Table of Contents

Stereopsis

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 minimizes $C(\vec{d})$

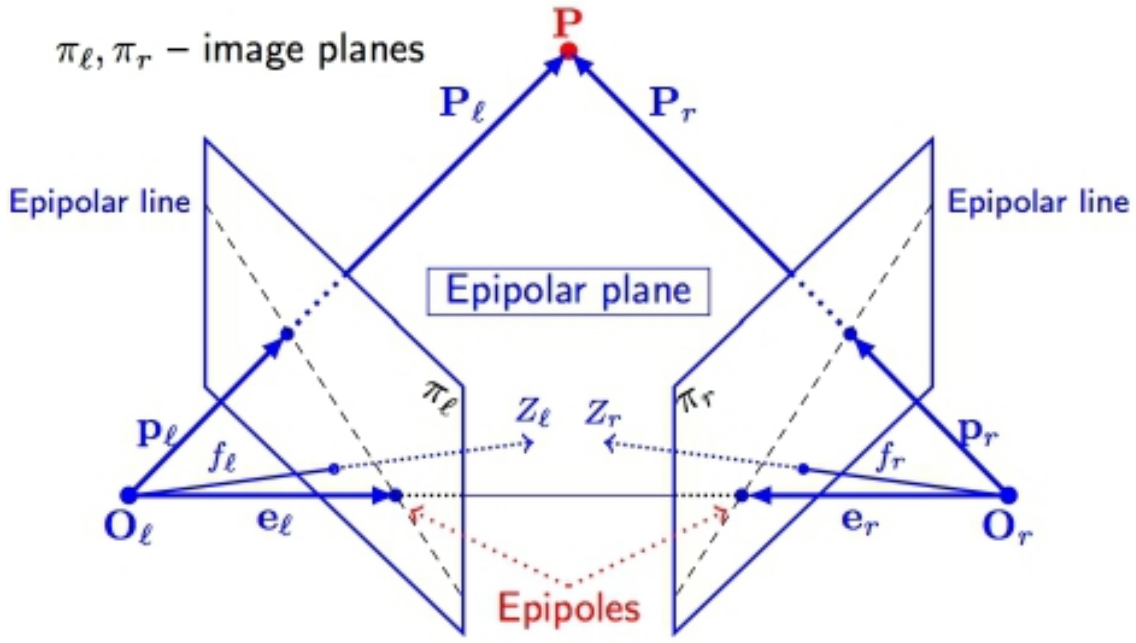
Usually, the function ψ is a Sum-of-Squared-Differences (SSD).

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

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

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

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

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 M_l and M_r have full rank and E has rank 2

Estimating the Essential and Fundamental Matrices

Assume we have established n correspondences between two images. Each yields a homogeneous linear equation of the form:

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

for the nine values of F . If one establishes at least 8 correspondences that do not form a degenerate configuration, the elements of F can be estimated as the non-trivial solution of the system. Since the system is homogeneous, the solution is unique up to a signed scale factor. When more correspondences are obtained, one may resort to Singular Value Decomposition (SVD) techniques.

If A is the system's matrix and $A = UDV^T$, the solution is the column of V corresponding to the only null singular value of A . However, due to noise, the matrix A is more than likely to be of full rank, and in this case, the solution is the column of V associated with the least singular value of A . Also, the estimated fundamental matrix is almost certainly non-singular. Singularity can be enforced by computing the SVD of the fundamental matrix and setting the smallest singular value of matrix D to zero, and then recomputing F .

Algorithm

The input to the algorithm is n correspondences where $n \geq 8$.

- Construct the system of equations from the correspondences. Let A be

the $n \times 9$ matrix of the coefficients of the system and $A = UDV^T$.

- The elements of F (up to a signed scale factor) are the components of the column of V corresponding to the least singular value of A .
- To enforce singularity, compute the SVD of the fundamental matrix $F = UDV^T$.
- Set the smallest singular value in the diagonal of D to zero and let D' be the corrected matrix.
- The corrected estimate of F is thus given by $F' = UD'V^T$.

An important note on how to implement this algorithm is given by Trucco and Verri:

- *The most important action to take is to normalize the coordinates of the corresponding points so that the entries in matrix A are of comparable size. Typically, the first two coordinates (in pixel units) of an image point are referred to the top left corner of the image and can vary between a few pixels to a few hundreds. This difference can make the matrix seriously ill-conditioned. To make things worse, the third (homogenous) coordinate of image points is usually set to one. A simple procedure to avoid numerical instability is to translate the first 2 coordinates of each point to the centroid of each data set, and scale the norm of each point so that the average norm over the data set is one. This can be accomplished by multiplying each left (and right) point by two suitable 3 by 3 matrices H_r and H_l , and then use the 8-point algorithm to estimate the matrix $\tilde{F} = H_r F H_l$, where F is obtained as $H_r^{-1} \tilde{F} H_l^{-1}$.*

These matrices can be determined as follows:

- Given a set of n points $\boldsymbol{p_i} = (x_i, y_i, 1)^T$ define $\bar{x} = \sum_i \frac{x_i}{n}$, $\bar{y} = \sum_i \frac{y_i}{n}$ and $\tilde{d} = \frac{\sum_i \sqrt{(x_i - \bar{x})^2 + (y_i - \bar{y})^2}}{n\sqrt{2}}$
- Find the 3 by 3 matrix H such that $H\,\boldsymbol{p}_i = \hat{\boldsymbol{p}}_i$ with $\boldsymbol{p}_i = ((x_i - \bar{x})/d\,, (y_i - \bar{y})/d)^T$

Locating Epipoles

Consider the fundamental matrix F . We can write:

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

since the epipole $\bar{e}_l$ lies on the epipolar line of the left image for every $\bar{p}_r$. But since F is not identically zero, this is possible if and only if

$$F\bar{e}_l=0$$

and hence it follows that the epipole $\bar{e}_l$ is the null space of $F.$ Similarly, $\bar{e}_r$ is the null space of F^T . Finding the epipoles is then immediate:

- Find the SVD of the fundamental matrix $F=UVD^T$.
- The epipole $\bar{e}_l$ is the column of V corresponding to the null singular value.
- The epipole $\bar{e}_r$ is the column of U corresponding to the null singular value.

Note that singularity must be enforced on F (by the 8-point algorithm) in order to find a null singular value.

Rectification

Given a pair of stereo images, rectification determines a transformation of each image such that pairs of conjugate become collinear and parallel to one of the image axes, usually the horizontal one. Rectification is important because it reduces the correspondence problem from 2D to 1D, on a scan line that is trivially identified.

Assuming that for both cameras:

- the origin of the image reference frame is the principal point
- the focal length is f

Then we can perform rectification with the following steps:

- Rotate the left camera so that the epipole goes to infinity along the horizontal axis
- Apply the same rotation to the right camera to recover the original geometry
- Adjust the scale in both camera reference frames

We need to construct an orthonormal base of vectors $\{\vec{e}_1,\vec{e}_2,\vec{e}_3\}$. The first vector is given by the epipole. Since the image center is in the origin, $\vec{e}_1$ coincides with the translation:

$$\vec{e}_1=\frac{\vec{T}}{\|\vec{T}\|}$$

The only constraint we have on $\vec{e}_2$ is that it must be orthogonal to $\vec{e}_1$. We thus compute and normalize the cross product of $\vec{e}_1$ with the direction vector of the optical axis, to obtain:

$$\vec{e}_2 = \frac{1}{\sqrt{T_x^2 + T_y^2}} \left[-T_y, T_x, 0 \right]^T$$

The third vector is simply $\vec{e}_3 = \vec{e}_1 \times \vec{e}_2$. The orthogonal matrix defined as:

$$R' = \begin{bmatrix} \vec{e}_1^{\,T} \\ \vec{e}_2^{\,T} \\ \vec{e}_3^{\,T} \end{bmatrix}$$

rotates the left camera about the projection center in such a way that the epipolar lines become parallel to the horizontal axis. The complete algorithm follows:

- Build matrix R'
- Set $R_l = R'$ and $R_r = RR'$
- For each left-camera point $\vec{p}_l = (x, y, f)^T$ compute $R_l \vec{p}_l = (x', y', z')$ and the coordinates of the corresponding rectified point $\vec{p}\,'_l = \frac{f}{z}'(x', y', z')$
- Repeat the previous steps for the right camera using R_r and $\vec{p}_r$

The output is the pair of transformations to be applied the the two cameras in order to rectify the two input point sets.

Notice that the rectified coordinates are not integer in general. Therefore, rectification should be implemented backwards, starting from the new image plane and applying inverse transformations so that the pixel values in the new image can be computed as a bilinear interpolation of the pixel values in the original image.

3D Reconstruction

Once a disparity has been found for a pixel, we can reconstruct the 3D point in absolute coordinates. Let T be the norm of $\vec{T}$ (the baseline of the stereo system). Given a match pair $\bar{p}_l = (x_l, y_l)^T$ and $\bar{p}_r = (x_r, y_r)$ for a 3D point $P = (X_p, Y_p, Z_p)^T$, we have

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

And find with similar triangles:

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

where $d = x_l - x_r$ is the disparity.