1

Projective Camera Model

1.1 Projection in Metric Space

Consider a point light emitted from the top of the Eiffel tower as shown in Figure 1.1. The light is located at located at $\mathbf{X}_c = (X_c, Y_c, Z_c)^\mathsf{T}$ where the coordinate system is aligned with the camera optical axis, i.e., the origin is at the camera center (pinhole), and Z axis is the camera look-out vector (perpendicular to the CCD sensor plane). While the light travels over all directions, a particular light passes through the pinhole of the camera and is projected onto the camera's CCD plane. The CCD plane is centered at $Z_{\text{screen}} = -f_m$ where f_m is focal length of the camera, e.g., iPhone 7 camera has 29mm focal length.

With the projected point in the CCD plane, it forms two similar triangles:

$$\tan\theta_x = \frac{X_c}{Z_c} = \frac{u_{\text{ccd}}}{f_m}, \tag{1.1}$$

where θ_x is the vertical angle (opposite angles made by two intersecting lines), and u_{ccd} is the x-coordinate of the projected point. Similarly, the y-coordinate can be written as:

$$\tan\theta_y = \frac{Y_c}{Z_c} = \frac{v_{\text{ccd}}}{f_m}. \tag{1.2}$$

This is called *perspective projection* where a 3D point $(X_c, Y_c, Z_c)^\mathsf{T} \in \mathbb{R}^3$ is

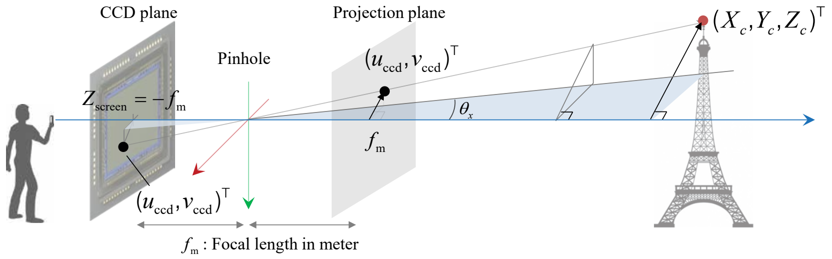


Figure 1.1: A 3D point $(X_c, Y_c, Z_c)^\mathsf{T}$ is projected onto the CCD plane at f_m to form $(u_{\text{ccd}}, v_{\text{ccd}})^\mathsf{T}$. One dimension (depth) is lost during this metric projection. To simplify the representation, we will use the mirror image of the CCD plane.

mapped to a 2D point $(u_{\rm ccd}, v_{\rm ccd})^{\mathsf{T}} \in \mathbb{R}^2$ using the following equations:

$$\begin{bmatrix} u_{\rm ccd} \\ v_{\rm ccd} \end{bmatrix} = -f_m \begin{bmatrix} \tan\theta_x \\ \tan\theta_y \end{bmatrix} = -\frac{f_m}{Z_c}\begin{bmatrix} X_c \\ Y_c \end{bmatrix}. \tag{1.3}$$

Note 1.1. *This projection produces an upside-down image due to the negative sign.*

Note 1.2. *There is one dimensional information loss ($\mathbb{R}^3 \rightarrow \mathbb{R}^2$) due to the projection.*

This projection is equivalent to the projection onto the virtual screen in front of the pinhole at f_m, which can drop the negative sign and correct the upside-down image:

$$\begin{bmatrix} u_{\rm ccd} \\ v_{\rm ccd} \end{bmatrix} = \frac{f_m}{Z_c}\begin{bmatrix} X_c \\ Y_c \end{bmatrix}. \tag{1.4}$$

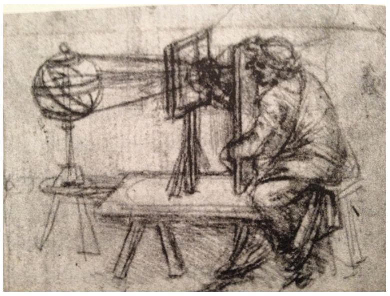


Figure 1.2: Camera obscura sketched by Leonardo da Vinci in Codex Atlanticus (1515), preserved in Biblioteca Ambrosiana, Milan (Italy).

This representation is similar to da Vinci's camera obscura sketch as shown in Figure 1.2 illustrating a perspective painting tool.

Example 1.1 (Subject size). *Consider a person (1.8m height) at 4m away from a photographer using an iPhone 7. How big does this person appear in the image?*
Answer iPhone 7 has 3.99mm focal length[1] and $1/3$ inch sensor size (4.8mm×3.6mm). The person will occupy the half of the image because:

$$v_{\rm ccd} = 3.99\text{mm} \times \frac{1.8\text{m}}{4\text{m}} \approx 1.8\text{mm},$$

which is a half of the CCD sensor's height (3.6mm).

1.2 Projection in Pixel Space

Equation (1.4) is the projection of a 3D point onto the CCD plane in metric scale. As a result of the projection, the light intensity is recorded in a form of an image represented by pixels, i.e., the CCD contains an array of photo-receptors where each receptor corresponds to a pixel in the image. The projected location in the CCD plane is directly related with the pixel location in the image: the relative locations in the CCD sensor with respect to the sensor size and in the image with respect to the image size are the same:

$$\frac{u_{\rm img}}{W_{\rm img}} = \frac{u_{\rm ccd}}{W_{\rm ccd}}, \quad \frac{v_{\rm img}}{H_{\rm img}} = \frac{v_{\rm ccd}}{H_{\rm ccd}} \tag{1.5}$$

where $(W_{\rm ccd}, H_{\rm ccd})$ is width and height of the CCD sensor, e.g., (4.8mm×3.6mm) for iPhone 7, and $(W_{\rm img}, H_{\rm img})$ is width and height of the image, e.g., (4k pix×3k pix). $(u_{\rm img}, v_{\rm img})^{\mathsf{T}}$ is the equivalent location of $(u_{\rm ccd}, v_{\rm ccd})^{\mathsf{T}}$ in the image as shown in Figure 1.2.

[1] 28mm with 7.21 crop factor in 35mm equivalent focal length.

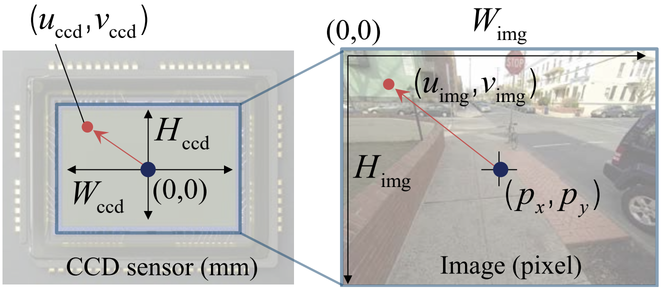


Figure 1.3: Photoreceptive sensors in CCD plane (metric space) is mapped to generate an image (pixel space).

In machine vision convention, the left top corner of an image has been used for the origin $(0,0)^\mathsf{T}$. This differs from the metric representation in Equation (1.4) where the center of the CCD plane is used for the origin, i.e., the pixel coordinate needs to be shifted. This results in introducing a notion of image center called *principal point* in pixel $(p_x, p_y)^\mathsf{T}$ that can change the pixel coordinate as follows:

$$\frac{u_{\text{img}} - p_x}{W_{\text{img}}} = \frac{u_{\text{ccd}}}{W_{\text{ccd}}}, \quad \frac{v_{\text{img}} - p_y}{H_{\text{img}}} = \frac{v_{\text{ccd}}}{H_{\text{ccd}}} \tag{1.6}$$

Note 1.3. *The principal point is often located near the center of an image, i.e.,* $(640, 480)^\mathsf{T}$ *for 1280×960 size image, as it is the origin of the CCD sensor where the CCD plane and Z axis intersect.*

By combining Equation (1.4) and (1.6), the projection of a 3D point can be represented in pixel:

$$\begin{aligned} u_{\text{img}} &= f_m \frac{W_{\text{img}}}{W_{\text{ccd}}} \frac{X_c}{Z_c} + p_x = f_x \frac{X_c}{Z_c} + p_x \\ v_{\text{img}} &= f_m \frac{H_{\text{img}}}{H_{\text{ccd}}} \frac{Y_c}{Z_c} + p_y = f_y \frac{Y_c}{Z_c} + p_y \end{aligned} \tag{1.7}$$

where

$$f_x = f_m \frac{W_{\text{img}}}{W_{\text{ccd}}}, \quad f_y = f_m \frac{H_{\text{img}}}{H_{\text{ccd}}}. \tag{1.8}$$

f_x and f_y are focal lengths in pixel. If the aspect ratio of CCD is consistent with that of image, $\frac{W_{\text{img}}}{W_{\text{ccd}}} = \frac{H_{\text{img}}}{H_{\text{ccd}}}$, then, $f_x = f_y$.

Equation (1.7) directly relates a 3D point in metric space with 2D projection in pixel space bypassing CCD mapping. In a matrix form, Equation (1.7) can be re-written as:

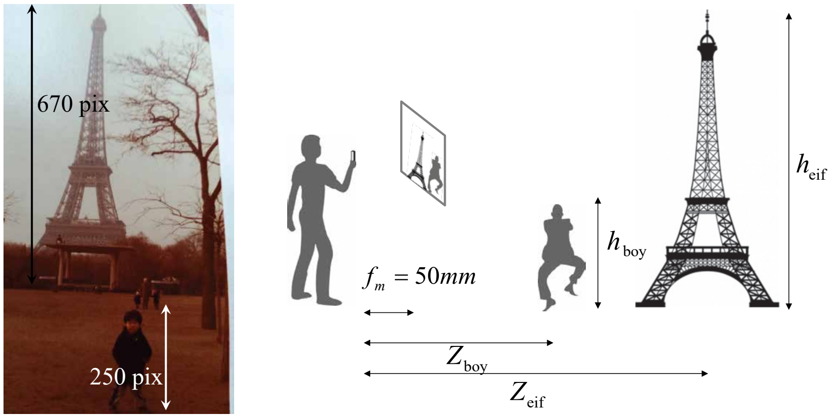
$$\begin{aligned} Z_c \begin{bmatrix} u_{\text{img}} \\ v_{\text{img}} \\ 1 \end{bmatrix} &= \begin{bmatrix} f_x & 0 & p_x \\ 0 & f_y & p_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} X_c \\ Y_c \\ Z_c \end{bmatrix}, \\ \text{or,} \quad Z_c \begin{bmatrix} \mathbf{u}_{\text{img}} \\ 1 \end{bmatrix} &= \mathbf{K}\mathbf{X}_c, \end{aligned} \tag{1.9}$$

where $\mathbf{u}_{\rm img} = (u_{\rm img}, v_{\rm img})^{\mathsf{T}}$ and $\mathbf{X}_c = (X_c, Y_c, Z_c)^{\mathsf{T}}$. $\mathbf{K}$ is called the camera's intrinsic parameter that transforms $\mathbf{X}_c$ in metric space to $\mathbf{u}_{\rm img}$ in pixel space. It includes focal length and principal point in pixel.

Z_c is the depth of the 3D point with respect to the camera, which is often unknown due to loss of dimension. We represent the unknown depth factor using λ:

$$\lambda \begin{bmatrix} \mathbf{u}_{\rm img} \\ 1 \end{bmatrix} = \mathbf{K}\mathbf{X}_c \tag{1.10}$$

where λ is an unknown scalar.



(a) Image

(b) Geometry

Figure 1.4: It is possible to predict the location of the boy (where am I?) with respect to the Eiffel Tower.

Example 1.2 (Where was I?). *Consider an old picture in Figure 1.4(a) taken from a film camera with approximately 50mm focal length. The 4 years old boy appears 250 pixels in the image where the height of image is 1280 pixels while the Eiffel Tower appears 670 pixels. Where was the boy?*

Answer In an old day, 35mm film has been widely used. The distance from the photographer to the boy (average height of 4 year old male: $h_{\rm boy}$ =102.3cm) is:

$$h^{\rm img}_{\rm boy} = f_m \frac{H_{\rm img}}{H_{\rm ccd}} \frac{h_{\rm boy}}{Z_{\rm boy}}$$
$$Z_{\rm boy} = 50{\rm mm} \times \frac{1280{\rm pix}}{35{\rm mm}} \times \frac{1.023{\rm m}}{250{\rm pix}} = 7.48{\rm m}.$$

Similarly, the distance from the photographer to the Eiffel Tower (324m) is:

$$h^{\rm img}_{\rm eif} = f_m \frac{H_{\rm img}}{H_{\rm ccd}} \frac{h_{\rm eif}}{Z_{\rm eif}}$$
$$Z_{\rm eif} = 50{\rm mm} \times \frac{1280{\rm pix}}{35{\rm mm}} \times \frac{324{\rm m}}{670{\rm pix}} = 884.2{\rm m}.$$

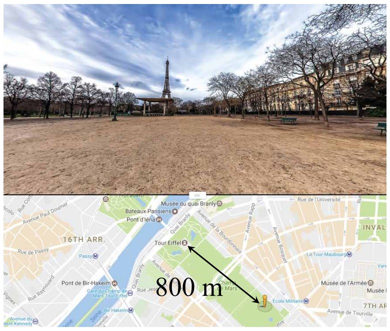


Figure 1.5: We verify the distance prediction using Google Streetview, which is similar to Figure 1.4(a).

The distance from the boy to the Eiffel Tower is approximately 877m. Figure 1.5 shows a Google Streetview image taken a similar location, which is approximately 800m away from the Eiffel Tower.

1.3 Property of Intrinsic Parameter

The focal length of a camera defines the field of view and zoom factor. Larger focal length makes the pixel coordinate larger in Equation (1.7), resulting in zooming in and reducing field of view as shown in Figure 1.8 and 1.3. This also makes an object appeared relatively flat because the camera focuses on large distant objects ($Z \gg 0$) where the object's thickness Δd is relatively smaller as shown in Figure 1.6:

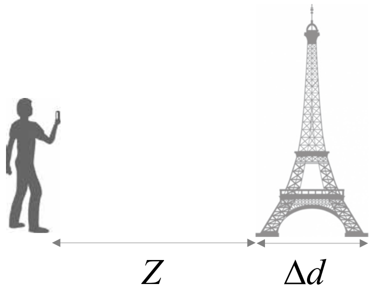


Figure 1.6: For large focal length, objects in the image are seen relatively flat.

$$u_{\mathrm{img}} = f_x \frac{X}{Z+\Delta d} + p_x \approx f_x \frac{X_c}{Z} + p_x \qquad (1.11)$$

if $Z \gg \Delta d$. Conversely, smaller focal length produces wide field of view and generates greater perspective sensation.

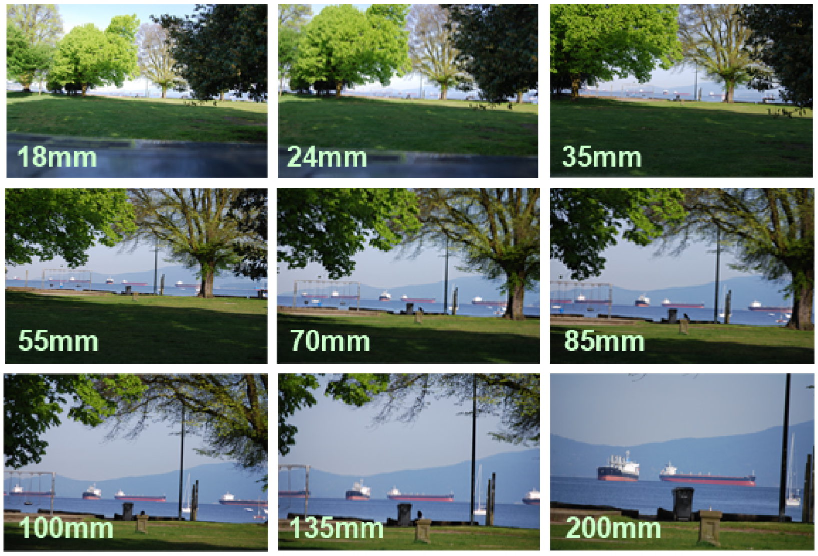


Figure 1.7: An image with larger focal length produces zoom-in effect.

As noted earlier, the principal point $(p_x, p_y)^\mathsf{T}$ in the intrinsic parameter is often aligned with the center of image. However, the mechanical misalignment can occur due to lens and sensor geometric configuration where the center of lens is shifted with a few pixels caused by errors in camera manufacturing. Also physical force applied to modern cameras mounted on cellphone, laptop, and tablet can result in mis-alignment.

In some case where the sensor and lens are not parallel, a skewed image

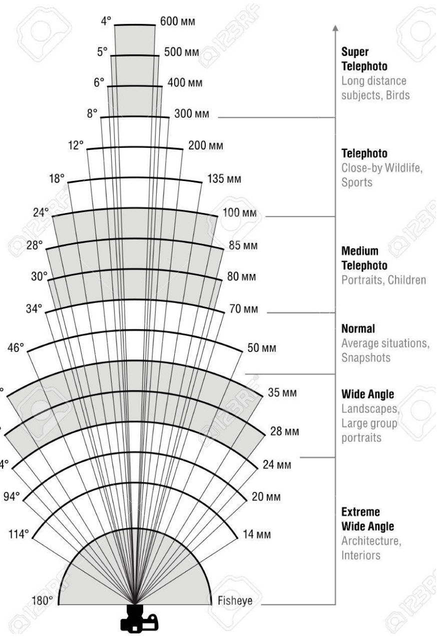


Figure 1.8: The larger focal length, the smaller field of view.

can be produced:

$$Z_c \begin{bmatrix} u_{\text{img}} \\ v_{\text{img}} \\ 1 \end{bmatrix} = \begin{bmatrix} f_x & s & p_x \\ 0 & f_y & p_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} X_c \\ Y_c \\ Z_c \end{bmatrix}, \qquad (1.12)$$

where s is the skew parameter that x-coordinate of image is dependent on Y_c. Equation (1.12) is the general linear projective camera model, which can be calibrated in Section **??**.

1.4 Dolly Zoom

Dolly-zoom or Vertigo effect is an in-camera visual effect that produces a 3D sensation. The camera moves along with a dolly where the focal length of the camera is precisely controlled to keep the size of the subject constant. This induces background motion while the foreground stationary, which produces various perceptual experiences. Alfred Hitchcock first introduced the dolly-zoom effect in his thriller movie, *Vertigo (1958)* to convey the dizziness of the actor, and many subsequent modern films such as *Pycho (1960)*, *Jaws (1975)*, and *Goodfellas (1990)* have employed this effect.

Figure 1.9: Alfred Hitchcock's movie, *Vertigo (1958)*

The key insight of the dolly-zoom effect is to describe the focal length, f_m, as a function of the depth, Z_c, using Equation (1.7):

$$h_{\text{img}} = f\frac{h_o}{Z_o} = f^*\frac{h_o}{Z_o + \Delta Z}, \qquad (1.13)$$

where h_{img} is the size (height or width) of the focused subject in an image, h_o is the size of the focused object in metric space at Z_o distant from the camera. ΔZ is the distance that the camera moves along the dolly and f_m^* is the controlled focal length. Note that h_{img} needs to keep constant as ΔZ changes. The controlled focal length is

$$f^* = f\frac{Z_o + \Delta Z}{Z_o}. \qquad (1.14)$$

Note that the size of the subject is canceled.

Example 1.3 (Dolly zoom effect)**.** *Consider foreground object* A *(*h_{A} *=4m) and background object* B *(*h_{B} *=6m) where the distance between them is* d *=2m as shown in Figure 1.3. They appear* $h_{\mathsf{A}}^{\text{img}}$ *=400 pix and* $h_{\mathsf{B}}^{\text{img}}$ *=120 pix, respectively, as shown in Figure 1.11. How far does the camera need to step back to create a dolly zoom effect where* $\times 2$ *zoom factor is applied? How high will the background object be after the dolly zoom?*

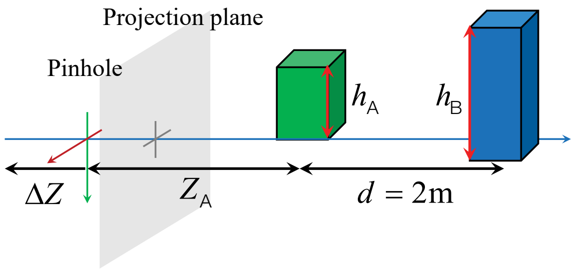


Figure 1.10: Dolly zoom geometry.

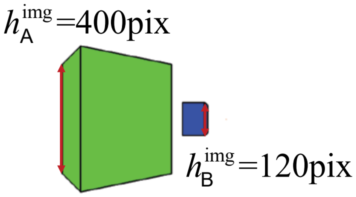

$h_{\mathsf{A}}^{\text{img}}$=400pix
$h_{\mathsf{B}}^{\text{img}}$=120pix

Figure 1.11: Image at Z_{A}.

Solution From Equation (1.13), the foreground object A appears constant:

$$h_{\mathsf{A}}^{\text{img}} = f\frac{h_{\mathsf{A}}}{Z_{\mathsf{A}}} = 2f\frac{h_{\mathsf{A}}}{Z_{\mathsf{A}} + \Delta Z}, \tag{1.15}$$

$$\Rightarrow \Delta Z = Z_{\mathsf{A}}.$$

Also, background object B satisfies the following:

$$h_{\mathsf{B}}^{\text{img}} = f\frac{h_{\mathsf{A}}}{Z_{\mathsf{A}} + d}. \tag{1.16}$$

Z_{A} and f are unknown. From Equation (1.15) and (1.16),

$$Z_{\mathsf{A}} = \frac{d}{\frac{h_{\mathsf{A}}^{\text{img}}}{h_{\mathsf{B}}^{\text{img}}}\frac{h_{\mathsf{B}}}{h_{\mathsf{A}}} - 1} = 0.5\text{m},$$

and therefore, ΔZ =0.5m. f can be computed by back-substitution, which is 50pix.

Finally, the background object B will appear:

$$\widehat{h}_{\mathsf{B}}^{\text{img}} = 2f\frac{h_{\mathsf{B}}}{Z_{\mathsf{A}} + \Delta Z + d} = 100\text{pix}.$$

1.5 *First Person Camera in World Coordinate System*

The coordinate of the 3D world point, $\mathbf{X}_c$ in Equation (1.10) aligns with the first person camera. If the camera moves, the point representation also changes as the camera coordinate system moves although it is stationary point. Alternatively, it can be represented by the fixed world coordinate, $\mathbf{X}_w$ with an Euclidean transform to the camera:

$$\mathbf{X}_c = \mathbf{R}\mathbf{X}_w + \mathbf{t}, \tag{1.17}$$

where $\mathbf{R} \in SO(3)$ is a rotation matrix which is orthogonal, i.e., $\mathbf{R}^{\mathsf{T}}\mathbf{R} = \mathbf{R}\mathbf{R}^{\mathsf{T}} = \mathbf{I}_3$ and $\det(\mathbf{R}) = 1$, that transforms the third person world coordinate to first person camera coordinate. $\mathbf{t} \in \mathbb{R}^3$ is a translation vector which is the world origin seen from the camera coordinate system at the camera origin, i.e., $\mathbf{t}$ is represented in the camera coordinate.

Equivalently, the Euclidean transformation can be written as:

$$\mathbf{X}_c = \mathbf{R}\left(\mathbf{X}_w - \mathbf{C}\right), \tag{1.18}$$

where $\mathbf{C}$ is the camera origin seen from the world coordinate where $\mathbf{C} = -\mathbf{R}^\mathsf{T}\mathbf{t}$. Note that Equation (1.17) translates after the rotation while Equation (1.18) translates before the rotation.

By combining Equation (1.10) and (1.17),

$$\lambda \begin{bmatrix} \mathbf{u}_{\text{img}} \\ 1 \end{bmatrix} = \mathbf{K} \begin{bmatrix} \mathbf{R} \mid \mathbf{t} \end{bmatrix} \begin{bmatrix} \mathbf{X}_w \\ 1 \end{bmatrix} \tag{1.19}$$

$$= \mathbf{K}\mathbf{R} \begin{bmatrix} \mathbf{I} \mid -\mathbf{C} \end{bmatrix} \begin{bmatrix} \mathbf{X}_w \\ 1 \end{bmatrix} \tag{1.20}$$

$$= \mathbf{P} \begin{bmatrix} \mathbf{X}_w \\ 1 \end{bmatrix}, \tag{1.21}$$

where $\mathbf{P} \in \mathbb{R}^{3\times 4}$ is called camera projection matrix that maps a 3D world coordinate to the first person image (metric space to pixel space). It includes intrinsic parameter (focal length and principal point) and extrinsic parameter (rotation and translation).

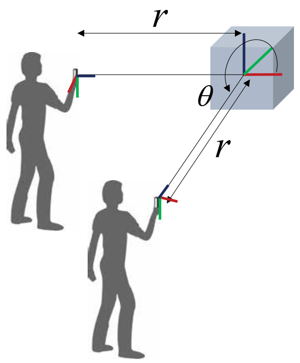


Figure 1.12: Camera motion orbiting around a cubic.

MATLAB 1.1 (Orbiting camera)**.** *Draw 3D polygons such as cubics similar to Figure 1.12 and animate the sequence of images while the camera orbiting around the polygon.*

Answer The camera motion can be written as:

$$\mathbf{C} = \begin{bmatrix} r\cos\theta \\ r\sin\theta \\ 0 \end{bmatrix}, \mathbf{R} = \begin{bmatrix} -\sin\theta & \cos\theta & 0 \\ 0 & 0 & -1 \\ -\cos\theta & -\sin\theta & 0 \end{bmatrix} \tag{1.22}$$

RotateCamera.m

```
K = [200 0 100;
    0 200 100;
    0 0 1];
r = 5;
theta = 0:0.02:2*pi;
for i = 1 : length(theta)
   C = [r*cos(theta(i)); r*sin(theta(i)); 0];
   R = [-cos(theta(i)) -sin(theta(i)) 0;
       0 0 -1;
       -sin(theta(i)) cos(theta(i)) 0];
   P = K * R * [eye(3) -C];
   proj = [];
   for j = 1 : size(sqaure_point,1)
      u = P * [sqaure_point(j,:)';1];
      proj(j,:) = u'/u(3);
   end
end
```

1.6 *Inverse Projection*

Can we predict the location of a 3D world point $\mathbf{X}_w$ given a 2D point $\mathbf{u}_{\text{img}}$ in image? Without an assumption about the scene, it is impossible because of dimensional loss of projection, i.e., depth is unknown. Then, what does the 2D point mean in 3D?

A 2D point in in an image is equivalent to a 3D ray $\mathbf{L}$ emitted from the pinhole passing through the 2D point in the projection plane:

$$\mathbf{L} = \lambda \mathbf{K}^{-1} \begin{bmatrix} u_{\text{img}} \\ v_{\text{img}} \\ 1 \end{bmatrix}, \quad \text{for} \quad \lambda \geq 0. \tag{1.23}$$

This ray also passes through the 3D point, i.e., $\mathbf{X}_c \in \mathbf{L}$.

For the world coordinate representation,

$$\mathbf{X}_w \in \mathbf{R}^{\mathsf{T}}\mathbf{L} + \mathbf{C}, \tag{1.24}$$

$$\text{or,} \quad \mathbf{X}_w = \lambda \mathbf{R}^{\mathsf{T}} \mathbf{K}^{-1} \begin{bmatrix} u_{\text{img}} \\ v_{\text{img}} \\ 1 \end{bmatrix} + \mathbf{C}, \quad \text{for some } \lambda. \tag{1.25}$$

The RHS of Equation (1.25) transforms the ray $\mathbf{L}$ in the camera coordinate to the world coordinate, i.e., it is emitted from the optical center of the camera and oriented to $\mathbf{L}$ with respect to the camera orientation.

1.7 *Geometric Interpretation of Projection Matrix*

We introduce a camera projection matrix $\mathbf{P}$ that encapsulates intrinsic and extrinsic parameters. The 12 elements in the matrix have physical meaning.

First, the columns of the projection matrix, $\mathbf{P} = \begin{bmatrix} \mathbf{p}_1 & \mathbf{p}_2 & \mathbf{p}_3 & \mathbf{p}_4 \end{bmatrix}$ indicate the projection of the world origin and directions of X, Y, and Z axes onto the image.

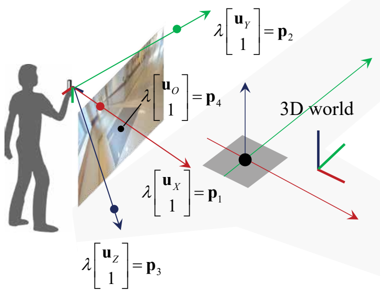

$\lambda \begin{bmatrix} \mathbf{u}_Y \\ 1 \end{bmatrix} = \mathbf{p}_2$
$\lambda \begin{bmatrix} \mathbf{u}_O \\ 1 \end{bmatrix} = \mathbf{p}_4$
3D world
$\lambda \begin{bmatrix} \mathbf{u}_X \\ 1 \end{bmatrix} = \mathbf{p}_1$
$\lambda \begin{bmatrix} \mathbf{u}_Z \\ 1 \end{bmatrix} = \mathbf{p}_3$

Figure 1.13: The columns of the camera projection matrix represents 3D world coordinate system.

The projection of the world origin, $(0, 0, 0)^{\mathsf{T}}$, is

$$\lambda \begin{bmatrix} \mathbf{u}_O \\ 1 \end{bmatrix} = \begin{bmatrix} | & | & | & | \\ \mathbf{p}_1 & \mathbf{p}_2 & \mathbf{p}_3 & \mathbf{p}_4 \\ | & | & | & | \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} = \mathbf{p}_4. \tag{1.26}$$

The projection of the direction of the world X axis, i.e., a point at infinity along the X axis $(\infty, 0, 0)^{\mathsf{T}}$, is:

$$\lambda \begin{bmatrix} \mathbf{u}_X \\ 1 \end{bmatrix} = \lim_{X_w \to \infty} \begin{bmatrix} | & | & | & | \\ \mathbf{p}_1 & \mathbf{p}_2 & \mathbf{p}_3 & \mathbf{p}_4 \\ | & | & | & | \end{bmatrix} \begin{bmatrix} X_w \\ Y_w \\ Z_w \\ 1 \end{bmatrix} = \mathbf{p}_1 \tag{1.27}$$

Similarly, $\mathbf{u}_Y$ and $\mathbf{u}_Z$ can be represented by $\mathbf{p}_2$ and $\mathbf{p}_3$, respectively, as shown in Figure 1.13.

Second, the rows of the projection matrix, $\mathbf{P} = \begin{bmatrix} \mathbf{p}_X^\mathsf{T} & \mathbf{p}_Y^\mathsf{T} & \mathbf{p}_Z^\mathsf{T} \end{bmatrix}^\mathsf{T}$ indicates the planes defined by the camera axes, i.e., YZ, ZX, and XY planes.

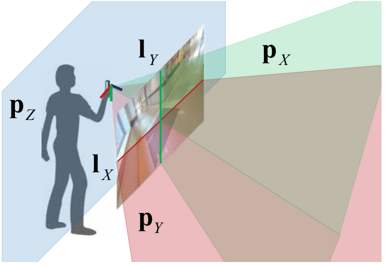


Figure 1.14: The rows of the camera projection matrix represents the 3 planes of camera axes.

Among all world 3D points, $\mathbf{X}_w$, particular points project onto the line $\mathbf{l}_X$ that aligns with the X axis of the image:

$$\lambda \begin{bmatrix} u \\ 0 \\ 1 \end{bmatrix} = \begin{bmatrix} - & \mathbf{p}_X & - \\ - & \mathbf{p}_Y & - \\ - & \mathbf{p}_Z & - \end{bmatrix} \begin{bmatrix} \mathbf{X}_w \\ 1 \end{bmatrix}, \tag{1.28}$$

which indicates $\mathbf{p}_Y \mathbf{X}_w = 0$, or $\mathbf{p}_{Y1} X_w + \mathbf{p}_{Y2} Y_w + \mathbf{p}_{Y3} Z_w + \mathbf{p}_{Y4} = 0$ where $\mathbf{p}_Y = (\mathbf{p}_{Y1}, \mathbf{p}_{Y2}, \mathbf{p}_{Y3}, \mathbf{p}_{Y4})$. Such points are the points on the plane that passes the camera optical center and $\mathbf{l}_X$ in 3D. This plane is represented by $\mathbf{p}_Y$, which is perpendicular to the camera's Y axis (the plane spanned by the camera's X and Z axes). Similarly, $\mathbf{p}_X$ represents the plane spanned by the camera's Y and Z axes as shown in Figure 1.14.

Notably, $\mathbf{p}_Z$ represents the set of points in the line at infinity (see Section **??**), i.e.,

$$\lambda \begin{bmatrix} \mathbf{u}_Z \\ 0 \end{bmatrix} = \begin{bmatrix} - & \mathbf{p}_X & - \\ - & \mathbf{p}_Y & - \\ - & \mathbf{p}_Z & - \end{bmatrix} \begin{bmatrix} \mathbf{X}_w \\ 1 \end{bmatrix}. \tag{1.29}$$

These points lie in the plane parallel to the projection plane at the camera optical center. Therefore, $\mathbf{p}_Z$ represents the plane spanned by X and Y axes of the camera.

1.8 Approximated Camera Model

The dolly zoom effect in Section 1.4 can induce strong depth sensation as controlling the focal length. For instance, in Figure 1.15, the depth sensation, or perspectiveness in the left image is less stronger than the right image. The left image was taken by farther away from the camera with larger focal length.

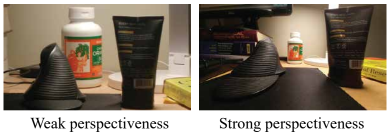


Figure 1.15: The rows of the camera projection matrix represents the 3 planes of camera axes.

Two distances play a role: (1) distance, Z, between an object and camera; (2) distance, d, between objects. If $d/Z \approx 0$ where the camera is far from the objects, the perspectiveness becomes less powerful, i.e., satellite image. This creates a special instance of the projective camera called *affine camera*. Consider an object $\mathbf{X}$ moves away from the camera along the camera's optical axis $\mathbf{r}_z^\mathsf{T}$, i.e., $\mathbf{X}_{\text{affine}} = \mathbf{X} + \mu \mathbf{r}_z$. While moving away, the camera focal length is adjusted as the dolly zoom effect such that it maintains the size of the object $f_{\text{affine}} = f(d+\mu)/d$ from Equation (1.14) where d and f are the distance of $\mathbf{X}$ and focal length before moving, respectively.

The affine projection of $\mathbf{X}$ can be written as:

$$\begin{aligned} u_{\text{affine}} &= \lim_{\mu \to \infty} \frac{f(\mu + d)(\mathbf{r}_x\mathbf{X} + t_x) + p_x}{d(\mathbf{r}_z\mathbf{X} + t_z + \mu)} \\ &= \frac{f}{d}\left(\mathbf{r}_x\mathbf{X} + t_x\right), \end{aligned} \tag{1.30}$$

where the subscript xy indicates the first two rows of matrix, i.e., $\mathbf{R}_{xy} = \begin{bmatrix} \mathbf{r}_x^\mathsf{T} & \mathbf{r}_y^\mathsf{T} \end{bmatrix}^\mathsf{T}$. Likewise, the Y coordinate of the affine projection can be expressed, which results in:

$$\mathbf{u}_{\text{affine}} = \begin{bmatrix} f/d & 0 \\ 0 & f/d \end{bmatrix} \begin{bmatrix} -\mathbf{r}_x- & t_x \\ -\mathbf{r}_y- & t_y \end{bmatrix} \begin{bmatrix} \mathbf{X} \\ 1 \end{bmatrix} \tag{1.31}$$

$$= \mathbf{K}_{\text{affine}}\left(\mathbf{R}_{xy}\mathbf{X} + \mathbf{t}_{xy}\right) \tag{1.32}$$

Note that this projection is linear, i.e., no scalar on LHS in Equation (1.32), which simplifies the projective geometry yet a good approximation of objects at far distance.

When the image coordinate is normalized by the intrinsic parameter, i.e., $\mathbf{K}_{\text{affine}}^{-1}\mathbf{u}_{\text{affine}}$, it also produces a special instance of the affine camera called *orthographic camera*:

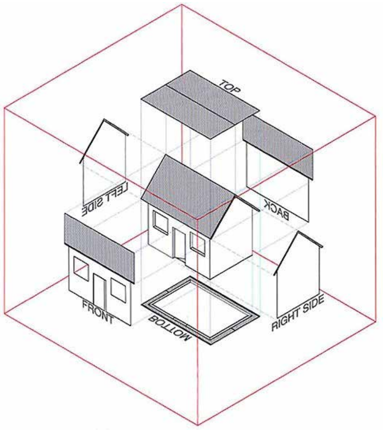


Figure 1.16: Orthographic camera

$$\mathbf{u}_{\text{orth}} = \begin{bmatrix} -\mathbf{r}_x- & t_x \\ -\mathbf{r}_y- & t_y \end{bmatrix} \begin{bmatrix} \mathbf{X} \\ 1 \end{bmatrix} \tag{1.33}$$

$$= \left(\mathbf{R}_{xy}\mathbf{X} + \mathbf{t}_{xy}\right). \tag{1.34}$$

Note that the orthographic projection maps from metric to metric space where there is no notion of pixel space as shown in Figure 1.16.