A Simple Camera Model

Carlo Tomasi

The images we process in computer vision are formed by light bouncing off surfaces in the world and into the lens of the camera. 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.

The study of what happens to the light that leaves surfaces in the world and makes it to the camera can be divided into considering what happens up to the moment when the light hits the sensor, and what happens thereafter. The first part occurs in the realm of optics, and is often encapsulated into what computer vision calls the *pinhole camera model*, a very much simplified description of camera optics. The first section below introduces this model. The section thereafter explains enough differences between the pinhole camera model and real lenses that you will know what to watch out for as you run experiments on 3D reconstruction.

The second part of image formation is called *sensing*, and is a matter of electronics. Key aspects of sensing, including how cameras handle color, are briefly outlined in the Appendix.

1 The Pinhole Camera Model

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.

A pinhole camera is a box with five opaque faces and a translucent one. 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 Figure 1(a), 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 when the plane of the screen intersects the star of rays through the pinhole. 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).

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(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* (Figure 1(c)). The front screen is the *image plane*. The distance between center of projection and image plane is the *focal distance*,

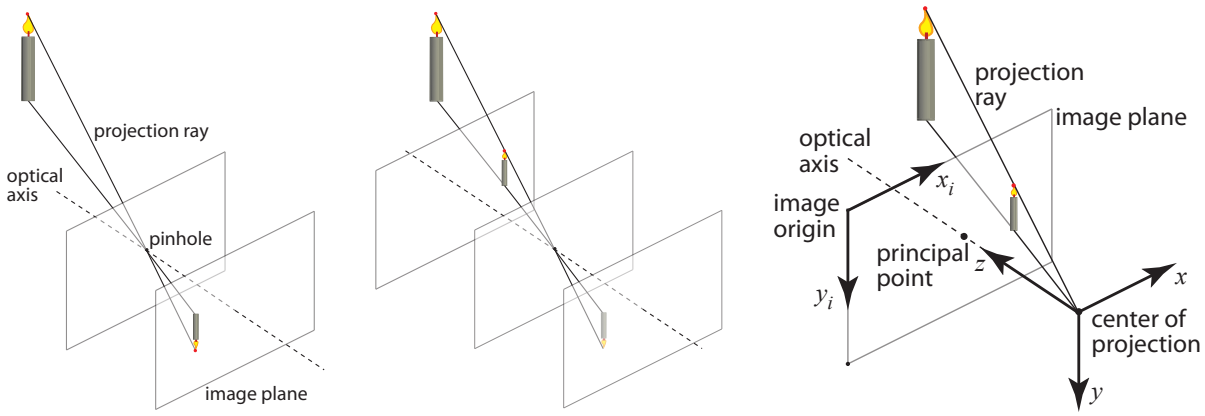


Figure 1: (a) Projection geometry for a pinhole camera. (b) If a screen could be placed in front of the pinhole, rather than behind, without blocking the projection rays, then the image would be upside-up. (c) What is left is the so-called *pinhole camera model*. The camera coordinate frame (x, y, z) is right-handed.

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*.

The origin of the *image coordinate system* (x_i, y_i) is placed in the top 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(c), the camera reference system is right-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. 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 (or *vice versa*) in the two directions.

Every point on the image plane has a z coordinate equal to 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 millimeter.

The *projection equations* relate the camera-system coordinates $\mathbf{P} = (X, Y, Z)$ of a point in space to the camera-system coordinates $\mathbf{p} = (x, y)$ of the projection of $\mathbf{P}$ 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 in Figure 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 then 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.

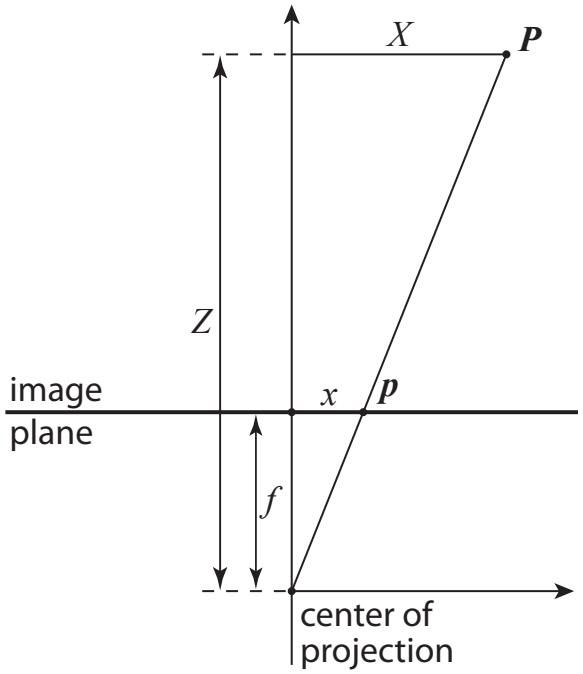


Figure 2: A top view of figure 1 (c).

2 Lenses and Discrepancies from the Pinhole Model

The pinhole camera is a useful and simple reference system for talking about the geometry of image formation. As pointed out above, however, this device 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 relate to focusing and distortion.

2.1 Focusing

Figure 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.

For the image of the candle tip to be sharply focused, it is necessary for the lens to funnel onto a single point in the image all of the rays from the candle tip that the aperture lets through. 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.

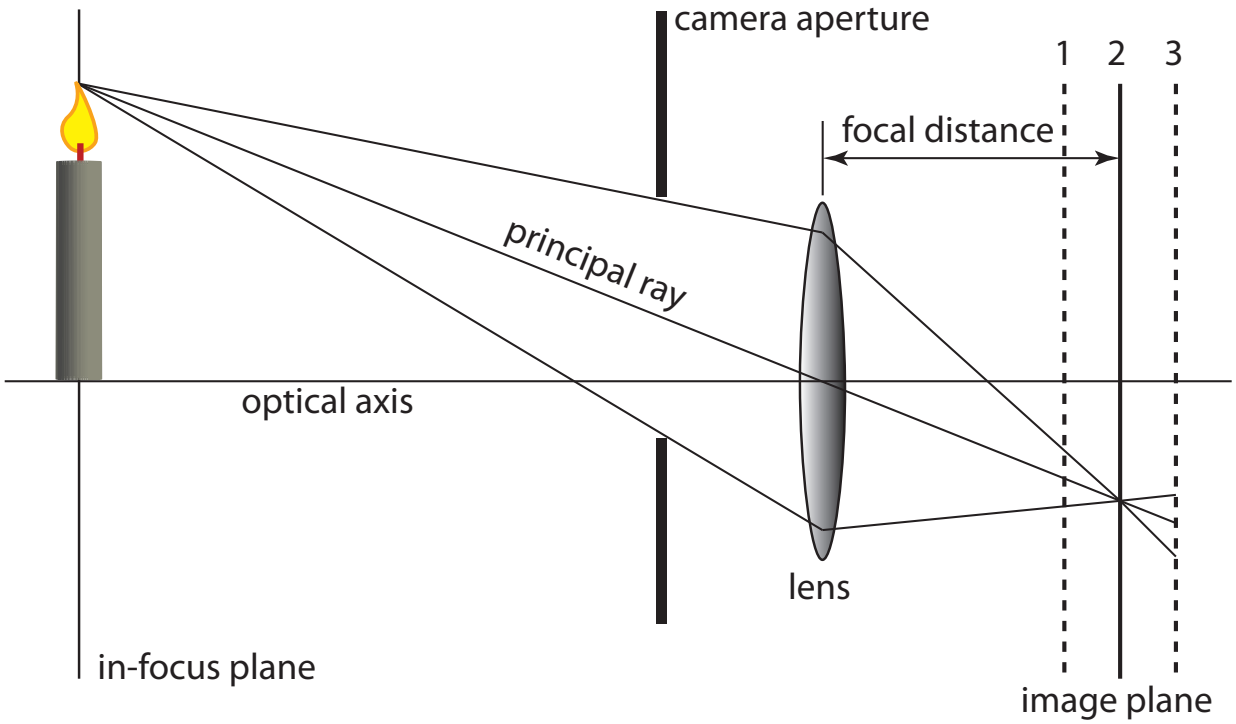
Since the correct focal distance depends on the distance of the world point from the lens, for any fixed focal distance, only the points on a single plane in the world are in focus. An image plane in position 1 in the Figure would focus points that are farther away than the candle, and an image plane in position 3 would focus points that are closer by. The dependence of focus on distance is visible in Figure 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 3(b) has a *shallow 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 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).

The focal distance at which a given lens focuses objects at infinite distance from the camera is called

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

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



(a)

(b) (c)

Figure 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.

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]

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 3 (b) was obtained with a relatively wide aperture $f/4.2$, while the greater depth of field of the image in Figure 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 4, from

`http://www.hp.com/united-states/consumer/digital_photography/take_better_photos/tips/depth.html` .

2.2 Distortion

Even the high quality lens[5] used for the images in Figure 3 exhibits distortion. For instance, if you place a ruler along the vertical edge of the blue book on the far left of the 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 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.

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

Figure 4: A shallow depth of field draw attention to what is in focus, at the expense of what is not.

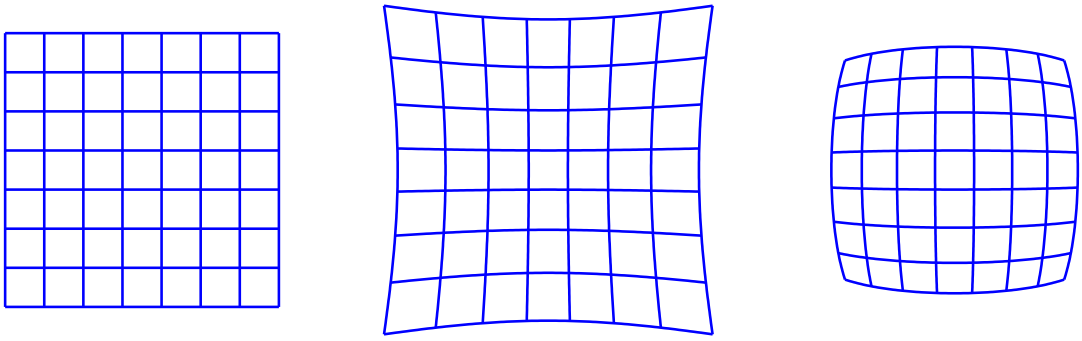


Figure 5: (a) An undistorted grid. (b) The grid in (a) with pincushion distortion. (c) The grid in (a) with barrel distortion.

Practical Aspects: Achieving Low Distortion. 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 3 are designed for a 23.6 by 15.8 millimeter sensor. Distortion is small but not negligible (see Figure 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 Appendix).

Appendix A: 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.

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), Apple Thunderbolt, or Gigabit Ethernet.

The next Section 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.

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.

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 6: low input voltages are spread out at the expense of high voltages, thereby increasing the dynamic range[6] 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.

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

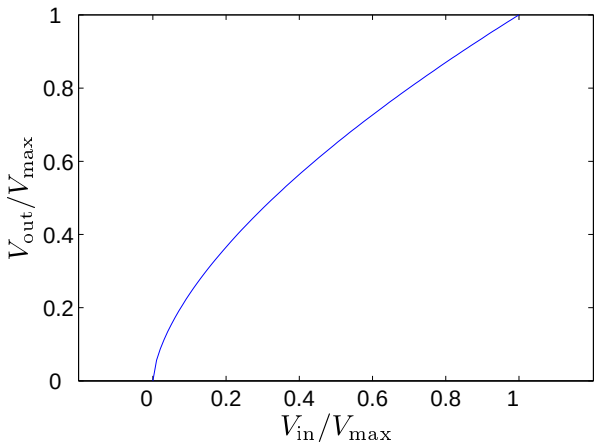


Figure 6: Plot of the normalized gamma correction curve for $\gamma = 1.6$.

Every camera has *gain control* circuitry, either manually adjustable or automatic, 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 7 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 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 8 shows this model in diagram form.

Color Sensors

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

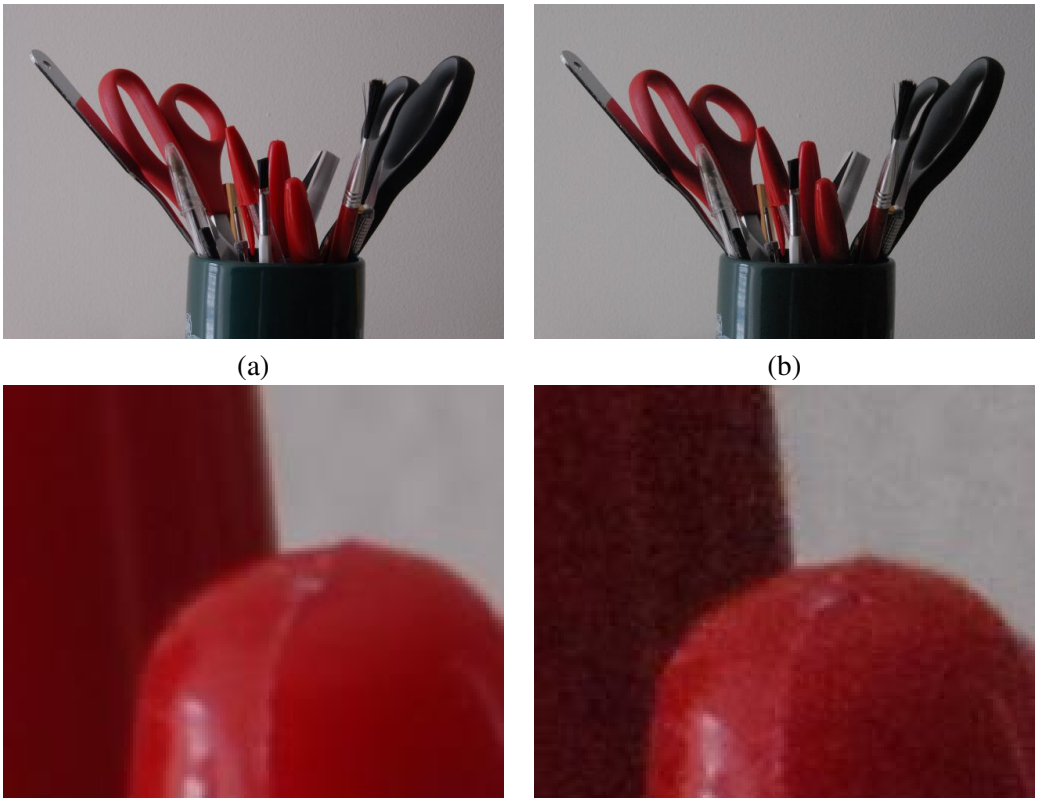


Figure 7: These two images were taken with the same lens aperture of f/20. However, (a) was taken with a low gain setting, corresponding to sensitivity ISO 100, and a one-second exposure, while (b) was taken with a high gain setting of ISO 1600, and an exposure of 1/15 of a second. (c) and (d) show the same detail from (a) and (b), respectively.

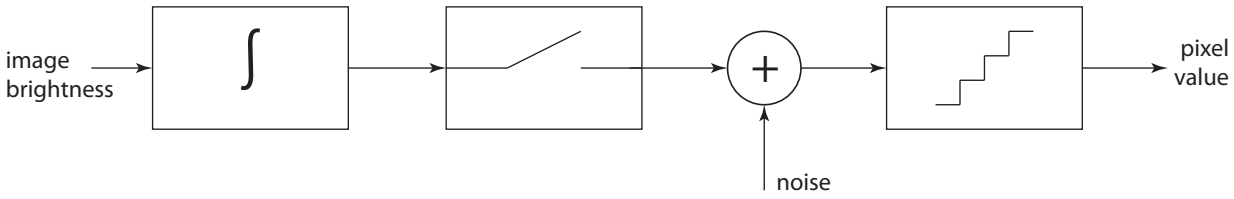
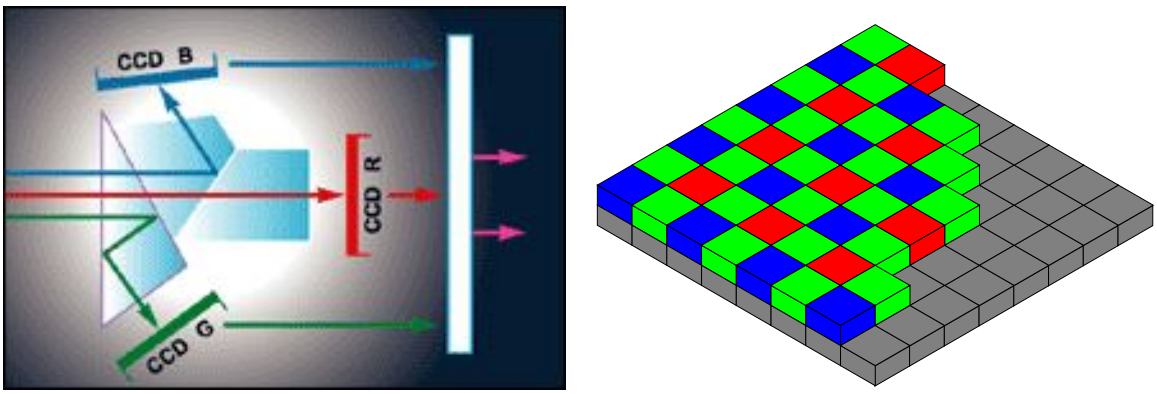


Figure 8: A simple sensor model. The three rectangular boxes are an integrator, a sampler, and a quantizer. Both integrator and samplers are in two dimensions. Noise statistics depend on input brightness and on camera settings.

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 9 (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 colors 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 9 (b).



(a) (b)

Figure 9: (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`.

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 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 than the trebled sensor cost. 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.