent intensity of the light in the image will be diminished by a factor of $\cos^4 \alpha$ when the source is placed at angle α from the optical axis.

As illustrated in Figure 1.6, a drop of $\cos^2 \alpha$ is created because the source location is at distance D from the lens, rather than the distance $D \cos \alpha$ for a source on the same frontal plane but along the optical axis. Since brightness decays with the square of distance, this yields a first factor of $\cos^2 \alpha$. An additional factor of $\cos \alpha$ is introduced because the cone of light rays from the off-axis source enters the lens at an angle α, rather than hitting the lens head-on as light from the on-axis source would do. Finally, light from the off-axis source hits the image plane at an angle α, for an additional factor $\cos \alpha$.

If the lens has a 90 degree field of view, this drop-off means that the edges of the image will be only one fourth as bright as the center: $\alpha = 90/2 = 45$ degrees, $\cos \alpha = \sqrt{2}/2$, and $\cos^4 \alpha = 1/4$.

Practical Aspects: Good Images with Poor Lenses. Real lenses can cause further variation in illumination, called *vignetting*, for other reasons. A common solution computer vision researchers have used to sidestep both geometric and radiometric problems is to use only narrow-angle lenses, with fields of view less than 50 degrees, or, even better, to use only

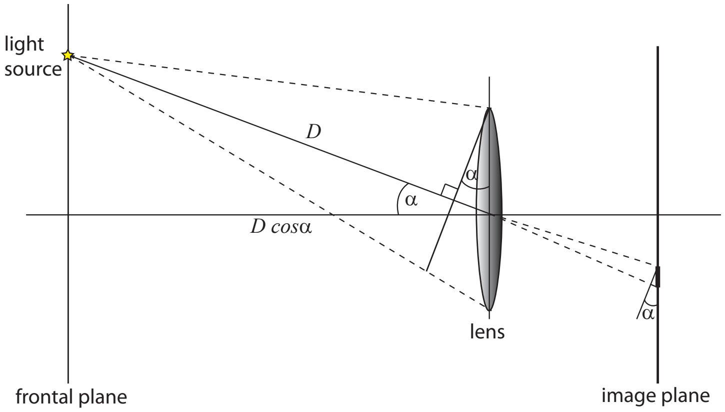


Figure 1.6: Factors that lead to a $\cos\alpha^4$ image brightness drop-off at an angle α away from the optical axis. See text for details.

a central part of an oversized lens with a small sensor. Both radial distortion and radiometric drop-off are then often insignificant. However, the lack of peripheral vision is a handicap for visual searching, navigation, and the detection of objects moving towards the observer, for which a wide field of view is desirable. In these cases, and if intensity values are of importance, the $\cos^4\alpha$ drop-off must be accounted for through calibration.

1.3 Sensing

In a digital camera, still or video, the light that hits the image plane is collected by one or more *sensors*, that is, rectangular arrays of sensing elements. Each element is called a *pixel* (for "picture element"). The finite overall extent of the sensor array, together with the presence of diaphragms in the lens, limits the cone (or pyramid) of directions from which light can reach pixels on the sensor. This cone is called the *field of view* of the camera-lens combination, described next more quantitatively. This Section then describes how pixels convert light intensities into voltages, and how these are in turn converted into numbers within the camera circuitry. This involves processes of integration (of light over the sensitive portion of each pixel), sampling (of the integral over time and at each pixel location), and addition of noise at all stages. These processes, as well as solutions for recording images in color, are then described in turn.

1.3.1 The Field of View

The field of view of a lens-sensor combination is determined by the focal distance f and by the size of the sensor. Figure 1.7 (a) shows a top view of the geometry. We have

$$\tan\frac{\phi}{2} = \frac{w/2}{f} \tag{5}$$

so that the horizontal angular width of the field of view is

$$\phi = 2 \arctan \frac{w}{2f} \ .$$

A similar expression holds for the vertical field of view:

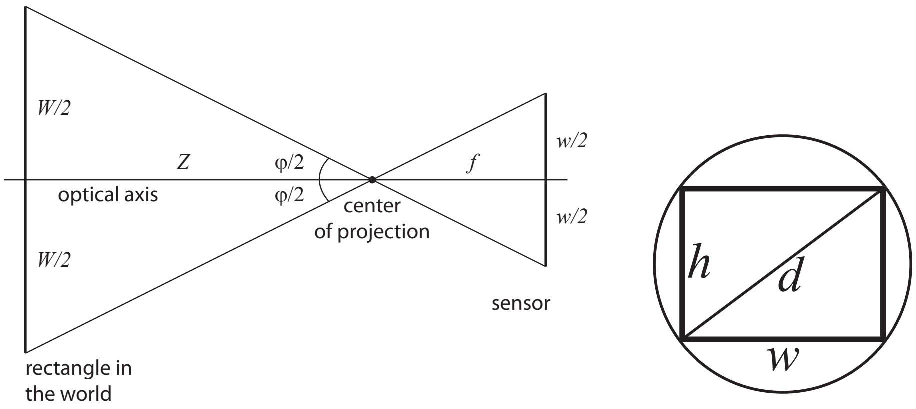
$$\phi' = 2 \arctan \frac{h}{2f} \tag{6}$$

where h is the height of the sensor. Since using two fields of view is inconvenient, the field of view is often specified as the vertex angle of the cone corresponding to the smallest circle that contains the sensor. The diameter of that circle is (see Figure 1.7 (b)):

$$d = \sqrt{w^2 + h^2}$$

so that the diagonal field of view is

$$\phi_d = 2 \arctan \frac{d}{2f} \ .$$



(a) (b)

Figure 1.7: (a) A sensor of width w at a focal distance f from the center of projection sees a rectangle of width W at distance Z. (b) The smallest circle containing a sensor of width w and height h.

Practical Aspects: Sensor Sizes. The aspect ratio

$$a = \frac{w}{h}$$

of most consumer and surveillance-grade cameras, both still and video, is 4:3. The high definition television standard specifies an aspect ratio of 16:9, which is comparatively wider.

Format (inches)	d (mm)	w (mm)	h (mm)
1/4	4	3.2	2.4
1/3	6	4.8	3.6
1/2	8	6.4	4.8
2/3	11	8.8	6.6
1	16	12.8	9.6

Table 1: Approximate dimensions of standard CCTV camera sensors. The diagonal d, width w, and height h of the sensor chip are shown in Figure 1.7 (b).

> Computer vision experimentation in the 20-th Century typically used surveillance-grade cameras, also known as Closed Circuit Television (CCTV) cameras. These were chosen for their low cost, for the availability of a large selection of lenses, and for the existence of frame grabbers, that is, computer peripheral devices that convert the analog signal from these cameras into an array of digital pixel intensities and copies the array to computer main memory. The sizes of CCTV sensors are specified in a rather arcane way, by giving the diameter in inches of the cathode-ray tube that a particular sensor was designed to replace. Table 1 lists typical CCTV sensor sizes.
>
> In the 21-st Century, digital cameras have become pervasive in both the consumer and professional markets as well as in computer vision research. SLR (Single-Lens Reflex) still cameras are the somewhat bulkier cameras with an internal mirror that lets the photographer view the exact image that the sensor will see once the shutter button is pressed (hence the name: a single lens with a mirror (reflex)). These have larger sensors than CCTV cameras have, typically about 24 by 16 millimeters, although some very expensive models have sensors as large as 36 by 24 millimeters. More modern CCTV cameras are similar to the old ones, but produce a digital rather than analog signal directly. This signal is transferred to computer through a digital connection such as USB, or, for high-bandwidth video, IEEE 1394 (also known as Firewire), or a Gigabit Ethernet connection.

As an example of use of the formulas and sizes above, suppose that we have a 1/2-inch camera sensor and we want a horizontal field of view ϕ of, say 50 degrees. We then find the horizontal sensor dimension, 4.8 mm, in Table 1, and solve equation (5) for the focal distance f:

$$f = \frac{w}{2\tan\phi/2} \; .$$

This yields a focal distance

$$f = \frac{4.8}{2\tan(50/2 \times \pi/180)} \approx 5.15 \text{ mm.}$$

Since focal length is focal distance when the lens is focused at infinity, we select a lens with a 5mm focal length (or, more practically, the nearest one we have to this value). When we focus at a finite distance, the focal distance increases somewhat (see equation (4)), and the field of view shrinks accordingly. However, this effect is very slight, and can usually be ignored. For instance, if we bring the subject to $D = 2$ meters (2000 millimeters) from the camera with an $f = 5$ millimeter lens, the thin-lens equation (4) yields a new focal distance

$$d = \frac{1}{\frac{1}{f} - \frac{1}{D}} = \frac{1}{\frac{1}{5} - \frac{1}{2000}} \approx 5.012 \text{ mm},$$

which is only about one quarter of one percent longer than 5 millimeters.

As another example, if we want to take a picture of someone's face at a distance of about 10 meters (10,000 millimeters, say, for a surveillance application), we want that person's face to fill the image for maximum resolution. If a head (with some margin) is $H = 25$ centimeters (250 millimeters) tall, we see from Figure 1.7 (a) (or rather its vertical analog) that we need a field of view

$$\phi' = 2\arctan\frac{H}{2Z} = 2\arctan\frac{250}{2 \times 10000} \approx 1.4 \text{ degrees.}$$

This is a very narrow field of view. With a standard SLR camera that has a 24 by 16 millimeter sensor, this would require a lens with a focal length that equation (6) shows to be

$$f = \frac{h}{2\tan\phi'/2} = \frac{16}{2\tan 0.7 \times \pi/180} \approx 655 \text{ mm},$$

a very long telephoto lens indeed (Canon makes a hugely expensive 1,200 mm lens for wildlife and sport photography; a more standard lens for these applications is 500 mm long). The lens would be proportionally shorter with one of the smaller CCTV sensors in Table 1. For instance, a 1/4 inch sensor would require a lens with focal length of about 100 millimeters.

1.3.2 Pixels

A *pixel* on a digital camera sensor is a small rectangle that contains a photosensitive element and some circuitry. The photosensitive element is called a *photodetector*, or light detector. It is a semiconductor junction placed so that light from the camera lens can reach it. When a photon strikes the junction, it creates an electron-hole pair with approximately 70 percent probability (this probability is called the *quantum efficiency* of the detector). If the junction is part of a polarized electric circuit, the electron moves towards the positive pole and the hole moves towards the negative pole. This motion constitutes an electric current, which in turn causes an accumulation of charge (one electron) in a capacitor. A separate circuit discharges the capacitor at the beginning of the *shutter* (or *exposure*) interval. The charge accumulated over this interval of time is proportional to the amount of light that struck the capacitor during exposure, and therefore to the brightness of the part of the scene that the lens focuses on the pixel in question. Longer shutter times or greater image brightness both translate to more accumulated charge, until the capacitor fills up completely ("saturates").

Practical Aspects: CCD and CMOS Sensors. Two methods are commonly used in digital cameras to read these capacitor charges: the CCD and the CMOS active sensor. The Charge-Coupled Device (CCD) is an electronic, analog shift register, and there is typically one shift

> register for each column of a CCD sensor. After the shutter interval has expired, the charges from all the pixels are transferred to the shift registers of their respective array columns. These registers in turn feed in parallel into a single CCD register at the bottom of the sensor, which transfers the charges out one row after the other as in a bucket brigade. The voltage across the output capacitor of this circuitry is proportional to the brightness of the corresponding pixel. A Digital to Analog (D/A) converter finally amplifies and transforms these voltages to binary numbers for transmission. In some cameras, the A/D conversion occurs on the camera itself. In others, a separate circuitry (a frame grabber) is installed for this purpose on a computer that the camera is connected to.
>
> The photodetector in a CMOS camera works in principle in the same way. However, the photosensitive junction is fabricated with the standard Complementary-symmetry Metal-Oxide-Semiconductor (CMOS) technology used to make common integrated circuits such as computer memory and processing units. Since photodetector and processing circuitry can be fabricated with the same process in CMOS sensors, the charge-to-voltage conversion that CCD cameras perform serially at the output of the CCD shift register can be done instead in parallel and locally at every pixel on a CMOS sensor. This is why CMOS arrays are also called Active Pixel Sensors (APS).
>
> Because of inherent fabrication variations, the first CMOS sensors used to be much less consistent in their performance, both across different chips and from pixel to pixel on the same chip. This caused the voltage measured for a constant brightness to vary, thereby producing poor images at the output. However, CMOS sensor fabrication has improved dramatically in the recent past, and the two classes of sensors are now comparable to each other in terms of image quality. Although CCDs are still used where consistency of performance if of prime importance, CMOS sensors are eventually likely to supplant CCDs, both because of their lower cost and because of the opportunity to add more and more processing to individual pixels. For instance, “smart” CMOS pixels are being built that adapt their sensitivity to varying light conditions and do so differently in different parts of the image.

1.3.3 Pixel Size, Resolution, and Focal Length

It was shown in Section 1.1.1 that it is often useful to express the focal distance in pixels rather than in millimeters, because then quantities like x/f become dimensionless if x is expressed in pixels. However, if pixels are not square, there is both a horizontal and a vertical pixel size, so there are two focal distances. These can be determined in pixels either by calculation or by calibration.

The calculation is simple: a sensor of width w and height h with m rows and n columns of pixels is said to have *resolution* m by n. One pixel is then w/n millimeters wide and h/m millimeters tall. If the focal distance in millimeters is f_{mm}, then the horizontal and vertical focal distances in pixels are

$$f = f_{\text{mm}} \frac{n}{w} \quad \text{and} \quad f' = f_{\text{mm}} \frac{m}{h} \; .$$

Fortunately, many camera sensors are made with square pixels, so the two calculations return the same value. If f_{mm} is read from the lens barrel, then it is a focal length (focal distance at infinity).

Measuring (rather than computing) the focal distances in pixels is simple as well. Place a target of known size at a known distance from the center of projection of the camera. Let H, W, and Z

be height, width, and distance from the camera, respectively. Let h and w be the height and width of the image of the object. Then, similar triangles (see Figures 1.2 and 1.7(a)) yield

$$W/Z = w/f \quad \text{and} \quad H/Z = h/f'$$

so that

$$f = w\frac{Z}{W} \quad \text{and} \quad f' = h\frac{Z}{H} \; .$$

Practical Aspects: Measurement Accuracy. Both calculation and calibration methods for determining the focal lengths require a bit of care.

For the calculations to return accurate values one must make sure that the sensor size and the resolution reported in the specifications refer to the same part of the array. Sometimes, a rim of unexposed pixels is added around the active part of the sensor, for a combination of packaging and mounting reasons. In that case, the sensor size given in the specifications often measures the complete rectangle of pixels on the sensor, but the resolution only counts the pixels that are actually exposed.

In the calibration method, it is usually difficult to know the exact location of the center of projection on a given camera, let alone measure Z from it. Because of this, the calibration object should be large and far from the camera. In this way, errors in where the depth Z is measured from have a small effect. In addition, the two lengths W and H should be measured frontally, that is, orthogonally to the optical axis of the camera lens. However, this 90 degree angle need not be exact: W and H change with the cosine of the error in this angle, so the effects of a wrong orientation are of the second order.

1.3.4 A Simple Sensor Model

Not all of the area dedicated to a pixel is necessarily photosensitive, as part of it is occupied by circuitry. The fraction of pixel area that collects light that can be converted to current is called the pixel's *fill factor*, and is expressed in percent. A 100 percent fill factor is achievable by covering each pixel with a properly shaped droplet of silica (glass) or silicon on each pixel. This droplet acts as a *micro-lens* that funnels photons from the entire pixel area onto the photo-detector. Not all cameras have micro-lenses, nor does a micro-lens necessarily work effectively on the entire pixel area. So different cameras can have very different fill factors. In the end, the voltage output from a pixel is the result of integrating light intensity over a pixel area determined by the fill factor.

The voltage produced is a nonlinear function of brightness. An approximate linearization is typically performed by a transformation called *gamma correction*,

$$V_{\text{out}} = V_{\max}\left(\frac{V_{\text{in}}}{V_{\max}}\right)^{1/\gamma}$$

where $V_{\max}$ is the maximum possible voltage and γ is a constant. Values of gamma vary, but are typically between 1.5 and 3, so V_{out} is a concave function of V_{in}, as shown in Figure 1.8: low input

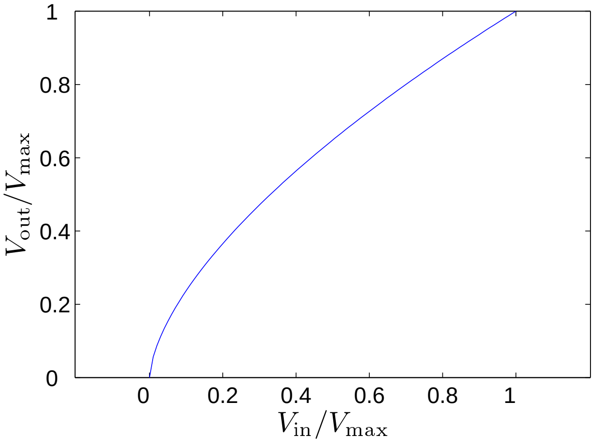


Figure 1.8: Plot of the normalized gamma correction curve for $\gamma = 1.6$.

voltages are spread out at the expense of high voltages, thereby increasing the dynamic range[9] of the darker parts of the output image.

Noise affects all stages of the conversion of brightness values to numbers. First, a small current flows through the photodetectors even if no photons hit its junction. This source of imaging noise is called the *dark current* of the sensor. Typically, the dark current cannot be canceled away exactly, because it fluctuates somewhat and is therefore not entirely predictable. In addition, *thermal noise*, caused by the agitation of molecules in the various electronic devices and conductors, is added at all stages of the conversion, with or without light illuminating the sensor. This type of noise is well modeled by a Gaussian distribution. A third type of noise is the *shot noise* that is visible when the levels of exposure are extremely low (but nonzero). In this situation, each pixel is typically hit by a very small number of photons within the exposure interval. The fluctuations in the number of photons are then best described by a Poisson distribution.

Every camera has *gain control* circuitry, either manually or automatically adjustable, which modifies the gain of the output amplifier so that the numerical pixel values occupy as much of the available range as possible. With dark images, the gain is set to a large value, and to a small value for bright ones. Gain is typically expressed in ISO values, from the standard that the International Standardization Organization (ISO) has defined for older film cameras. The ISO scale is linear, in the sense that doubling the ISO number corresponds to doubling the gain.

If lighting in the scene cannot be adjusted, a dark image can be made brighter by either (i) opening the lens aperture or (ii) by increasing exposure time, or (iii) by increasing the gain. The effects, however, are very different. As discussed earlier, widening the aperture decreases the depth of field. Increasing exposure time may result into blurry images if there is motion in the scene.

Figure 1.9 shows the effect of different gains. The two pictures were taken with constant lighting and aperture. However, the one in (a) (and the detail in (c)) was taken with a low value of gain, and the one in (b) (and (d)) was taken with a gain value sixteen times greater. From the

[9] Dynamic range: in this context, this is the range of voltages available to express a given range of brightnesses.

image as a whole ((a) and (b)) one can notice some greater degree of "graininess" corresponding to a higher gain value. The difference is more obvious when details of the images are examined ((c) and (d)).

So there is no free lunch: more light is better for a brighter picture. That is, brightness should be achieved by shining more light on the scene or, if depth of field is not important, by opening the aperture. Increasing camera gain will make the picture brighter, but also noisier.

In summary, a digital sensor can be modeled as a light integrator over an area corresponding to the pixel's fill factor. This array is followed by a sampler, which records the values of the integral at the centers of the pixels. At the output, an adder adds noise, which is an appropriate combination of dark current, Gaussian thermal noise, and shot noise. The parameters of the noise distribution typically depend on brightness values and camera settings. Finally, a quantizer converts continuous voltage values into discrete pixel values. The gamma correction can be ignored if the photodetectors are assumed to have an approximately linear response. Figure 1.10 shows this model in diagram form.

1.3.5 Color

The photodetectors in a camera sensor are only sensitive to light brightness, and do not report color. Two standard methods are used to obtain color images. The first, the 3-sensor method, is expensive and high quality. The second, the Bayer mosaic, is less expensive and sacrifices resolution for color. These two methods are discussed in turn.

The 3-Sensor Method In a 3-sensor color camera, a set of glass prisms uses a combination of internal reflection and refraction to split the incoming image into three. The three beams exit from three different faces of the prism, to which three different sensor arrays are attached. Each sensor is coated with a die that lets only light in a relatively narrow band go through in the red, green, and blue parts of the spectrum, respectively. Figure 1.11 (a) shows a schematic diagram of a beam splitter.

The Bayer Mosaic A more common approach to color imaging is the *sensor mosaic*. This scheme uses a single sensor, but coats the micro-lenses of individual pixels with red, green, or blue die. The most common pattern is the so-called *Bayer mosaic*, shown in Figure 1.11 (b).

With this arrangement, half of the pixels are sensitive to the green band of the light spectrum, and one quarter each to blue and red. This is consistent with the distribution of color-sensitive cones in the human retina, which is more responsive to the green-yellow part of the spectrum than to its red or blue components.

The raw image produced with a Bayer mosaic contains one third of the information that would be obtained with a 3-sensor camera of equal resolution on each chip. While each point in the field of view is seen by three pixels in a 3-sensor camera, *no* point in the world is seen by more than one pixel in the Bayer mosaic. As a consequence, the blue and red components of a pixel that is sensitive only to the green band must be inferred, and an analogous statement holds for the other two types of pixels. After properly normalizing and gamma-correcting each pixel value, this

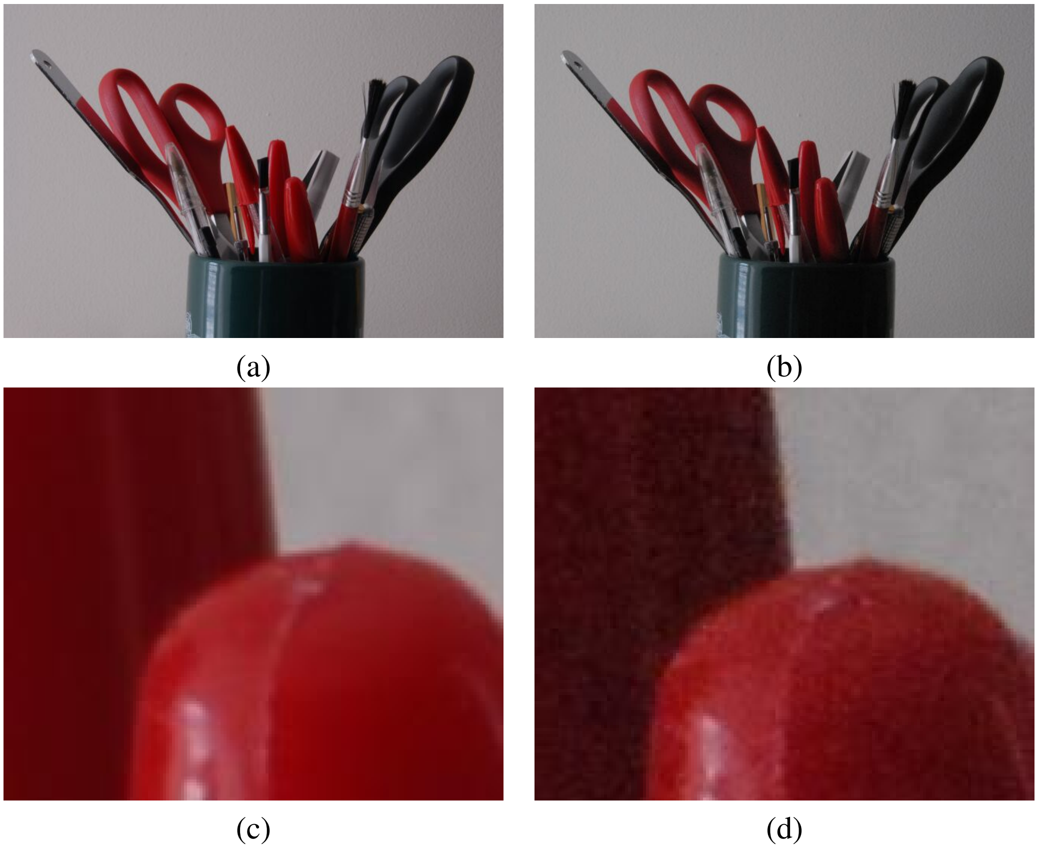


Figure 1.9: These two images were taken with the same lens aperture of f/20. However, (a) was taken with a low gain setting, corresponding to sensitivity ISO 100, and a one-second exposure, while (b) was taken with a high gain setting of ISO 1600, and an exposure of 1/15 of a second. (c) and (d) show the same detail from (a) and (b), respectively.

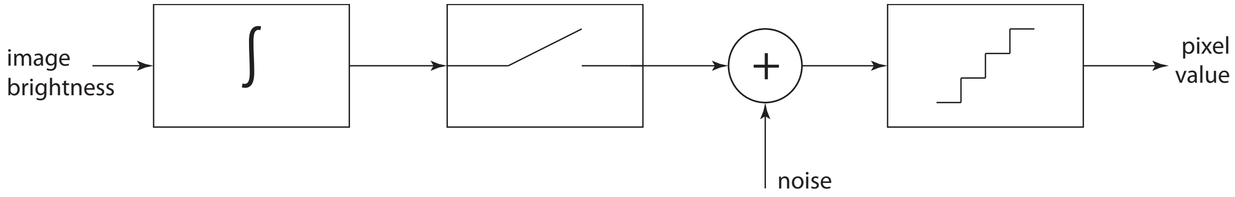
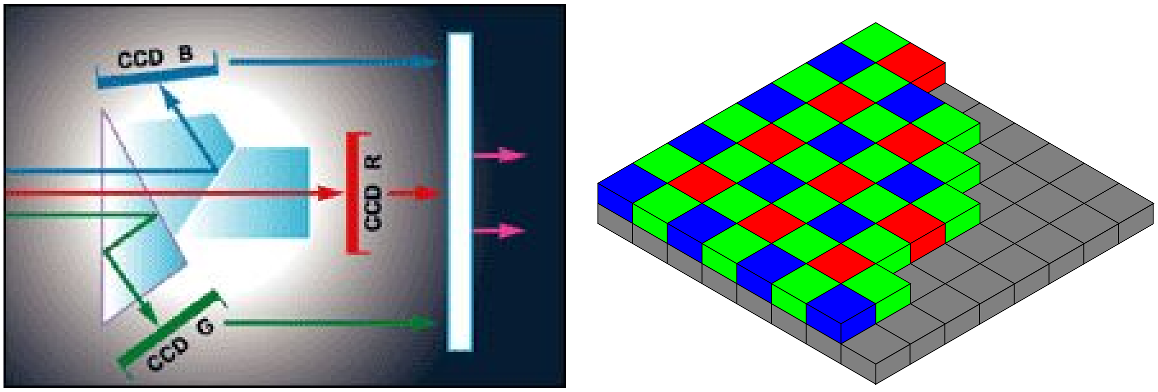


Figure 1.10: A simple sensor model. The three rectangular boxes are an integrator, a sampler, and a quantizer. Both integrator and samplers are in two dimensions. Noise statistics depend on input brightness and on camera settings.



(a) (b)

Figure 1.11: (a) Schematic diagram of a 3-sensor beam splitter for color cameras. From `http://www.usa.canon.com/`. (b) The Bayer color pattern. From `http://en.wikipedia.org/wiki/Image:Bayer_pattern_on_sensor.svg`.

inference proceeds by interpolation, under the assumption that nearby pixels usually have similar colors.

> **Practical Aspects: 3-Sensor Versus Bayer.** Of course, the beam splitter and the additional two sensors add cost to a 3-sensor camera. In addition, the three sensors must be aligned very precisely on the faces of the beam splitter. This fabrication aspect has perhaps an even greater impact on final price. Interestingly, even high end SLR cameras use the Bayer mosaic for color, as the loss of information caused by mosaicing is usually satisfactorily compensated by sensor resolutions in the tens of millions of pixels.

References

[1] Max Born and Emil Wolf. *Principles of Optics*. Pergamon Press, Oxford, 1975.

[2] B. K. P. Horn. *Robot Vision*. Mc Graw-Hill, New York, New York, 1986.

[3] B.K.P. Horn. Understanding image intensities. *Artificial Intelligence*, 8(11):201–231, 1977.

[4] E. Krotkov. Focusing. *International Journal of Computer Vision*, 1:223–237, 1987.

[5] S.K. Nayar, M. Watanabe, and M. Noguchi. Real-time focus range sensor. In *IEEE International Conference on Computer Vision*, pages 824–831, 1995.

[6] E. Trucco and A. Verri. *Introductory techniques for 3-D computer vision*. Prentice Hall, Upper Saddle River, NJ, 1998.