Structure from motion

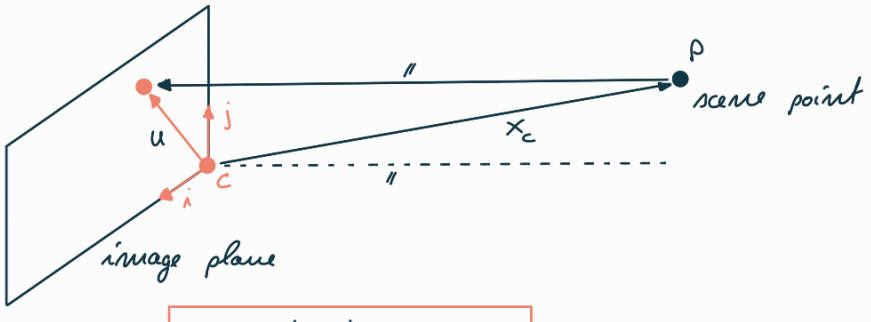
Compute the 3D scene structure and camera motion from a sequence of frames.

INPUT : video of a scene, sequence of frames.
OUTPUT: 3D scene structure

① Detect feature points : corners, SIFT points, ...
② Track the features through the entire video : template matching, optical flow, comparing SIFT descriptors, ...
→ set of tracked features
N points in each frame $1, \dots, F$.

INPUT : set of corresponding image points (2D) : $(u_{f,p}, v_{f,p}) \quad \forall p \in 1, \dots, N$
OUTPUT: find scene points (3D)

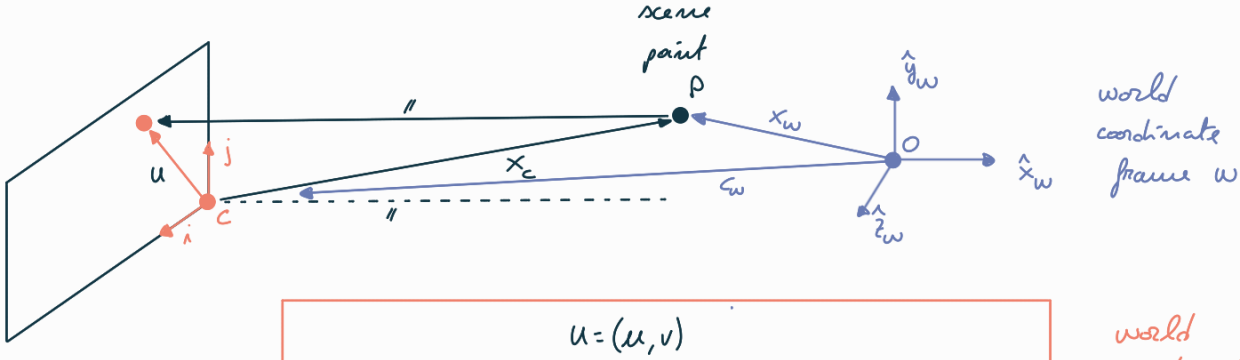
A set is a sequence of points ∀ feature (in time)
Assume we are using an orthographic camera : distance between points in the scene is small compared to the distance of the scene from the camera: the magnification of the camera is constant for all points in the scene and between scenes.



$u = (u, v)$

$u = i \cdot x_c = i^T x_c$
$v = j \cdot x_c = j^T x_c$

camera coordinate frame



world coordinate frame w

$u = (u, v)$

$u = i \cdot x_c = i^T x_c = i^T (x_w - c_w) = i^T (P - C)$
$v = j \cdot x_c = j^T x_c = j^T (x_w - c_w) = j^T (P - C)$

world coordinate frame

Given corresponding image points (2D) $(u_{f,p}, v_{f,p})$
Find scene points $\{P_p\}$.
Camera positions $\{C_f\}$ and camera orientations $\{(i_f, j_f)\}$ are unknown.

Image of point P_p in camera frame f:

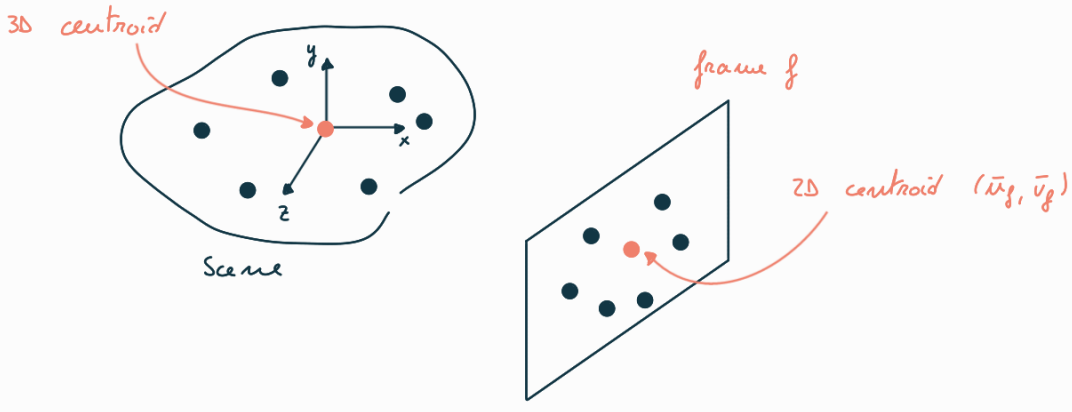
$$\begin{cases} u_{f,p} = i_f^T (P_p - C_f) \\ v_{f,p} = j_f^T (P_p - C_f) \end{cases}$$

$u_{f,p}$, $v_{f,p}$: known — i_f, j_f: unknown

We remove C_f from the equations to simplify the problem using the centering trick.

Centering trick: assuming origin of the world at centroid of scene points:

$$\frac{1}{N} \sum_{p=1}^{N} P_p = \bar{P} = 0$$



Centroid $(\bar{u}_f, \bar{v}_f)$ of the image points in frame f:

$$\bar{u}_f = \frac{1}{N} \sum_{p=1}^{N} u_{f,p} = \frac{1}{N} \sum_{p=1}^{N} i_f^T (P_p - C_f)$$

$$= \frac{1}{N} i_f^T \sum_{p=1}^{N} P_p - \frac{1}{N} \sum_{p=1}^{N} i_f^T C_f$$

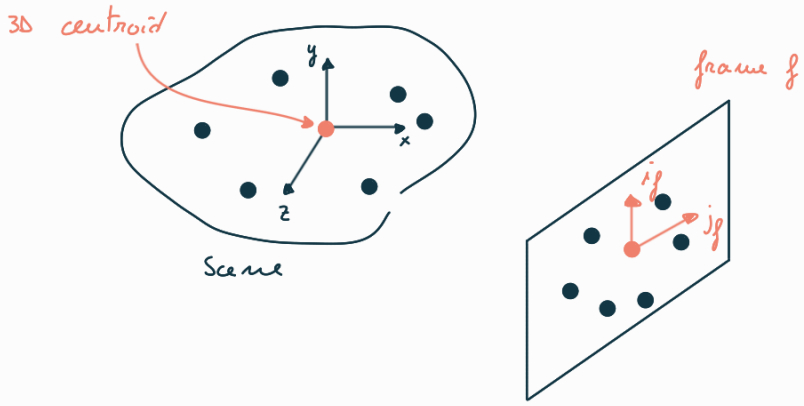
we set the origin on the centroid — constant

$$\bar{u}_f = - i_f^T C_f$$

The centroid of the feature points is not a function of the locations of the scene points

the $\bar{v}_f$ component can be computed the same way

$$\bar{v}_f = - j_f^T C_f$$



Shift camera origin to the centroid $(\bar{u}_f, \bar{v}_f)$

New image points:

$$\tilde{u}_{f,p} = u_{f,p} - \bar{u}_f = i_f^T (P_p - C_f) - i_f^T C_f = i_f^T P_p$$
$$\tilde{v}_{f,p} = v_{f,p} - \bar{v}_f = j_f^T (P_p - C_f) - j_f^T C_f = j_f^T P_p$$

Now we have the image coordinates of same points but we no longer have the camera center in the expression

Now for each point in a particular frame, we get this expression:

$$\begin{aligned} \tilde{u}_{f,p} &= i_f^T P_p \\ \tilde{v}_{f,p} &= j_f^T P_p \end{aligned} \longrightarrow \begin{bmatrix} \tilde{u}_{f,p} \\ \tilde{v}_{f,p} \end{bmatrix} = \begin{bmatrix} i_f^T \\ j_f^T \end{bmatrix} P_p$$

We have N points and F frames. We can organize them into a single matrix:

$$\begin{array}{c|cccc} & \text{point 1} & \text{point 2} & \cdots & \text{point N} \\ \hline \text{image 1} & \tilde{u}_{1,1} & \tilde{u}_{1,2} & \cdots & \tilde{u}_{1,N} \\ \text{image 2} & \tilde{u}_{2,1} & \tilde{u}_{2,2} & \cdots & \tilde{u}_{2,N} \\ \vdots & \vdots & \vdots & & \vdots \\ \text{image F} & \tilde{u}_{F,1} & \tilde{u}_{F,2} & \cdots & \tilde{u}_{F,N} \\ \text{image 1} & \tilde{v}_{1,1} & \tilde{v}_{1,2} & \cdots & \tilde{v}_{1,N} \\ \text{image 2} & \tilde{v}_{2,1} & \tilde{v}_{2,2} & \cdots & \tilde{v}_{2,N} \\ \vdots & \vdots & \vdots & & \vdots \\ \text{image F} & \tilde{v}_{F,1} & \tilde{v}_{F,2} & \cdots & \tilde{v}_{F,N} \end{array} = \begin{array}{c} \\ i_1^T \\ i_2^T \\ \vdots \\ i_F^T \\ j_1^T \\ j_2^T \\ \vdots \\ j_F^T \end{array} \begin{array}{c|ccc} \text{point 1} & \text{point 2} & & \text{point N} \\ \hline P_1 & P_2 & \cdots & P_N \end{array}$$

$W_{2F \times N}$ — Centroid-Subtracted Feature points (Known)

$M_{2F \times 1}$ — Camera motion (unknown)

$S_{3 \times N}$ — Scene structure (unknown)

Observation matrix

Can we find M and S from W?

Yes because the observation matrix W has a special property: it has a very low rank.

linear independence of vectors

A set of vectors $\{v_1, v_2, \dots, v_n\}$ is said to be linearly independent if no vector can be represented as a weighted linear sum of the others.

Column rank

The column rank of a matrix is the number of linearly independent columns of the matrix

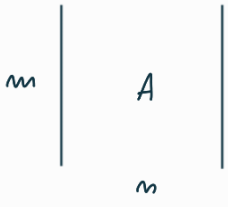
Row rank

The row rank of a matrix is the number of linearly independent rows of the matrix

Rank

For any matrix:

$$\text{columnRank}(A) = \text{rowRank}(A) = \text{rank}(A)$$
$$\text{rank}(A) \leq \min(m, n)$$

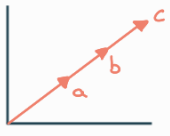


$\text{columnRank}(A) \leq n$
$\text{rowRank}(A) \leq m$

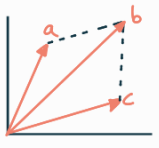
Rank is the dimensionality of the space spanned by column or row vectors of the matrix.

$$A = \begin{vmatrix} a_1 & b_1 & c_1 \\ a_2 & b_2 & c_2 \\ a_3 & b_3 & c_3 \end{vmatrix} = | \; a \; b \; c \; |$$

$\text{rank}(A) = 1$ ⟷



$\text{rank}(A) = 2$ ⟷

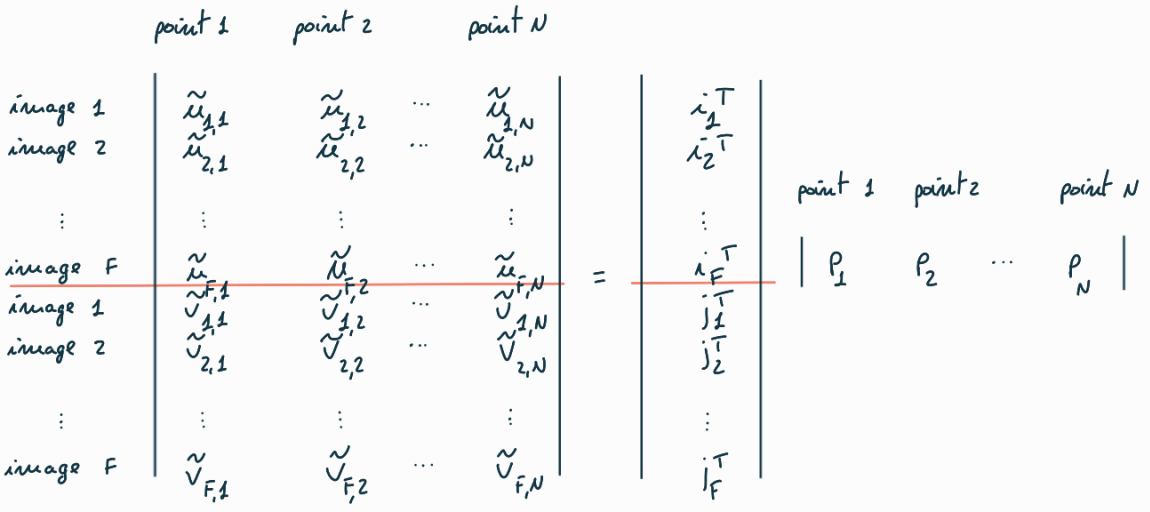


$\text{rank}(A) = 3$ ⟷ in a 3-dimensional space we need three vectors to express any arbitrary vector

The rank has some properties :

- $\text{rank}(A^T) = \text{rank}(A)$
- $\text{rank}(A_{m\times n}\, B_{n\times p}) = \min(\text{rank}(A_{m\times n}), \text{rank}(B_{n,p})) \leq \min(m, n, p)$
- $\text{rank}(AA^T) = \text{rank}(A^TA) = \text{rank}(A^T) = \text{rank}(A)$
- $A_{m\times m}$ is invertible iff $\text{rank}(A_{m\times m}) = m$

Back to the Observation matrix



$W_{2F\times N}$ — Centroid-Subtracted Feature points (known)

$M_{2F\times 3}$ — Camera motion (unknown)

$S_{3\times N}$ — Scene structure (unknown)

Observation matrix

$$W = M \times S$$
$$2F\times N \quad 2F\times 3 \quad 3\times N$$

$\text{rank}(MS) \leq \text{rank}(M) \longrightarrow \text{rank}(MS) \leq \min(3, 2F)$
$\text{rank}(MS) \leq \text{rank}(S) \longrightarrow \text{rank}(MS) \leq \min(3, N)$

$$\text{rank}(W) = \text{rank}(MS) \leq \min(3, 2F, N) = 3$$

Singular Value Decomposition (SVD)

For any matrix A there exists a factorization:

$$A_{m\times n} = U_{m\times m} \cdot \Sigma_{m\times n} \cdot V^T_{n\times n}$$

where U and V are orthonormal and Σ is diagonal.

$$\Sigma_{m\times n} = \begin{vmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_N \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & 0 \end{vmatrix}$$

$\sigma_1 \le \sigma_2 \le \cdots \le \sigma_N$ are the *singular values*

If $\text{rank}(A) = r$ then A has r non-zero singular values.

Tomasi-Kanade Factorization Method

We can "factorize" W into R and S.
We will use a technique called Singular Value Decomposition (SVD)

Using SVD: $W = U\Sigma V^T$

$$= \begin{vmatrix} \\ U \\ \\ \end{vmatrix} \quad \begin{vmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_N \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & 0 \end{vmatrix} \quad \begin{vmatrix} \\ V^T \\ \\ \end{vmatrix}$$

$2F \times 2F$ $\qquad$ $2F \times N$ $\qquad$ $N \times N$

since $\text{rank}(W) \le 3$, $\text{rank}(\Sigma) \le 3$ (Σ has at most 3 non-zero singular values).

3 $\qquad$ $2F-3$

$$= \begin{vmatrix} \\ U_1 \\ \\ \end{vmatrix} \quad \begin{vmatrix} U_2 \end{vmatrix} \quad \begin{vmatrix} \sigma_1 & 0 & 0 & \cdots & 0 \\ 0 & \sigma_2 & 0 & \cdots & 0 \\ 0 & 0 & \sigma_3 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 \end{vmatrix} \quad \begin{vmatrix} V_1^T \\ \\ V_2^T \end{vmatrix} \quad \begin{matrix} 3 \\ \\ N-3 \end{matrix}$$

$2F \times 2F$ $\qquad$ $2F \times N$ $\qquad$ $N \times N$

$$W = U_1 \quad \Sigma_1 \quad V_1^T$$
$$(2F\times 3) \quad (3\times 3) \quad (3\times N)$$

Factorization (Finding M, S)

$W = U_1 \quad \Sigma_1 \quad V_1^T$

$W = U_1(\Sigma_1)^{\frac{1}{2}} \quad (\Sigma_1)^{\frac{1}{2}} V_1^T$ Two possible factors for W

$W = U_1(\Sigma_1)^{\frac{1}{2}} Q \quad Q^{-1}(\Sigma_1)^{\frac{1}{2}} V_1^T$ Also two possible valid factors for W

now we can use the orthonormality of M

$$M = \begin{vmatrix} i_1^T \\ i_2^T \\ \vdots \\ i_F^T \\ j_1^T \\ j_2^T \\ \vdots \\ j_F^T \end{vmatrix} = \underbrace{U_1(\Sigma_1)^{\frac{1}{2}}}_{\text{computed}} \underbrace{Q}_{\text{unknowns}} = \begin{vmatrix} \hat{i}_1^T \\ \vdots \\ \hat{i}_F^T \\ \hat{j}_1^T \\ \vdots \\ \hat{j}_F^T \end{vmatrix} Q = \begin{vmatrix} \hat{i}_1^T Q \\ \vdots \\ \hat{i}_F^T Q \\ \hat{j}_1^T Q \\ \vdots \\ \hat{j}_F^T Q \end{vmatrix}$$

orthonormality constraints:

$$\left.\begin{aligned} i_f \cdot i_f = i_f^T i_f = 1 \\ j_f \cdot j_f = j_f^T j_f = 1 \\ i_f \cdot j_f = i_f^T j_f = 0 \end{aligned}\right\} \longrightarrow \left\{\begin{aligned} \hat{i}_f^T Q Q^{-1} \hat{i}_f = 1 \\ \hat{j}_f^T Q Q^{-1} \hat{j}_f = 1 \\ \hat{i}_f^T Q Q^{-1} \hat{j}_f = 0 \end{aligned}\right.$$

for each frame we get these 3 equations

Q is 3×3, thus it has 9 variables
we have $3F$ of these constraints
Q can be solved with 3 or more images ($F \geq 3$) using Newton's method.

After finding Q:

$M = U_1(\Sigma_1)^{\frac{1}{2}} Q$ camera motion

$S = Q^{-1}(\Sigma_1)^{\frac{1}{2}} V_1^T$ scene structure

Summary : orthographic structure from motion

① Detect and track feature points

② Create the centroid-subtracted matrix W of corresponding feature points

③ Compute SVD of W and enforce rank constraint

$$W = U \Sigma V^T = \underset{(2F \times 3)}{U_1} \quad \underset{(3 \times 3)}{\Sigma_1} \quad \underset{(3 \times N)}{V_1^T}$$

④ Set $M = U_1 (\Sigma_1)^{\frac{1}{2}} Q$ and $S = Q^{-1} (\Sigma_1)^{\frac{1}{2}} V_1^T$

⑤ Find Q by enforcing the orthonormality constraint