Lecture 11 - 3D imaging geometry

Gidon Rosalki

2026-01-04

1 Reminder

We will begin by discussing models from the first lecture. The pinhole model has all the rays of the scene pass through a single point. This point is called the Centre of Projection (COP), or focal point. It results in a 2D image being formed on the Image Plane, and the focal length f is distance from COP to the Image Plane.
Perspective projection involves transforming the 3D world (X, Y, Z) into a 2D image (x, y). This may be done through continuous perspective projection (optics), and all rays must pass through one focal point, where f =focal length.

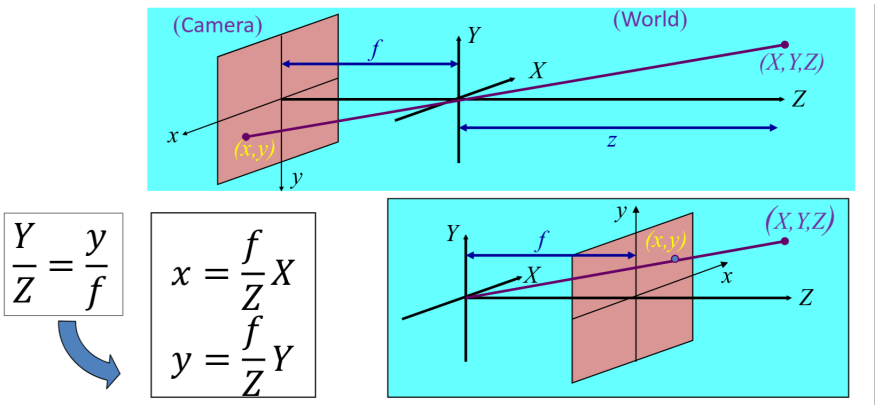

Figure 1: Perspective projection

When world and camera use same X, Y axes, and origin of world axes (X=0, Y=0, Z=0) is at camera's optical centre, then the camera matrix is given by

$$\begin{bmatrix} fX \\ fY \\ Z \end{bmatrix} = \begin{bmatrix} f & - & - & 0 \\ 0 & f & - & 0 \\ - & - & 1 & 0 \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix}$$

A camera in a general position, and not at the world origin There is always a 1 in the bottom right hand corner, and there are 11 degrees of freedom.

In general, the camera matrix M has 11 degrees of freedom (there are 12 parameters, but the scale is arbitrary). 5 are *intrinsic*, these are the camera parameters, and are comprised of f_x, f_y, Centre (c_x, c_y), and the skew. 6 are *extrinsic*, demarking the camera location, 3 for rotation, and 3 for translation. One correspondence between a 2D image point (x, y) to a 3D world point (X, Y, Z) gives two independent linear equations (m_{ij} are unknowns), one for x and one for y. At least 6 correspondences between 3D points and image points needed to compute M. To calibrate a camera, given $n \geq 6$ points with known 3D coordinates X_i, and known image projections x_i, estimate the camera parameters by solving the equations (think of looking at a known scene, with dots on it, sort of vaguely like motion capture of actors).

Since we know that

$$\begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \sim \begin{bmatrix} m_{00} & m_{01} & m_{02} & m_{03} \\ m_{10} & m_{11} & m_{12} & m_{13} \\ m_{20} & m_{21} & m_{22} & 1 \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix}$$

So the camera maps each of the 3D points (X_i, Y_i, Z_i) to an image point (x_i, y_i):

$$x_i = \tag{1}$$

$$y_i = \tag{2}$$

2 Time to collision

A bus is moving towards the camera. When will it collide with the camera? Well, if we put a ruler across the ruler, we may see that the bus width is doubled after 2 seconds. Therefore, in 2 seconds, it has travelled half the distance between us, so it will hit the camera in another 2 seconds. So, given 2 images, one taken at T_1, and the seconds at T_2, we may compute the time to collision without knowing the size, the speed, or the distance of the bus, or event he camera matrix.

At frame 1, the image width of the bus is x_1. At frame 2, the image widht of the bus is x_2. Therefore

$$\begin{aligned} TTC &= \frac{Z_1}{Z_1 - Z_2} \\ &= \frac{\frac{1}{Z_2}}{\frac{1}{Z_2} - \frac{1}{Z_1}} \\ &= \frac{f\frac{X}{Z_2}}{f\frac{X}{Z_2} - f\frac{X}{Z_1}} \\ &= \frac{x_2}{x_2 - x_1} \end{aligned}$$

Where the time is the number of frames.
If $x_1 = x_2 \implies TTC = \infty$, and if $x_2 = 2x_1 \implies TTC = 2$.

3 Epipolar geometry

In epipolar geometry, we have 2 cameras. This leaves us with 3 questions:

1. Correspondence geometry: Given an image point x in the first image, how does this constrain the position of the corresponding point x' in the second image?
2. Camera geometry (camera motion): Given a set of corresponding image points between images, $\{x_i \leftrightarrow x_i'\}$, where $i \in [n]$, what are the cameras matrices M and M for the two views?
3. Scene geometry (structure, stereo): Given corresponding image points between images, $x_i \leftrightarrow x_i'$ and cameras M, M, what is the position of the 3D point X in world coordinates? Or: what is the geometric transformation between the views?

Consider the epipolar line: Given an image point x, it may originate from any 3D point X along a straight line through C_L and x, where C_L is the centre of the left camera. Given image point x, the corresponding point x in the second image can be anywhere along the epipolar line l, which is the projection of the line connecting C_L and x.

We may also consider the epipolar plane: Two cameras with centres C, C' view a 3D point X in image points x, x'. All these points are on the epipolar plane π. The epipolar plane π is determined by the points C, C', X.

The epipolar geometry is the camera baseline, connecting the two camera centres, intercepts the image planes at the epipoles e, e'. An plane π containing the baseline is an epipolar plane. The epipole is the image at hte centre of the **other** camera.

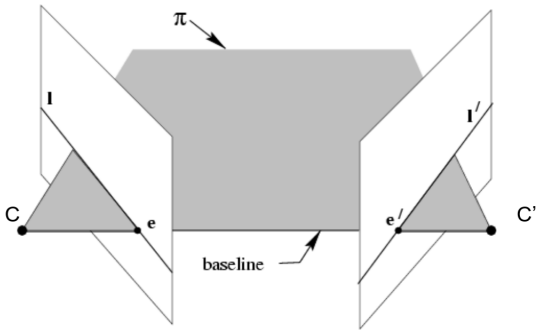

Figure 2: Epipolar geometry

All points on the epipolar plane project to the epipolar lines I and I'. Therefore, it cannot be a physical plane, like the floor, since the floor will not interact with the baseline.
so, the epipoles are the intersection of the baseline with an image plane, or the image projection centre (pinhole), of the other camera. The epipolar plane is a plane containing the baseline, and an epipolar line is the intersection of an epipolar plane with the image plane (always come in corresponding pairs).

When the cameras move in parallel to the image plane, then the baseline intersects the image planes at infinity, with epipoles at infinity, and epipolar lines parallel to the x axis. So, two cameras a horizontal distance apart will have epipolar lines that are parallel in the direction of translation, and two corresponding epipolar lines will go through the same scene points.

3.1 Fundamental matrix

For any two cameras viewing the same scene, there exists a 3 by 3 "Fundamental Matrix" F, such that for any two corresponding image points x and x' :

$$(x')^T Fx = 0$$

Given x and F, all points on the epipolar line l' satisfy the above equation.

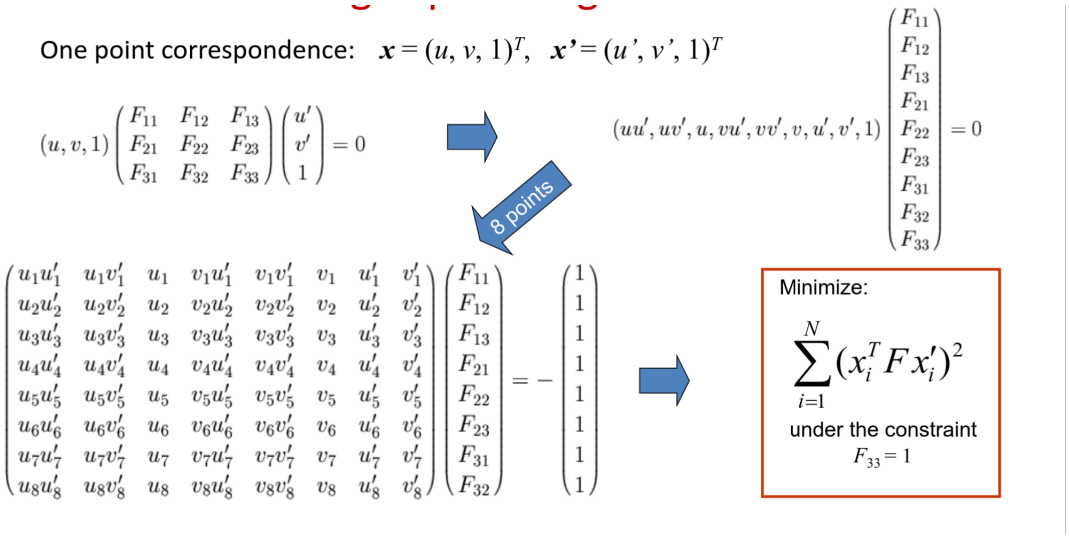

Figure 3: Compute the fundamental matrix

4 3D stereo reconstruction

Assuming that we know the viewing geometry (extrinsic parameters) and the intrinsic parameters (Calibrated camera), we want to find point correspondences exploiting epipolar geometry, search for correspondences only on epipolar lines, and compute depth by triangulation.
So, two cameras with the same image plane have parallel optical axis and horizontal epipolar lines.

4.1 Stereo

To compute the depth from stereo, we will begin with some definitions. We will have x_l, x_r represent the same point, in the left and right image accordingly. The baseline b will represent the distance between the centre of the two cameras, and the disparity d is the difference in image location of the same 3D point when viewed by two different cameras. It is computed

$$d = x_l - x_r$$

The basic stereo algorithm is as follows: For each pixel in the left image, take a window around the pixel. Compare this window with every pixel on the epipolar line in the right image. Pick the pixel in the right image that has the window with the highest correlation, and use ordering constraint (left-right relationships are mostly preserved).

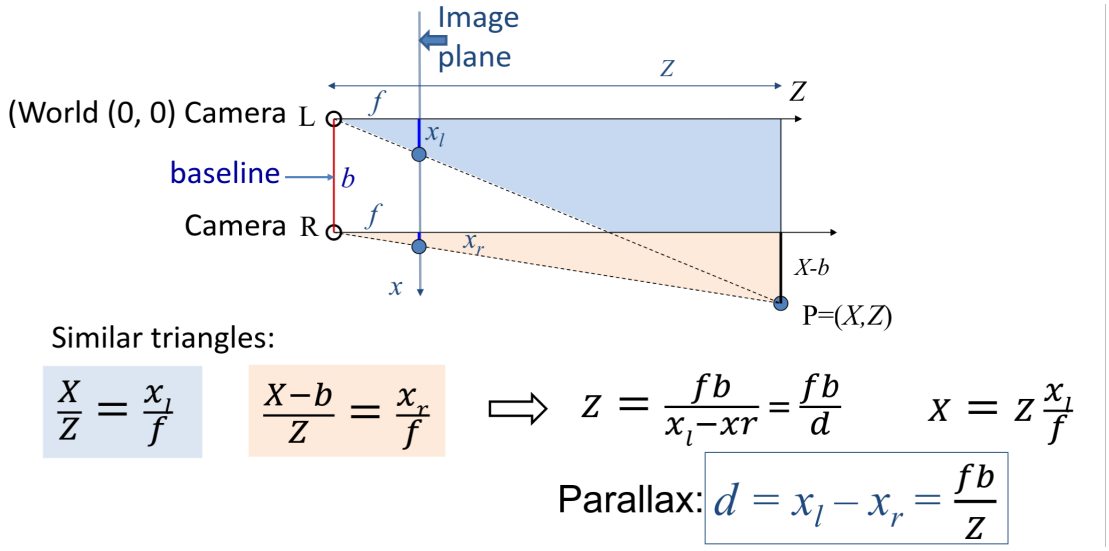

Figure 4: Parallel optic angles

So, to determine correspondences, we use block matching, or the Sum of Squared Differences:

$$E(x, y; d) = \sum_{x', y' \in N(x,y)} \left[I_L(x' + d, y') - I_R(x', y') \right]^2$$

Where d is the disparity (horizontal motion). We are left with the problem of choosing the size of a neighbourhood. A small neighbourhood will have more details, but a larger neighbourhood will have fewer isolated mistakes.

4.2 Kinect

The kinect game controller (by prime sense / used by xbox), had 3 main sensors. An IR emitter, and IR sensor, and a colour camera. The IR emitter emitted lots of small dots across the scene, the IR sensor would create a depth perception map, and the colour sensor (camera) would work in conjunction with the IR sensor (camera) for parallax perception. Since the distance from the IR emitter, and the IR camera increases at the same rate, the emitted dots will always appear to be the same size to the camera, no matter the distance.
The emitter emits a coded dot pattern that is similar to "blue noise", that is generated by a laser LED passing through a grating. Each 9 by 9 block in the epipolar line gives a unique pattern, allowing the sensor to identify locations. In every image location, it searches for a disparity of up to 34 pixels along the epipolar line, so a disparity becomes a lookup issue. This has the problem of not working well outside, since the active emitted light is insufficiently strong compared to the sun.

5 3D from a single camera

We have seen how to compute depth from 2 cameras using stereo methods, also known as binocular cues. 3D cues are also available from a single image, from one camera. Like how dots on a sphere look warped according to their location on the sphere, or how railroad lines tend towards each other with distance.
The horizon is the projection of a point at infinity. In a horizontal camera, the horizon is at the centre of the image, and the horizon intersects objects at camera height. Any two parallel lines have the same vanishing point v. The ray from the camera centre through v is parallel to these lines. An image may have more than one vanishing point.