About triangulation (conventionals and mine)

Nobuaki Tanaka

November 15, 2025

1 Basics of triangulation and camera motion parameters

We fix a world coordinate system XYZ and call it the *world coordinate system*. For each camera, we define a *camera coordinate system* whose origin is at the optical center O_{lens} and whose z–axis is aligned with the optical axis of the lens.

The pose (position and orientation) of a camera is described by rotation and translation parameters. We denote by $R \in SO(3)$ the rotation matrix, and by $\boldsymbol{t} \in \mathbb{R}^3$ the translation of the camera center in the world frame. The pair

$$\{R, \boldsymbol{t}\}$$

is called the *motion parameters* of the camera.

For a 3D point in the world

$$\boldsymbol{X} = \begin{pmatrix} X \\ Y \\ Z \\ 1 \end{pmatrix} \in \mathbb{P}^3,$$

its image on the camera sensor is denoted by the (homogeneous) image point

$$\boldsymbol{u} = \begin{pmatrix} x \\ y \\ 1 \end{pmatrix} \in \mathbb{P}^2.$$

Assuming a pinhole camera model with focal length f_0 and the principal point at the origin, the projection from $\boldsymbol{X}$ to $\boldsymbol{u}$ is written as

$$\lambda \begin{pmatrix} x \\ y \\ 1 \end{pmatrix} = P \begin{pmatrix} X \\ Y \\ Z \\ 1 \end{pmatrix}, \tag{1}$$

where $P \in \mathbb{R}^{3\times 4}$ is the camera matrix and $\lambda \neq 0$ is a scalar projective factor.

Writing

$$P = \begin{pmatrix} P_{11} & P_{12} & P_{13} & P_{14} \\ P_{21} & P_{22} & P_{23} & P_{24} \\ P_{31} & P_{32} & P_{33} & P_{34} \end{pmatrix},$$

each row of (1) gives one scalar equation. Eliminating the scale factor λ yields linear equations in X, Y, Z.

Linear triangulation with explicit matrices T and $\boldsymbol{p}$

We consider two calibrated cameras observing the same 3D point. Let their camera matrices be

$$P^{(1)} = \begin{pmatrix} P_{11}^{(1)} & P_{12}^{(1)} & P_{13}^{(1)} & P_{14}^{(1)} \\ P_{21}^{(1)} & P_{22}^{(1)} & P_{23}^{(1)} & P_{24}^{(1)} \\ P_{31}^{(1)} & P_{32}^{(1)} & P_{33}^{(1)} & P_{34}^{(1)} \end{pmatrix}, \quad P^{(2)} = \begin{pmatrix} P_{11}^{(2)} & P_{12}^{(2)} & P_{13}^{(2)} & P_{14}^{(2)} \\ P_{21}^{(2)} & P_{22}^{(2)} & P_{23}^{(2)} & P_{24}^{(2)} \\ P_{31}^{(2)} & P_{32}^{(2)} & P_{33}^{(2)} & P_{34}^{(2)} \end{pmatrix}.$$

Let the image coordinates of the point in the two images be

$$(x_1, y_1) \quad \text{in the first image,} \qquad (x_2, y_2) \quad \text{in the second image,}$$

and let f_0 be the (known) focal length of both cameras.[1]

From the projection equation (1) for each camera, and after eliminating the scale factors, we obtain four linear equations in (X, Y, Z). These can be written compactly as

$$T\,\boldsymbol{X}_{\text{xyz}} + \boldsymbol{p} \approx \boldsymbol{0},$$

where

$$\boldsymbol{X}_{\text{xyz}} = \begin{pmatrix} X \\ Y \\ Z \end{pmatrix}, \qquad T \in \mathbb{R}^{4\times 3}, \quad \boldsymbol{p} \in \mathbb{R}^4.$$

Using the component-wise pattern you specified, the entries of T and $\boldsymbol{p}$ are explicitly given by

$$T = \begin{pmatrix} f_0 P_{11}^{(1)} - x_1 P_{31}^{(1)} & f_0 P_{12}^{(1)} - x_1 P_{32}^{(1)} & f_0 P_{13}^{(1)} - x_1 P_{33}^{(1)} \\ f_0 P_{21}^{(1)} - y_1 P_{31}^{(1)} & f_0 P_{22}^{(1)} - y_1 P_{32}^{(1)} & f_0 P_{23}^{(1)} - y_1 P_{33}^{(1)} \\ f_0 P_{11}^{(2)} - x_2 P_{31}^{(2)} & f_0 P_{12}^{(2)} - x_2 P_{32}^{(2)} & f_0 P_{13}^{(2)} - x_2 P_{33}^{(2)} \\ f_0 P_{21}^{(2)} - y_2 P_{31}^{(2)} & f_0 P_{22}^{(2)} - y_2 P_{32}^{(2)} & f_0 P_{23}^{(2)} - y_2 P_{33}^{(2)} \end{pmatrix},$$

$$\boldsymbol{p} = \begin{pmatrix} f_0 P_{14}^{(1)} - x_1 P_{34}^{(1)} \\ f_0 P_{24}^{(1)} - y_1 P_{34}^{(1)} \\ f_0 P_{14}^{(2)} - x_2 P_{34}^{(2)} \\ f_0 P_{24}^{(2)} - y_2 P_{34}^{(2)} \end{pmatrix}.$$

The least–squares solution of this overdetermined system is obtained by solving the normal equations

$$(T^\top T)\,\boldsymbol{X}_{\text{xyz}} = -\,T^\top \boldsymbol{p}, \tag{2}$$

which yields the estimate of the 3D position (X, Y, Z) of the point.

[1] Here we assume a simple intrinsic model with a single focal length f_0 and the principal point at the origin. More general intrinsics can be absorbed into $P^{(1)}$, $P^{(2)}$, but the following form is convenient for our derivation.

2 Conventional triangulation and pattern–projector-based triangulation

Camera matrices P, P' from motion parameters $\{R, \boldsymbol{t}\}$

We consider a two-view setup and use the first camera as the world reference frame. The world coordinate system coincides with the first camera coordinate system.

Let the first camera have focal length f and intrinsic matrix

$$K = \begin{pmatrix} f & 0 & 0 \\ 0 & f & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Since the first camera is placed at the origin with identity orientation, its camera matrix is

$$P = K[\, I_3 \mid \mathbf{0}\,]. \tag{3}$$

The pose of the second camera is described by the motion parameters

$$R \in SO(3), \qquad \boldsymbol{t} \in \mathbb{R}^3,$$

which we define as the rigid transform from the *second camera coordinate system to the world coordinate system*:

$$\boldsymbol{X}_w = R\,\boldsymbol{X}'_c + \boldsymbol{t}, \tag{4}$$

where $\boldsymbol{X}'_c$ denotes coordinates in the second camera frame and $\boldsymbol{X}_w$ denotes coordinates in the world (first-camera) frame.

From (4), the inverse transform (world to second-camera coordinates) is

$$\boldsymbol{X}'_c = R^\top(\boldsymbol{X}_w - \boldsymbol{t}). \tag{5}$$

In homogeneous coordinates this can be written as

$$\boldsymbol{X}'_c = \begin{bmatrix} R^\top & -R^\top\boldsymbol{t} \end{bmatrix} \boldsymbol{X}^{\mathrm{h}}_w, \qquad \boldsymbol{X}^{\mathrm{h}}_w = \begin{pmatrix} X \\ Y \\ Z \\ 1 \end{pmatrix}.$$

Let the second camera have focal length f' and intrinsic matrix

$$K' = \begin{pmatrix} f' & 0 & 0 \\ 0 & f' & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Then the camera matrix of the second view is given by

$$P' = K'\,[\, R^\top \mid -R^\top\boldsymbol{t}\,]. \tag{6}$$

In this parametrization, the relative transform from the first camera frame (world frame) to the second camera frame is

$$\boldsymbol{X}'_c = R_{\mathrm{rel}}\boldsymbol{X}_c + \boldsymbol{t}_{\mathrm{rel}}, \qquad R_{\mathrm{rel}} = R^\top, \quad \boldsymbol{t}_{\mathrm{rel}} = -R^\top\boldsymbol{t}.$$

The pair of camera matrices (P, P') is thus uniquely determined by the intrinsics (K, K') and the motion parameters $\{R, \boldsymbol{t}\}$.

Epipolar geometry and optimal correction of correspondences

Let $\boldsymbol{x} = (u, v, 1)^\top$ and $\boldsymbol{x}' = (u', v', 1)^\top$ be homogeneous image points in the first and second images, respectively. We define the normalized image coordinates

$$\tilde{\boldsymbol{x}} = K^{-1}\boldsymbol{x}, \qquad \tilde{\boldsymbol{x}}' = K'^{-1}\boldsymbol{x}',$$

which live in the respective camera coordinate systems up to scale.

Given the relative motion parameters

$$R_{\mathrm{rel}} = R^\top, \qquad \boldsymbol{t}_{\mathrm{rel}} = -R^\top \boldsymbol{t},$$

the *essential matrix* is

$$E = [\boldsymbol{t}_{\mathrm{rel}}]_\times R_{\mathrm{rel}} \in \mathbb{R}^{3\times 3}, \tag{7}$$

where $[\boldsymbol{t}_{\mathrm{rel}}]_\times$ is the 3×3 skew-symmetric matrix

$$[\boldsymbol{t}_{\mathrm{rel}}]_\times = \begin{pmatrix} 0 & -t_{\mathrm{rel},z} & t_{\mathrm{rel},y} \\ t_{\mathrm{rel},z} & 0 & -t_{\mathrm{rel},x} \\ -t_{\mathrm{rel},y} & t_{\mathrm{rel},x} & 0 \end{pmatrix}.$$

The normalized points satisfy the epipolar constraint

$$\tilde{\boldsymbol{x}}'^\top E\, \tilde{\boldsymbol{x}} = 0. \tag{8}$$

In pixel coordinates, the corresponding *fundamental matrix* F is

$$F = K'^{-\top} E K^{-1}, \tag{9}$$

and any noise-free correspondence $(\boldsymbol{x}, \boldsymbol{x}')$ satisfies

$$\boldsymbol{x}'^\top F\, \boldsymbol{x} = 0. \tag{10}$$

In practice, detected correspondences are noisy and do not exactly satisfy (10). To obtain correspondences that are consistent with the epipolar geometry, we apply an optimal correction based on the epipolar line equation.

Projection onto the epipolar line. Let $\boldsymbol{x}$ be fixed and consider correcting $\boldsymbol{x}'$. The epipolar line in the second image corresponding to $\boldsymbol{x}$ is

$$\boldsymbol{\ell}' = F\boldsymbol{x} = \begin{pmatrix} a \\ b \\ c \end{pmatrix},$$

so that any point $\boldsymbol{y}' = (u', v', 1)^\top$ on this line satisfies

$$\boldsymbol{\ell}'^\top \boldsymbol{y}' = au' + bv' + c = 0.$$

Given a noisy correspondence $\boldsymbol{x}' = (u', v', 1)^\top$, the closest point $\boldsymbol{x}'_{\mathrm{corr}}$ on the epipolar line $\boldsymbol{\ell}'$ in the Euclidean sense is obtained by orthogonal projection:

$$d' = au' + bv' + c, \tag{11}$$

$$u'_{\mathrm{corr}} = u' - a\,\frac{d'}{a^2 + b^2}, \tag{12}$$

$$v'_{\mathrm{corr}} = v' - b\,\frac{d'}{a^2 + b^2}. \tag{13}$$

The corrected homogeneous point is

$$\boldsymbol{x}'_{\text{corr}} = \begin{pmatrix} u'_{\text{corr}} \\ v'_{\text{corr}} \\ 1 \end{pmatrix},$$

which satisfies the epipolar constraint up to numerical precision:

$$\boldsymbol{x}'^{\top}_{\text{corr}} F \, \boldsymbol{x} \approx 0.$$

Symmetric correction. Similarly, one can project $\boldsymbol{x}$ onto the epipolar line in the first image

$$\boldsymbol{\ell} = F^{\top} \boldsymbol{x}' = \begin{pmatrix} a' \\ b' \\ c' \end{pmatrix},$$

and obtain a corrected point $\boldsymbol{x}_{\text{corr}}$ in the same way. A symmetric correction moves both $\boldsymbol{x}$ and $\boldsymbol{x}'$ minimally so that the corrected pair $(\boldsymbol{x}_{\text{corr}}, \boldsymbol{x}'_{\text{corr}})$ lies exactly on the epipolar curve defined by (10). In practice, the first-order *Sampson approximation* is often used as a computationally efficient measure of epipolar consistency, but in this note we focus on the projection-based correction described above.

Conventional stereo triangulation

Consider a standard horizontal stereo camera with baseline length b and focal lengths f and f' for the left and right cameras, respectively. Let u_L and u_R be the horizontal image coordinates of a corresponding point in the left and right images. The disparity d is defined by

$$d = u_L - u_R.$$

If the two cameras are rectified and share the same focal length $f = f'$, the depth Z of the point in front of the camera is given by the classic formula

$$Z = \frac{fb}{d}. \tag{14}$$

This conventional triangulation requires reliable feature detection and matching between the two images. On flat, textureless surfaces (such as floors or tabletops), feature points are scarce and outliers are likely to occur, which makes robust stereo matching difficult.

Pinhole camera model and pattern–projector-based triangulation

For completeness, we restate the pinhole camera model. For a 3D point $\boldsymbol{X} = (X, Y, Z)^{\top}$ and its image point $\boldsymbol{u} = (u, v, 1)^{\top}$, we have

$$\lambda \begin{pmatrix} u \\ v \\ 1 \end{pmatrix} = K \begin{bmatrix} R & \boldsymbol{t} \end{bmatrix} \begin{pmatrix} X \\ Y \\ Z \\ 1 \end{pmatrix}, \tag{15}$$

where K is the intrinsic parameter matrix, R is a rotation matrix, and $\boldsymbol{t}$ is a translation vector.

In the proposed method, we replace one of the cameras by a structured-light pattern projected from a smartphone display, and treat the grid intersections of the projected pattern as virtual feature points on the surface. This enables triangulation even on weakly textured surfaces. The detected grid intersections are further refined using the epipolar-based correction described above, and the resulting corrected correspondences are fed into the linear triangulation framework based on (2).

3 Smartphone-based structured-light triangulation

In this section we describe a triangulation method that uses a smartphone both as a projector and as a geometric reference. The smartphone is placed above a flat surface (e.g. the floor or a tabletop) with its display facing downward. The display shows a grid pattern, and each grid intersection is treated as a virtual light source on the display plane.

Figure 1 illustrates the geometry in a two-dimensional cross-section. The rectangular shape represents the smartphone display, which is modeled as a planar surface at a fixed height above the ground plane. The point labeled *depth* is the orthogonal projection of a scene point onto the display normal, and the slanted segment labeled p represents one of the rays from the scene point to a particular pixel on the display.

We introduce a world coordinate system whose z–axis is perpendicular to the ground plane. The ground plane is written as

$$\Pi_{\text{ground}} : z = 0,$$

and the smartphone display is modeled as a parallel plane

$$\Pi_{\text{disp}} : z = h,$$

where $h > 0$ is the distance ("depth") between the display and the ground. A pixel on the display with 2D display coordinates (u_d, v_d) is mapped to a 3D point

$$\boldsymbol{S}(u_d, v_d) = (X_d(u_d, v_d),\, Y_d(u_d, v_d),\, h)^\top$$

on the display plane, where (X_d, Y_d) are obtained from the known physical size and resolution of the screen.

For a point $\boldsymbol{X} = (X, Y, 0)^\top$ on the ground plane, the line segment connecting $\boldsymbol{S}(u_d, v_d)$ and $\boldsymbol{X}$ is one of the rays shown in Fig. 1. In the actual system, the intersection points of the projected grid with the ground are extracted from the captured image (after LoG filtering and skeletonization), and they are matched to the corresponding grid intersections on the display. Thus, for each detected grid intersection on the ground we obtain a pair

$$\big(\boldsymbol{S}(u_d, v_d),\, \boldsymbol{X}\big),$$

which forms a triangle together with the display normal. The height h corresponds to the distance labeled *depth* in the figure.

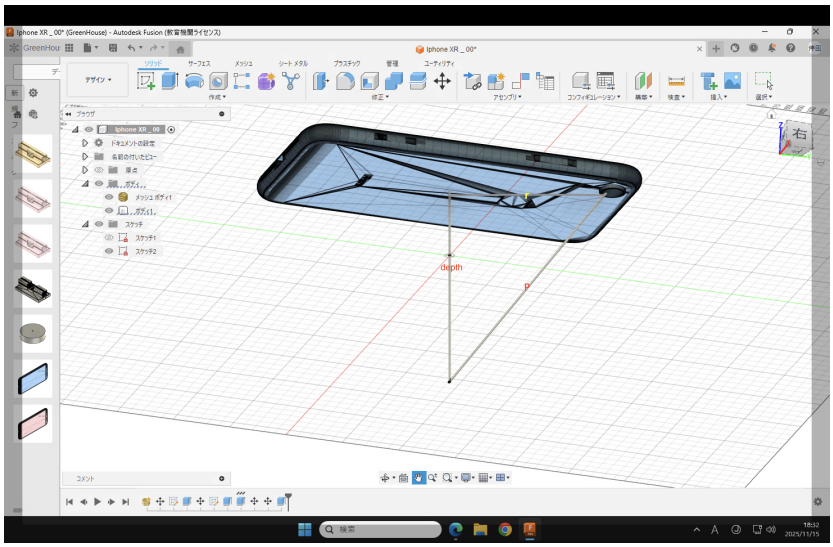

Figure 1: Cross-sectional view of the smartphone-based triangulation setup. The smartphone display is modeled as a plane above the ground. The vertical segment (depth) represents the distance h between the display and the ground, and the slanted segment p is a ray connecting a display pixel to a point on the surface.

Once the relative pose between the smartphone (display + front camera) and the global coordinate system is calibrated, these triangles can be used in the same way as conventional stereo triangulation: each grid intersection provides a pair of rays (one from the camera center and one from the corresponding display point), and the 3D position of the surface point is obtained as the least–squares intersection of these rays. In this way, the smartphone display acts as a structured-light projector that enables triangulation even on textureless surfaces.