Notes on Structure-from-Motion (SfM)

Simone Milani

A.A. 2020/2021

1 Preliminaries

These notes focus on the Structure-from-Motion (SfM) 3D reconstruction. Given a single camera moving around a static scene, it is possible to reconstruct a 3D model of the environment.

Then, two methods of uncalibrated reconstruction are considered.

- Method 1: perspective reconstruction
  - perspective reconstruction;
  - Euclidean promotion
- Method 2: self-calibration (8-points algorithm);
- incremental and hierarchical reconstruction.

2 Estimation and factorization of the essential matrix

Let us assume that we have a single camera, whose intrinsic parameters are known, moving around a static scene.

At time instant t, the camera is defined by the projective matrix $P = [Q|\mathbf{q}]$ while at time instant $t+1$ the camera is defined by the matrix $P' = [Q'|\mathbf{q}']$.

Matching points (Fig. 1) can be connected by the Longuet-Higgins equation

$$\mathbf{m}'^T [\mathbf{e}']_\times Q' Q^{-1} \mathbf{m} = \mathbf{m}'^T F \mathbf{m} = 0$$

Let us assume that the reference system of world coordinates correspond to the system of camera P. Then, it is possible to write

$$P = K[I|\mathbf{0}] \qquad P' = K[R|\mathbf{t}]$$

Assuming that K is known, it is possible to normalize the coordinates, i.e., $\mathbf{p} = K^{-1}\mathbf{m}$. Then, the related camera projection matrices can be written as

$$K^{-1}P = [I|\mathbf{0}] \qquad K^{-1}P' = [R|\mathbf{t}]$$

where $\mathbf{t}$ and R are the relative translation and rotation of camera P' w.r.t. to the reference system of P.

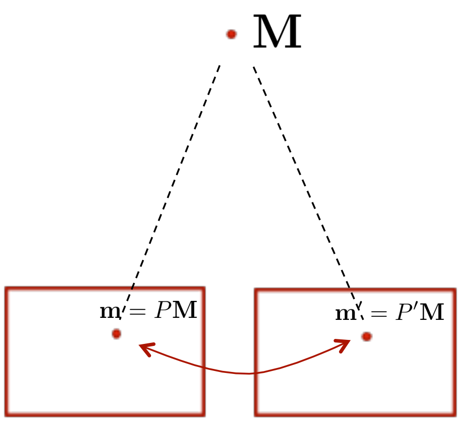

Figure 1: Matching point

The Longuet-Higgins equation becomes

$$\mathbf{p}'^T E \mathbf{p} = \mathbf{p}'^T [\mathbf{t}]_\times R \mathbf{p} = 0 \tag{1}$$

where the essential matrix is $E \triangleq [\mathbf{t}]_\times R$.

3 Perspective reconstruction

3.1 Projectivities

Let us assume that the 3D point $\mathbf{M}^j$ is projected on the image plane of the i-th camera in the pixel $\mathbf{m}_i^j$, i.e.,

$$\mathbf{m}_i^j \simeq P_i \mathbf{M}^j.$$

Given a set of n points $\mathbf{m}_i^j$ projected on h cameras, reconstruct P_i and $\mathbf{M}^j$ w.r.t. a transformation T, i.e.,

$$\text{if } \{P_i\} \text{ and } \{M^j\} \text{ are solutions} \qquad \Rightarrow \qquad \{P_i\ T\} \text{ and } \{T^{-1}\ M^j\} \text{ as well.}$$

If we consider the scaling factor ζ_i^j, we can write

$$\zeta_i^j \mathbf{m}_i^j = P_i \mathbf{M}^j, \qquad i = 1, \ldots, h \qquad j = 1, \ldots, n;$$

it is possible to gather all the equation in a single matrix equation

$$\begin{bmatrix} \zeta_1^1 \mathbf{m}_1^1 & \zeta_1^2 \mathbf{m}_1^2 & \ldots & \zeta_1^n \mathbf{m}_1^n \\ \zeta_2^1 \mathbf{m}_2^1 & \zeta_2^2 \mathbf{m}_2^2 & \ldots & \zeta_2^n \mathbf{m}_2^n \\ \vdots & \vdots & \ddots & \vdots \\ \zeta_h^1 \mathbf{m}_h^1 & \zeta_h^2 \mathbf{m}_h^2 & \ldots & \zeta_h^n \mathbf{m}_h^n \end{bmatrix} = \begin{bmatrix} P_1 \\ P_2 \\ \vdots \\ P_n \end{bmatrix} \begin{bmatrix} \mathbf{M}^1 & \mathbf{M}^2 & \ldots & \mathbf{M}^n \end{bmatrix} \tag{2}$$

which can be written more synthetically

$$W_{h\times n} = P_{h\times 4} M_{4\times n}.$$

The matrix W can be factorized into P and M.

Let us suppose that the factors ζ_i^j are known; then, W is completely defined and it is possible to apply the SVD

$$W = U\ D\ V^T. \tag{3}$$

Since W is defined by P and M and P has rank 4, W must have rank 4. Only the first 4 singular values are different from 0, i.e.,

$$D = \begin{bmatrix} \sigma_1 & 0 & 0 & 0 & 0 & \dots & 0 \\ 0 & \sigma_2 & 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & \sigma_3 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & \sigma_4 & 0 & \dots & 0 \\ \vdots & & \dots & & 0 & & \vdots \\ \vdots & & & & & \ddots & \\ 0 & & \dots & & & & 0 \end{bmatrix}$$

which leads to the simplified equation

$$W = U_{3h\times 4} \begin{bmatrix} \sigma_1 & 0 & 0 & 0 \\ 0 & \sigma_2 & 0 & 0 \\ 0 & 0 & \sigma_3 & 0 \\ 0 & 0 & 0 & \sigma_4 \end{bmatrix} V_{4\times n}^T \tag{4}$$

This leads to the factorization

$$P = U_{3h\times 4} \begin{bmatrix} \sigma_1 & 0 & 0 & 0 \\ 0 & \sigma_2 & 0 & 0 \\ 0 & 0 & \sigma_3 & 0 \\ 0 & 0 & 0 & \sigma_4 \end{bmatrix} \qquad \text{and} \qquad M = V_{4\times n}^T. \tag{5}$$

N.B. This solution minimizes the Frobenius norm $\|W - PM\|_2^2$.

What if $rank(W) \neq 4$ (because of noisy data)? It is possible to regularize W by zeroing all the singular values after σ_4. In this way, we force W to have only 4 non-zero singular values.

Scales are still unknown!

In case we know P and M, we can write

$$P\mathbf{M}^j = \begin{bmatrix} \zeta_1^j \mathbf{m}_1^j \\ \zeta_2^j \mathbf{m}_2^j \\ \vdots \\ \zeta_h^j \mathbf{m}_h^j \end{bmatrix} = \begin{bmatrix} \mathbf{m}_1^j & 0 & \dots & 0 \\ 0 & \mathbf{m}_2^j & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & \dots & 0 \end{bmatrix} \begin{bmatrix} \zeta_1^j \\ \zeta_2^j \\ \vdots \\ \zeta_h^j \end{bmatrix} = Q^j \boldsymbol{\zeta}^j. \tag{6}$$

So, from P and M we can have $\boldsymbol{\zeta}^j$; from $\boldsymbol{\zeta}^j$, it is possible to find P and M: Chicken-egg problem !

It is possible to solve it via an iterative optimization.

1. Set initially $\zeta_i^j = 1$; it is possible to generate matrix W.
2. Normalize W s.t. $\|W\|_F = 1$ (needed to avoid the ill-posed case $\zeta_i^j = 0$).
3. Apply SVD on W finding P and M.
4. If $\|W - PM\|_2^2$ is small enough, go to 8.
5. Find $\boldsymbol{\zeta}^j$ from $Q^j \boldsymbol{\zeta}^j = PM^j$, $j = 1, \ldots, n$.
6. Update W.
7. Go to 2.
8. End

3.2 Euclidean promotion

We have already stated that the reconstruction is performed w.r.t. a transformation T.

How to compute T ?

Let us make two assumptions:

- we have enough images/cameras with enough corresponding points.
- intrinsic parameters are fixed, i.e., $K_i = K = const$.

Projective reconstruction permits obtaining camera projection matrices $\{P_i^p\}$, $i = 1, \ldots, h$. Let us take the camera $i = 1$ as reference, i.e.,

$$P_1^p = [I|\mathbf{0}] \qquad P_i^p = [Q_i|\mathbf{q}_i]\,; \tag{7}$$

as for Euclidean projection matrices, we have

$$P_1^e = K\,[I|\mathbf{0}] \qquad P_i^e = K\,[R_i|\mathbf{t}_i]\,.$$

The target is finding T such that

$$\mathbf{m}_i^j = P_i^p \; T \; T^{-1}\mathbf{M}^j,$$

i.e.,

$$P_i^e \simeq P_i^p T. \tag{8}$$

Since $P_1^e = [K \mid \mathbf{0}] = P_1^p \; T = [I \; |\mathbf{0} \;]T$, the matrix T can be defined as

$$T = \begin{bmatrix} K & \mathbf{0} \\ \mathbf{r}^T & s \end{bmatrix} \qquad \text{where } \mathbf{r} = \begin{bmatrix} r_1 \\ r_2 \\ r_3 \end{bmatrix}.$$

T is characterized by 8 parameters: 5 from K, and 3 from $\mathbf{r}$. The parameter s can be set to 1 since all the relations are defined w.r.t. to a scale.

This leads to the equation

$$P_i^e \simeq P_i^p \; T = \left[Q_i \; K + \mathbf{q}_i \; \mathbf{r}^T \mid \mathbf{q}_i\right] \tag{9}$$

which can be compared with $P_i^e = K\,[R_i|\mathbf{t}_i]$ leading to the equation

$$Q_i \; K + \mathbf{q}_i \; \mathbf{r}^T \simeq K \; R_i \qquad \text{(Keyden-Anstrom '96).} \tag{10}$$

The parameters Q_i and $\mathbf{q}_i$ are known from perspective reconstruction; K, $\mathbf{r}$, and R_i are not known (R_i is a rotation matrix).

The relation

$$P_i^p \begin{bmatrix} K \\ \mathbf{r}^T \end{bmatrix} \simeq K \; R_i$$

permits writing

$$\begin{aligned} P_i^p \begin{bmatrix} K \\ \mathbf{r}^T \end{bmatrix} \left(P_i^p \begin{bmatrix} K \\ \mathbf{r}^T \end{bmatrix} \right)^T &= P_i^p \begin{bmatrix} K \\ \mathbf{r}^T \end{bmatrix} \begin{bmatrix} K \\ \mathbf{r}^T \end{bmatrix}^T P_i^{pT} \\ &= P_i^p \begin{bmatrix} KK^T & K\mathbf{r} \\ \mathbf{r}^T K^T & \mathbf{r}^T \mathbf{r} \end{bmatrix} P_i^{pT} \\ &\simeq KR_i \, (KR_i)^T = KR_i R_i^T K^T \\ &= KK^T \end{aligned}$$

which can be synthesized in the equation

$$P_i^p \begin{bmatrix} KK^T & K\mathbf{r} \\ \mathbf{r}^T K^T & \mathbf{r}^T \mathbf{r} \end{bmatrix} P_i^{pT} \simeq KK^T \qquad \textbf{(Kruppa's bound).} \tag{11}$$

Note that P_i^p are known, while K and $\mathbf{r}$ are to be determined. In this case, we have 8 unknowns: α_u, α_v, u_0, v_0, r_1, r_2, and r_3. The number of equations obtained from eq. (11) is 5. We have 3×3 matrices that are symmetric: therefore, the $3 \times 3 = 9$ equations reduces to 6. Moreover, the relations is defined w.r.t. a scale factor ($\simeq$): the number of useful equations is utterly reduced to 5.

As a matter of fact, we need at least 3 cameras (two couple of cameras) to find the unknowns. Camera 1 always satisfy the relation ($Q_1 = I$, $\mathbf{q}_1 = \mathbf{0}$).

It is possible to express the problem as a zero-crossing point search for the function

$$0 = f_i(K, \mathbf{r}, \lambda_i) = \lambda_i^2 KK^T - P_i^p \begin{bmatrix} KK^T & K\mathbf{r} \\ \mathbf{r}^T K^T & \mathbf{r}^T \mathbf{r} \end{bmatrix} P_i^{pT} \tag{12}$$

where we have replaced the relation $\simeq$ with an equality by including the scale factor λ_i, i.e., using

$$P_i^p \begin{bmatrix} K \\ \mathbf{r}^T \end{bmatrix} = \lambda_i K \; R_i.$$

Note that three cameras are still sufficient since we have 10 equations in 10 unknowns (the previous ones + two λ factors).

But what if K is not constant?

4 Self-calibration: the 8 points algorithm

Let us go back to Longuet-Higgins equation.

$$\mathbf{m}'^T \ F \ \mathbf{m} = 0.$$

Given a sufficient number of corresponding points, it is possible to estimate F.

Remind that $\mathbf{m}'^T \ F \ \mathbf{m} = 0$. Note that the epipolar line equation allows us to write

$$[\mathbf{e}']_\times \ \mathbf{m}' \simeq \lambda[\mathbf{e}']_\times Q' Q^{-1}\mathbf{m}.$$

Multiplying by $\mathbf{m}'^T$ on the left, we have

$$\mathbf{m}'^T[\mathbf{e}']_\times \ \mathbf{m}' = 0 \simeq \lambda\mathbf{m}'^T[\mathbf{e}']_\times Q' Q^{-1}\mathbf{m}.$$

which allows us to write

$$F = \mathbf{m}'^T[\mathbf{e}']_\times Q' Q^{-1}. \tag{13}$$

Note that F is defined w.r.t. a scale factor; note also that $det([\mathbf{e}']_\times) = 0$ and therefore $det(F) = 0$. This imply that F has 7 d.o.f.

Remember that

$$F = K'^{-T} E K^{-1} = K'^{-T} \ ([\mathbf{t}]_\times R) \ K^{-1}, \tag{14}$$

where E has 5 d.o.f. (due to the fact that two singular values must be equal and the third is 0). These are called rigidity bounds. The difference depends on K and K'. These two extra bounds are useful to compute K and K'.

Since the unknowns are 5, we need more bounds to find K, i.e., we need more couple of cameras. If $K = K' = const$, 3 cameras (3 independent couples) are enough.