1 Theory

1.1 Pinhole camera model

The most common imaging model used in computer vision is the pinhole camera model, which is illustrated in Fig. 1.

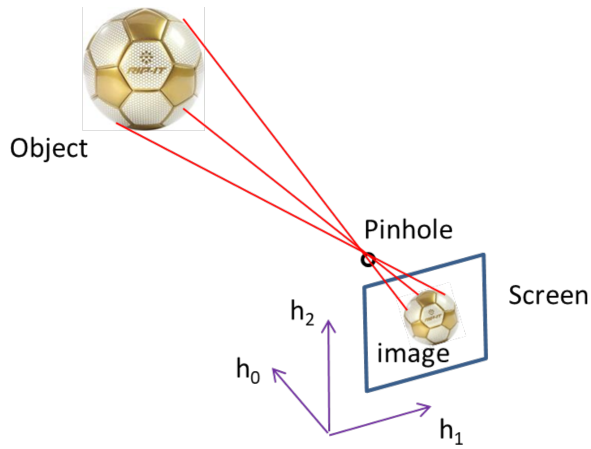

Figure 1: Pinhole camera model

Without loss of generality, suppose the pinhole is located at $(0, 0, 0)$. Denote the position of an object by $\vec{r} = (x, y, z)$, its velocity by $\vec{v} = (v_x, v_y, v_z)$, its acceleration by $\vec{a} = (a_x, a_y, a_z)$. The screen, on which the image is captured, is located at a vector $\vec{h_s} = h_s \hat{h_0}$ from the origin. The screen coordinates basic vectors are $\hat{h_0}$, $\hat{h_1}$ and $\hat{h_2}$. That is, $< \hat{h_0}, \hat{h_1}, \hat{h_2} >$ also form a right-hand coordinate system. Note that

$$\hat{h_0} \perp \hat{h_1} \perp \hat{h_2} \tag{1}$$

$$|\hat{h_0}| = |\hat{h_1}| = |\hat{h_2}| = 1 \tag{2}$$

The screen plane is given by

$$h_s \hat{h_0} + a\hat{h_1} + b\hat{h_2}, \quad a, b \in R \tag{3}$$

Then the projection of $\vec{r}$ on the screen,which is the image of the object, $\vec{p} = (a, b)$, is given by

$$-k\vec{r} = h_s \hat{h_0} + a\vec{h_1} + b\vec{h_2} \tag{4}$$

where k is the distance of the object from the pinhole, a and b are the “screen coordinates” of its image.

$$\left(\vec{r}, \hat{h_1}, \hat{h_2}\right) \begin{pmatrix} k \\ a \\ b \end{pmatrix} = -h_s \hat{h_0} \tag{5}$$

or

$$\left(\hat{h_1}, \hat{h_2}, \vec{r}\right)\begin{pmatrix} a \\ b \\ k \end{pmatrix} = -h_s\hat{h_0}$$

Define

$$C^{-1} = \left(\hat{h_1}, \hat{h_2}, \vec{r}\right)^{-1} \tag{6}$$

as the inverse camera matrix. The screen coordinates of the object (a, b), as well as its distance k from the pinhole, is given by

$$\underline{\begin{pmatrix} a \\ b \\ k \end{pmatrix} = -C^{-1}h_s\hat{h_0}} \tag{7}$$

1.1.1 The characteristics of the screen coordinate system

The world coordinates is denoted by $(\hat{x}, \hat{y}, \hat{z})$, where $\hat{z}$ is the upward vertical direction and $(\hat{x}, \hat{y})$ spans the ground. In the real world, usually not only do the objects move, but also the camera moves simultaneously. So $< \hat{h_0}, \hat{h_1}, \hat{h_2} >$ translates and rotates around the pinhole, and the pinhole also moves. In typical scenarios, the camera stands vertically. So $\vec{h_2}$ is close to vertical, the angle between them being θ. Then

$$\hat{h_2}.\hat{z} = \cos\theta \approx 1 \tag{8}$$

1.1.2 The solution to the imaging equation

In Eq. (4)

$$k\vec{r} + h_s\hat{h_0} + a\vec{h_1} + b\vec{h_2} = 0$$

Multiplying the equation by $\hat{h_i}$,$i = 0, 1, 2$, utilizing the property

$$\hat{h}_i.\hat{h}_j = \begin{cases} 0 & i \neq j \\ 1 & i = j \end{cases}$$

Then

$$k\left(\vec{r}.\hat{h_0}\right) = -h_s$$

$$k\left(\vec{r}.\hat{h_1}\right) + a = 0$$

$$k\left(\vec{r}.\hat{h_2}\right) + b = 0$$

Therefore

$$\underline{k = -\frac{h_s}{\vec{r}.\hat{h_0}}}, \quad a = -k\left(\vec{r}.\hat{h_1}\right) = \underline{h_s\frac{\vec{r}.\hat{h_1}}{\vec{r}.\hat{h_0}}}, \quad \underline{b = h_s\frac{\vec{r}.\hat{h_2}}{\vec{r}.\hat{h_0}}} \tag{9}$$

The image is at (a, b) on the capturing device, typically a CCD or CMOS image sensor. Note that the final image shown in an video will be translated and zoomed according to the frame configuration. So the final location of the pixel is

$$\left(a', b'\right) = (c_x, c_y) + m\,(a, b)$$

where (c_x, c_y) are the offset of the image sensor center to the video center, m is the zoom factor.

Note that if $\vec{r} \perp \hat{h_0}$,this object is out of the FOV (field-of-view) and not on the screen.

Similarly, if $\vec{r}.\hat{h}_0 < 0$, the object is "behind" the camera, and can not form an image.

1.2 The bounding box of an object

Asusming its diameter is D. Then the whole shape can be approximated by a box whose vertexes are given by

$$\vec{r_i} = \vec{r} + \left(\pm\frac{D}{2}, \pm\frac{D}{2}, \pm\frac{D}{2}\right)$$

where $1 \le i \le 8$. Then for each $\vec{r_i}$, its projection on the camera screen is $\vec{p_i}$, also obtained by Eq. (9).

Then the bounding box of the object's image is the range of the 8 projections:

$$ll_x = \min(p_i(x)), \quad ll_y = \min(p_i(y)) \tag{10}$$

$$ur_x = \max(p_i(x)), \quad ur_y = \max(p_i(y)) \tag{11}$$

where ll, ur mean lower-left and upper-right corner, respectively.

1.3 Motion model

In a short time Δt, all three frameworks could move and rotate.

1. The object moves to $\vec{r} + \vec{v}\Delta t$,
2. The pinhole moves to $(0, 0, 0) + \vec{u}\Delta t$. We can safely assume $\vec{u}$ is very small, $\vec{u} \approx 0$.
3. The screen rotates from T to $T.\Delta T$, where ΔT is a "small" orthonormal matrix

then its projection on the screen is given by

$$\left(\hat{h_1}, \hat{h_2}, \vec{r} + (\vec{v} - \vec{u})\,\Delta t\right) \begin{pmatrix} a + \Delta a \\ b + \Delta b \\ k + \Delta k \end{pmatrix} = -h_s \hat{h_0}$$

According to Eq. (9), the new projection point is

$$a + \Delta a = h_s \frac{(\vec{r} + \Delta\vec{r}) \, . \left(\hat{h_1} + \Delta\hat{h_1}\right)}{(\vec{r} + \Delta\vec{r}) \, . \left(\hat{h_0} + \Delta\hat{h_0}\right)} \tag{12}$$

$$b + \Delta b = h_s \frac{(\vec{r} + \Delta\vec{r}) \, . \left(\hat{h_2} + \Delta\hat{h_2}\right)}{(\vec{r} + \Delta\vec{r}) \, . \left(\hat{h_0} + \Delta\hat{h_0}\right)} \tag{13}$$

Let $\Delta t \to 0$, the velocities of a, b are given by

$$\frac{da}{dt} = h_s \frac{\left(\vec{r}'.\hat{h_1} + \vec{r}.\hat{h}_1'\right)\left(\vec{r}.\hat{h_0}\right) - \left(\vec{r}.\hat{h_1}\right).\left(\vec{r}'.\hat{h_0} + \vec{r}.\hat{h}_0'\right)}{\left(\vec{r}.\hat{h_0}\right)^2} = h_s \frac{U}{\left(\vec{r}.\hat{h_0}\right)^2} \tag{14}$$

$$U = \vec{r}.\left[\left(\vec{r}'.\hat{h_1} + \vec{r}.\hat{h}_1'\right)\hat{h_0} - \left(\vec{r}'.\hat{h_0} + \vec{r}.\hat{h}_0'\right).\hat{h_1}\right]$$

If we simply let $\hat{h}_1' = \hat{h}_0' = 0$, i.e, the screen and the pinhole do not move, then

$$U = \vec{r}.\left[\left(\vec{r}'.\hat{h_1}\right)\hat{h_0} - \left(\vec{r}'.\hat{h_0}\right)\hat{h_1}\right]$$

$$U = \vec{r}.\left[\left(\vec{v}.\hat{h_1}\right)\hat{h_0} - \left(\vec{v}.\hat{h_0}\right)\hat{h_1}\right] \tag{15}$$

Recall the *vector triple product* identity, $\vec{a} \times \left(\vec{b} \times \vec{c}\right) = (\vec{a}.\vec{c})\,\vec{b} - \left(\vec{a}.\vec{b}\right)\vec{c}$, then (15) becomes

$$U = \vec{r}.\left(\vec{v} \times \left(\vec{h_0} \times \vec{h_1}\right)\right) = \vec{r}.\left(\vec{v} \times \vec{h_2}\right) \tag{16}$$

$$\boxed{\frac{da}{dt} = h_s \frac{\vec{r}.\left(\vec{v} \times \vec{h_2}\right)}{\left(\vec{r}.\hat{h_0}\right)^2}} \tag{17}$$

$$\boxed{\frac{db}{dt} = -h_s \frac{\vec{r}.\left(\vec{v} \times \vec{h_1}\right)}{\left(\vec{r}.\hat{h_0}\right)^2}} \tag{18}$$

Note that

$$\vec{A}.\left(\vec{B} \times \vec{C}\right) = \left(\vec{A}, \vec{B}, \vec{C}\right) = \begin{vmatrix} A_x & A_y & A_z \\ B_x & B_y & B_z \\ C_x & C_y & C_z \end{vmatrix}$$

1.4 Network

1. Yolo: gives the bounding box of the ball in a frame, (a_1, b_1), (a_2, b_2), denoted by

$$B = \begin{pmatrix} a_1 & b_1 \\ a_2 & b_2 \end{pmatrix} = \begin{pmatrix} a_1 \\ b_1 \\ a_2 \\ b_2 \end{pmatrix}$$

2. Physical network (PNN):
   (a) input: a series of bounding boxes in consecutive frames, B_1, B_2, ..., B_m
   (b) output: the predicted bounding boxes in the following frames, B_{m+1}, ...,B_{m+L}, where L is the prediction sequence length.
   (c) intermediate variables (hidden variables): the position of the

2 Appendix

2.1 Matrix op

Let A^{-1} be the inverse matrix of A. Consider the inverse matrix of $A + dA$, where $|dA| \ll |A|$.

$$A + dA = A(I + A^{-1}dA)$$

Let $dB = A^{-1}dA$, then

$$(I - dB)\, A^{-1}.A\,(I + dB) = (I - dB)\,(I + dB) = I - (dB)^2 \approx I$$

Therefore

$$(A + dA)^{-1} \approx \boxed{A^{-1} - A^{-1}\,(dA)\,A^{-1}}$$

2.2 Differential vectors

For a dot product of two vectors $u = \vec{x}.\vec{y}$, its derivative is

$$\frac{du}{dt} = \frac{d}{dt}\,(\vec{x}.\vec{y}) = \frac{d\vec{x}}{dt}.\vec{y} + \vec{x}.\frac{d\vec{y}}{dt}$$