3.1 Camera Models

Assume $OXYZ$ is the global coordinate system in the 3D scene, while $O_cX_cY_cZ_c$ is the camera coordinate system. Note that, in general these two systems may not be identical. Let $\widetilde{\mathbf{X}}$ be a (non-homogeneous) point in the scene (in global coordinates), and $\widetilde{\mathbf{X}}_{cam}$ be the coordinates for the same point in the camera coordinate system. It holds that

$$\widetilde{\mathbf{X}}_{cam} = R(\widetilde{\mathbf{X}} - \widetilde{\mathbf{C}}) \tag{2}$$

where $\widetilde{\mathbf{C}}$ is the center of the camera in the global coordinate system and R is a 3×3 rotation matrix which describes the orientation of the camera coordinate system with respect to the global coordinate system.

If $\mathbf{X}$ is the 3D point in homogeneous coordinates and $\mathbf{x}$ is the corresponding 2D point in the image, then

$$\mathbf{x} = KR[I| - \widetilde{\mathbf{C}}]\mathbf{X} \tag{3}$$

where K is the camera's calibration matrix which is defined by the intrinsic parameters of the camera. We can use the above equation to describe the camera matrix, P: $P = K[R|\mathbf{t}]$, where $\mathbf{t} = -R\widetilde{\mathbf{C}}$.

In case of two cameras, it is often easier to assume the global coordinate system to be same as the first camera's coordinate system. In that case, $P_1 = K_1[I|\mathbf{0}]$ and $P_2 = K_2[R|\mathbf{t}]$.

3.2 Fundamental Matrix

The fundamental matrix F is a 3×3 matrix which relates corresponding points in stereo images. Assume $\mathbf{x}_1 \leftrightarrow \mathbf{x}_2$ are corresponding points in an image pair, then the following relation holds:

$$\mathbf{x}_2^t F \mathbf{x}_1 = 0 \tag{4}$$

Note that the fundamental matrix is of rank 2.

3.3 The Essential Matrix

The essential matrix E relates calibrated corresponding image points, and it can be defined from the fundamental matrix as follows:

$$E = K_2^t F K_1 \tag{5}$$

If $P_1 = K_1[I|\mathbf{0}]$ and $P_2 = K_2[R|\mathbf{t}]$, the essential matrix can also be written as

$$E = [\mathbf{t}]_x R \tag{6}$$

where $[\mathbf{t}]_x$ is the representation of the cross product with $\mathbf{t}$. We can also show that the essential matrix is of rank 2 and has two identical real singular values while the third one is zero (as an optional exercise, you can try to prove it!). For more details refer to Chapter 7 in [1].

Eqn. 6 also tells us that knowing the essential matrix can allow us to estimate the camera orientation and center location as is described below.

3.4 Algorithms

In this section, we describe algorithms that are used to solve various problems related to the task of 3D reconstruction.

3.4.1 Eight-Point Algorithm

This algorithm fits the fundamental matrix defined by corresponding points $\mathbf{x}_1 \leftrightarrow \mathbf{x}_2$ in an image pair. In particular, assume points $\mathbf{x}_1^{(i)} = (x_1^{(i)}, y_1^{(i)})$ in the first image corresponds to points $\mathbf{x}_2^{(i)} = (x_2^{(i)}, y_2^{(i)})$ in the second image for $i = 1, ..., N$. Note that, $N \geq 8$, we need at least 8 corresponding points.

Normalization. Usually, the point correspondences are not completely accurate. To avoid numerical errors due to noise, we can normalize the points in the image, by translating them by μ (so that the mean of the points is at the origin), and scaling them by σ (so that the mean distance from the origin is a constant (e.g. $\sqrt{2}$)). Assume that the linear transformation described above is represented by the matrices T_1 and T_2 for the two images respectively.

Optimization. Eq. 4 can be rewritten equivalently as $Af = 0$, where f is formed from the entries of F stacked to a 9-vector row-wise and A is a $N \times 9$ dimensional matrix. In particular, the i-th row of A is equal to

$$A_i = [x_1^{(i)} x_2^{(i)} \quad y_1^{(i)} x_2^{(i)} \quad x_2^{(i)} \quad x_1^{(i)} y_2^{(i)} \quad y_i^{(i)} y_2^{(i)} \quad y_2^{(i)} \quad x_1^{(i)} \quad y_1^{(i)} \quad 1] \tag{7}$$

where the point coordinates have already been normalized, i.e. $\mathbf{x} \leftarrow T\mathbf{x}$.

The linear system$Af = 0$ has an exact non-zero solution if $rank(A) = 8$, which would happens if the point correspondences are exact. But usually due to noise, $rank(A) > 8$. In that case, there is no exact solution to the linear system and an approximate solution has to be found.

An approximate solution can be found by solving the following optimization problem

$$\min_f \quad ||Af||_2 \quad s.t. \;\; ||f||_2 = 1 \tag{8}$$

which can be solved by using the singular value decomposition of matrix A.

The solution F^* found by the approximately solving the problem $Af = 0$, does not guarantee that the fundamental matrix will be of rank 2. We can enforce this constraint, by solving another optimization problem:

$$\min_F \quad ||F - F^*||_F \quad s.t. \;\; rank(F) = 2 \tag{9}$$

Again, this problem can be solved using the singular value decomposition of F^*. In particular, if $F^* = USV^t$, where $S = diag(s_1, s_2, s_3)$ and $s_1 \geq s_2 \geq s_3$ then $F = U\hat{S}V^t$, where $\hat{S} = diag(s_1, s_2, 0)$, is the matrix that solves the above optimization problem.

Denormalization. After enforcing the rank-2 constraint, we need to remove the normalization such that the fundamental matrix corresponds to points in the actual 2D image space, i.e. $F \leftarrow T_2^t F T_1$.

3.4.2 Estimating Extrinsic Camera Parameters from the Essential Matrix

Another problem is to estimate the extrinsic camera parameters, i.e. $R, \mathbf{t}$ for the second camera, when only the essential matrix E is known.

The operation $[\mathbf{t}]_x$ can also be written as

$$[\mathbf{t}]_x = SZR_{90^o}S^t \tag{10}$$

where $S = [\mathbf{s}_0 \;\; \mathbf{s}_1 \;\; \mathbf{t}]$ is an orthogonal matrix, $Z = diag(1, 1, 0)$ and R_{90^o} is the rotation matrix for rotation angle $\theta = 90^o$.

From Eq. 6

$$E = [\mathbf{t}]_x R = SZR_{90^o}S^t R = U\Sigma V^t \tag{11}$$

Since we know that $\Sigma = diag(1, 1, 0)$ (remember in homogeneous coordinates we define everything up to a multiplication factor), it holds that $U = S$ and $V = R^t U R_{90^o}^t$.

Thus, the direction of $\mathbf{t}$ as well as R can be determined by the SVD analysis of E. Notice that there is no way we can determine the actual norm of the vector $\mathbf{t}$ from just image point correspondences. Absolute values can only be determined if we have a known reference distance in the scene. In addition, the signs of both $\mathbf{t}$ and R also can not be determined from SVD decomposition. Therefore, this algorithm generates numerous candidate positions and orientations for the camera.

Different combinations of $\mathbf{t}$ and R result in different camera matrix $P_2 = K_2[R|\mathbf{t}]$. After triangulation (3D reconstruction of the point cloud), the combination that gives the largest number of points in front of the image planes is selected.

3.4.3 Triangulation

Triangulation refers to the estimation of the 3D position of a point when the corresponding points in the two image planes are given as well as the parameters of the camera. In particular, assume $\mathbf{x}_1 \leftrightarrow \mathbf{x}_2$ are two corresponding points in the image plane and the camera matrices are P_1 and P_2 respectively. We want to find the 3D point $\mathbf{X} = (X, Y, Z, W)$ such that

$$\mathbf{x}_1 = P_1\mathbf{X} \quad and \quad \mathbf{x}_2 = P_2\mathbf{X} \tag{12}$$

Eq. 12 can also be written as:

$$x_j = \frac{P_{11}^{(j)} X + P_{12}^{(j)} Y + P_{13}^{(j)} Z + P_{14}^{(j)} W}{P_{31}^{(j)} X + P_{32}^{(j)} Y + P_{33}^{(j)} Z + P_{34}^{(j)} W} \tag{13}$$

$$y_j = \frac{P_{21}^{(j)} X + P_{22}^{(j)} Y + P_{23}^{(j)} Z + P_{24}^{(j)} W}{P_{31}^{(j)} X + P_{32}^{(j)} Y + P_{33}^{(j)} Z + P_{34}^{(j)} W} \tag{14}$$

where $j = 1, 2$ are the two camera indices and $\mathbf{x}_j = (x_j, y_j)$.

The above system of non linear equations can be turned into a linear system, if both equations are multiplied by the denominators. The system will be homogeneous if the point in 3D is represented in homogeneous coordinates and thus can be solved via SVD. Alternatively, if we set $W = 1$ then the system is no longer homogenous and can be solved using least squares.