Exercise 4 – Image Formation and Stereo Vision

Overview

In this exercise, we are first going to revisit traditional camera models used to represent the image formation process. In a second step, we will look at how a stereo camera setup can recover depth information from a specific scene.

Q1 Image Formation

1. Figure 1 shows a camera with a single *thin lens* looking at an object of height h. Assume that there is some light source so that reflections from the object pass through the thin lens to create an image on the right side. Show *geometrically* in Figure 1 where we have to place the image plane so that the object's image will appear in focus. Which properties of the thin lens did you use in this process? Denote the distance between the image plane and the thin lens as b and the height of the image as B.
2. The above question shows that given the distance to the object a and the focal length f, we can geometrically tell at which distance b we have to place the image plane. Please derive an algebraic relation to relate the parameters $\{a, b, f\}$. What is this equation called?

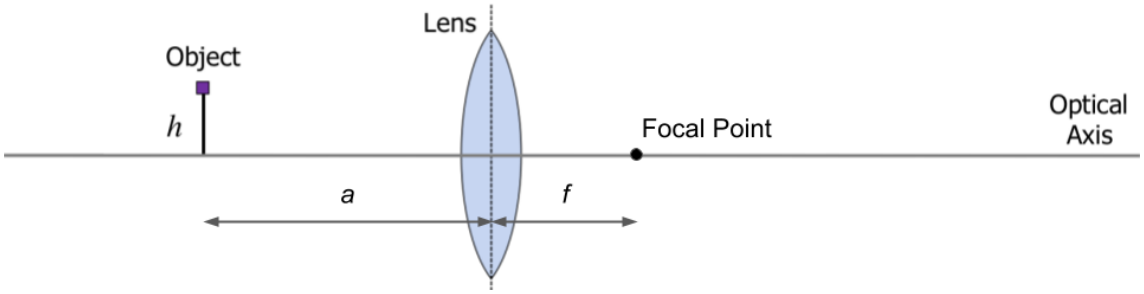


Figure 1: Thin lens model.

3. For mapping a 3D point P to the image plane, we have introduced the *intrinsic parameter matrix* K. Recall that the K matrix uses the *pinhole approximation* and can be used to project a point ${}_{\mathcal{C}}P$ represented in the *camera* frame $\mathcal{C}$ to pixel coordinates (u, v) on the image plane (see Figure 2). Please show the structure of the K matrix, describe its entries, and explain how we can use it to project point ${}_{\mathcal{C}}P$ to pixel coordinates (u, v).
4. Assume now that the coordinates of point P are not given in the *camera* frame $\mathcal{C}$ as ${}_{\mathcal{C}}P$ but instead in the *world* coordinate frame ${}_{\mathcal{W}}P$ (see Figure 2). Which additional computation step do we need to perform now?
5. Now, consider a camera with the specifications as below:
   (a) Field of view along the horizontal axis (X_c/u) : 60°
   (b) Field of view along the vertical axis (Y_c/v) : 45°
   (c) Size of the image plane: 640 pixels along the X_c/u axis and 480 pixels along the Y_c/v axis

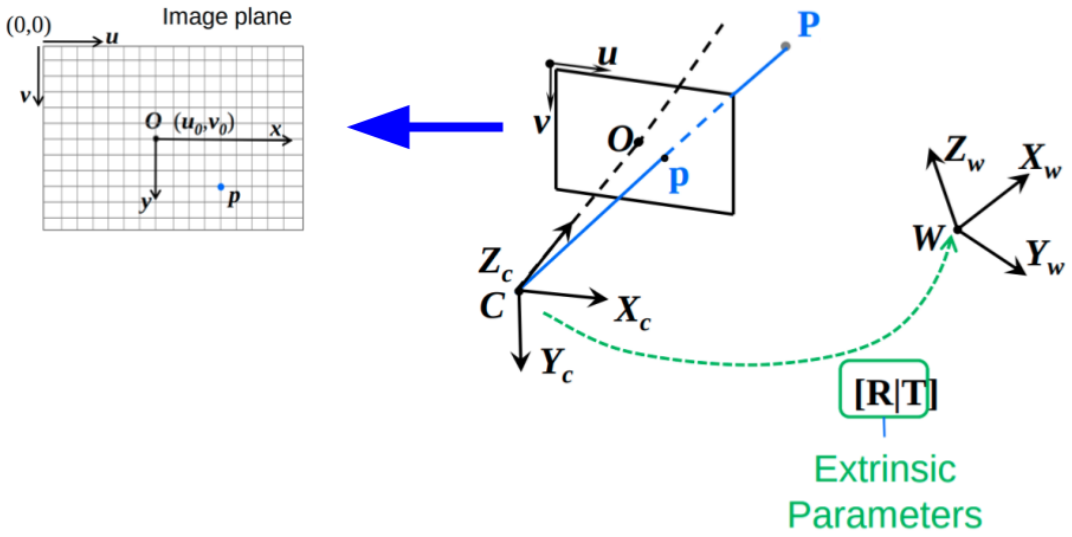


Figure 2: Coordinate system convention.

(d) The *camera* frame $\mathcal{C}$ is centered at $(0.0, -5.0, 1.5)^T$ in the *world* frame $\mathcal{W}$ and is rotated relative to the latter as shown in Figure 3.

(e) There is no offset between the optical axis and the center of the image plane (principal point O).

Figure 3: Q1.5: Rotation between world frame $\mathcal{W}$ and camera frame $\mathcal{C}$.

Given this description, compute the intrinsic camera matrix K, the rotation matrix $R_{\mathcal{CW}}$, and the translation vector ${}_{\mathcal{C}}t_{CW}$ so that we can transform points from the *world* frame $\mathcal{W}$ to image coordinates (u, v). Recall the introduced coordinate frame convention depicted in Figure 2.

6. Please indicate if the following statements are true or false.

   (a) Depth information can be retrieved from a monocular camera.

   (b) A *barrel distortion* stretches the image towards the edges.

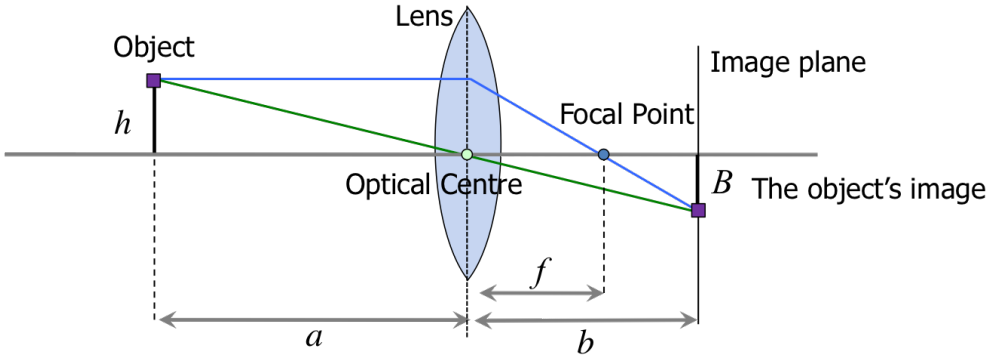
   (c) All catadioptric cameras have a single effective viewpoint (also called a *central* camera).

7. (MATLAB) We will now apply the computed intrinsic and extrinsic calibration matrices to form an image from a given *3D structure*. Please refer to `Ex4.m` for further instructions.

Answer:

1. (a) Rays parallel to the optical axis converge at the *focal point*.

   (b) Rays passing through the *optical centre* are not deviated.



2.

Similar Triangles:

$$\left.\begin{aligned} \frac{B}{h} &= \frac{b}{a} \\ \frac{B}{h} &= \frac{b-f}{f} \quad = \frac{b}{f} - 1 \end{aligned}\right\} \quad \frac{b}{f} - 1 = \frac{b}{a} \Rightarrow \boxed{\frac{1}{f} = \frac{1}{a} + \frac{1}{b}}$$

"Thin lens equation"

3.

$$\tilde{p} = \begin{bmatrix} \alpha_x & 0 & u_0 \\ 0 & \alpha_y & v_0 \\ 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} X_c \\ Y_c \\ Z_c \end{bmatrix} = K \cdot {}_{\mathcal{C}}P$$

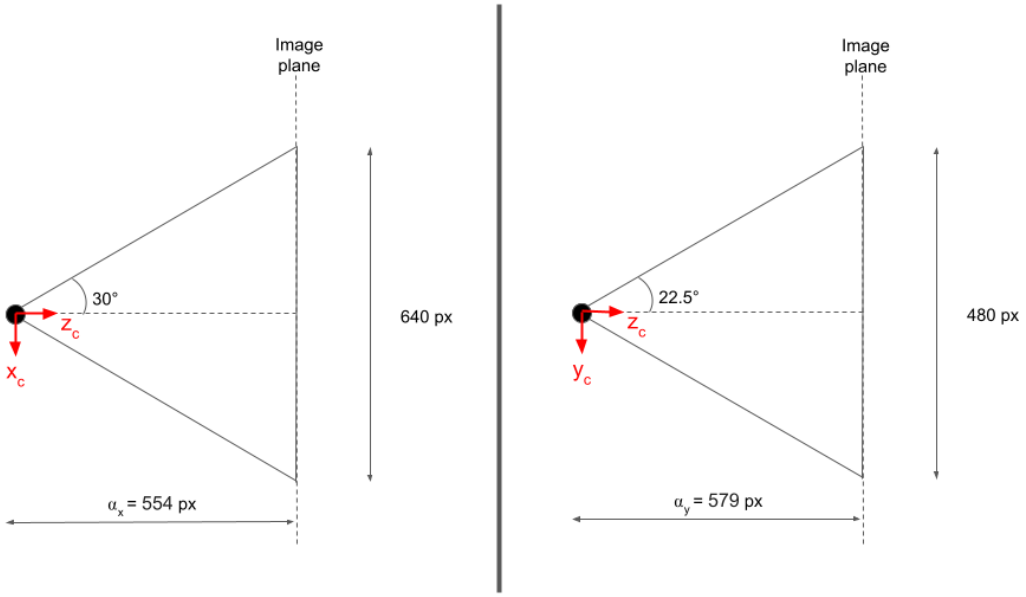
The K matrix contains the focal lengths (in pixels) α_x and α_y, where the first represents the projection along the X_c/u axis and the latter represents the projection along the Y_c/v axis. The pixel offset between the *principal point* O (where the Z_c axis intersects the image plane) and the origin of the pixel coordinate system $u - v$ is denoted by u_0 and v_0. Recall from the lecture that the linear mapping shown above is defined in *homogenous coordinates* $\tilde{p}$ and that we need to divide $\tilde{p}$ by the scaling factor λ to retrieve the actual pixel coordinates (u, v):

$$\tilde{p} = \begin{bmatrix} \tilde{u} \\ \tilde{v} \\ \tilde{w} \end{bmatrix} = \lambda \cdot \begin{bmatrix} u \\ v \\ 1 \end{bmatrix}$$

4. In this case, we need to first convert point P from *world* coordinates ${}_{\mathcal{W}}P$ to *camera* coordinates ${}_{\mathcal{C}}P$ by considering the relative rotation $R_{\mathcal{CW}}$ and translation ${}_{\mathcal{C}}t_{CW}$. The last two quantities are also referred to as *extrinsic parameters*.

$${}_{\mathcal{C}}P_C = R_{\mathcal{CW}} \cdot {}_{\mathcal{W}}P_W + {}_{\mathcal{C}}t_{CW}$$

5. From basic trigonometry we can compute $\alpha_x = \frac{320px}{tan(30°)} \approx 554px$ and similarly $\alpha_y = \frac{240px}{tan(22.5°)} \approx 579px$



Then, we can write the K, $R_{\mathcal{CW}}$, and ${}_{\mathcal{C}}t_{CW}$ matrices as:

$$K = \begin{bmatrix} 554 & 0 & 320 \\ 0 & 579 & 240 \\ 0 & 0 & 1 \end{bmatrix}$$

$$R_{\mathcal{CW}} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{bmatrix}$$

$${}_{\mathcal{C}}t_{CW} = \begin{bmatrix} 0 \\ 1.5 \\ 5.0 \end{bmatrix}$$

which we can plug into the equations we have derived earlier:

$$\begin{aligned} \tilde{p} &= \lambda \cdot \begin{bmatrix} \mathrm{u} \\ \mathrm{v} \\ 1 \end{bmatrix} \\ &= K \cdot (R_{\mathcal{CW}} \cdot {}_{\mathcal{W}}P + {}_{\mathcal{W}}t_{CW}) \end{aligned}$$

6. (a) False, monocular cameras measure angles, not distances (see Perception II – Camera Image Formation, Perspective Projection).
   (b) True, a *pincushion* distortion on the other hand distorts the image towards its centers (see Perception II – Camera Image Formation, Perspective Projection).
   (c) False, it depends on the mirror shape of the catadioptric cameras (see Perception II - Omnidirectional Projection, Camera Calibration, Unified Model).
7. Please refer to `Ex4_solutions.m` for further instructions.

Q2 Stereo Vision

This exercise will consider the stereo triangulation problem. We will see how a stereo camera setup can be used to recover depth information from images.

1. Figure 4 depicts a simple stereo setup with two identical cameras that are aligned along the x-axis and are offset by the baseline b and both have focal length f for the projection along the X axis. The 3D point P projects onto the pixel locations u_l and u_r, in the respective images. Given the parameters $\{u_l, u_r, b, f\}$, please derive an expression for Z_p.

2. Qualitatively explain the effect of the baseline of the stereo setup? Especially, consider the cases of a very small and a very large baseline.

3. Please indicate if the following statements are true or false.

   (a) Foreground objects experience a bigger disparity than background objects.

   (b) The disparity map holds the metric distance in each pixel.

4. (MATLAB) We will now see how the derived equations can be used to reconstruct a 3D structure with a stereo camera setup. Please refer to `Ex4.m` for further instructions.

5. We will now consider the more generic stereo setup shown in Figure 5. Assume for a moment that we are only given the location of the left and right camera frames $\mathcal{C}_l$ and $\mathcal{C}_r$ as well as the projection of an *unknown* 3D point P_w in the right image p_r. We assume that the point P_w also projects into the left image plane but don't know exactly where. Qualitatively show in Figure 5 how we can reduce the search space in the left image to a line. What is this line called?

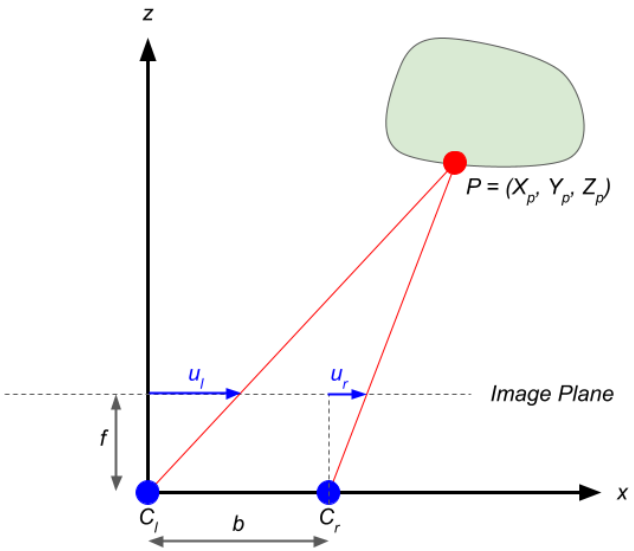


Figure 4: Simple stereo camera setup.

Answer:

1. From similar triangles we can write:

   (a) $\frac{f}{Z_p} = \frac{u_l}{X_p}$

   (b) $\frac{f}{Z_p} = \frac{-u_r}{b - X_p}$

   From which it follows that $Z_p = \frac{bf}{u_l - u_r}$

2. (a) A very large baseline b could lead to point P_W not being visible in both images anymore. This problem gets worse, the closer the point P_W is to the camera.

   (b) A very small baseline would yield small disparity values which leads to a worse signal-to-noise ratio. This in turn means less accurate depth estimates.

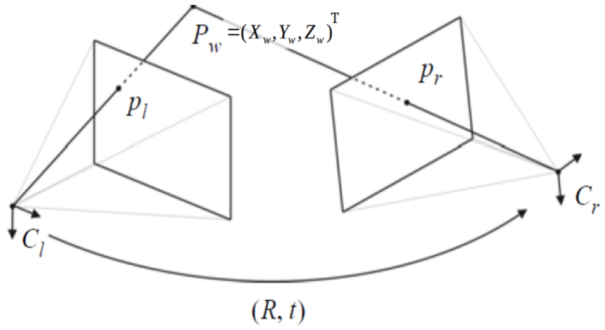


Figure 5: Generic stereo camera setup.

3. (a) True, this becomes clear when we look at the equation derived in Q2.1.
   (b) False, disparity refers to how far apart the projected point are in the image. The distance can then be triangulated by knowing the disparity and the parameters of the stereo setup (see Perception II - Stereo Vision).

4. Please refer to `Ex4_answers.m`

5. It's called the *epipolar line* (dashed blue) and can be constructed by intersecting the triangle $\Delta C_l C_r p_r$ (red) with the left image plane (or vice versa if p_l is given instead).