Table of Contents

Stereo Geometry

In stereopsis, two visual sensors are used in order to obtain the depth of scene points, as an attempt the reconstruct the observed scene. Image features from one image must correlate with the features observed in the other image. This is commonly known as the correspondence problem. Once correspondences are obtained, it is possible to reconstruct the scene by computing the 3D coordinates of the feature points.

The Correspondence Problem

We assume that most scene points are visible from both viewpoints, and that corresponding image regions are similar. Given an element in the left image, we search for the corresponding element in the right image. We may use a correlation-based stereo approach, which attempts to match image neighboring image regions between the two images, leading to dense disparity fields. Alternatively, we may use a feature-based stereo approach, which yields sparser disparity fields.

In correlation-based methods, the tokens to be matched are image regions of a fixed size:

- $\vec{p}_l$, $\vec{p}_r$: pixels in the left and right images
- $2W+1$: width of correlation window
- $R(\vec{p}_l)$: search region in the right images associated with $\vec{p}_l$
- $\psi(u,v)$: a function of two pixel values

Here is a typical correlation-based stereo algorithm, based on the simple assumption that there are no occlusions:

- For each pixel $\vec{p}_l(i,j)^T$ in the left image
  - For each displacement $\vec{d}=(d_1,d_2)^T \in R(\vec{p}_1)$
  - Compute $C(\vec{d})=\sum_{k=-W}^{W}\sum_{l=-W}^{W}\psi(\vec{p}_l(i+k,j+l),\vec{p}_r(i+k-d_1,j+l-d_2))$

- The disparity vector for $\vec{p}_l$ is $\vec{d}(d_1, d_2)^T$ which maximizes $C(\vec{d})$

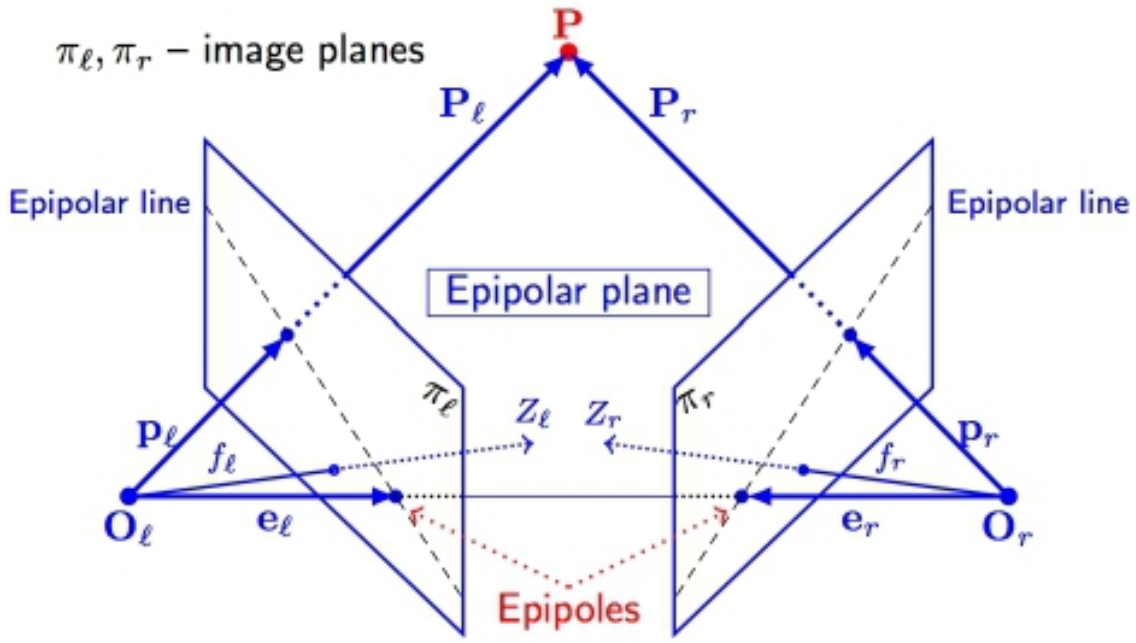
Usually, the function ψ is a Sum-of-Squared-Differences (SSD).

In feature-based methods, we search for correspondences within sparse sets of features. Most algorithms narrow down possible matches with constraints such as those derived from epipolar geometry.

Epipolar Geometry

Let the following variables be:

- O_l , O_r : projection centers
- π_l , π_r : image planes
- f_l , f_r : focal lengths
- $\vec{P}_l=(X_l, Y_l, Z_l)^T$ and $\vec{P}_r=(X_r, Y_r, Z_r)^T$: a 3D point, viewed from the left and right cameras
- $\vec{p}_l=(x_l, y_l)^T$ and $\vec{p}_r=(x_r, y_r)^T$: projections of the same point



The frames of reference for the cameras are related via the extrinsic parameters $\vec{P}_r=R(\vec{P}_l-\vec{T})$. The projections of the 3D point on the two cameras are given by:

$$\vec{p}_l = \frac{f_l}{Z_l}\vec{P}_l \qquad \vec{p}_r = \frac{f_r}{Z_r}\vec{P}_r$$

The epipolar constraint states that the correct stereo match for the point must lie on the epipolar line, and thus reduces the search to a one-dimensional problem. The equation of the epipolar plane can be written as a coplanarity condition on vectors $\vec{P}_l$, $\vec{T}$, and $\vec{P}_l - \vec{T}$ (using the triple scalar product):

$$(\vec{P}_l - \vec{T})^T \vec{T} \times \vec{P}_l = 0$$

which can be rewritten as:

$$(R^T \vec{P}_r)^T \vec{T} \times \vec{P}_l = 0$$

since $R^T \vec{P}_r = \vec{P}_l - \vec{T}$. The cross product can be expressed as a matrix multiplication in the following way:

$$\vec{T} \times \vec{P}_l = S\vec{P}_l$$

where

$$S = \begin{bmatrix} 0 & -T_z & T_y \\ T_z & 0 & -T_x \\ -T_y & T_x & 0 \end{bmatrix}$$

Hence, the coplanarity condition becomes

$$(R^T \vec{P}_r)^T S \vec{P}_l = \vec{P}_r{}^T RS \vec{P}_r{}^T = \vec{P}_r{}^T E \vec{P}_l = 0$$

where $E = RS$ is the essential matrix, as it establishes a natural link between the epipolar constraint and the extrinsic parameters of the stereo cameras.

Using the perspective projection equations in the following way

$$\vec{P}_l = \frac{Z_l}{f_l}\vec{p}_l \qquad \vec{P}_r = \frac{Z_r}{f_r}\vec{p}_r$$

and substituting in the coplanarity condition equation results in

$$\frac{Z_r}{f_r}\vec{p}_r{}^T E \frac{Z_l}{f_l}\vec{p}_l = 0$$

Multiplying both sides by $\frac{f_r f_l}{Z_r Z_l}$ yields $\vec{p}_r{}^T E \vec{p}_l = 0$. Hence, the coplanarity constraint holds under perspective projection. Note that $\vec{u}_r = E\vec{p}_l$ is the epipolar

line on the right image (conversely $\vec{u}_l = E^T \vec{p}_r$ is the epipolar line on the left image).

In addition to the essential matrix, there exists the fundamental matrix. The fundamental matrix is defined in terms of pixel coordinates, as opposed to sensor coordinates, and if one estimates the fundamental matrix from point matches in pixel coordinates, then we can reconstruct the epipolar geometry without the knowledge of the intrinsic and extrinsic parameters of the stereo sensors. In other words, calibration is unnecessary in this context.

Suppose we have:

- M_l , and M_r : matrices of the intrinsic parameters of the left and right cameras
- $\bar{p}_l$, $\bar{p}_r$: image points in pixel coordinates, corresponding to $\vec{p}_l$ and $\vec{p}_r$

It is then possible to write $\vec{p}_l = M_l^{-1} \bar{p}_l$ and $\vec{p}_r = M_r^{-1} \bar{p}_r$. By substitution, we obtain

$$\bar{p}_r^{\,T} F \bar{p}_l = 0$$

where $F = (M_r^{-1})^T E M_l^{-1}$ is the fundamental matrix, and

$$M = \begin{bmatrix} \frac{-f}{s_x} & 0 & o_x \\ 0 & \frac{-f}{s_y} & o_y \\ 0 & 0 & 0 \end{bmatrix}$$

As before we have $\vec{u}_r = F \bar{p}_l$. The most important difference between the essential and fundamental matrices is that the fundamental matrix is defined in terms of pixel coordinates while the essential matrix is defined in terms of camera coordinates. Consequently, it is possible from a set of image matches to reconstruct the epipolar geometry, without using intrinsic or extrinsic calibration parameters.

In summary:

- For each pair of corresponding points $\vec{p}_l$ and $\vec{p}_r$ in camera coordinates, the essential matrix satisfies the equation $\vec{p}_r^{\,T} E \vec{p}_l = 0$.
- For each pair of corresponding points $\bar{p}_l$ and $\bar{p}_r$ in pixel coordinates, the fundamental matrix satisfies the equation $\bar{p}_r^{\,T} F \bar{p}_l = 0$.
- Both matrices enable the reconstruction of the epipolar geometry. If M_l

and M_r are the matrices of the intrinsic parameters, then the relation between the essential and fundamental matrices is given by $F = (M_r^{-1})^T E M_l^{-1}$.

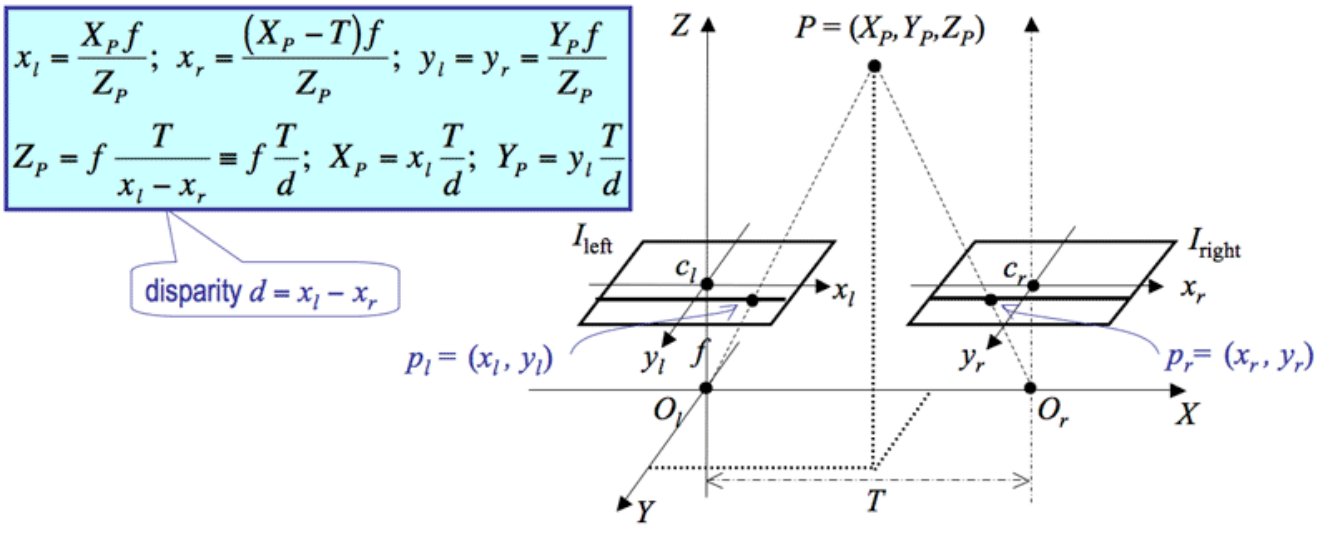
- The essential matrix:
  - encodes information on the extrinsic parameters only
  - has rank 2, since S has rank 2 and R has full rank
  - its 2 non-zero singular values are equal
- The fundamental matrix:
  - encodes information on both the intrinsic and extrinsic parameters
  - has rank 2, since $\vec{T}_l$ and $\vec{T}_r$ have full rank and E has rank 2

Image Rectification

Rectification defines a transformation such that epipolar lines become collinear and parallel to the horizontal image axis. After such a transformation is applied to both the left and right images. The process of finding image disparities on the epipolar lines (now rows of pixels) can begin.

Disparity and Depth

Once a disparity has been found for a pixel, we can reconstruct the 3D point in absolute coordinates.



Let T be the norm of $\vec{T}$, or the baseline of the stereo system. Then it is easy to show that given a match pair $\bar{p}_l = (x_l, y_l), \bar{p}_r = (x_r, y_r)$ for the 3D point

$P(X_p, Y_p, Z_p)$:

$$x_l = \frac{X_p f}{Z_p} \qquad x_r = \frac{(X_p - T) f}{Z_p} \qquad y_l = y_r = \frac{Y_p f}{Z_p}$$

By using similar triangles, we find that

$$Z_p = f \frac{T}{x_l - x_r}$$

and we can compute the 3D coordinates of point P .

Camera Parameters

Extrinsic parameters describe position and orientation of a camera with respect to a world coordinate system, or to another camera. Intrinsic parameters establish the relationship between pixel coordinates of an image point with its coordinates the camera frame of reference. The typical extrinsic parameters are R and $\vec{T}$, a rotation matrix and a translation vector. The expression of a point in world coordinate system in the camera coordinate system is given by

$$\vec{P}_c = R(\vec{P}_w - \vec{T})$$

The intrinsic parameters are (o_x, o_y) , the pixel coordinates of the image center, (s_x, s_y) , the effective pixel size in millimeters, and f , the focal length. The $(x, y)^T$ camera coordinates are given by the following relationship with the intrinsic parameters:

$$\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} -(x_i - o_x) s_x \\ -(y_i - o_y) s_y \end{pmatrix}$$

where $(x_i, y_i)^T$ are the pixel coordinates of (x, y) . We can put these equations together to obtain

$$-(x_i - o_x) s_x = f \frac{\vec{R}_1^T (\vec{P}_w - \vec{T})}{\vec{R}_3^T (\vec{P}_w - \vec{T})}$$

and

$$-(y_i - o_y) s_y = f \frac{\vec{R}_2^T (\vec{P}_w - \vec{T})}{\vec{R}_3^T (\vec{P}_w - \vec{T})}$$

Note that $\vec{R}_i$ is a vector formed with the i^{th} row of matrix R . These two equations can be expressed as a simple matrix product:

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = M_i M_e \vec{P}_w$$

where

$$M_e = \begin{pmatrix} r_{11} & r_{12} & r_{13} & -\vec{R}_1^T \vec{T} \\ r_{21} & r_{22} & r_{23} & -\vec{R}_2^T \vec{T} \\ r_{31} & r_{32} & r_{33} & -\vec{R}_3^T \vec{T} \end{pmatrix}$$

and

$$M_i = \begin{pmatrix} \frac{-f}{s_x} & 0 & o_x \\ 0 & \frac{-f}{s_y} & o_y \\ 0 & 0 & 1 \end{pmatrix}$$

Note that $\frac{x}{z} = x_i$ and $\frac{y}{z} = y_i$, the pixel coordinates of the point.