1 Introduction

In computer vision, the camera matrix is a 3×4 matrix that encapsulates the mapping performed by a pinhole camera, projecting 3D points in the real world onto 2D points in an image. This mini-project focuses on understanding and implementing the simplest camera model, the basic pinhole camera.

We will begin by describing the projection of points in 3D space onto a 2D plane using a series of matrices that we will define. The majority of this report is based on the explanations provided in *Multiple View Geometry in Computer Vision* by Richard Hartley and Andrew Zisserman. Finally, we will discuss the implementation of a specific algorithm in MATLAB to compute the camera matrix.

1.1 TODO: Notational note

To remain consistent with the textbook by Hartley and Zisserman, we will keep our notations largely similar. To emphasize a couple of matrices, I am going to

2 The Pinhole Camera Model

The pinhole camera model is a simplified representation of how a camera projects three-dimensional (3D) points in space onto a two-dimensional (2D) image plane. Under this model, we aim to map a 3D point, denoted by $\mathbf{X} = \begin{pmatrix} x & y & z \end{pmatrix}^\top$, to its corresponding point on the image plane. This point is the intersection of the line passing through $\mathbf{X}$ and the center of projection (the camera's optical center) with the image plane.

The geometry of the setup defines the center of projection at the origin of the camera coordinate system, with the image plane located at a fixed distance f (the focal length) from the origin. The projection follows the principle of similar triangles.

[[TODO: INSERT DIAGRAM HERE]]

The equations governing this projection are given by:

$$u = \frac{fx}{z}, \quad v = \frac{fy}{z}.$$

Now, note that these perspective projections are nonlinear due to the division by z. To simplify our operations, we want to transform our system using a linear model. In our case, we can accomplish this by using homogeneous coordinates.

2.1 Linearizing and using homogenous coordinates

In the homogeneous form, the coordinates of our point $\mathbf{X}$ are expressed as $\begin{pmatrix} x & y & z & 1 \end{pmatrix}^\top$, and the corresponding point on the image plane becomes $\begin{pmatrix} u & v & z \end{pmatrix}^\top$. Then, we can write

our projection in matrix form as:

$$\begin{pmatrix} u \\ v \\ z \end{pmatrix} = \begin{pmatrix} f & 0 & 0 & 0 \\ 0 & f & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \\ 1 \end{pmatrix}.$$

This homogeneous representation serves as the foundation for deriving the complete camera matrix, which incorporates additional parameters to account for intrinsic and extrinsic camera properties.

2.2 Intrinsic Matrix

In real-world cameras, the optical center does not necessarily coincide with the center of the image plane. Instead, the projection may be offset by a certain amount. This offset is characterized by the principal point, denoted by p_x and p_y for the horizontal and vertical displacements, respectively.

Thus, the projection equations are modified to account for this offset:

$$u = \frac{fx}{z} + p_x, \quad v = \frac{fy}{z} + p_y.$$

In the homogeneous representation, this is expressed as:

$$\begin{pmatrix} u \\ v \\ z \end{pmatrix} = \begin{pmatrix} f & 0 & p_x & 0 \\ 0 & f & p_y & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \\ 1 \end{pmatrix}.$$

Here, the 3×3 submatrix

$$\mathbf{K} = \begin{pmatrix} f & 0 & p_x \\ 0 & f & p_y \\ 0 & 0 & 1 \end{pmatrix}$$

is known as the *camera calibration matrix*, which encapsulates the *intrinsic parameters* of the camera, such as the focal length f and the principal point offsets, p_x and p_y.

The entire matrix is referred to as the *intrinsic matrix*. Now, our 2D coordinate $\mathbf{x}$ can be expressed as

$$\mathbf{x} = \mathbf{K} \begin{bmatrix} I & | & \vec{\mathbf{0}} \end{bmatrix} \mathbf{x}_{\text{cam}}$$

where $\mathbf{x}_{\text{cam}} = \begin{pmatrix} x & y & z & 1 \end{pmatrix}^\top$ is our camera at the origin of a Euclidean coordinate system.

2.3 The Extrinsic Matrix

In addition to the intrinsic properties of a camera, we must account for its position and orientation in the world. These are described by the *extrinsic parameters*, which define the relationship between the *camera coordinate frame* (CCF) and the *world coordinate frame* (WCF).

The transformation between the WCF and the CCF involves two components:

- A 3×3 rotation matrix $\mathbf{R}$, which describes the orientation of the camera with respect to the WCF. Note that this matrix is orthogonal.
- A 3×1 translation vector $\mathbf{t}$, which specifies the position of the camera's optical center in the WCF.

The matrix $\mathbf{R}$ has three rows, each representing a basis vector of the camera coordinate frame expressed in the world coordinate frame:

- The first row of $\mathbf{R}$ represents the direction of the camera's x-axis in the WCF.
- The second row of $\mathbf{R}$ represents the direction of the camera's y-axis in the WCF.
- The third row of $\mathbf{R}$ represents the direction of the camera's optical axis (or z-axis) in the WCF.

We again express the system in terms of a homogeneous coordinate to transform a point $\mathbf{X}_{\text{w}}$ in the WCF to the CCF, we apply the rotation and translation as follows:

$$\mathbf{x}_{\text{cam}} = \mathbf{R}\mathbf{X}_{\text{w}} + \mathbf{t} = \begin{pmatrix} \mathbf{R} & \mathbf{t} \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x_{\text{w}} \\ y_{\text{w}} \\ z_{\text{w}} \\ 1 \end{pmatrix}.$$

Here, the 4×4 matrix $\begin{pmatrix} \mathbf{R} & \mathbf{t} \\ 0 & 1 \end{pmatrix}$ is known as the *extrinsic matrix*, combining rotation and translation into a single transformation.

3 The Camera Matrix