1 Numerical Exercise

1. Explain the Dolly-Zoom Effect.

   Solution
   The Dolly-Zoom effect (see `https://youtu.be/u5JBlwlnJX0?t=44` for a video) is the result of a change in focal length combined with a camera translation. In plain English, the camera is zoomed out while being moved forward, which results in the background becoming more compressed (seeming more distance) compared to the foreground.
   To understand the effect better, consider the two sketches shown in Figure 1. In the first (zoomed) position, all points p_1 to p_4 are visible in the image plane, with the background points being roughly three times further apart than the foreground points p_1 and p_2. When the camera is moved closer to Position 2 and zoomed out, the foreground points are still visible at the same points in the image plane, the background point p_4 has now moved much closer to p_2, while p_3 disappeared completely. Hence, the foreground remained the same while the background got compressed.

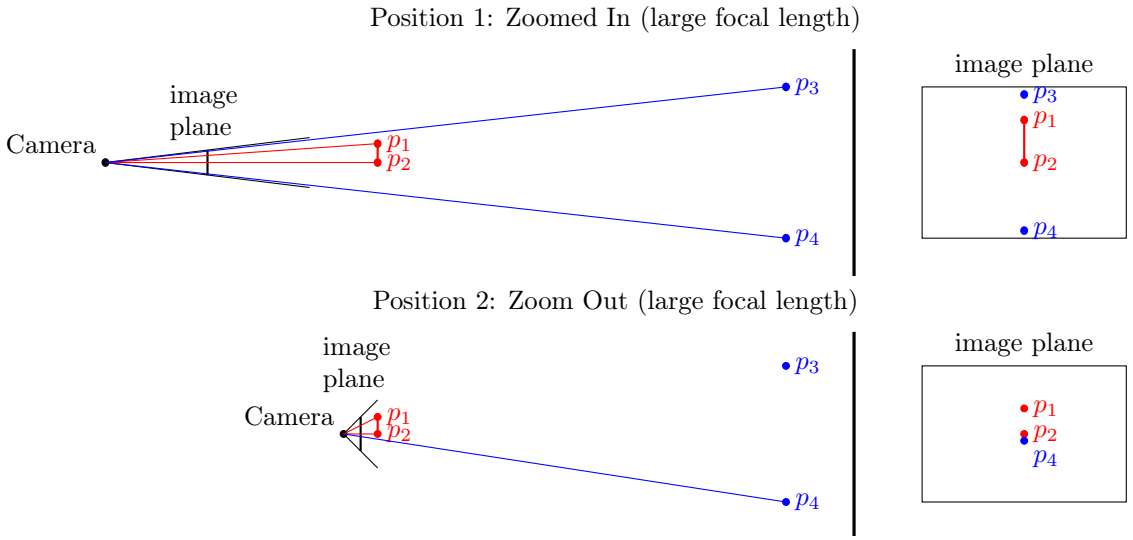


Figure 1: Illustration of the Dolly Zoom

2. What is the condition on the elements of a projective transformation H such that parallel lines remain parallel? Recall, a projective transformation relates the projected coordinates of the same 3D point between two different image planes/viewpoints.

   Solution
   Parallel lines remaining parallel implies that, for all the directions, points at infinity (which are the intersections of the parallel lines) are again mapped to points at infinity. In homogeneous coordinetes, the points at infinity are defined as $[a \;\; b \;\; 0]$[1]. Hence, $[a \;\; b \;\; 0]$ is mapped to $[a' \;\; b' \;\; 0]$ under the projective transformation H. Consequently $H \; [a \;\; b \;\; 0]^\top$ should have the third entry zero.

$$H = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{bmatrix}, \qquad h_{31}a + h_{32}b = 0 \;\forall a, b \Longrightarrow h_{31} = h_{32} = 0$$

3. Write the projective transformation H that corresponds to doubling the focal length.

   Solution

[1] `https://en.wikipedia.org/wiki/Homogeneous_coordinates`

The components h_{11} and h_{22} correspond to the focal length. Therefore,

$$H = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

4. Consider a single camera C with the following intrinsic parameters:

   - f the focal length $[m]$
   - k_x and k_y the resolution in $[\frac{pixels}{m}]$ of the camera in x and y direction
   - c_x and c_y in $[pixels]$ as the x and y coordinates of the camera central point

   and the following extrinsic parameters:

   - $\mathbf{R}_{CW}$ and ${}_C\mathbf{t}_{CW}$, forming the world-to-camera extrinsics $[\mathbf{R}_{CW}|{}_C\mathbf{t}_{CW}]$ (equivalently $\mathbf{T}_{CW}$).

   In homogeneous coordinates the projection $\tilde{\mathbf{q}}$ of a world point ${}_W\tilde{\mathbf{P}}$ onto the image plane is computed as

   $$\lambda\tilde{\mathbf{q}} = K[\mathbf{R}_{CW}|{}_C\mathbf{t}_{CW}]\,{}_W\tilde{\mathbf{P}}$$

   with the camera calibration matrix K

   $$K = \begin{pmatrix} fk_x & 0 & c_x \\ 0 & fk_y & c_y \\ 0 & 0 & 1 \end{pmatrix}$$

   - Which parameters can be calibrated given a planar calibration pattern?
   - How can they be estimated?
   - Which parameters cannot be calibrated? How can this problem be tackled?

   Solution
   The DLT-method discussed in the lectures can be used to estimate the complete calibration matrix K from a planar pattern. However, the products fk_x and fk_y can only be estimated jointly as the focal length in pixels. To actually obtain f, k_x and k_y seperately one could, for example, get the pixel distance (pixel spacing) from the manufacturer and use this information to convert the focal length in pixels into a metric focal length.

5. Discuss the use of regular squares and regular circles as calibration patterns.

   - Which points would you use as calibration points?
   - How would you detect them?
   - Which approach is more accurate?

   Solution
   As we have not discussed feature detection in class yet, we will limit ourselves to basic ways on how to detect possible calibration points. If the size of the circles is known, a simple yet effective strategy could be to detect a circle by maximizing the correlation of the image with a circular kernel of the correct size. The correlation would be maximal, if the kernel is perfectly aligned with a circle in the image. By "shifting" a circlular pattern over the image and measuring the correlation at each position, we can detect the circles. For the square pattern, the "Harris Corner" detector (explained in Lecture 04) would be suitable. For an accurate calibration, accurate point detection is the key ingredient. For the circles, their center is the only thing that can be reliably detected. However, in constrast to the checkerboard pattern, the center is not as clearly identifyable as the corners of the black and white squares. Thus the positional accuracy of the circle-center detection will be much lower than the accuracy of the squares-corner detection. Consequently, one would prefer to use squares.

6. You are tasked with finding the camera matrix P using the Direct Linear Transform (DLT) method. Given a set of 3D world points $X_i = (X_i, Y_i, Z_i, 1)$ and their corresponding 2D image points $x_i = (u_i, v_i, 1)$, use the following data to estimate the camera matrix P. Given world coordinates (in homogeneous coordinates):

$$\begin{pmatrix} X_1 & Y_1 & Z_1 & 1 \\ X_2 & Y_2 & Z_2 & 1 \\ X_3 & Y_3 & Z_3 & 1 \\ X_4 & Y_4 & Z_4 & 1 \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 \end{pmatrix}$$

Given corresponding 2D image coordinates (in homogeneous coordinates):

$$\begin{pmatrix} u_1 & v_1 & 1 \\ u_2 & v_2 & 1 \\ u_3 & v_3 & 1 \\ u_4 & v_4 & 1 \end{pmatrix} = \begin{pmatrix} 150 & 120 & 1 \\ 330 & 120 & 1 \\ 330 & 300 & 1 \\ 150 & 300 & 1 \end{pmatrix}$$

Set up the linear system $A \cdot p = 0$, where A is constructed using the world coordinates and image coordinates, and p is the vector of unknown camera parameters. (Don't need to solve the system)

Solution:

To solve this, we first use the relationship between 3D world points $X_i = (X_i, Y_i, Z_i, 1)$ and 2D image points $x_i = (u_i, v_i, 1)$:

$$x_i \sim P \cdot X_i$$

This leads to two equations for each point:

$$u_i(p_{31}X_i + p_{32}Y_i + p_{33}Z_i + p_{34}) = p_{11}X_i + p_{12}Y_i + p_{13}Z_i + p_{14}$$
$$v_i(p_{31}X_i + p_{32}Y_i + p_{33}Z_i + p_{34}) = p_{21}X_i + p_{22}Y_i + p_{23}Z_i + p_{24}$$

For each point, we generate two equations. For example, for $X_1 = (0, 0, 0, 1)$ and $(u_1, v_1) = (150, 120)$:

$$-150(p_{31} \cdot 0 + p_{32} \cdot 0 + p_{33} \cdot 0 + p_{34}) + p_{14} = 0$$
$$-120(p_{31} \cdot 0 + p_{32} \cdot 0 + p_{33} \cdot 0 + p_{34}) + p_{24} = 0$$

We repeat this process for all points to set up the system of equations.

For $X_2 = (1, 0, 0, 1)$ and $(u_2, v_2) = (330, 120)$:

$$-330(p_{31} \cdot 1 + p_{32} \cdot 0 + p_{33} \cdot 0 + p_{34}) + p_{11} + p_{14} = 0$$
$$-120(p_{31} \cdot 1 + p_{32} \cdot 0 + p_{33} \cdot 0 + p_{34}) + p_{21} + p_{24} = 0$$

We proceed similarly for the remaining points, $X_3 = (1, 1, 0, 1)$, $X_4 = (0, 1, 0, 1)$, generating the linear system $A \cdot p = 0$.

The system can be written in matrix form as:

$$A = \begin{pmatrix} 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -150 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & -120 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & -330 & 0 & 0 & -330 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & -120 & 0 & 0 & -120 \\ 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & -330 & -330 & 0 & -330 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & -300 & -300 & 0 & -300 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & -150 & 0 & -150 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & -300 & 0 & -300 \end{pmatrix}$$

7. A video camera looks onto a table of variable height, illuminated by a projector which projects a pattern, comprising a single cross, onto the scene. The system is calibrated by adjusting the table in height to 50, 110, and 200 mm. For these table positions, the center of the cross is observed at the following image positions (in pixels):

$$(u, v) = \{(100, 250), (140, 330), (200, 450)\}$$

Try to determine the height of the table when the cross is observed at:

(a) (u, v) = (130, 310)

(b) (u, v) = (170, 390)

Solution

The transformation between table height Z and cross position (u, v) can be written in the following form:

$$\begin{bmatrix} su \\ sv \\ s \end{bmatrix} = \begin{bmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \\ p_{31} & p_{32} \end{bmatrix} \begin{bmatrix} Z \\ 1 \end{bmatrix}$$

The u and v coordinates of the calibration observations provide six equations:

For $Z = 50, u = 100, v = 250$:

$$100 = \frac{su}{s} = \frac{50p_{11} + p_{12}}{50p_{31} + p_{32}}, \quad 250 = \frac{sv}{s} = \frac{50p_{21} + p_{22}}{50p_{31} + p_{32}}$$

For $Z = 110, u = 140, v = 330$:

$$140 = \frac{su}{s} = \frac{110p_{11} + p_{12}}{110p_{31} + p_{32}}, \quad 340 = \frac{sv}{s} = \frac{110p_{21} + p_{22}}{110p_{31} + p_{32}}$$

For $Z = 200, u = 200, v = 450$:

$$200 = \frac{su}{s} = \frac{200p_{11} + p_{12}}{200p_{31} + p_{32}}, \quad 450 = \frac{sv}{s} = \frac{200p_{21} + p_{22}}{200p_{31} + p_{32}}$$

Writing these equations in matrix form gives:

$$\begin{bmatrix} 50 & 1 & 0 & 0 & -5000 & -100 \\ 110 & 1 & 0 & 0 & -15400 & -140 \\ 200 & 1 & 0 & 0 & -40000 & -200 \\ 0 & 0 & 50 & 1 & -12500 & -250 \\ 0 & 0 & 110 & 1 & -36300 & -330 \\ 0 & 0 & 200 & 1 & -90000 & -450 \end{bmatrix} \begin{bmatrix} p_{11} \\ p_{12} \\ p_{21} \\ p_{22} \\ p_{31} \\ p_{32} \end{bmatrix} = 0$$

We solve this by orthogonal least squares (see question 3 for the method, and use Matlab to calculate the eigenvector via SVD):

$$\begin{bmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \\ p_{31} & p_{32} \end{bmatrix} = \begin{bmatrix} 0.6667 & 66.6667 \\ 1.3333 & 183.3333 \\ 0 & 1.0 \end{bmatrix}$$

where the solution has been arbitrarily scaled to give a 1 in the lower right-hand corner.

Hence, to determine the value of Z of different (u, v), these equations can be solved for Z using a pseudo-inverse (e.g., use Matlab) and the calibration parameters found earlier.

We can now use this calibration to find the height of the table Z for new observations of the cross. Each of the u and v observations provides a single equation in Z:

$$u = \frac{su}{s} = \frac{Zp_{11} + p_{12}}{Zp_{31} + 1}, \quad v = \frac{sv}{s} = \frac{Zp_{21} + p_{22}}{Zp_{31} + 1}$$

Hence we could solve the Z values using the projection eqations with the p values solved.

(a) $(u, v) = (130, 310) \Rightarrow Z = 95\,\text{mm}$.

(b) $(u, v) = (170, 390) \Rightarrow Z = 155\,\text{mm}$.

8. A drone is equipped with a calibrated camera and uses PnP (Perspective-n-Point) for localization and navigation in an urban environment. Consider the following scenario: The drone needs to navigate through a city to deliver a package. It detects 4 corners of a distinctive building, whose 3D coordinates are stored in its map database. The drone's camera captures these corners in its 2D image.

   a) Explain how PnP could be used in this scenario to aid the drone's navigation. What specific information would PnP provide?

   b) What are two potential challenges or sources of error when using PnP in this drone navigation scenario? How might these affect the drone's decision-making?

   c) Apart from building corners, name two other types of urban features that could be useful for PnP-based localization in drone navigation. Why are they suitable choices?

   Solution
   (a) PnP could be used in this scenario to determine the exact pose (position and orientation) of the drone in the world frame (determined by the map database). Specifically, PnP would-provide (i) The drone's 3D position and (ii) the drone's 3D orientation (rotation) in the world coordinate system. This information is crucial for precise localization of the drone in the urban environment.

   (b) Two potential challenges or sources of error: (i) Dynamic environments: Construction or changes in the urban landscape might alter the appearance or position of landmarks, leading to mismatches between the database and reality. (ii) Weather conditions: Strong winds or precipitation could cause camera shake or obscure visibility, affecting the accuracy of detecting building corners in the 2D image.

   These errors could lead to inaccurate pose estimation, potentially causing the drone to misjudge its position or altitude, affecting decision-making about navigation routes or obstacle avoidance.

   (c) Two other suitable urban features for PnP-based localization: (i) Distinctive rooftop patterns: Many buildings have unique rooftop structures or patterns that are easily visible from above and can serve as reliable landmarks. (ii) Street intersections: The corners where streets meet create distinct patterns that can be recognized from aerial views and used for localization.