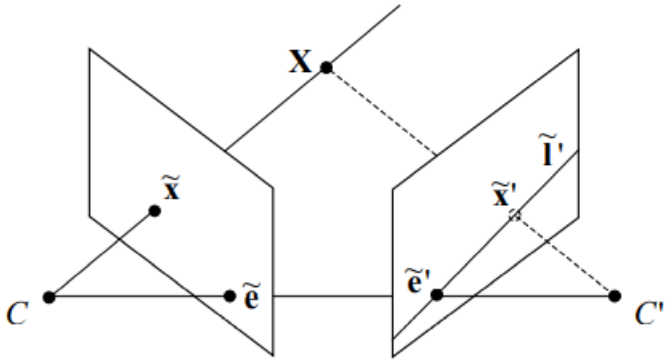
Epipolar Geometry

We have seen how to compute the intrinsic and extrinsic parameters of a camera using one or more views using a typical camera calibration procedure or single view metrology. This process culminated in deriving properties about the 3D world from one image. However, in general, it is not possible to recover the entire structure of the 3D world from just one image. This is due to the intrinsic ambiguity of the 3D to the 2D mapping: some information is simply lost.

Having knowledge of geometry when multiple cameras are present can be extremely helpful. Specifically, we will first focus on defining the geometry involved in two viewpoints and then present how this geometry can aid in further understanding the world around us.

Setup



The standard epipolar geometry setup involves two cameras observing the same 3D point $\mathbf{X}$, whose projection in each of the image planes is located at $\tilde{\mathbf{x}}$ and $\tilde{\mathbf{x}}'$ respectively. The camera centers are located at C_1 and C', and the line between them is referred to as the *baseline*. We call the plane defined by the two camera centers and $\mathbf{X}$ the *epipolar plane*. The locations of where the baseline intersects the two image planes are known as the *epipoles* $\tilde{\mathbf{e}}$ and $\tilde{\mathbf{e}}'$. Finally, the lines defined by the intersection of the epipolar plane and the two image planes are known as the *epipolar lines*. The epipolar lines have the property that they intersect the baseline at the respective epipoles in the image plane.

The Essential Matrix

Given the projection of a 3D point in one image, its projection in a second image is restricted to the corresponding epipolar line. The epipolar constraint can be formulated algebraically using the essential matrix $\mathbf{E}$, which relates corresponding image points in two views. Consider two pinhole cameras with projection matrices $\mathbf{P}$ and $\mathbf{P}'$. Given a Euclidean point $\mathbf{X}'$ in the coordinate system of camera C', and its position $\mathbf{X}$ in the coordinate system of C is given by:

$$\mathbf{X} = \mathbf{R}\mathbf{X}' + \mathbf{T}$$

where $\mathbf{R}$ is a 3×3 rotation matrix and $\mathbf{T}$ is a 3-vector. Pre-multiplying both sides by $\mathbf{X}^\top [\mathbf{T}]_\times$ gives:

$$\mathbf{X}^\top [\mathbf{T}]_\times \mathbf{R}\mathbf{X}' = \mathbf{X}^\top \mathbf{E}\mathbf{X}' = 0 \qquad (*)$$

where 3×3 *essential matrix* $\mathbf{E} \sim [\mathbf{T}]_\times \mathbf{R}$ and $[\mathbf{T}]_\times$ is the cross product matrix. For $\mathbf{T} = [t_x, t_y, t_z]^\top$

$$[\mathbf{T}]_\times = \begin{bmatrix} 0 & -t_z & t_y \\ t_z & 0 & -t_x \\ -t_y & t_x & 0 \end{bmatrix}$$

Equation $(*)$ also holds for image points $\tilde{\mathbf{x}}$ and $\tilde{\mathbf{x}}'$, which gives the epipolar constraint:

$$\tilde{\mathbf{x}}^\top \mathbf{E} \tilde{\mathbf{x}}' = 0$$

It could be seen that $l' = \mathbf{E}^\top \tilde{\mathbf{x}}$ gives the epipolar line in the image plane of camera 2. Similarly $l = \mathbf{E}\tilde{\mathbf{x}}'$ gives the epipolar line the image plane of camera 1.

Note that $\mathbf{E}$ depends only on $\mathbf{R}$ and $\mathbf{T}$ and is defined only up to an arbitrary scale factor. Thus it has 5 parameters.

The Fundamental Matrix

Image points $\tilde{\mathbf{x}}$ may be related to pixel position $\tilde{\mathbf{u}}$ by the inverse camera calibration matrix $\mathbf{K}^{-1}$:

$$\tilde{\mathbf{x}} \sim \mathbf{K}^{-1}\tilde{\mathbf{u}}$$

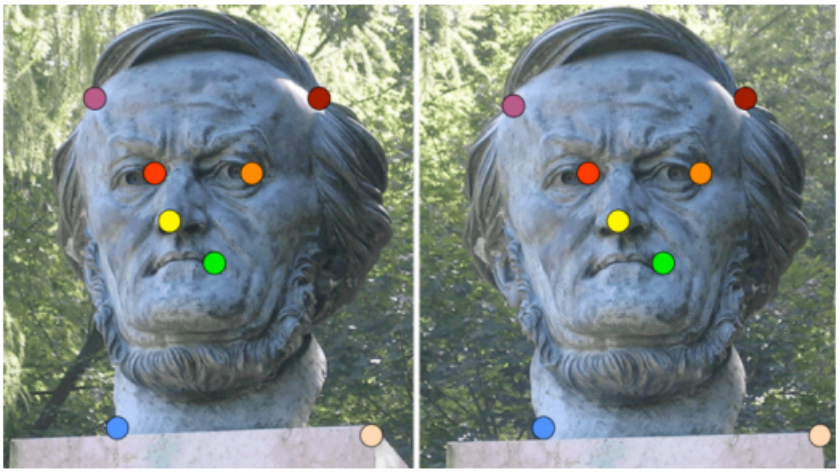
Substitute the expression in equation $(*)$ we have

$$\begin{aligned} \left(\mathbf{K}^{-1}\tilde{\mathbf{u}}\right)^\top \mathbf{E} \left(\mathbf{K}'^{-1}\tilde{\mathbf{u}}'\right) &= 0 \\ \tilde{\mathbf{u}}^\top \left(\mathbf{K}^{-\top}\mathbf{E}\mathbf{K}'^{-1}\right)\tilde{\mathbf{u}}' &= 0 \\ \tilde{\mathbf{u}}^\top \mathbf{F} \tilde{\mathbf{u}}' &= 0 \end{aligned}$$

where $\mathbf{F} \sim \mathbf{K}^{-\top}\mathbf{E}\mathbf{K}'^{-1}$ is the *fundamental matrix*. $\mathbf{F}$ is a 3×3 matrix that has rank 2(the epipole $\tilde{\mathbf{e}}$ is the null space of $\mathbf{F}$). It can be estimated linearly given 8 or more corresponding points.

Estimating the Fundamental Matrix

It is possible to estimate the fundamental matrix given two images of the same scene and *without knowing the extrinsic or intrinsic parameters of the camera*. The method we discuss for doing so is known as the **Eight-Point Algorithm**.



Each point correspondence, $\tilde{\mathbf{u}}_i \sim [u_i, v_i, 1]^\top$ and $\tilde{\mathbf{u}}'_i \sim [u'_i, v'_i, 1]^\top$, generates one constraint on the elements of the fundamental matrix $\mathbf{F}$:

$$\begin{bmatrix} u'_i & v'_i & 1 \end{bmatrix} \begin{bmatrix} f_{11} & f_{12} & f_{13} \\ f_{21} & f_{22} & f_{23} \\ f_{31} & f_{32} & f_{33} \end{bmatrix} \begin{bmatrix} u_i \\ v_i \\ 1 \end{bmatrix} = 0$$

Rearrange it:

$$\begin{pmatrix} u'_i\tilde{\mathbf{u}}_i^\top & v'_i\tilde{\mathbf{u}}_i^\top & \tilde{\mathbf{u}}_i{}^\top \end{pmatrix} \begin{pmatrix} \mathbf{f}_1 \\ \mathbf{f}_2 \\ \mathbf{f}_3 \end{pmatrix} = \mathbf{0}$$

or explicitly for n pairs of correspondences,

$$\begin{bmatrix} u'_1u_1 & u'_1v_1 & u'_1 & v'_1u_1 & v'_1v_1 & v'_1 & u_1 & v_1 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ u'_nu_n & u'_nv_n & u'_1 & v'_nu_n & v'_nv_n & v'_n & u_n & v_n & 1 \end{bmatrix} \begin{bmatrix} f_{11} \\ f_{12} \\ f_{13} \\ f_{21} \\ f_{22} \\ f_{23} \\ f_{31} \\ f_{32} \\ f_{33} \end{bmatrix} = \mathbf{0}$$

in short

$$\mathbf{A}\mathbf{f} = 0$$

where $\mathbf{A}$ is an $n \times 9$ measurement matrix, and $\mathbf{f}$ represents the elements of the fundamental matrix f_{ij} as a 9-vector. Given 8 or more correspondences a least squares solution can be found as the unit eigenvector (f is defined up to an arbitrary scale) corresponding to the minimum eigenvalue of $\mathbf{A}^\top\mathbf{A}$.

Note that the fundamental matrix has only 7 degrees of freedom since its determinant must be zero. A non-unique solution can be obtained from only 7 point correspondences.

The Normalized Eight-Point Algorithm

The main problem of the standard Eight-Point Algorithm stems from the fact that $\mathbf{A}$ is ill-conditioned for SVD. For SVD to work properly, $\mathbf{A}$ should have one singular value equal to (or near) zero, with other singular values being nonzero. However, the correspondences $\tilde{\mathbf{u}}_i$ will often have extremely large values in the first and second coordinates due to the pixel range of a modern camera (i.e. $\tilde{\mathbf{u}}_i = (1832, 1023, 1)$).

To solve this problem, we will normalize the points in the image before constructing $\mathbf{A}$. This means we pre-condition $\mathbf{A}$ by applying both a translation and scaling on the image coordinates such that two requirements are satisfied:

- the origin of the new coordinate system should be located at the centroid of the image points (translation).
- the mean square distance of the transformed image points from the origin should be 2 pixels (scaling).

We can compactly represent this process by transformation matrices H, H' that translate by the centroid and scale by the scaling factor

$$\sqrt{\frac{2N}{\sum_{i=1}^{N} \|\mathbf{u}_i - \overline{\mathbf{u}}\|^2}}$$

for each image.

Using the new, normalized coordinates, we can compute the new $\mathbf{F}_n$ using the regular least-squares Eight Point Algorithm. And after denormalization, we have

$$\mathbf{F} = H^\top \mathbf{F}_n H'$$

Recovering Projection Matrices

If the camera calibration matrices $\mathbf{K}$ and $\mathbf{K}'$ are known, we can transform the recovered fundamental matrix into an essential matrix

$$\mathbf{E} \sim \mathbf{K}^\top \mathbf{F} \mathbf{K}'$$

and decompose this matrix into a skew-symmetric matrix corresponding to translation and an orthonormal matrix corresponding to the rotation between the views;

$$\mathbf{E} \sim [\mathbf{T}]_\times \mathbf{R}$$

The latter is in fact only possible if the essential matrix has rank 2 and two equal singular values.

This decomposition can be achieved by computing the singular value decomposition of the essential matrix

$$\mathbf{E} = \mathbf{U} \Lambda \mathbf{V}^\top$$

where $\Lambda = \operatorname{diag}(\sigma_1, \sigma_2, \sigma_3)$ and the matrices $\mathbf{U}$ and $\mathbf{V}$ are orthogonal. The decomposition into a translation vector and the rotation between the two views requires that $\sigma_1 = \sigma_2 \neq 0$ and $\sigma_3 = 0$. The nearest essential matrix (in the sense of minimizing the Frobenius norm between the two matrices) with the correct properties can be obtained by setting the two largest singular values to be equal to their average and the smallest one to zero.

The translation and axis and angle of rotation can then be obtained directly up to arbitrary signs and unknown scale for the translation:

$$[\mathbf{T}]_\times = \mathbf{U} \begin{bmatrix} 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \mathbf{U}^\top$$

$$\mathbf{R} = \mathbf{U} \begin{bmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \mathbf{V}^\top$$

The projection matrices follow directly from the recovered translation and rotation by aligning the reference coordinate system with the first camera to give:

$$\mathbf{P} = \mathbf{K}[\mathbf{I} \mid \mathbf{0}]$$
$$\mathbf{P}' = \mathbf{K}'[\mathbf{R} \mid \mathbf{T}]$$

where $\mathbf{T}$ is typically scaled such that $|\mathbf{T}| = 1$. Four solutions are still possible due to the arbitrary choice of signs for the translation $\mathbf{T}$ and rotation $\mathbf{R}$, however the correct one is easily disambiguated by ensuring that the reconstructed points lie in front of the cameras.

References

Structure from motion
03-epipolar-geometry_2022.pdf (stanford.edu)