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).

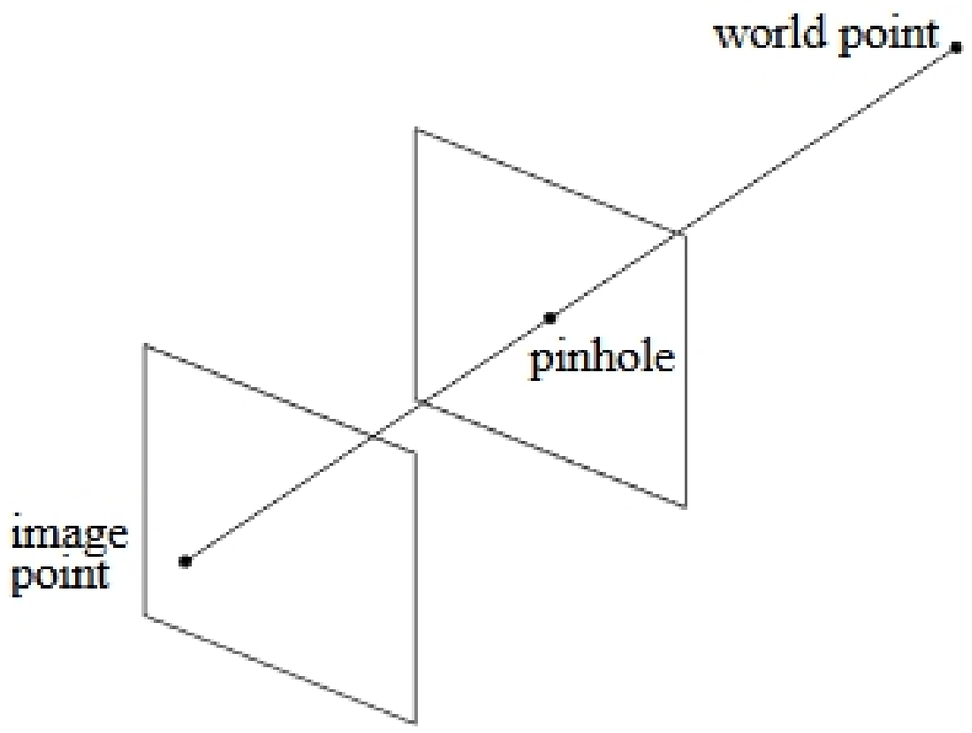

Figure 1: Model for a pinhole camera.

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,

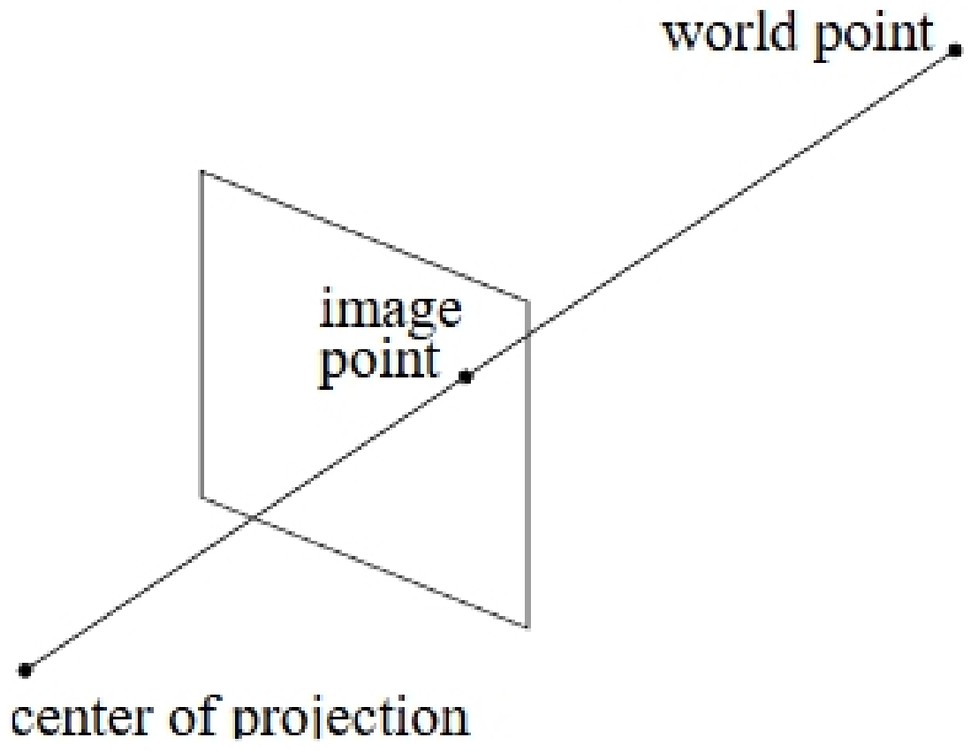

Figure 2: A mathematically more convenient projection model.

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

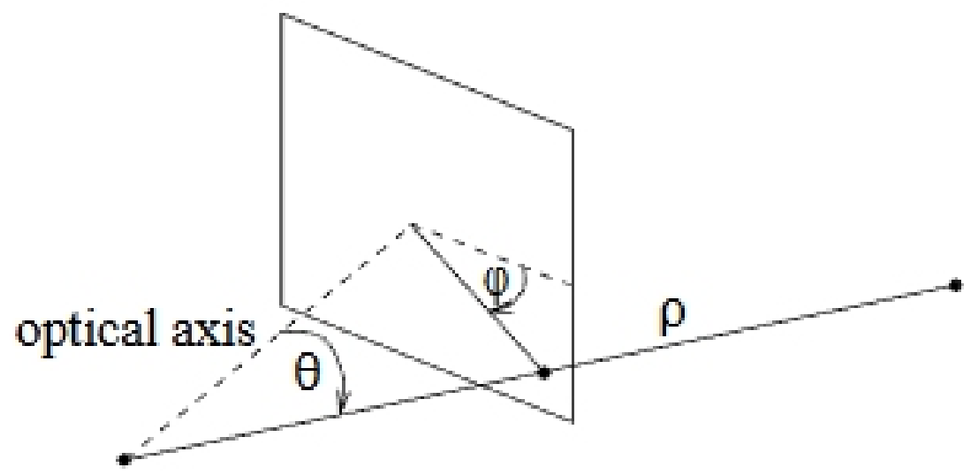

Figure 3: A natural spherical reference system.

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

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