Image Formation

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 system. The light then hits a sensor inside the camera and 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 frame grabber) 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.

In fact, it does not cease to amaze me 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 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.

Any model is a simplified description of reality. In image formation, it is convenient to take the extreme approach of defining a very simple model, and call everything else an "error". Calibration is the process whereby the errors are determined for a given camera so they can be undone. This is a very useful approach. In fact, as a result of it, all of the theory of computer vision can assume a mathematically simple imaging model, and the cameras are made to conform to it through calibration.

To summarize, we will now study the optics of image formation, some aspects of sensing, and a few simple calibration techniques. The calibration methods we study are not accurate enough for photogrammetric applications like drawing geographic maps from aerial imagery. However, they are good enough for removing gross discrepancies between ideal and real images.

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, *how much* of it lands on the sensor is the photometric, or radiometric, aspect.

1.1 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 form therefore a star of lines.

For ever more distant scenes, 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 often 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.

1.1.1 Perspective Projection

A pinhole camera is a box with a pinhole on one, opaque face and a translucent screen on the opposite face. All other faces are opaque. A cardboard pinhole camera is easy and instructive to build. Figure 1 shows what happens in the box. Only a thin beam from a narrow set of directions hits any given point on the 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).

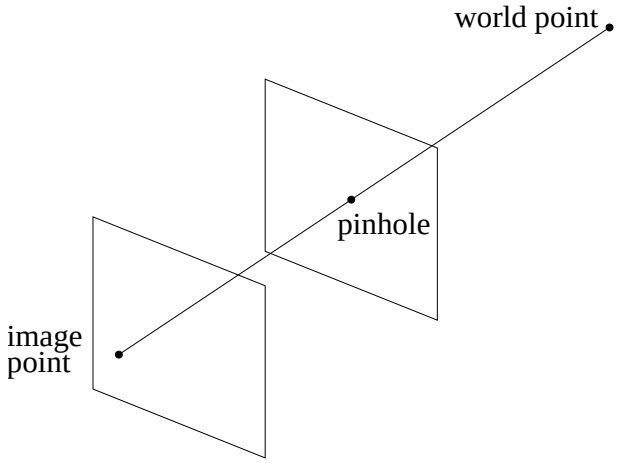

Figure 1: Model for a pinhole camera.

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 2. This is of course an idealization, since a screen in this position would block the light rays. In this model, the pinhole is called more appropriately the *center of projection*. The new image is isomorphic to the old one. The new plane is often placed at unit distance from the center of projection to simplify the projection equations.

Mathematically, the easiest way to describe this situation is to select a spherical coordinate system with its origin at the pinhole, as done in figure 3. The choice of reference directions is not critical, but it is natural to select one axis as the *optical axis*, defined as the line through the pinhole orthogonal to the screen, and the other axis as being parallel to the horizontal lines on the screen. Horizontal, here, is either an arbitrary direction or the direction on the screen that is orthogonal to gravity. In this reference system, also depicted in figure 3, the world point with coordinates (ρ, θ, ϕ) projects to the image point (ρ_i, θ, ϕ) where $\rho_i = \sqrt{1 + \tan^2 \theta}$ is univocally determined by θ and therefore provides redundant information. In this sense,

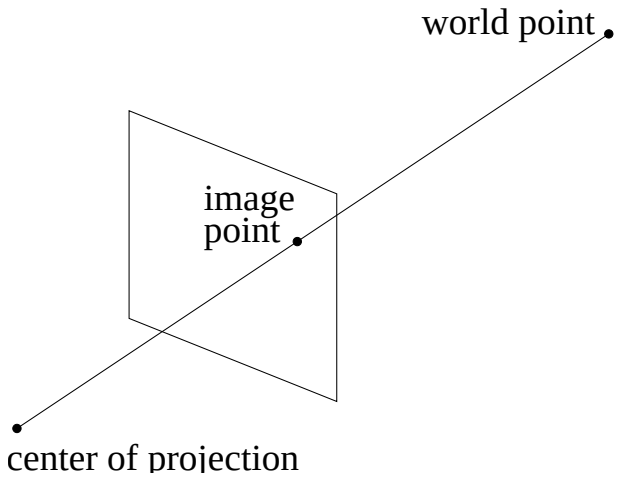

Figure 2: A mathematically more convenient projection model.

under spherical perspective, the world point (ρ, θ, ϕ) projects to image point (θ, ϕ).

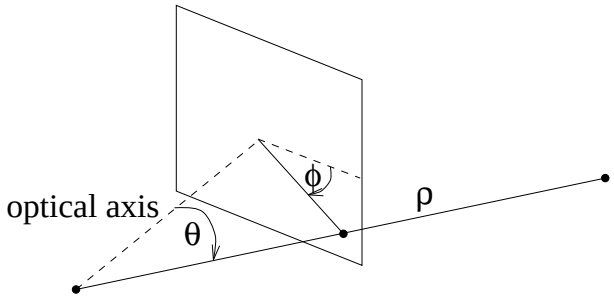

Figure 3: A natural spherical reference system.

The mechanics of the world, on the other hand, is more easily expressed in Cartesian coordinates. The reference system of figure 4 is therefore more popular.

Authors vary in their choice of x, y, z labels for the axes and of the axes' positive directions. The choice in figure 4 was made to have positive z coordinates for objects in front of the camera and a right-handed system at the same time. The z coordinate of a point in the world is called the point's *depth*.

In this system of reference, the image axes are considered to be parallel to the world's x and y coordinates, and their origin is at the *principal point*, defined as the intersection of the optical axis with the image plane.

The Cartesian projection equations can be easily derived for the x coordinate from the top view of figure 5. In fact, the triangle with orthogonal sides of length X and Z (two of the world point coordinates) is similar to that with orthogonal sides of length x (an image point coordinate) and f (the focal length), so that $X/Z = x/f$. Similarly, for the Y coordinate, one gets $Y/Z = y/f$. In conclusion,

under planar perspective, the world point with coordinates (X, Y, Z) projects to the image point with coordinates

$$\begin{aligned} x &= f\frac{X}{Z} \\ y &= f\frac{Y}{Z}\,. \end{aligned} \tag{1}$$

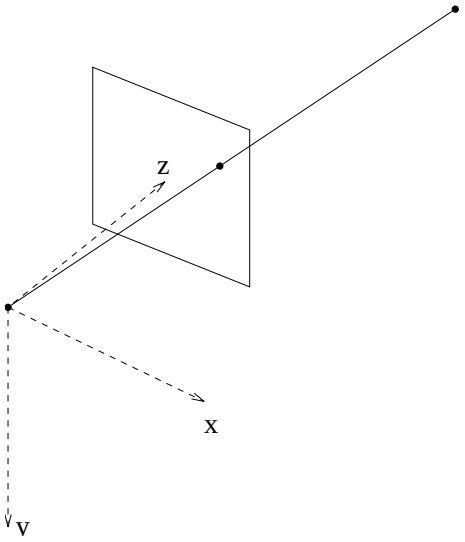

Figure 4: A Cartesian reference system.

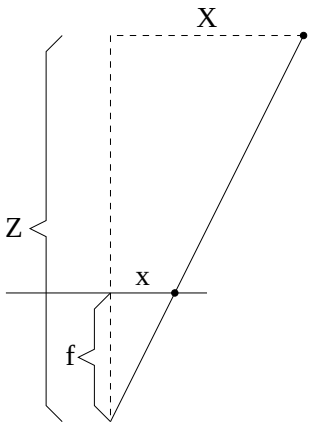

Figure 5: A top view of figure 4.

The relation between Cartesian and spherical reference systems is given by the following relations:

$$\begin{aligned} X &= \rho \sin\theta \cos\phi \\ Y &= \rho \sin\theta \sin\phi \\ Z &= \rho \cos\theta \end{aligned}$$

and their inverses:

$$\begin{aligned} \rho &= \sqrt{X^2 + Y^2 + Z^2} \\ \theta &= \arctan_2(\sqrt{X^2 + Y^2}, Z) \\ \phi &= \arctan_2(Y, X) \ . \end{aligned}$$

The two-argument function $\arctan_2$ is defined as follows:

$$\arctan_2(y,x) = \begin{cases} \arctan(\frac{y}{x}) & \text{if } x > 0 \\ \pi + \arctan(\frac{y}{x}) & \text{if } x < 0 \\ \frac{\pi}{2} & \text{if } x = 0 \text{ and } y > 0 \\ -\frac{\pi}{2} & \text{if } x = 0 \text{ and } y < 0 \\ 0 & \text{if } x = 0 \text{ and } y = 0 \end{cases} .$$

It returns the arctangent of y/x (notice the order of the arguments) in the proper quadrant, and extends the function by continuity along the y axis. The convention $\arctan_2(0,0) = 0$ is mathematically awkward (the function should be left undefined there), but convenient in a program.

1.1.2 Orthographic Projection

As the camera recedes and gets farther away from the scene, 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. One then gets the scaled image coordinates

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

As the camera recedes to infinity, Z and Z_0 grow together, 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 orthography, the world point with coordinates (X,Y,Z) projects to the image point with coordinates**
>
> $$\begin{aligned} x &= X \\ y &= Y . \end{aligned} \tag{2}$$

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

1.1.3 Lenses and Discrepancies from the Pinhole Model

As we 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. To obtain a good approximation, one can buy the extremely expensive imaging devices used in photogrammetry. Alternatively, one can try to measure the discrepancy between a real and an ideal lens, a process called calibration, and then compute the ideal image from the real one. To be done well, every point in the field of view, in a sense, must be calibrated individually. This point-by-point calibration of a camera's field of view used to be prohibitively laborious in traditional photogrammetry, where most of the computation used to be manual. This is the reason for the expensive equipment. Point-by-point calibration, on the other hand, is a viable option for computer vision. In other words, computer vision can replace expensive lenses with good calibration algorithms. With good camera calibration, and image digitization, the fundamental limit to image quality are optical aberrations which reduce image sharpness.

Problems related to projection trough a lens can be of the following types:

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

1. Images may be blurred or, to use the technical term, *astigmatic*, even with the best possible focusing adjustment.

2. Images may be distorted, that is, straight lines may project onto curves other than straight lines.

3. The apparent brightness of a point in the world may vary depending on where that point appears in the field of view, even if the center of projection does not move.

A lens that produces images without the first two defects is said to be *perfect* in geometrical optics. Thus, a perfect lens gives perfectly focused and undistorted images.

Blurring in good lenses is usually so small that it becomes negligible with respect to the size of the pixels on the camera sensor. Consequently, it can often be ignored. Distortion, on the other hand, 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. A mathematical theory of distortions is beyond the scope of this class. One needs to be concerned with it when defining a mathematical model for it, or when building a lens. For vision, we can content ourselves by knowing that distortion can be modeled by an odd, low-order (fifth-order) polynomial. We will see later how to estimate the coefficients of this polynomial during calibration. Typical distortions deform a grid in front of the camera into a barrel-like or a pincushion-like shape.

1.2 Radiometry

The other aspect of image formation, besides the geometric one, is radiometry, which describes how light is attenuated in different parts of the field of view. The considerations that follow are adapted from [1].

Ideally, if we place a uniformly illuminated piece of paper in front of a camera, an image of constant brightness should form. This, however, is not the case even for ideal lenses. In fact, with an ideal perspective lens, the apparent intensity of a patch of surface at angle α from the optical axis will be diminished by a factor of $\cos^4 \alpha$. A drop of $\cos^2 \alpha$ is created because the image location is $1/\cos\alpha$ further from the lens than a location on the optical axis and intensity drops as the square of distance. Two more factors of α are introduced because the cone of light rays enters the lens at an angle and light hits the image plane at an angle. If the lens has a 90 degree field of view, this means that the edges of the image will be only one fourth as bright as the center. This is for an ideal lens. Real lenses can cause further variation in illumination for other reasons.

A common trick 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. For these lenses, both radial distortion and radiometric drop-off are not substantial. However, the lack of peripheral vision is a handicap for visual searching, navigation, and detecting objects moving towards the observer. Although a human eye has good resolution for only about 50 degrees, the full visual field of each eye extends at least 90 degrees. The peripheral area delivers only low-resolution information, but this is useful for detecting motion and finding large objects. Inexpensive 35mm and C-mount lenses can deliver a field of view as large as 110 degrees. Wide-angle lenses have a very large depth of field, an asset for identifying objects when the camera and/or objects may change position.

Furthermore, people can assemble wide-angle images of the world by rotating their eyes and combining information from both eyes. Without moving your head, your field of view is nearly 180 degrees. This panoramic view provides a context for deciding where to direct your high-resolution foveal vision. It also allows you to detect unexpected events (e.g., an attacking lion, an approaching car, a softball) coming from many directions. The human field of view is limited by the fact that both eyes point forward, so as to provide stereo depth information. Animals for whom panoramic view is more important than stereo (e.g, many herbivores) have eyes that point sideways and an even wider field of view. It has been observed that predators have good stereo vision and parallel eyes, while preys have a good peripheral vision and eyes on the sides of their head.

2 Sensing

During digitization, problems can occur both in the camera and in the frame grabber. More specifically:

1. The assumed geometric mapping from sensor to image coordinates may be wrong.

2. The mapping may be a function of time and image position.

3. The mapping may be spatially coarse.

4. The image intensity values may be incorrectly measured.

The most important consequence of the first type of problem is that the position of the image center and the orientation of the camera sensor with respect to the optical axis may need to be determined experimentally during calibration. Problem number 2 is called *line jitter* in the digitizer literature, and causes adjacent scan lines to be shifted relatively to each other. This problem can be minimized by the use of appropriate circuitry for analog cameras, and is absent in digital cameras. Problems of type 4 are due to electronic distortions, intensity quantization, and noise in the electronic circuits, both in the camera and in the digitizer. These defects can be reduced by proper circuit design and cooling of the devices.

3 Calibration

Precise camera calibration is a complex and expensive procedure. However, for low-precision applications, it is sufficient to remove visible, gross distortions from the image. Margaret Fleck suggested a simple, low-tech procedure for doing this "to ensure that no one thinks they have an excuse for tolerating gross errors in images (e.g., visible radial distortion) when they have physical access to the camera" [1]. Here is the procedure.

The first question is, where is the center of the image? The center of the image is defined as the point where the optical axis of the lens, that is, the axis of symmetry of the glass, pierces the sensor plane. Typically, the sensor is simply glued into the camera, so the position of the camera center can vary considerably not only from model to model, but from one specific specimen to another of the same model. It has been shown that the image center can be relatively far (25 pixels out of 512 pixels of a scanline) from the middle of the digitized image. However, it has also been shown that the errors in scene reconstruction that are caused by a poor estimate of the center of projection are small, so this parameter can usually be left uncalibrated. In other words, one can take the middle of the digitized image to be the optical center of the image.

The second parameter one needs to calibrate is the aspect ratio of the pixels. If pixels are not square, a given angular distance between two points can cover some number of pixels horizontally and a different number of pixels vertically. In usual CCD cameras, the aspect ratio is often different from one. A simple way to measure this ratio is to place a table-tennis ball in the center of the field of view and compute its elongation. For instance, suppose that the ball measures 30 pixels in width and 45 in height. Then, since the image of the ball is circular, there are 45 pixels vertically in the same space in which 30 pixels fit horizontally. The aspect ratio is therefore $45/30 = 1.5$, in the sense that pixels are 1.5 wider than they are tall.

The field of view is the angular distance between two points at the opposite edges of the image, either horizontally or vertically. Thus, there are two fields of view. To measure this, simply place a pad with thick vertical lines in front of the camera and move it until two of the lines are at the edges of the image. Then measure the distance between the camera and the pad, as well as the distance between the two lines. Simple trigonometry yields the horizontal field of view. The vertical field can be measured similarly. Notice that the "distance to the camera" is supposed to be the distance from the center of projection of the camera, whose position within the camera case is hard to establish. However, if the pad is far from the camera (or perhaps markings on the wall are used), a small error on the distance to the camera has little effect on the final angle.

To clarify this computation, if the two lines on the pad are d cm apart and the distance between pad and camera is D cm, then the field of view is

$$\phi = 2 \arctan(\frac{d}{2D}) \ .$$

This leaves the radial distortion parameters to be calibrated from scratch. A sheet of radial graph paper, held flat and perpendicular to the viewing direction, is photographed approximately centered in the digitized image. The center of the pattern is marked by hand in the image, as well as a few points on two of the circles along some radial lines. If radial distortion is modeled by an odd-symmetric, fifth order polynomial, the coefficients of this polynomial can be computed from the average distance from the center to each circle.

The radiometric drop-off can be measured by using a light table to provide uniform lighting across the field of view, and recording the corresponding image brightness values. This part of the calibration, however, is not too reliable, unless the light table is known to be really uniform. Here is where more expensive professional equipment would be preferable.

References

[1] Margaret Fleck. Shape and wide-angle image. Technical Report 04, University of Iowa, 1994.