Theory

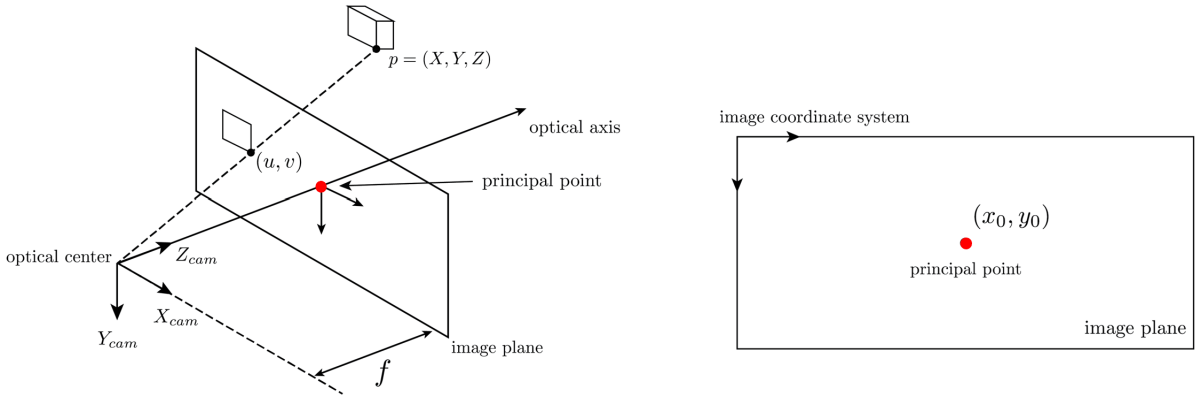

Figure 2: Geometry behind a pinhole camera

Homogeneous coordinates

In class, we have encountered homogeneous coordinates for 3D points, which work by appending a 1 to the end of the 3 vector to get a vector in 4D. When writing the coordinates of a point in the image plane (which is 2 dimensional) we will use *2D homogeneous coordinates*. This means we will represent the point $x = (u, v) \in \mathbb{R}^2$ as $(u, v, 1)$ in homogeneous coordinates. Note that the homogeneous representation of a 2D point is a 3D vector.

Image formation

We consider a pinhole model of image formation. See Figure 2. We denote the center of the perspective projection (the point in which all the rays intersect) as the optical center or camera center and the line perpendicular to the image plane passing through the optical center as the optical axis. Additionally, the intersection point of the image plane with the optical axis is called the principal point.

We always associate a reference frame with each camera as shown. By convention, we center this reference frame at the optical center, take the $X - Y$ plane of this reference frame to be parallel to the image plane, and take the $Z-$axis to be perpendicular to the image plane, pointing in the direction of viewing. Additionally, there is a 2D reference frame attached to the image plane, with respect to which the "image coordinates" of any point are measured. A discretized version of this reference frame give us the familiar "pixel coordinates" of any points (in columns and rows).

The figure shows a point p with spatial coordinates $\bar{X} = (X, Y, Z)$ in the camera reference frame, and image coordinates $x = (u, v)$. As we stated above, we will default to representing image coordinates in homogeneous form as $x = (u, v, 1)$, and usually we will overload this notation wherever it is obvious if we are using homogeneous or regular coordinates.

The camera parameters (such as the focal length f and others; see Lab 6) are specified in the form of a 3×3 *camera matrix* K. This matrix K is always invertible. The spatial coordinates $\bar{X}$ (in the camera reference frame) and the image (homogeneous) coordinates $x = (u, v, 1)$ of a given point p

are related via the K matrix as

$$x = \frac{1}{Z} K \bar{X} \tag{3}$$

where Z is the $Z-$coordinate of the point in the camera reference frame. Observe that this $Z-$coordinate has a significant geometric meaning. It is the distance from the camera's $X - Y$ plane to the point, along the direction of viewing. In other words, it is the "depth" of the point as seen from the camera. So, we give this depth its own symbol λ and move it to the LHS to get the less unwieldy expression

$$\lambda x = K \bar{X} \tag{4}$$

Note that given the depth λ, the camera matrix K, and the image coordinates x, the spatial coordinates $\bar{X}$ of the point can be recovered by inverting equation (4). On the other hand, without knowing the depth, the spatial coordinates $\bar{X}$ can only be recovered up to a scale factor. This makes geometric sense, since we can see from figure (2) that any point along the line conecting the optical center to p gets projected to the same image coordinates as p, and hence knowing only the image coordinates, we can at best specify a line along which p must lie.