Robot Vision

Introduction

Some slides were adapted/taken from various sources, including 3D Computer Vision of Prof. Hee, NUS, Air Lab Summer School, The Robotic Institute, CMU, Computer Vision of Prof. Mubarak Shah, UCF, Computer Vision of Prof. William Hoff, Colorado School of Mines and many more. We thankfully acknowledge them. Students are requested to use this material for their study only and **NOT** to distribute it.

Syllabus

- 3D Computer Vision:
  - Pinhole Camera projection model,
  - Epi-polar geometry, Essential and Fundamental matrix,
  - RANSAC Algorithm,
  - Solve camera pose from essential matrix,
  - Feature detector and descriptor,
  - Optical Flow: Lucas-Kanade Algorithm,
  - Camera Pose and depth estimation.
- Visual SLAM
  - Feature based: ORB SLAM, Dense direct method: DTAM, Semi dense direct method: LSD SLAM
- Learning based visual odometry
  - Supervised Method: Pose-Net, Deep-VO, Self-supervised Method: SfM Learner, Hybrid method: DSVO
- Object Tracking:
  - Introduction, Correlation Filters and MOSSE, Median Flow, Tracking-Learning-Detection, Case Study: Tracking moving objects from a UAV

Books

- **Text Books:**
  - D. A. Forsyth and J. Ponce, Computer Vision, A Modern Approach, Pearson Education, 2003.
  - Richard Hartley and Andrew Zisserman, Multiple View Geometry in Computer Vision, Cambridge University Press, 2011.
- **Reference Books**:
  - Milan Sonka, Vaclav Hlavac and Roger Boyle, Image Processing, Analysis and Machine Vision, Cengage, Third Edition (2013)

Module I:
3D Computer Vision:

- Pinhole Camera projection model
- Epi-polar geometry, Essential and Fundamental matrix
- RANSAC Algorithm
- Solve camera pose from essential matrix
- Feature detector and descriptor
- Optical Flow: Lucas-Kanade Algorithm
- Camera Pose and depth estimation

What is camera?

- A camera is a mapping between the 3D world (object space) and a 2D image.

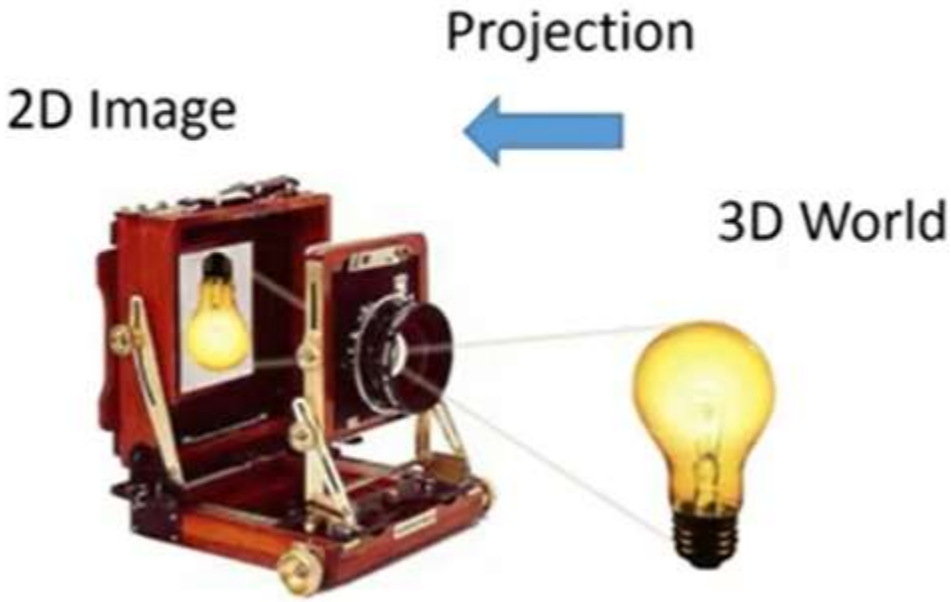


Image source: http://www.shortcourses.com/guide/guide1-3.html

Pinhole camera projection model

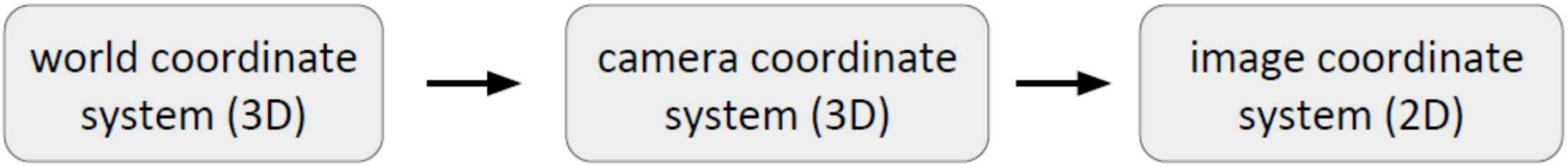
Camera projection is a transformation (mapping) between 3D world and 2D image

This mapping is described as:

$$x = PX$$

x: 2D Image point, P: Projection matrix, X: 3D world point

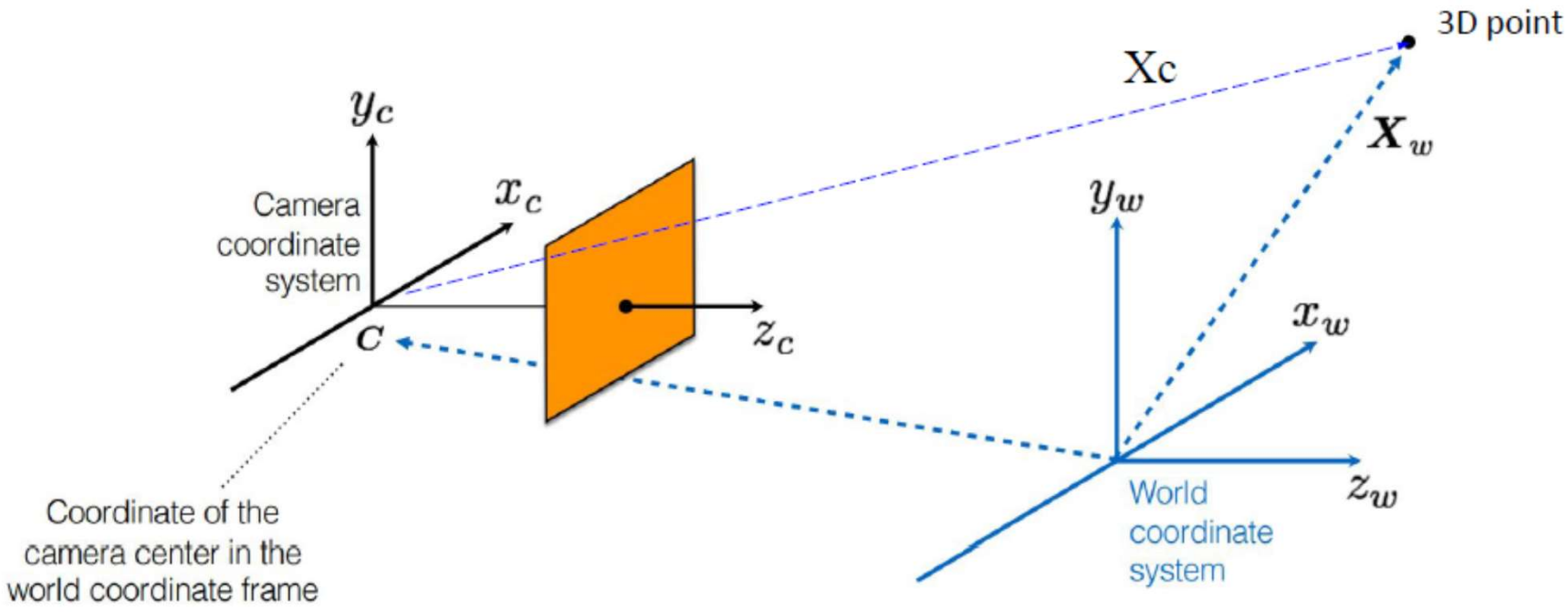
The projection consists of two parts:



Camera Models

- In this lecture, we will look at camera models with central projection.
- Camera models with central projection fall into two major classes: those with a finite centre, and those with a centre "at infinity".
- We will see more details of the projective camera with a finite centre and affine camera with a centre "at infinity".

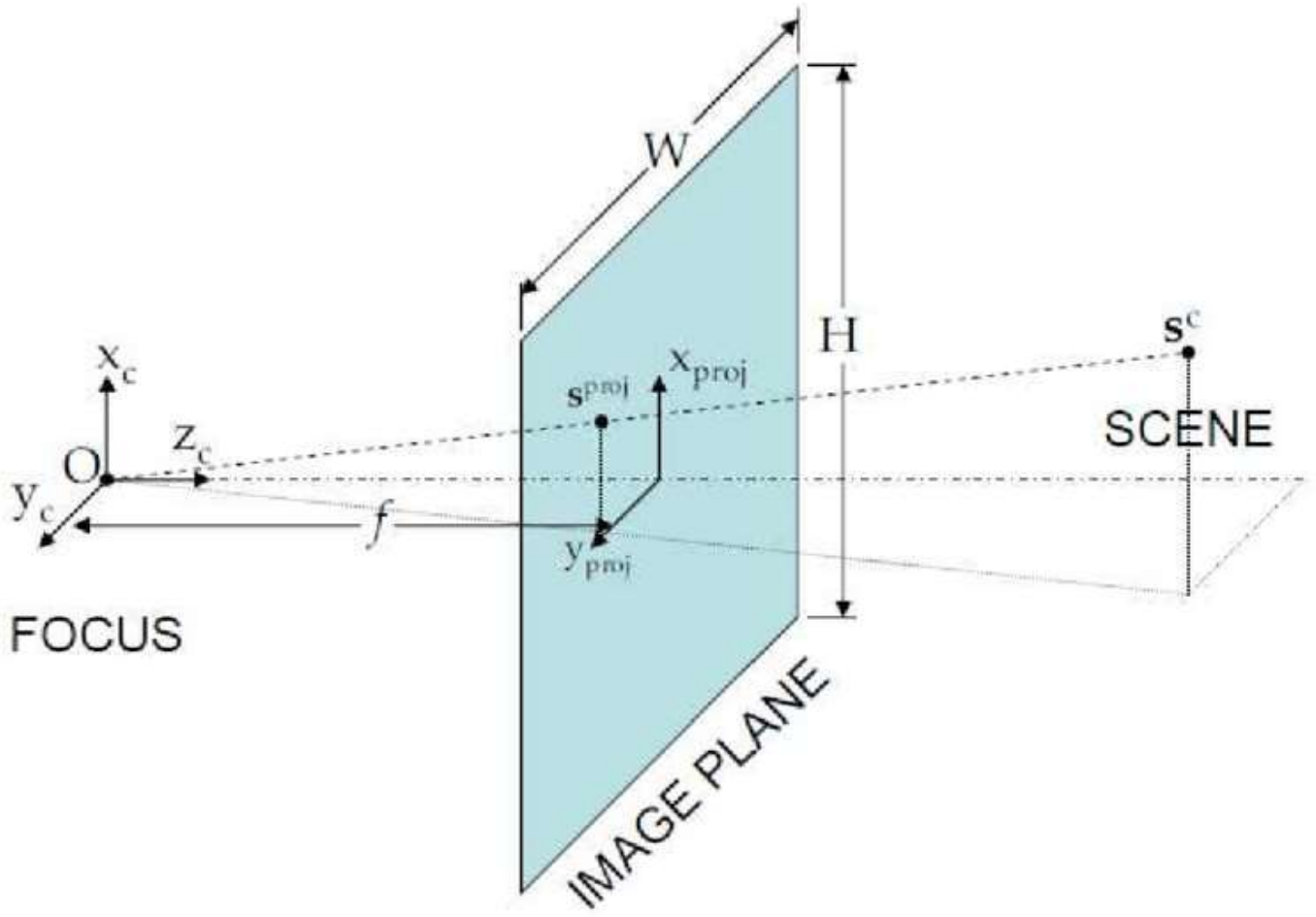
Camera Models



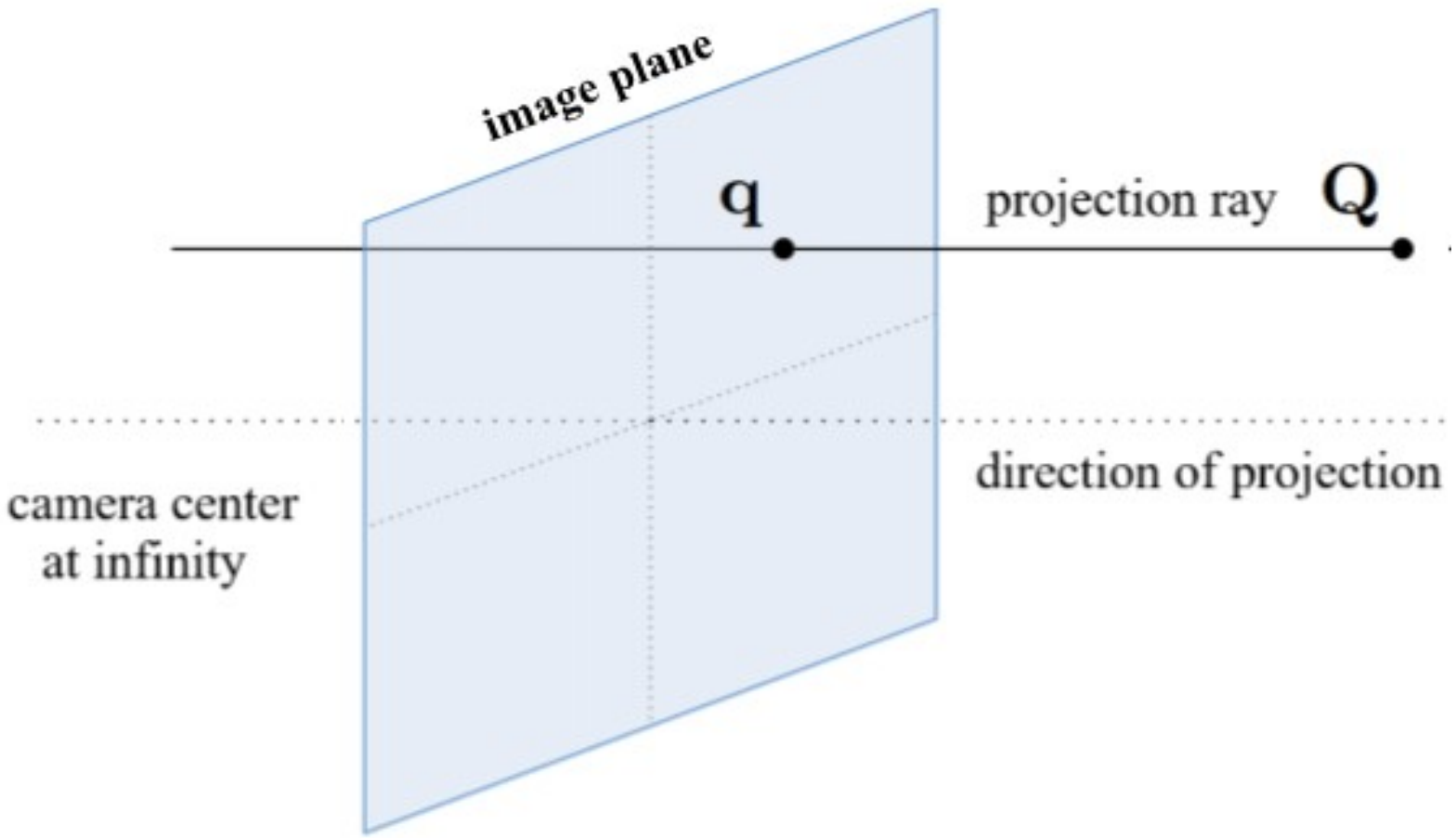
$$X_C = X_W - C$$

C is also the camera translation in relative to the world coordinate system

Projective camera with finite center



Affine camera with center at infinity



The basic pinhole model

- The projective camera is based on the basic pinhole camera.
- Let the centre of projection be the origin of a Euclidean coordinate system.
- And consider the plane Z = *f* as the image plane or focal plane.

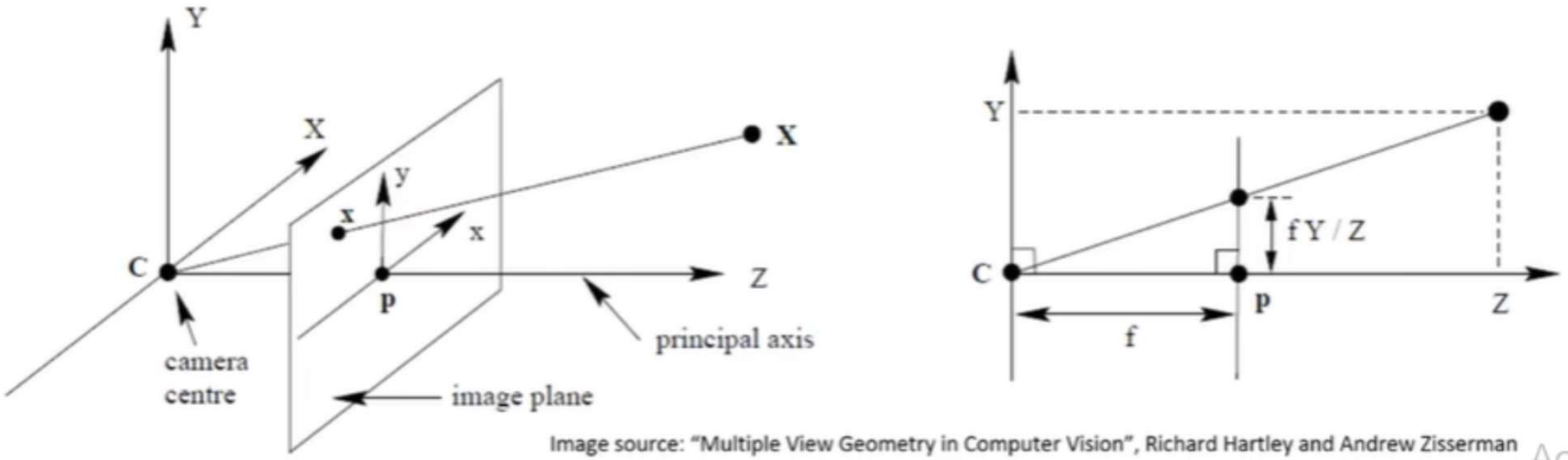


Image source: "Multiple View Geometry in Computer Vision", Richard Hartley and Andrew Zisserman

The basic pinhole model

- Using similar triangle, we can see that the point $(X, Y, Z)^T$ is mapped to the point $(fX/Z, fY/Z, f)^T$ on the image plane.
- Ignoring the final coordinate, we get the central projection mapping from world to image coordinates :

$$(\mathrm{X}, \mathrm{Y}, \mathrm{Z})^{\mathsf{T}} \mapsto (f\mathrm{X}/\mathrm{Z}, f\mathrm{Y}/\mathrm{Z})^{\mathsf{T}}, \text{ i.e. } \mathbb{R}^3 \mapsto \mathbb{R}^2$$

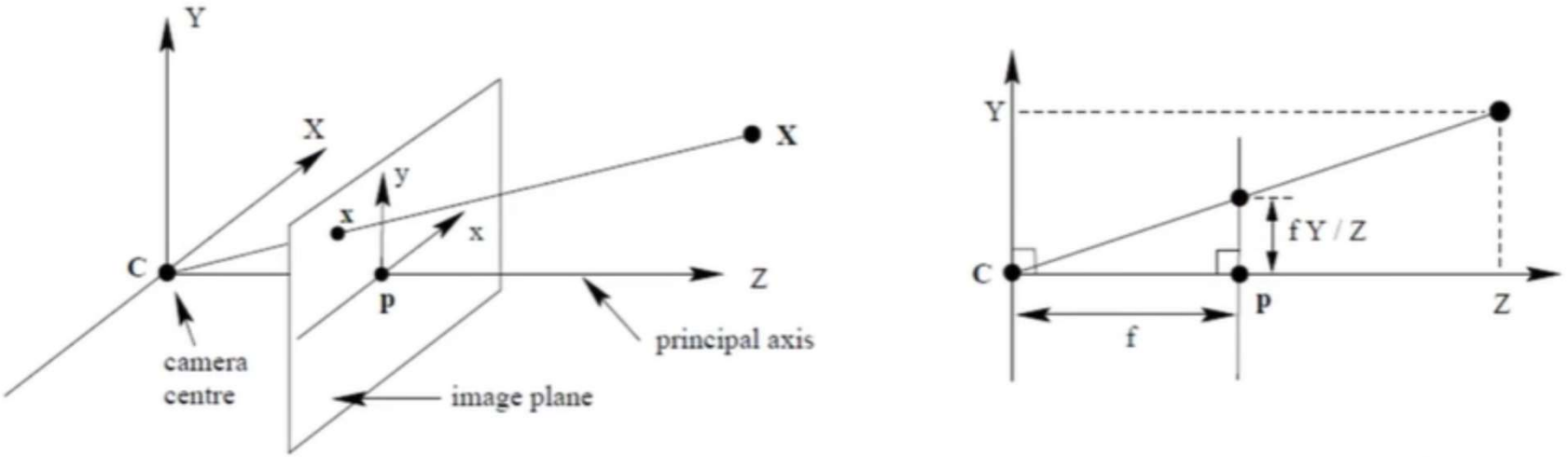


Image source: "Multiple View Geometry in Computer Vision", Richard Hartley and Andrew Zisserman

The basic pinhole model

- Principal Axis or Principal Ray: Line from camera centre perpendicular to image plane.
- Principal Point: Point where principal axis meets the image plane.
- Principal Plane: Plane through the camera centre parallel to the image plane.

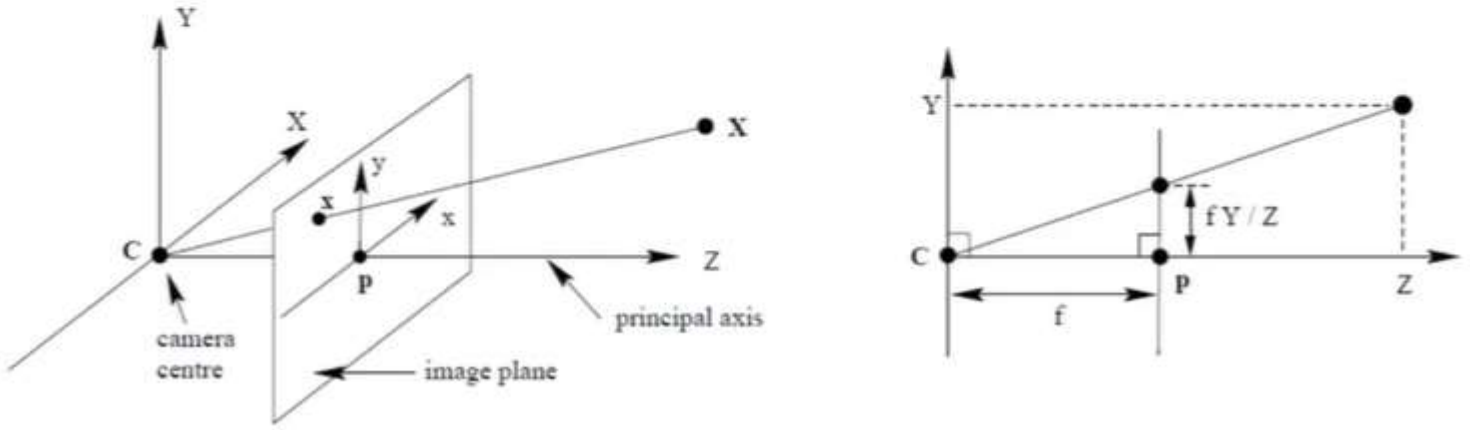


Image source: "Multiple View Geometry in Computer Vision", Richard Hartley and Andrew Zisserman

Central projection using the homogeneous coordinates

- The world and image points becomes a linear mapping in homogeneous coordinates:

$$\begin{pmatrix} X \\ Y \\ Z \\ 1 \end{pmatrix} \mapsto \begin{pmatrix} fX \\ fY \\ Z \end{pmatrix} = \underbrace{\begin{bmatrix} f & & & 0 \\ & f & & 0 \\ & & 1 & 0 \end{bmatrix}}_{\text{diag}(f,f,1)[\mathrm{I} \mid \mathbf{0}]} \begin{pmatrix} X \\ Y \\ Z \\ 1 \end{pmatrix}.$$

- Letting $\mathrm{P} = \text{diag}(f, f, 1)[\mathrm{I} \mid 0]$, $\mathbf{x} = (f\mathrm{X}, f\mathrm{Y}, \mathrm{Z})^T$ and $\mathbf{X} = (\mathrm{X}, \mathrm{Y}, \mathrm{Z}, 1)^T$, we get:

$$\mathbf{x} = \mathrm{P}\mathbf{X},$$

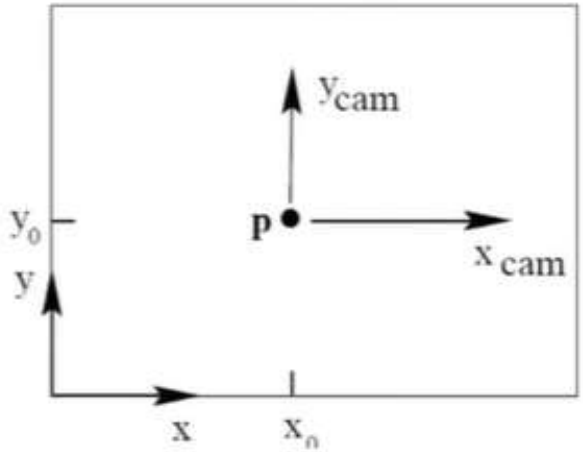
- P is the 3x4 homogeneous camera projection matrix.

Principal Point Offset

- In practice, the origin of coordinates in the image plane might not be at the principal point, i.e.

$$(\mathrm{X}, \mathrm{Y}, \mathrm{Z})^{\mathsf{T}} \mapsto (f\mathrm{X}/\mathrm{Z} + p_x, f\mathrm{Y}/\mathrm{Z} + p_y)^{\mathsf{T}}.$$

- $(p_x, p_y)^T$ are the coordinates of the principal point.



- Expressing in homogeneous coordinates, we get:

$$\begin{pmatrix} \mathrm{X} \\ \mathrm{Y} \\ \mathrm{Z} \\ 1 \end{pmatrix} \mapsto \begin{pmatrix} f\mathrm{X} + \mathrm{Z}p_x \\ f\mathrm{Y} + \mathrm{Z}p_y \\ \mathrm{Z} \end{pmatrix} = \begin{bmatrix} f & & p_x & 0 \\ & f & p_y & 0 \\ & & 1 & 0 \end{bmatrix} \begin{pmatrix} \mathrm{X} \\ \mathrm{Y} \\ \mathrm{Z} \\ 1 \end{pmatrix}.$$

Image source: "Multiple View Geometry in Computer Vision", Richard Hartley and Andrew Zisserman

Camera Calibration Matrix

- Now, writing:

$$\mathrm{K} = \begin{bmatrix} f & & p_x \\ & f & p_y \\ & & 1 \end{bmatrix},$$

- We can rewrite

$$\begin{pmatrix} f\mathrm{X} + \mathrm{Z}p_x \\ f\mathrm{Y} + \mathrm{Z}p_y \\ \mathrm{Z} \end{pmatrix} = \begin{bmatrix} f & & p_x & 0 \\ & f & p_y & 0 \\ & & 1 & 0 \end{bmatrix} \begin{pmatrix} \mathrm{X} \\ \mathrm{Y} \\ \mathrm{Z} \\ 1 \end{pmatrix} \quad \text{as} \quad \mathbf{x} = \mathrm{K}[\mathrm{I} \mid \mathbf{0}]\mathbf{X}_{\mathrm{cam}}.$$

- The matrix K is called the camera calibration matrix.

Camera Rotation and Translation

- $\mathbf{X}_{\text{cam}} = (\text{X}, \text{Y}, \text{Z}, 1)^T$ is expressed in the camera coordinate frame, where the camera is at the origin and principal axis points in the z-axis.
- In general, 3D points are expressed in a different Euclidean coordinate frame, known as the world coordinate frame.
- The two frames are related via a rigid transformation (R, t).

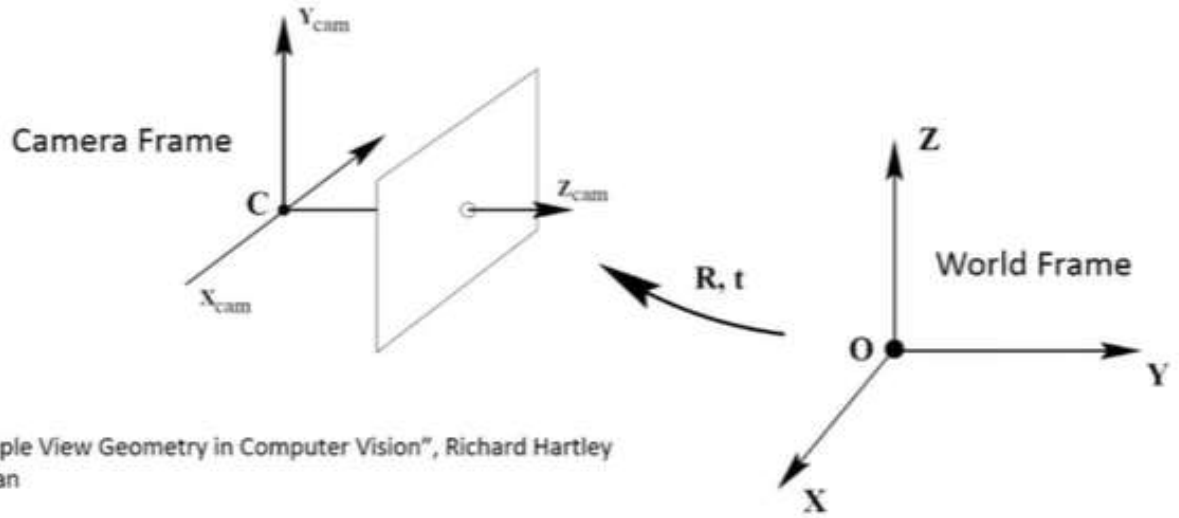
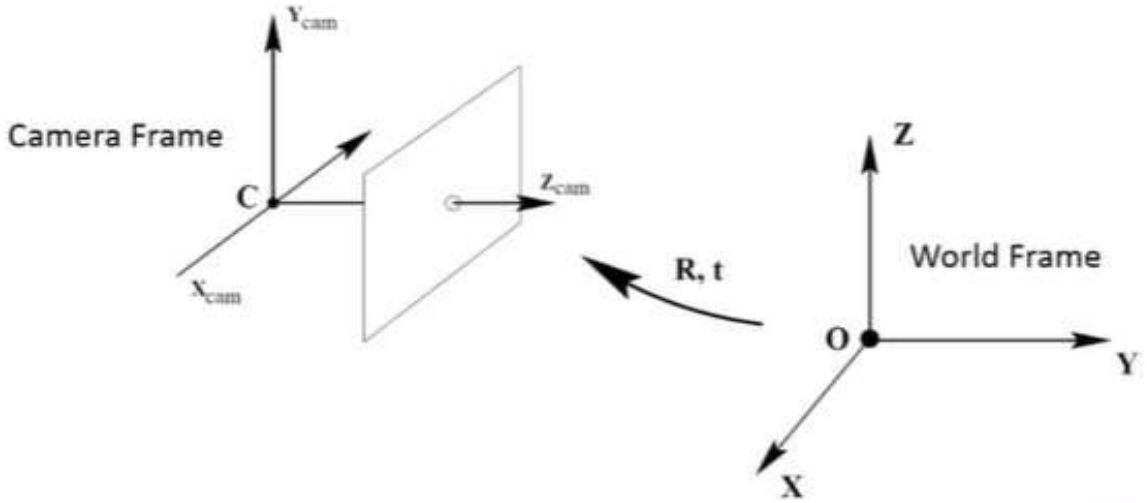


Image source: "Multiple View Geometry in Computer Vision", Richard Hartley and Andrew Zisserman

Camera Rotation and Translation

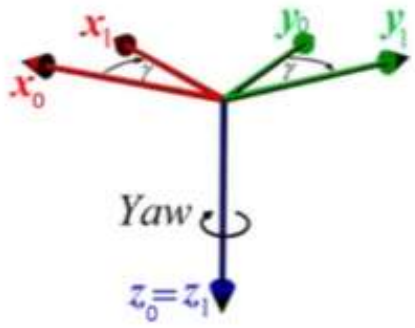


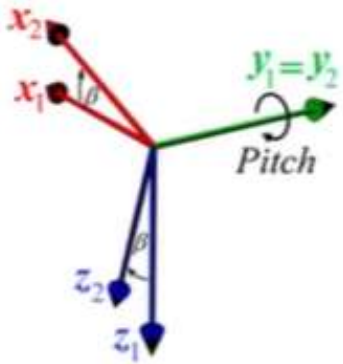
- Denoting the coordinates of the camera centre in the world frame as $\tilde{C}$, we write:

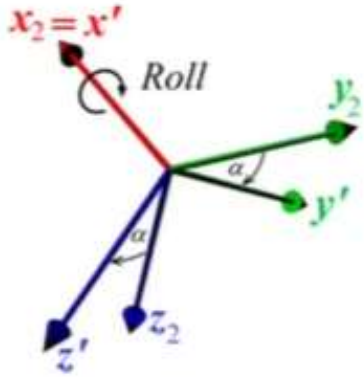
$$\mathbf{X}_{\text{cam}} = \begin{bmatrix} \mathrm{R} & -\mathrm{R}\tilde{\mathbf{C}} \\ 0 & 1 \end{bmatrix} \begin{pmatrix} \mathrm{X} \\ \mathrm{Y} \\ \mathrm{Z} \\ 1 \end{pmatrix} = \begin{bmatrix} \mathrm{R} & -\mathrm{R}\tilde{\mathbf{C}} \\ 0 & 1 \end{bmatrix} \mathbf{X}.$$

Image source: "Multiple View Geometry in Computer Vision", Richard Hartley and Andrew Zisserman

Euler Angles to Rotation Matrix







$$R_z(\gamma) = R_1^0 = \begin{bmatrix} \cos\gamma & -\sin\gamma & 0 \\ \sin\gamma & \cos\gamma & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$$\Rightarrow X_0 = R_1^0 X_1$$

$$R_y(\beta) = R_2^1 = \begin{bmatrix} \cos\beta & 0 & \sin\beta \\ 0 & 1 & 0 \\ -\sin\beta & 0 & \cos\beta \end{bmatrix}$$

$$\Rightarrow X_1 = R_2^1 X_2$$

$$R_x(\alpha) = R_3^2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\alpha & -\sin\alpha \\ 0 & \sin\alpha & \cos\alpha \end{bmatrix}$$

$$\Rightarrow X_2 = R_3^2 X_3$$

$$R_3^0 = R_1^0 R_2^1 R_3^2 = \begin{bmatrix} \cos\gamma & -\sin\gamma & 0 \\ \sin\gamma & \cos\gamma & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \cos\beta & 0 & \sin\beta \\ 0 & 1 & 0 \\ -\sin\beta & 0 & \cos\beta \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\alpha & -\sin\alpha \\ 0 & \sin\alpha & \cos\alpha \end{bmatrix}$$

$$\Rightarrow X_0 = R_3^0 X_3$$

Image Source: http://www.mdpi.com/1424-8220/15/3/7016/htm

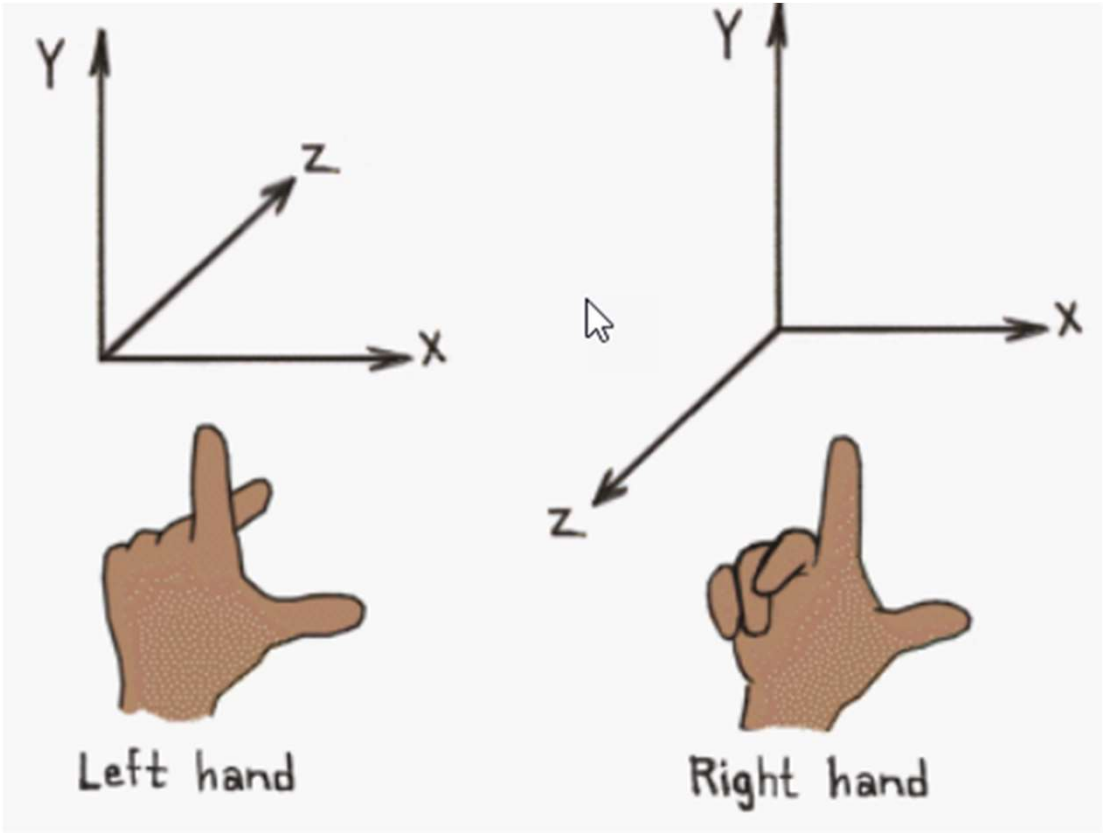
Properties of Rotation Matrix

Rotation matrices are:

- Square matrices 2x2 (2 dimensional) or 3x3 (3 dimensional) with real entries.
- Orthonormal matrices with the following properties:

1. $\det(R) = \begin{cases} +1, & \text{Right-Hand coordinate frame} \\ -1, & \text{Left-Hand coordinate frame} \end{cases}$
2. $R^T = R^{-1}$,
3. $r_i \times r_j = r_k$, (third column is the cross-product of the other two columns)
4. $r_i^T r_j = 0$, where r_i is column i of the rotation matrix
5. $\|r_1\| = \|r_2\| = \|r_3\| = 1.$

Left and right Hand Coordinate System



The Basic Pinhole Model

- Putting $\mathbf{X}_{cam}$ back into $\mathrm{x} = \mathrm{K}[\mathrm{I} \mid 0]\mathrm{X}_{\mathrm{cam}}$, we get the general mapping of a pinhole camera:

$$\mathbf{x} = \mathrm{KR}[\mathrm{I} \mid -\tilde{\mathbf{C}}]\mathbf{X}$$

  where $\mathbf{X}$ is now in a world coordinate frame.

- We write the camera projection matrix as:

$$\mathrm{P} = \mathrm{KR}[\mathrm{I} \mid -\tilde{\mathbf{C}}],$$

- P has 9 degrees of freedom: 3 for K (the elements f, p_x, p_y), 3 for R, and 3 for $\tilde{\mathbf{C}}$.

The Basic Pinhole Model

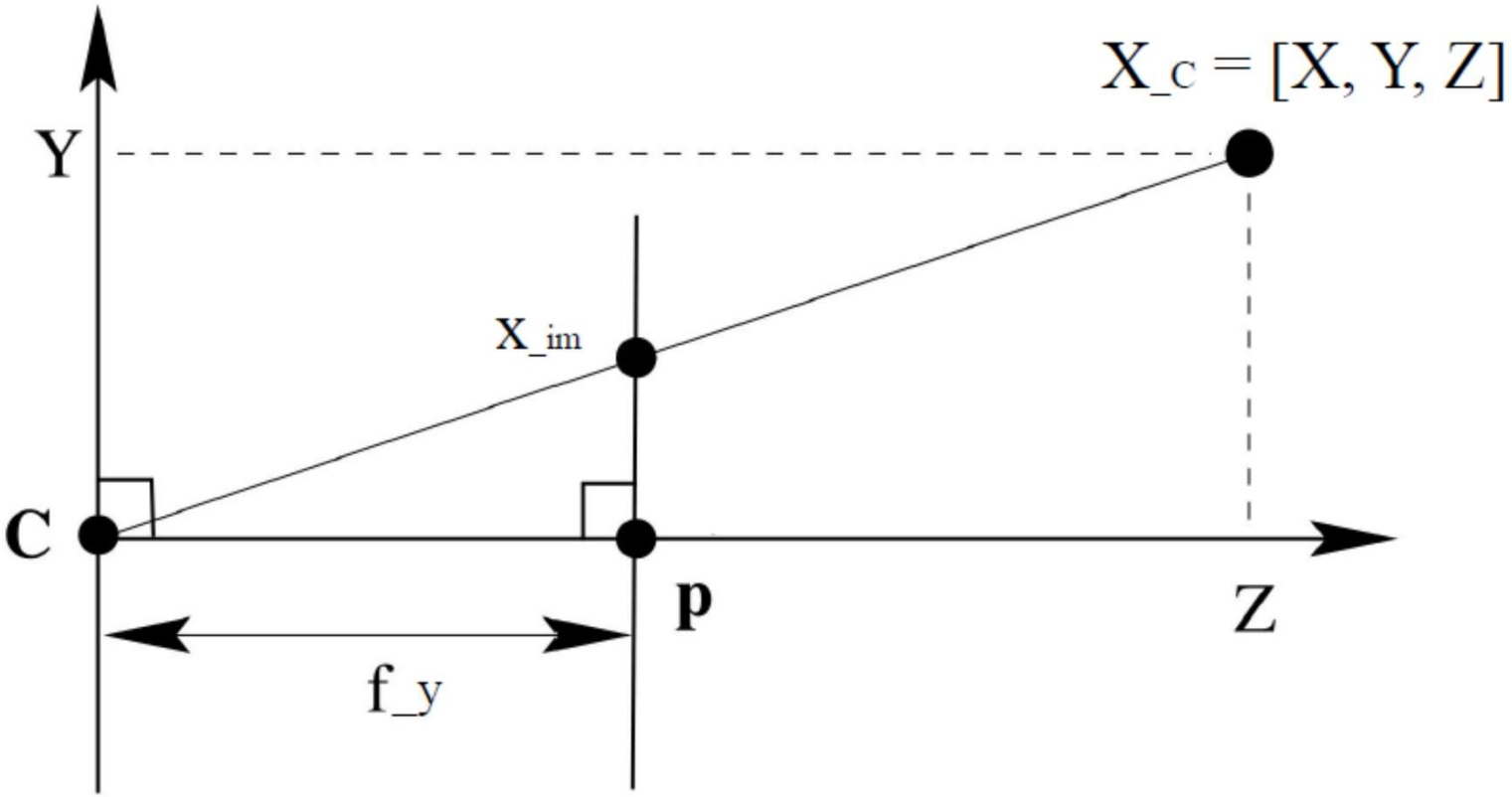
- The parameters contained in K are called the internal camera parameters, or the intrinsic of the camera.
- The parameters of R and $\tilde{\mathbf{C}}$ are called the external parameters or the extrinsic of the camera.
- It is often more convenient to represent the extrinsics in terms of (R, t):

$$\mathrm{P} = \mathrm{K}[\mathrm{R} \mid \mathbf{t}]$$

By rewriting $\mathbf{t} = -\mathrm{R}\tilde{\mathbf{C}}$.

Summary: Pinhole Camera Projection Model

Take one plane of the camera coordinate system as example:



Pinhole Camera Projection Model

What is the coordinate x, in image coordinate system?
(Hint: use properties of Similar Triangles)

$$\frac{Z}{f_y} = \frac{Y}{y} \qquad \frac{Z}{f_x} = \frac{X}{x}$$

$$y = \frac{f_y Y}{Z} \qquad x = \frac{f_x X}{Z}$$

Which is a mapping from 3D Euclidean space to 2D Euclidean space

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} \longrightarrow \begin{bmatrix} \frac{f_x X}{Z} \\ \\ \frac{f_y Y}{Z} \end{bmatrix}$$

Is it linear?
What does this remind you of?

Pinhole Camera Projection Model

Considering the principal point p,
whose coordinate in 2D image coordinate system is [c_x, c_y],

$$x_{im} = \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} \frac{f_x X}{Z} + c_x \\ \frac{f_y Y}{Z} + c_y \end{bmatrix}$$

$$\tilde{x_{im}} = \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \begin{bmatrix} \frac{f_x X}{Z} + c_x \\ \frac{f_y Y}{Z} + c_y \\ 1 \end{bmatrix}$$

This gives us the coordinate of a projected 3D point, in 2D image coordinate system
How to write this mapping in matrix form?

Pinhole Camera Projection Model

$$\tilde{x_{im}} = K X_C$$

$$\begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \begin{bmatrix} f_x X + c_x Z \\ f_y Y + c_y Z \\ Z \end{bmatrix} = Z \begin{bmatrix} f_x X/Z + c_x \\ f_y Y/Z + c_y \\ 1 \end{bmatrix}$$

K is called **camera intrinsic matrix**, **camera intrinsic parameters,** or **calibration matrix**

$$K = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix}$$

Pinhole Camera Projection Model

Combine the 3D transformation together:

$$\tilde{X_C} = T\tilde{X_W} = \begin{bmatrix} R & t \\ \mathbf{0}^T & 1 \end{bmatrix} \tilde{X_W}$$

$$\tilde{x_{im}} = K\tilde{X_C} = K \begin{bmatrix} R & t \\ \mathbf{0}^T & 1 \end{bmatrix} \tilde{X_W}$$

Sometimes we write in this form,

$$\tilde{x_{im}} = K\tilde{X_C} = K[R|t]\tilde{X_W}$$

[R|t] is called **camera extrinsic parameters**

… to continue