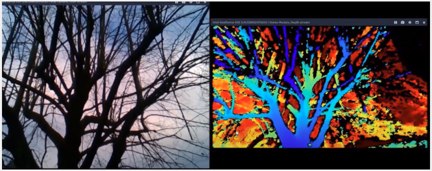
Autonomous Mobile Robotics

Multi-view Geometry

Geesara Kulathunga

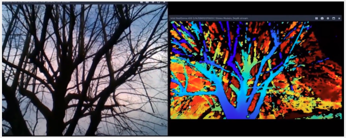


October 27, 2022

Multi-view Geometry

Contents

- Monocular Vision
  - Pinhole Camera Model
  - Image Plane, Camera Plane, Projection Matrix
  - Projective transformation
  - Finding Projection Matrix using Direct Linear Transform (DLT)
  - Camera Calibration
- Stereo Vision
  - Simple Stereo, General Stereo
  - Some homogeneous properties
  - Epipolar Geometry
  - Essential matrix, Fundamental matrix
- Depth Estimation



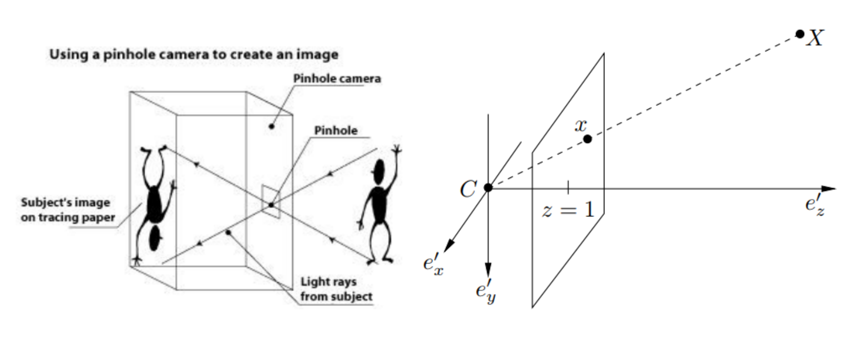


Figure: Pinhole camera model [1]

- This model allows to enter light rays through a small hole (the pinhole) and project as an image on the back of the camera wall
- Let define camera coordinate system as $\{e'_x, e'_y, e'_z\}$. The coordinate of camera center or pinhole of the camera(C) is at $(0, 0, 0)$
- The projection of $\mathbf{X} = (X_w, Y_w, Z_w)$ scene point into the image plane $\mathbf{x}' = (x', y', z')$ while assuming z' = 1 has the normal e_z lies at the distance 1 from the camera center. e_z can be defined as the viewing direction since the $\mathbf{X} - \mathbf{C}$ is the direction vector of viewing ray

$$\mathbf{C} + s(\mathbf{X} - \mathbf{C}) = s\mathbf{X}, s \in \mathbb{R} \tag{1}$$

- Thus, where will the intersection of this vector be , if $e_z = 1$?

$$\mathbf{x}' = \begin{bmatrix} X_w/Z_w \\ Y_w/Z_w \\ 1 \end{bmatrix} \quad (2)$$

Example 01

Compute the projection of the cube with corners: $(\pm 1, \pm 1, 2)$ and $(\pm 1, \pm 1, 4)$ in image plane?

Image Plane

In the real world examples, camera can undergo series of rotations and translations. Hence, it is required to transform world coordinate system into camera coordinate system.

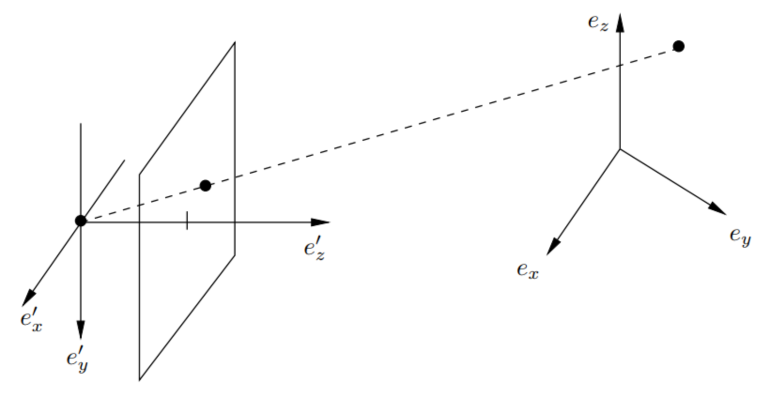


Figure: Global coordinate system and camera coordinate system [1]

For a given point in the global coordinate system can be represented with respect to camera coordinate system:

$$\begin{bmatrix} X_c \\ Y_c \\ Z_c \\ 1 \end{bmatrix} = [R\, t] \begin{bmatrix} X_w \\ Y_w \\ Z_w \\ 1 \end{bmatrix} \tag{3}$$

Example 02

Compute the projection of $\mathbf{X} = (0, 0, 1)$ in the cameras coordinate system if R is

$$\begin{bmatrix} 1 & 0 & -1 \\ 0 & \sqrt{2} & 0 \\ 1 & 0 & 1 \end{bmatrix} \tag{4}$$

and t vector equals to $[0, 0, \sqrt{2}]$. Also, how do you assume for a given point is in the front of the camera or not?

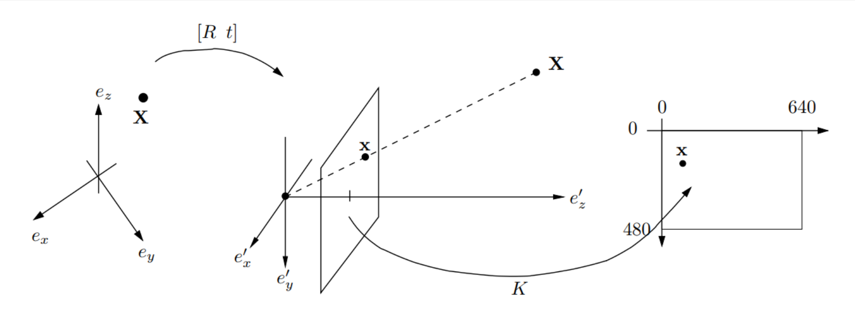


Figure: Global coordinate system and camera coordinate system[1]

In the image plane, center of the image is located in, i.e., 0,0,? $(c_x, c_y, 0)$. But when we take a photo with a real camera, coordinate of image is measured from the upper left corner as shown in Fig. 3. This is where we need inner parameters of camera in order to transfer pixel coordinate system into image coordinate system. This transformation matrix is denoted as K where it is invertible. In general, K is expressed as:

$$K = \begin{bmatrix} f_x & s & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} = \underbrace{\begin{bmatrix} 1 & 0 & c_x \\ 0 & 1 & c_y \\ 0 & 0 & 1 \end{bmatrix}}_{\textit{2d translation}} \times \underbrace{\begin{bmatrix} f_x & 0 & 0 \\ 0 & f_y & 0 \\ 0 & 0 & 1 \end{bmatrix}}_{\textit{2d scaling}} \times \underbrace{\begin{bmatrix} 1 & s/f_y & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}}_{\textit{2d shear}} \tag{5}$$

where f is called focal length, c_x and c_y is denoted the principle point of the camera, γ is the aspect ratio.

When the pixels are not square values, γ will not be equal to one. Otherwise it will be equal to 1. The final parameter is s which is defined as skew. This parameter is used to tilted the pixels as shown in Fig. 4.

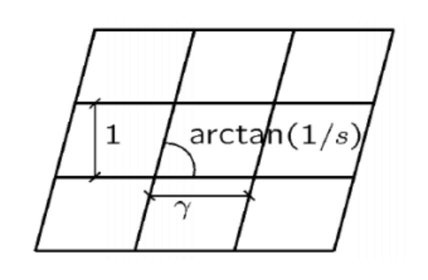


Figure: The skew parameter (s) corrects non-rectangular pixels and γ is used correct the aspect ratio issue. [1]

The relationship between point in the camera and world frame can be given as follows:

$$\lambda \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = K[R|t]X = PX \tag{6}$$

where $R|t$ is the homogeneous transformation which is composed out of a rotation matrix R, and a translation vector t.

$$\begin{bmatrix} X_c \\ Y_c \\ Z_c \\ 1 \end{bmatrix} = \begin{bmatrix} r_{11} & r_{12} & r_{13} & t_x \\ r_{21} & r_{22} & r_{23} & t_y \\ r_{31} & r_{32} & r_{33} & t_z \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} X_w \\ Y_w \\ Z_w \\ 1 \end{bmatrix} . \tag{7}$$

All in all, we can define the projective transformation that maps world coordinates points in $\mathbf{R}^3$ to $\mathbf{R}^2$ image coordinate system followed by normalized camera coordinate system.

The projective transformation that maps world coordinates points in $\mathbf{R}^3$ to $\mathbf{R}^2$ image coordinate system followed by normalized camera coordinate system.

$$
\begin{gathered}
Z_c \begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix} = \begin{bmatrix} X_c \\ Y_c \\ Z_c \\ 1 \end{bmatrix} = \begin{bmatrix} R | t \end{bmatrix} \begin{bmatrix} X_w \\ Y_w \\ Z_w \\ 1 \end{bmatrix} = \begin{bmatrix} r_{11} & r_{12} & r_{13} & t_x \\ r_{21} & r_{22} & r_{23} & t_y \\ r_{31} & r_{32} & r_{33} & t_z \end{bmatrix} \begin{bmatrix} X_w \\ Y_w \\ Z_w \\ 1 \end{bmatrix} \\
\lambda \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} r_{11} & r_{12} & r_{13} & t_x \\ r_{21} & r_{22} & r_{23} & t_y \\ r_{31} & r_{32} & r_{33} & t_z \end{bmatrix} \begin{bmatrix} X_w \\ Y_w \\ Z_w \\ 1 \end{bmatrix}, \\
\begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} f_x X_c / Z_c + c_x \\ f_y Y_c / Z_c + c_y \end{bmatrix}
\end{gathered} \tag{8}
$$

where $x' = X_c/Z_c$ and $y' = Y_c/Z_c$.

Example 03

Let's assume we know x and X, then how we are going to find the λ, K, R and t? Which one of these belongs to intrinsic parameters? As given in Equ. 6, task is to find camera matrix P which is a $? \times ?$ matrix. Thus, how may unknown we have and how many equations we need to solve in this problem?

Example 03

Let's say we have N number of points in which correspondence is known between world and camera frame.

$$\lambda_i x_i = PX_i, \quad i = 1, ..N \tag{9}$$

In order to find P, can you try to derive an expression for for minimum value for N to be satisfied? And prove that N should be equal or higher than 6.

Example 03

Let $p_i, i = 1, 2, 3$ be vectors containing the rows of P, that is,

$$P = \begin{bmatrix} p_1^T \\ p_2^T \\ p_3^T \end{bmatrix} \tag{10}$$

then, Equ. 9 can be reformulated as follows:

$$\begin{aligned} X_i^T p_1 - \lambda_i x_i = 0 \\ X_i^T p_2 - \lambda_i y_i = 0 \\ X_i^T p_3 - \lambda_i = 0 \end{aligned} \tag{11}$$

Example 03

Can you convert Eqs. 11 into matrix form?

$$\underbrace{\begin{bmatrix} X_1^T & 0 & 0 & -x_1 & 0 & 0 & \cdots \\ 0 & X_1^T & 0 & -y_1 & 0 & 0 & \cdots \\ 0 & 0 & X_1^T & -1 & 0 & 0 & \cdots \\ X_2^T & 0 & 0 & 0 & -x_2 & 0 & \cdots \\ 0 & X_2^T & 0 & 0 & -y_2 & 0 & \cdots \\ 0 & 0 & X_2^T & 0 & -1 & 0 & \cdots \\ X_3^T & 0 & 0 & 0 & 0 & -x_3 & \cdots \\ 0 & X_3^T & 0 & 0 & 0 & -y_3 & \cdots \\ 0 & 0 & X_3^T & 0 & 0 & -1 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}}_{M} \underbrace{\begin{bmatrix} p_1 \\ p_2 \\ p_3 \\ \lambda_1 \\ \lambda_2 \\ \lambda_3 \\ \vdots \end{bmatrix}}_{v} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ \vdots \end{pmatrix} \tag{12}$$

Example 03

In order to find vector v, we have to find the null space vector of M. Basically, need to solve system $Mv = 0$? Can we actually solve this? I would say no! what are you up to?
Thus, this can be solved as follows:

$$min_{|v|^2=1}|Mv|^2 = 0 \tag{13}$$

Camera calibration is all about finding project matrix (P = K [R t]). More information can be found here:`https://www.mathworks.com/help/vision/ug/single-camera-calibrator-app.html` or `http://wiki.ros.org/camera_calibration`.

Simple Stereo

If the camera matrices are known (the triangulation problem) Direct Linear Transformation (DLT) to find the projection matrix (P). On the contrary, if the scene points and camera matrices are not known problem get complicated. The main intuition is to find some similarities between considered two images where part of those are overlapping each other. The technique is used so as to solve this problem is called the **epipolar geometry**.

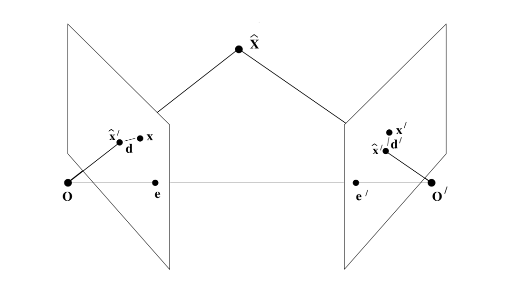


Figure: Simple stereo system

Simple Stereo

If both projection matrices, i.e., P_1 and P_2, are known, how can we estimate the $\hat{\mathbf{x}}$ and $\hat{\mathbf{x}}'$ for a known $\hat{\mathbf{X}}$?

In general, there are two types of problems that belongs to general stereo: matrix K is known, need to find [R t] matrix (Essential Matrix) or matrix K also unknown or has different focal lengths (Fundamental Matrix).

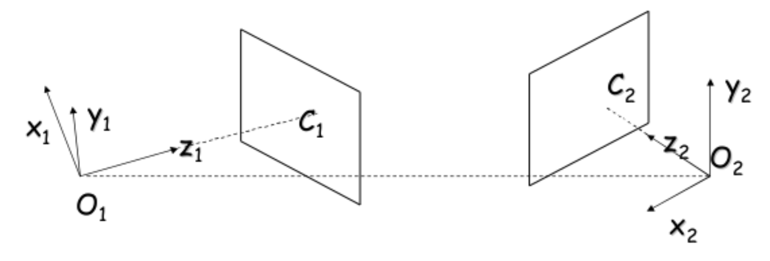


Figure: General stereo [3]

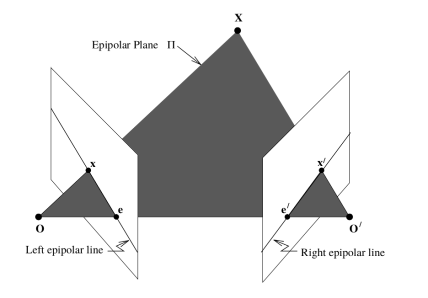


Figure: Epipolar Geometry

As shown in Fig 7, $\mathbf{e}$ and $\mathbf{e}'$ are considered as epipoles, epipolar plane is defined by points $\mathbf{O}'$, $\mathbf{O}$ and $\mathbf{X}$. Besides, assume f and f' are the focal lengths of left and right cameras, respectively

Some homogeneous properties

1. Point $\mathbf{x}$ on a line

$$\mathbf{l}^T\mathbf{x} = \mathbf{x}^T\mathbf{l} = 0,\ l_1x + l_2y + l_3 = 0 \tag{14}$$

2. Two points define a line

$$\mathbf{l} = \mathbf{x_1} \times \mathbf{x_2} \tag{15}$$

3. Intersection of two lines defines a point

$$\mathbf{x} = \mathbf{l_1} \times \mathbf{l_2} \tag{16}$$

   where cross product between two vectors can be written as a matrix multiplication

$$\mathbf{v} \times \mathbf{u} = [\mathbf{v}]_\times \mathbf{u},\ \mathbf{v}_\times = \begin{bmatrix} 0 & -v_z & v_y \\ v_z & 0 & -v_x \\ -v_y & v_x & 0 \end{bmatrix} \tag{17}$$

Epipolar Geometry

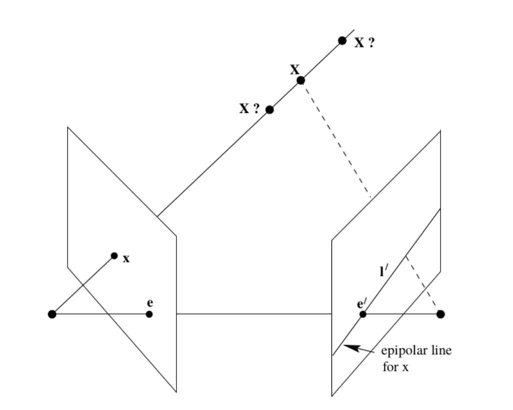


Figure: How can we see the location we see from left camera on the right camera

Since we have two cameras, two projection matrices with respect to left and right cameras have to be identified:

$$\begin{aligned} \mathbf{x} &= \lambda_1 P_1 \mathbf{X} \\ \mathbf{x}' &= \lambda_2 P_2 \mathbf{X}, \end{aligned} \tag{18}$$

where $P_1 = K_1(I|0)$ and $P_2 = K_2(R|t)$, and baseline between the two cameras is denoted by t. Let's start assuming K_1 and K_2 are known. Then if

$$\hat{\mathbf{x}}' = K_2^{-1}\mathbf{x}' = \lambda_2(R|t)\mathbf{X} \tag{19}$$

Now let's project $\mathbf{X}$ on the left and right images

$$\begin{aligned} \mathbf{x} = \lambda_1(I|0)\begin{pmatrix}\mathbf{X}\\1\end{pmatrix} \Rightarrow \mathbf{X} = \begin{pmatrix}\lambda_1^{-1}\mathbf{x}\\1\end{pmatrix} \\ \lambda_2(R|t)\begin{pmatrix}\lambda_1^{-1}\mathbf{x}\\1\end{pmatrix} = \lambda_2\lambda_1^{-1}R\mathbf{x} + \lambda_2 t = \lambda_2(\lambda_1^{-1}R\mathbf{x} + t) \end{aligned} \tag{20}$$

And this will be the epipolar line with respect to right camera in our setup. Let's take corresponding points when $\lambda_1 = 1$ and $\lambda_1 = \pm\infty$

$$\begin{gathered} t \\ (R\mathbf{x} + t) \end{gathered} \tag{21}$$

Thus, we can define the right epipolar line:

$$\begin{aligned} \mathbf{l}' = t \times (R\mathbf{x} + t) = t \times R\mathbf{x} + t \times t = t \times R\mathbf{x} \\ = [t]_\times R\mathbf{x} = E\mathbf{x} \end{aligned} \tag{22}$$

This matrix E is called the Essential matrix, which map point in the left image to a line in the right image. Thus, we can define the epipolar constraint that $\mathbf{x}'$ lies on $\mathbf{l}'$ can be written as

$$\mathbf{x}'^T \mathbf{l}' = \mathbf{x}'^T E\mathbf{x} = 0 \tag{23}$$

Some properties of Essential matrix

1. The epipolar line corresponding to $\mathbf{x}'$ is given by

$$\mathbf{l} = E^T \mathbf{x}' \tag{24}$$

2. The epipole $\mathbf{e}'$ by definition has

$$0 = \mathbf{e}'^T \mathbf{l}' \tag{25}$$

   where $\mathbf{e}'^T E = 0$ for all $\mathbf{x}$. Thus, $\mathbf{e}'$ is the left null space of E. Similarly, $E\mathbf{e} = 0$ that is the right null-space of E.

Heretofore, we assumed we have the camera parameters are known. What if we do not have those as well. In that sense, along with the Essential matrix can we calculate the correspondence. For that we have to go from image plane to camera plane and estimate the correspondence, namely Fundamental Matrix F, between camera planes.
Let's plug back-in the camera coordinates since we do not know the camera parameters.

$$\hat{\mathbf{x}}^\top E\hat{\mathbf{x}} = \mathbf{x}'^T K_2^{-T} E K_1^{-1}\mathbf{x} = \mathbf{x}'^T K_2^{-T}[t]_\times R K_1^{-1}\mathbf{x} = \mathbf{x}'^T F\mathbf{x} \tag{26}$$

where F is the fundamental matrix.

Some properties of Fundamental matrix:

1. F is a 3×3 rank 2 homogeneous matrix
2. $F^T \mathbf{e}' = 0$
3. 7 degree of freedom (-1 for scaling and -1 for det(F)=0)
4. Epipolar lines

$$\mathbf{l}' = F\mathbf{x},\ \mathbf{l} = F^T\mathbf{x}' \tag{27}$$

5. Epipoles:

$$F\mathbf{e} = 0,\ F^T\mathbf{e}' = 0 \tag{28}$$

Fundamental matrix

https://www.youtube.com/watch?v=EokL7E6o1AE

There are various techniques can be applied to calculate F. 8-point algorithm is the one of primitive techniques is used to find the matrix F. As we saw in the previous section, we have a epipolar constraint ($\mathbf{x'}^T F\mathbf{x} = 0$) for each corresponding points in right and left images. Let $\mathbf{x}' \sim (x'_i, y'_i, z'_i)$ and $\mathbf{x} \sim (x_i, y_i, z_i)$.

$$\mathbf{x'}^T F\mathbf{x} = 0 \tag{29}$$

Thus, if we have n number of correspondences in which each correspondence contributes with one linear constraint of F.

$$\begin{pmatrix} x'_1x_1 & x'_1y_1 ... z'_1z_1 \\ x'_2x_2 & x'_2y_2 ... z'_2z_2 \\ . & . \\ . & . \\ . & . \\ x'_nx_n & x'_ny_n ... z'_nz_n \end{pmatrix} \begin{pmatrix} F_{11} \\ F_{12} \\ . \\ . \\ . \\ F_{33} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ . \\ . \\ . \\ 0 \end{pmatrix} \tag{30}$$

These kind of linear homogeneous systems can be solved with SVD (Singular Value Decomposition). The matrix F has 9 entries. But image correspondence is taken in the image plane, $z_i' = 1$ and $z_i = 1$. Thus, this system has the 8 degree of freedom. One of the properties of F is det(F) = 0. However this constraint is not actually true due to the noise of the system. Therefor, it is required to minimize this ($\min_{det(F)=0} \left|\hat{F} - F\right|$) in order to find matrix F. Solution to $\hat{F}$ is given by SVD of it.

$$USV^t = \hat{F} \tag{31}$$

where $S = diag(\sigma_1, \sigma_2, \sigma_3)$. Then F can be found by setting the smallest singular value $\sigma_3 = 0$, that is

$$F = Udiag(\sigma_1, \sigma_2, 0)V^\top \tag{32}$$

As we discussed earlier, $F\mathbf{e} = 0$. Hence, e is the last column of V. Similarly, $F^T\mathbf{e}' = 0$ which implies $\mathbf{e}'$ is the last column of U.

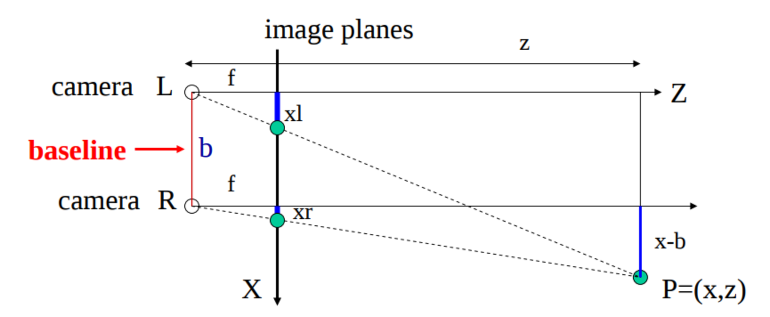


Figure: Stereo images triangulation [2]

Let's try to derive a formula for the depth estimation. As shown in the Fig 9, stereo cameras with parallel optical axes (assuming images are rectified), focal length f, baseline b, corresponding image points (xl,yl) and (xr,yr), the location of the 3D point can be estimated by using the following formula:

$$depth = \frac{f * b}{(xl - xr)} \tag{33}$$

where $xl - xr$ is also called as disparity.

References

http://www.ctr.maths.lu.se/media/FMA270/2015/alllectures.pdf.
2015.

https://courses.cs.washington.edu/courses/cse455/09wi/Lects/lect1

2018.

http://www.cse.psu.edu/ rtc12/CSE486/.
Lec18,Lec19.

Gregor Klancar, Andrej Zdesar, Saso Blazic, and Igor Skrjanc.
Wheeled mobile robotics: from fundamentals towards autonomous systems.
Butterworth-Heinemann, 2017.

Roland Siegwart, Illah Reza Nourbakhsh, and Davide Scaramuzza.

Introduction to autonomous mobile robots.
MIT press, 2011.

Sebastian Thrun.