ECE276A: Sensing & Estimation in Robotics
Lecture 10: Projective Geometry, Camera Model

Instructor:

Nikolay Atanasov: natanasov@ucsd.edu

Teaching Assistants:

Qiaojun Feng: qif007@eng.ucsd.edu

Tianyu Wang: tiw161@eng.ucsd.edu

Ibrahim Akbar: iakbar@eng.ucsd.edu

You-Yi Jau: yjau@eng.ucsd.edu

Harshini Rajachander: hrajacha@eng.ucsd.edu

UC San Diego

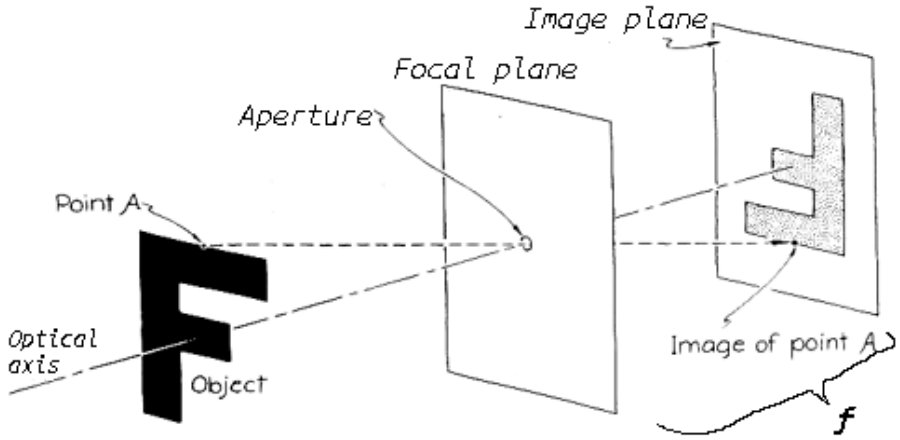
JACOBS SCHOOL OF ENGINEERING

Electrical and Computer Engineering

Image Formation

- **Image formation model**: must trade-off physical constraints and mathematical simplicity
- The values of an image depend on the shape and reflectance of the scene as well as the distribution of light
- **Image intensity/brightness/irradiance** $I(u, v)$ describes the energy falling onto a small patch of the imaging sensor (integrated both over the shutter interval and over a region of space) and is measured in power per unit area (W/m^2)
- A camera uses a set of lenses to control the direction of light propagation by means of *diffraction*, *refraction*, and *reflection*
- **Thin lens model**: a simple geometric model of image formation that considers only refraction
- **Pinhole model**: a thin lens model in which the lens aperture is decreased to zero and all rays are forced to go through the optical center and remain undeflected (diffraction becomes dominant).

Pinhole Camera Model



- **Focal plane**: perpendicular to the **optical axis** with a circular aperture at the **optical center**
- **Image plane**: parallel to the focal plane and a distance f (**focal length**) in **meters** from the optical center
- The pinhole camera model is described in an **optical frame** centered at the optical center with the optical axis as the z-axis:
  - optical frame: x = right, y = down, z = forward
  - world frame: x = forward, y = left, z = up
- **Ideal perspective projection**: relates the coordinates (X, Y, Z) of point A to its image coordinates (x, y) using similar triangles:

$$x = -f\frac{X}{Z} \qquad y = -f\frac{Y}{Z}$$

$$\begin{pmatrix} x \\ y \\ 1 \end{pmatrix} = \frac{1}{Z}\begin{bmatrix} -f & 0 & 0 & 0 \\ 0 & -f & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}\begin{pmatrix} X \\ Y \\ Z \\ 1 \end{pmatrix}$$

Pinhole Camera Model

- **Image flip**: the object appears upside down on the image plane. To eliminate this effect, we can simply flip the image $(x, y) \to (-x, -y)$, which corresponds to placing the image plane $\{z = -f\}$ in front of the optical center instead of behind $\{z = f\}$.
- **Field of view**: the angle subtended by the spatial extend of the image plane as seen from the optical center. If m is the side of the image plane in **meters**, then the field of view is $\boxed{\theta = 2 \arctan\left(\frac{m}{2f}\right)}$.
  - For a flat image plane: $\theta < 180°$.
  - For a spherical or ellipsoidal imaging surface, common in omnidirectional cameras, θ can exceed $180°$.
- **Ray tracing**: under assumptions of the pinhole model and Lambertian surfaces, image formation can be reduced to tracing rays from points on objects to pixels. A mathematical model associating 3-D points in the world frame to 2-D points in the image frame must account for:
  1. **Extrinsics**: world-to-camera frame transformation
  2. **Projection**: 3D-to-2D coordinate projection
  3. **Intrinsics**: scaling and translation of the image coordinate frame

Extrinsics

- Let $p_{wc} \in \mathbb{R}^3$ and $R_{wc} \in SO(3)$ be the camera position and orientation in the world frame
- Rotation from a regular to an optical frame: $R_{oc} := \begin{bmatrix} 0 & -1 & 0 \\ 0 & 0 & -1 \\ 1 & 0 & 0 \end{bmatrix}$
- Let (X_w, Y_w, Z_w) be the coordinates of point A in the world frame. The coordinates of A in the optical frame are then:

$$\begin{pmatrix} X_o \\ Y_o \\ Z_o \\ 1 \end{pmatrix} = \begin{bmatrix} R_{oc} & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} R_{cw} & p_{cw} \\ 0 & 1 \end{bmatrix} \begin{pmatrix} X_w \\ Y_w \\ Z_w \\ 1 \end{pmatrix} = \begin{bmatrix} R_{oc} R_{wc}^T & -R_{oc} R_{wc}^T p_{wc} \\ 0 & 1 \end{bmatrix} \begin{pmatrix} X_w \\ Y_w \\ Z_w \\ 1 \end{pmatrix}$$

Projection

- The 3D-to-2D projection in homogeneous coordinates from the optical frame to the image frame for a frontal pinhole camera model is:

$$\begin{pmatrix} x \\ y \\ 1 \end{pmatrix} = \frac{1}{Z_o} \begin{bmatrix} f & 0 & 0 & 0 \\ 0 & f & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{pmatrix} X_o \\ Y_o \\ Z_o \\ 1 \end{pmatrix}$$

- The above can be decomposed into:

$$\begin{pmatrix} x \\ y \\ 1 \end{pmatrix} = \underbrace{\begin{bmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{bmatrix}}_{\text{image flip: } R_f} \underbrace{\begin{bmatrix} -f & 0 & 0 \\ 0 & -f & 0 \\ 0 & 0 & 1 \end{bmatrix}}_{\text{focal scaling: } K_f} \underbrace{\frac{1}{Z_o} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}}_{\text{canonical projection: } \Pi_0} \begin{pmatrix} X_o \\ Y_o \\ Z_o \\ 1 \end{pmatrix}$$

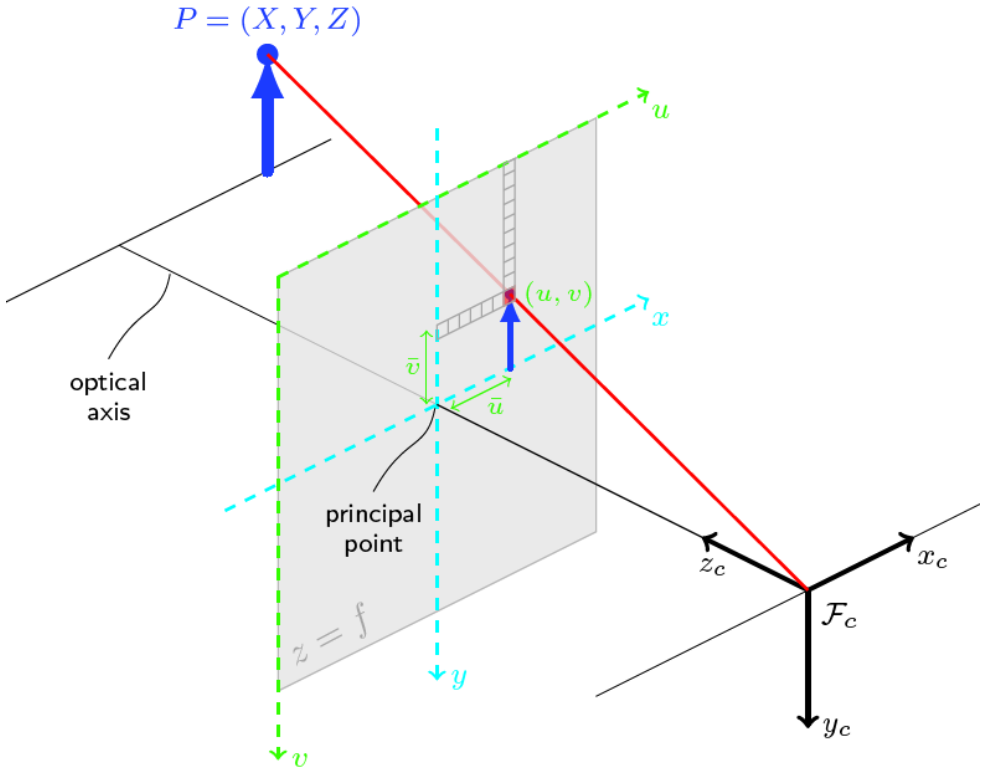
- The focal scaling K_f and image flip R_f are intrinsic parameters.

Intrinsics

- In practice, images are obtained in terms of pixels (u, v) with the origin of the pixel array typically in the upper-left corner of the image.
- The relationship between the image frame and the pixel array is specified via the following parameters:
  - (s_u, s_v) [pixels/meter]: define the **scaling** from meters to pixels and the **aspect ration** $\sigma = s_u/s_v$
  - (c_u, c_v) [pixels]: coordinates of the *principal point* used to translate the image frame origin, e.g., $(c_u, c_v) = (320.5, 240.5)$ for a 640×480 image
  - s_θ [pixels/meter]: **skew factor** that scales non-rectangular pixels and is proportional to $\cot(\alpha)$ where α is the angle between the coordinate axes of the pixel array.
- Normalized coordinates in the image frame are converted to pixel coordinates in the pixel array using the **intrinsic parameter matrix**:

$$\underbrace{\begin{bmatrix} s_u & s_\theta & c_u \\ 0 & s_v & c_v \\ 0 & 0 & 1 \end{bmatrix}}_{\text{pixel scaling: } K_s} \underbrace{\begin{bmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{bmatrix}}_{\text{image flip: } R_f} \underbrace{\begin{bmatrix} -f & 0 & 0 \\ 0 & -f & 0 \\ 0 & 0 & 1 \end{bmatrix}}_{\text{focal scaling: } K_f} = \underbrace{\begin{bmatrix} fs_u & fs_\theta & c_u \\ 0 & fs_v & c_v \\ 0 & 0 & 1 \end{bmatrix}}_{\text{calibration matrix: } K} \in \mathbb{R}^{3\times 3}$$

Pinhole Camera Model



- **Extrinsics**:

$$\begin{pmatrix} X_o \\ Y_o \\ Z_o \\ 1 \end{pmatrix} = \begin{bmatrix} R_{oc} R_{wc}^T & -R_{oc} R_{wc}^T p_{wc} \\ 0 & 1 \end{bmatrix} \begin{pmatrix} X_w \\ Y_w \\ Z_w \\ 1 \end{pmatrix}$$

- **Projection and Intrinsics**:

$$\underbrace{\begin{pmatrix} u \\ v \\ 1 \end{pmatrix}}_{\text{pixels}} = \underbrace{\begin{bmatrix} f s_u & f s_\theta & c_u \\ 0 & f s_v & c_v \\ 0 & 0 & 1 \end{bmatrix}}_{\text{calibration: } K} \underbrace{\frac{1}{Z_o} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}}_{\text{canonical projection: } \Pi_0} \begin{pmatrix} X_o \\ Y_o \\ Z_o \\ 1 \end{pmatrix}$$

Projection Functions

- **Canonical projection function**: for a vector $x \in \mathbb{R}^3$, define $\pi(x) := \frac{1}{x_3}x$. The pixel coordinates $z \in \mathbb{R}^2$ of a point $m \in \mathbb{R}^3$ in the world frame observed by a camera at position $p \in \mathbb{R}^3$ and orientation $R \in SO(3)$ with intrinsic parameters $K \in \mathbb{R}^{3\times 3}$ are:

$$z = K\pi(R_{oc}R^T(m-p))$$

- **Spherical perspective projection**: if the imaging surface is a sphere $\mathbb{S}^2 := \{x \in \mathbb{R}^3 \mid \|x\| = 1\}$ (motivated by retina shapes in biological systems), we can define a spherical projection $\pi_s(x) = \frac{x}{\|x\|_2}$. Similar to the planar perspective projection, the relationship between pixel coordinates z of a point and their 3-D metric counterpart m is:

$$z = K\pi_s(R_{oc}R^T(m-p))$$

- **Catadioptric model**: uses an ellipsoidal imaging surface

Radial distortion

- **Wide field of view camera**: in addition to linear distortions described by the intrinsic parameters K, one can observe distortion along radial directions.
- The simplest effective **model for radial distortion**:

$$x = x_d(1 + a_1 r^2 + a_2 r^4)$$
$$y = y_d(1 + a_1 r^2 + a_2 r^4)$$

where (x_d, y_d) are the pixel coordinates of distorted points and $r^2 = x_d^2 + y_d^2$ and a_1, a_2 are additional parameters modeling the amount of distortion.

Epipolar Geometry

- Let $m \in \mathbb{R}^3$ (world frame) be observed by two **calibrated** cameras
- Without loss of generality assume that the first camera frame coincides with the world frame. Let the position and orientation of the second camera be $p \in \mathbb{R}^3$ and $R \in SO(3)$
- The images of m in normalized image coordinates are:

$$\lambda_1 y_1 = m, \qquad \lambda_1 = \text{unknown scale}$$
$$\lambda_2 y_2 = R^T(m - p), \qquad \lambda_2 = \text{unknown scale}$$

- We obtain the following relationship between the image points:

$$\lambda_1 y_1 = R\lambda_2 y_2 + p$$

- To eliminate the unknown depths λ_i:
  - pre-multiply by $\hat{p}$
  - note that $\hat{p}y_1$ is perpendicular to y_1

$$\underbrace{\lambda_1 y_1^T \hat{p} y_1}_{0} = \lambda_2 y_1^T \hat{p} R y_2 + \underbrace{y_1^T \hat{p} p}_{0}$$

Essential Matrix

- Thus, $\lambda_2 y_1^T \hat{p} R y_2 = 0$ and since $\lambda_2 > 0$, we arrive at the following result
- **Epipolar constraint**: Consider observations $y_1 = K_1^{-1} z_1$, $y_2 = K_2^{-1} z_2$ in normalized image coordinates of the same point m from two calibrated cameras with relative pose (R, p) of camera 2 in the frame of camera 1. Then:

$$0 = y_1^T \hat{p} R y_2 = y_1^T E y_2$$

  where $E := \hat{p} R \in \mathbb{R}^{3\times 3}$ is the **essential matrix**.
- **Essential matrix characterization**: a non-zero $E \in \mathbb{R}^{3\times 3}$ is an essential matrix iff its singular value decomposition is $E = U\mathbf{diag}(\sigma, \sigma, 0)V^T$ for some $\sigma \geq 0$ and $U, V \in SO(3)$
- **Pose recovery from the Essential matrix**: There are exactly two relative poses corresponding to a non-zero essential matrix E:

$$(\hat{p}, R) = \left(U R_z \left(\frac{\pi}{2}\right) \mathbf{diag}(\sigma, \sigma, 0) U^T, U R_z^T \left(\frac{\pi}{2}\right) V^T \right)$$

$$(\hat{p}, R) = \left(U R_z \left(-\frac{\pi}{2}\right) \mathbf{diag}(\sigma, \sigma, 0) U^T, U R_z^T \left(-\frac{\pi}{2}\right) V^T \right)$$

Fundamental Matrix

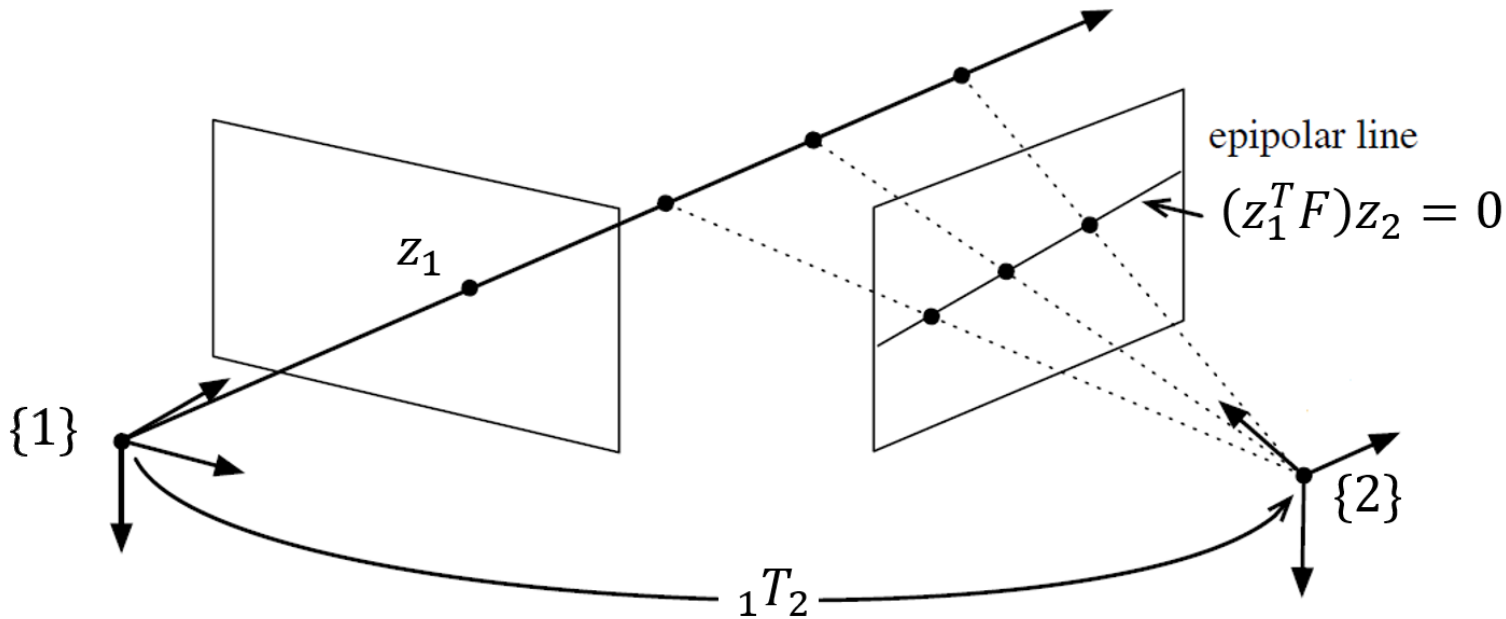
- The epipolar constraint holds even for two **uncalibrated** cameras
- Consider images $z_1 = K_1 y_1$ and $z_2 = K_2 y_2$ of the same point $m \in \mathbb{R}^3$ from two uncalibrated cameras with intrinsic parameter matrices K_1 and K_2 and relative pose (R, p) of camera 2 in the frame of camera 1:

$$0 = y_1^T \hat{p} R y_2 = y_1^T E y_2 = z_1^T K_1^{-T} E K_2^{-1} z_2 = z_1^T F z_2$$

- The matrix $F := K_1^{-T} \hat{p} R K_2^{-1}$ is called the **fundamental matrix**
- If a point m is observed in one camera z_1, and the fundamental matrix F between the two camera frames is known, the epipolar constraint desribes an **epipolar line**, along which the observation z_2 of m must lie
- The epipolar line is used to limit the search for matching points
- This is possible because the camera model is an affine transformation, i.e., a straight line in Euclidean space, projects to a straight line in image space

Epipolar Line

- If a point $m \in \mathbb{R}^3$ is observed as z_1 in one image and the fundamental matrix F is know, this can be used to define a line in the second image along which the observation z_2 must lie

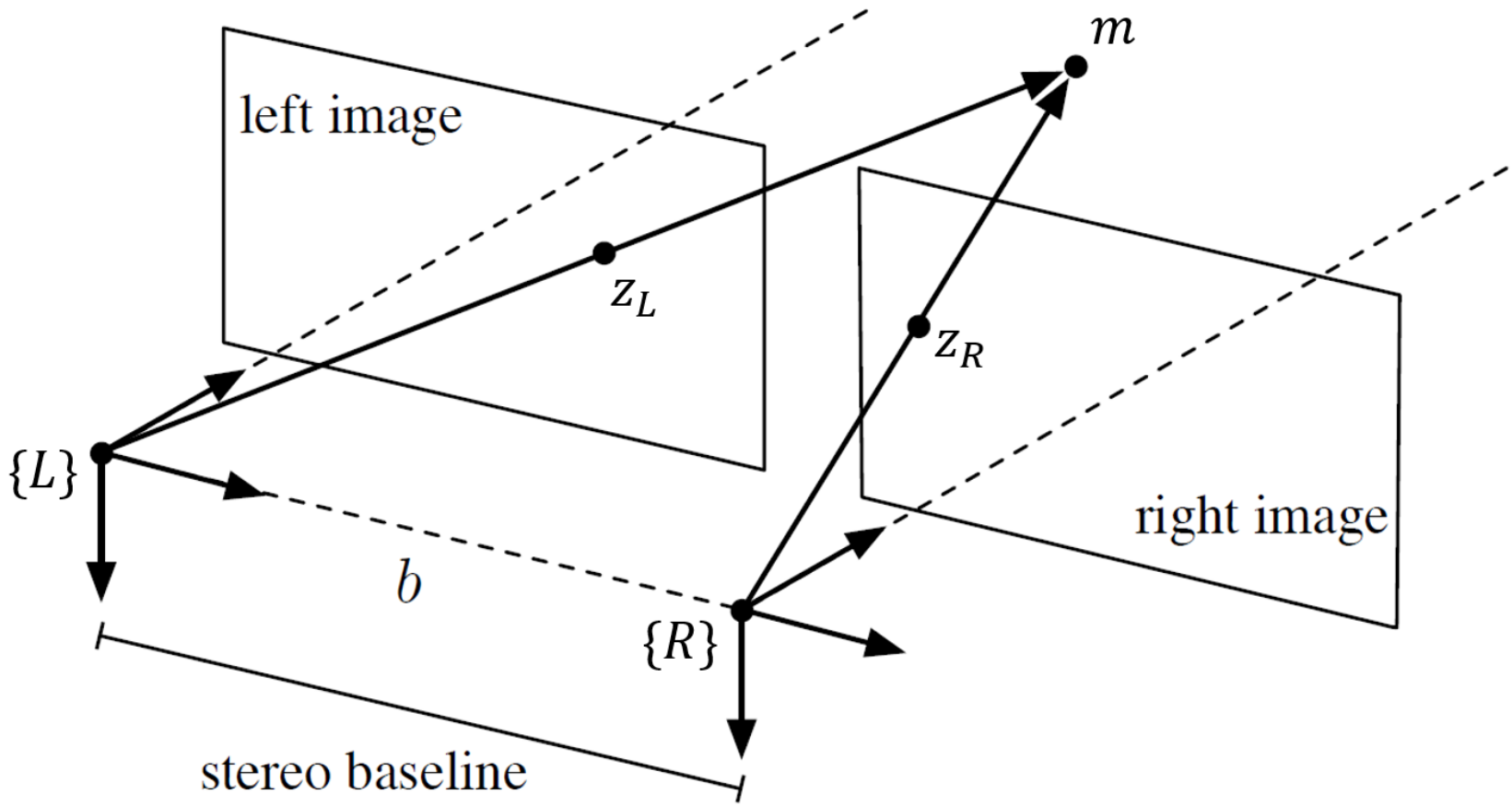


Stereo Camera Model

- **Stereo Camera**: two perspective cameras rigidly connected to one another with a known transformation
- Unlike a single camera, a stereo camera can determine the depth of a point from a single stereo observation
- **Stereo Baseline**: the transformation between the two stereo cameras is only a displacement along the x-axis (optical frame) of size b
- The pixel coordinates $z_L, z_R \in \mathbb{R}^2$ of a point $m \in \mathbb{R}^3$ in the world frame observed by a stereo camera at position $p \in \mathbb{R}^3$ and orientation $R \in SO(3)$ with intrinsic parameters $K \in \mathbb{R}^{3\times3}$ are:

$$z_L = K\pi\left(R_{oc}R^T(m-p)\right) \qquad z_R = K\pi\left(R_{oc}R^T(m-p) - b\mathbf{e}_1\right)$$

Stereo Camera Model



Stereo Camera Model

- Stacking the two observations together gives the stereo camera model:

$$\begin{bmatrix} u_L \\ v_L \\ u_R \\ v_R \end{bmatrix} = \underbrace{\begin{bmatrix} fs_u & 0 & c_u & 0 \\ 0 & fs_v & c_v & 0 \\ fs_u & 0 & c_u & -fs_u b \\ 0 & fs_v & c_v & 0 \end{bmatrix}}_{M} \frac{1}{z} \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix} \qquad \begin{bmatrix} x \\ y \\ z \end{bmatrix} = R_{oc} R^T (m - p)$$

- Because of the stereo steup, the last two rows of M are identical. The vertical coordinates of the two pixel observations are always the same because the epipolar lines in the stereo configuation are horizontal.

- The v_R equation is dropped, while the u_R equation is replaced with a **disparity** measurement $\boxed{d = u_L - u_R = \frac{1}{z} fs_u b}$ leading to:

$$\begin{bmatrix} u_L \\ v_L \\ d \end{bmatrix} = \begin{bmatrix} fs_u & 0 & c_u & 0 \\ 0 & fs_v & c_v & 0 \\ 0 & 0 & 0 & fs_u b \end{bmatrix} \frac{1}{z} \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix} \qquad \begin{bmatrix} x \\ y \\ z \end{bmatrix} = R_{oc} R^T (m - p)$$